# STATISTICS
## for Engineering and the Sciences
### SIXTH EDITION

William M. Mendenhall ▪ Terry L. Sincich

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

# STATISTICS

## for Engineering and the Sciences

### SIXTH EDITION

# STATISTICS

## for Engineering and the Sciences

### SIXTH EDITION

William M. Mendenhall

Terry L. Sincich

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

This book was previously published by Pearson Education, Inc.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742


Version Date: 20160302

International Standard Book Number-13: 978-1-4987-2887-4 (eBook - PDF)



**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

# Contents

## Chapter 4 Discrete Random Variables 133



## Chapter 5 Continuous Random Variables 186



## Chapter 6 Bivariate Probability Distributions and Sampling Distributions 234

## Chapter 7 Estimation Using Confidence Intervals 288



## Chapter 8 Tests of Hypotheses 368



## Chapter 9 Categorical Data Analysis 442

# Preface

## Overview

This text is designed for a two-semester introductory course in statistics for students majoring in engineering or any of the physical sciences. Inevitably, once these student graduate and are employed, they will be involved in the collection and analysis of data and will be required to think critically about the results. Consequently, they need to acquire knowledge of the basic concepts of data description and statistical inference and familiarity with statistical methods that will be required use on the job.

## Pedagogy

Chapters 1 through 6 identify the objectives of statistics, explain how we can describe data, and present the basic concepts of probability. Chapters 7 and 8 introduce the two methods for making inferences about population parameters: estimation with confidence intervals and hypothesis testing. These notions are extended in the remaining chapters to cover other topics that are useful in analyzing engineering and scientific data, including the analysis of categorical data (Chapter 9), regression analysis and model building (Chapters 10–12), the analysis of variance for designed experiments (Chapters 13–14), nonparametric statistics ((Chapter 15), statistical quality control (Chapter 16), and product and system reliability (Chapter 17).

## Features

Hallmark features of this text are as follows:

1. **Blend of theory and applications.** The basic theoretical concepts of mathematical statistics are integrated with a two-semester presentation of statistical methodology. Thus, the instructor has the option of presenting a course with either of two characteristics—a course stressing basic concepts and applied statistics, or a course that, while still tilted toward application, presents a modest introduction to the theory underlying statistical inference.
2. **Statistical software applications with tutorials.** The instructor and student have the option of using statistical software to perform the statistical calculations required. Output from three popular statistical software products — SAS, SPSS, and MINITAB—as well as Microsoft Excel are fully integrated into the text. Tutorials with menu screens and dialog boxes associated with the software are provided in Appendices C, D, and E. These tutorials are designed for the novice user; no prior experience with the software is needed.
3. **Blended coverage of topics and applications.** To meet the diverse needs of future engineers and scientists, the text provides coverage of a wide range of data analysis topics. The material on multiple regression and model building (Chapters 11–12), principles of experimental design (Chapter 13), quality control (Chapter 15), and reliability (Chapter 17) sets the text apart from the typical introductory statistics text. Although the material often refers to theoretical concepts, the presentation is oriented toward applications.
4. **Real data-based examples and exercises.** The text contains large number of applied examples and exercises designed to motivate students and suggest future uses of the methodology. Nearly every exercise and example is based on data or experimental results from actual engineering and scientific studies published in academic journals or obtained from the organization conducting the analysis. These applied exercises are located at the end of every section and at the ends of chapters.

5. **Statistics in Action case studies.** Each chapter begins with a discussion of an actual contemporary scientific study ("Statistics in Action") and the accompanying data. The analysis and inferences derived from the study are presented at key points in the chapter ("Statistics in Action Revisited"). Our goal is to show the students the importance of applying sound statistical methods in order to evaluate the findings and to think through the statistical issues involved.
6. **End-of-chapter summary material.** At the end of each chapter, we provide a summary of the topics presented via a "Quick Review" (key words and key formulas), "Language Lab" (a listing of key symbols and pronunciation guide), and "Chapter Summary Notes/Guidelines". These features help the student summarize and reinforce the important points from the chapter and are useful study tools.
7. **Standard mathematical notation for a random variable.** Throughout the chapters on random variables, we use standard mathematical notation for representing a random variable. Uppercase letters represent the random variable, and lowercase letters represent the values that the random variable can assume.
8. **Bootstrapping and Bayesian methods.** In optional sections, the text presents two alternative estimation methods (Section 7.12) and hypothesis testing methods (Section 8.13) that are becoming more popular in scientific studies—bootstrapping and Bayesian methods.
9. **All data sets provided online.** All of the data associated with examples, exercises, and Statistics in Action cases are made available online at www.crcpress.com/product/isbn/9781498728850. Each data file is marked with a icon and file name in the text. The data files are saved in four different formats: MINITAB, SAS, SPSS, and Excel. By analyzing these data using statistical software, calculations are minimized, allowing student to concentrate on the interpretation of the results.

## New to the Sixth Edition

Although the scope and coverage remain the same, the 6th edition of the text contains several substantial changes, additions, and enhancements:

1. **Over 1,000 exercises, with revisions and updates to 30%.** Many new and updated exercises, based on contemporary engineering and scientific-related studies and real data, have been added. Most of these exercises—extracted from scientific journals—foster and promote critical thinking skills.
2. **Updated technology.** Throughout the text, we have increased the number of statistical software printouts. All printouts from statistical software (SAS, SPSS, and MINITAB) and corresponding instructions for use have been revised to reflect the latest versions of the software.
3. **Statistics in Action Revisited.** For this edition, we introduce the "Statistics in Action" case (see above) at the beginning of each chapter. After covering the required methodology in the chapter, the solution (data analysis and inference) is then presented and discussed in a "Statistics in Action Revisited" at the end of the section.
4. **Chapter 1 (Collecting Data/Sampling).** Material on all basic sampling concepts (e.g., random sampling and sample survey designs) has been streamlined and moved to Section 1.4 to give the students an earlier introduction to key sampling issues.

5. **Chapter 7 (Matched Pairs vs. Independent Samples).** We have added an example (Example 7.12) that compares directly the analysis of data from matched pairs with a similar analysis of the data using an independent samples *t*-test.
6. **Chapter 8 (Hypothesis Test/*p*-values).** The section on *p*-values in hypothesis testing (Section 8.5) has been moved up to emphasize the importance of their use in engineering and scientific-related studies. Throughout the remainder of the text, conclusions from a test of hypothesis are based on *p*-values.
7. **Chapters 10 and 11 (Regression Residuals).** A new section (Section 10.8) has been added on using regression residuals to check the assumptions required in a simple linear regression analysis. A similar section (Section 11.10) in the multiple regression chapter has been modified to emphasize the different uses of regression residuals, including for assumption verification and for detecting outliers and influential observations.
8. **Chapter 13 (Experimental Design).** Two new examples (Examples 13.6 and 13.7) have been added on selecting the sample size for a designed experiment.
9. **Chapter 14 (Analysis of Variance).** Two new examples (Examples 14.8 and 14.10) have been added on analyzing a two-factor experiment with quantitative factors. The first employs the traditional ANOVA model and the second utilizes a regression model with higher-order terms.

Numerous, less obvious changes in details have been made throughout the text in response to suggestions by current users and reviewers of the text.

## Supplements

### *Student Solutions Manual*

Includes complete worked out solutions to the odd-numbered text exercises.

### *Instructor's Solutions Manual*

Solutions to all of the even-numbered text exercises are given in this manual. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text.

## Acknowledgments

This book reflects the efforts of a great many people over a number of years. First, we would like to thank the following professors, whose reviews and comments on this and prior editions have contributed to the 6th edition:

**Reviewers Involved with the Sixth Edition:**

Shyamaia Nagaraj (University of Michigan)
Stacie Pisano (University of Virginia)
Vishnu Nanduri (University of Wisconsin-Milwaukee)
Shuchi Jain (Virginia Commonwealth University)
David Lovell (University of Maryland)
Raj Mutharasan (Drexel University)
Gary Wasserman (Wayne State University)
Nasser Fard (Northeastern University)

**Reviewers of Previous Editions:**

Carl Bodenschatz (United States Air Force Academy)
Dharam Chopra (Wichita University)
Edward Danial (Morgan State University)
George C. Derringer (Battelle Columbus, Ohio, Division)
Danny Dyer (University of Texas-Arlington)
Herberg Eisenberg (West Virginia College of Graduate Studies)
Christopher Ennis (Normandale Community College)
Nasrollah Etemadi (University of Illinois-Chicago)
Linda Gans (California State Polytechnic University)
Carol Gattis (University of Arkansas)
Frank Guess (University of Tennessee)
Carol O'Connor Holloman (University of Louisville)
K. G. Janardan (Eastern Michigan University)
H. Lennon (Coventry Polytechnic, Coventry, England)
Nancy Matthews (University of Oklahoma)
Jeffery Maxey (University of Central Florida)
Curtis McKnight (University of Oklahoma)
Chand Midha (University of Akron)
Balgobin Nandram (Worcester Polytechnic Institute)
Paul Nelson (Kansas State University)
Norbert Oppenheim (City College of New York)
Giovanni Parmigiani (Duke University)
David Powers (Clarkson University)
Alan Rabideau (University of Bufffalo)
Charles Reilly (University of Central Florida)
Larry Ringer (Texas A&M University)
David Robinson (St. Cloud State University)
Shiva Saksena (University of North Carolina-Wilmington)
Arnold Sweet (Purdue University)
Paul Switzer (Stanford University)
Dennis Wackerly (University of Florida)
Donald Woods (Texas A&M Universtity)

**Other Contributors**

Special thanks are due to our supplements authors, including Nancy Boudreau, several of whom have worked with us for many years. Finally, the Taylor & Francis Publishing staff of David Grubbs, Jessica Vakili, and Suzanne Lassandro helped greatly with all phases of the text development, production, and marketing effort.

CHAPTER 1

# Introduction

**OBJECTIVE**

*To identify the role of statistics in the analysis of data from engineering and the sciences*

**CONTENTS**

## *STATISTICS IN ACTION*
### DDT Contamination of Fish in the Tennessee River

Chemical and manufacturing plants often discharge toxic waste materials into nearby rivers and streams. These toxicants have a detrimental effect on the plant and animal life inhabiting the river and the river's bank. One type of pollutant is dichlorodiphenyltrichloroethane, commonly known as DDT. DDT was used as an effective agricultural insecticide in the United States until it was banned for agricultural use in 1972 due to its carcinogenic properties. However, because DDT is often a by-product of certain manufactured materials (e.g., petroleum distillates, water-wettable powders, and aerosols), it remains an environmental hazard today.

This *Statistics in Action* case is based on a study undertaken to examine the level of DDT contamination of fish inhabiting the Tennessee River (in Alabama) and its tributaries. The Tennessee River flows in a west-east direction across the northern part of the state of Alabama, through Wheeler Reservoir, a national wildlife refuge. Ecologists fear that contaminated fish migrating from the mouth of the river to the reservoir could endanger other wildlife that prey on the fish. This concern is more than academic. A manufacturing plant was once located along Indian Creek, which enters the Tennessee River 321 miles upstream from the mouth. Although the plant has been inactive for a number of years, there is evidence that the plant discharged toxic materials into the creek, contaminating all of the fish in the immediate area.

The Food and Drug Administration sets the limit for DDT content in individual fish at 5 parts per million (ppm). Fish with a DDT content exceeding this limit are considered to be contaminated—that is, potentially hazardous to the surrounding environment. Are the fish in the Tennessee River and its tributary creeks contaminated with DDT? And if so, how far upstream have the contaminated fish migrated?

To answer these and other questions, members of the U.S. Army Corps of Engineers collected fish specimens at different locations along the Tennessee River and three tributary creeks: Flint Creek (which enters the river 309 miles upstream from the river's mouth), Limestone Creek (310 miles upstream), and Spring Creek (282 miles upstream). Six fish specimens were captured at each of the three tributary creeks, and 126 specimens at various locations (miles upstream) along the Tennessee River, for a total of 144 fish specimens. The location and species of each fish was determined as well as the weight (in grams) and length (in centimeters). Then the filet of each fish was extracted and the DDT concentration (ppm) measured. The data for the 144 captured fish are saved in the **DDT** file.

In the *Statistics in Action Revisited* at the end of this chapter, we discuss the type of data collected and the data collection method. Later in the text, we analyze the data for the purposes of characterizing the level of DDT contamination of fish in the Tennessee River, comparing the DDT contents of fish at different river locations, and to determine the relationship (if any) of length and weight to DDT content.

## 1.1 Statistics: The Science of Data

A successful engineer or scientist is one who is proficient at collecting information, evaluating it, and drawing conclusions from it. This requires proper training in statistics. According to *The Random House College Dictionary*, statistics is "the science that deals with the collection, classification, analysis, and interpretation of information or data." In short, **statistics** is the **science of data.**

**Definition 1.1**

**Statistics** is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data.

The science of statistics is commonly applied to two types of problems:

1. Summarizing, describing, and exploring data
2. Using sample data to infer the nature of the data set from which the sample was selected

As an illustration of the descriptive applications of statistics, consider the United States census, which involves the collection of a data set that purports to characterize the socioeconomic characteristics of the approximately 300 million people living in the United States. Managing this enormous mass of data is a problem for the computer software engineer, and describing the data utilizes the methods of statistics. Similarly, an environmental engineer uses statistics to describe the data set consisting of the daily emissions of sulfur oxides of an industrial plant recorded for 365 days last year. The branch of statistics devoted to these applications is called **descriptive statistics.**

**Definition 1.2**

The branch of statistics devoted to the organization, summarization, and description of data sets is called **descriptive statistics.**

Sometimes the phenomenon of interest is characterized by a data set that is either physically unobtainable or too costly or time-consuming to obtain. In such situations, we obtain a subset of the data—called a *sample*—and use the sample information to infer its nature. To illustrate, suppose the phenomenon of interest is the drinking-water quality on an inhabited, but remote, Pacific island. You might expect water quality to depend on such factors as temperature of the water, the level of the most recent rainfall, etc. In fact, if you were to measure the water quality repeatedly within the same hour at the same location, the quality measurements would vary, even for the same water temperature. Thus, the phenomenon "drinking-water quality" is characterized by a large data set that consists of many (actually, an infinite number of) water quality measurements—a data set that exists only conceptually. To determine the nature of this data set, we *sample* it—i.e., we record quality for $n$ water specimens collected at specified times and locations, and then use this sample of $n$ quality measurements to infer the nature of the large conceptual data set of interest. The branch of statistics used to solve this problem is called **inferential statistics.**

**Definition 1.3**

The branch of statistics concerned with using sample data to make an inference about a large set of data is called **inferential statistics.**

## 1.2 Fundamental Elements of Statistics

In statistical terminology, the data set that we want to describe, the one that characterizes a phenomenon of interest to us, is called a **population.** Then, we can define a **sample** as a subset of data selected from a population. Sometimes, the words *population* and *sample* are used to represent the objects upon which the measurements are taken (i.e., the *experimental units*). In a particular study, the meaning attached to these terms will be clear by the context in which they are used.

**Definition 1.4**

A statistical **population** is a data set (usually large, sometimes conceptual) that is our target of interest.

**Definition 1.5**

A **sample** is a subset of data selected from the target population.

**Definition 1.6**

The object (e.g., person, thing, transaction, specimen, or event) upon which measurements are collected is called the **experimental unit.** (*Note:* A population consists of data collected on many experimental units.)

In studying populations and samples, we focus on one or more characteristics or properties of the experimental units in the population. The science of statistics refers to these characteristics as **variables.** For example, in the drinking-water quality study, two variables of interest to engineers are the chlorine-residual (measured in parts per million) and the number of fecal coliforms in a 100-milliliter water specimen.

**Definition 1.7**

A **variable** is a characteristic or property of an individual experimental unit.

**Example 1.1**

Rate of Left-Turn Automobile Accidents

Engineers with the University of Kentucky Transportation Research Program have collected data on accidents occurring at intersections in Lexington, Kentucky. One of the goals of the study was to estimate the rate at which left-turn accidents occur at intersections without left-turn-only lanes. This estimate will be used to develop numerical warrants (or guidelines) for the installation of left-turn lanes at all major Lexington intersections. The engineers collected data at each of 50 intersections without left-turn-only lanes over a 1-year period. At each intersection, they monitored traffic and recorded the total number of cars turning left that were involved in an accident.

a. Identify the variable and experimental unit for this study.
b. Describe the target population and the sample.
c. What inference do the transportation engineers want to make?

Solution

a. Since the engineers collected data at each of 50 intersections, the experimental unit is an intersection without a left-turn-only lane. The variable measured is the total number of cars turning left that were involved in an accident.
b. The goal of the study is to develop guidelines for the installation of left-turn lanes at all major Lexington intersections; consequently, the target population consists of all major intersections in the city. The sample consists of the subset of 25 intersections monitored by the engineers.
c. The engineers will use the sample data to estimate the rate at which left-turn accidents occur at all major Lexington intersections. (We learn, in Chapter 7, that this estimate is the number of left-turn accidents in the sample divided by the total number of cars making left turns in the sample.)

The preceding definitions and example identify four of the five elements of an inferential statistical problem: a population, one or more variables of interest, a sample, and an inference. The fifth element pertains to knowing how good the inference is—that is, the **reliability** of the inference. The measure of reliability that accompanies an inference separates the science of statistics from the art of fortune-telling. A palm reader, like a statistician, may examine a sample (your hand) and make inferences about the population (your future life). However, unlike statistical inferences, the palm reader's inferences include no measure of how likely the inference is to be true.

To illustrate, consider the transportation engineers' estimate of the left-turn accident rate at Lexington, Kentucky, intersections in Example 1.1. The engineers are interested in the *error of estimation* (i.e., the difference between the sample accident rate and the accident rate for the target population). Using statistical methods, we can determine a *bound on the estimation error.* This bound is simply a number (e.g., 10%) that our estimation error is not likely to exceed. In later chapters, we learn that this bound is used to help measure our "confidence" in the inference. The reliability of statistical inferences is discussed throughout this text. For now, simply realize that an inference is incomplete without a measure of reliability.

**Definition 1.8**

A **measure of reliability** is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

A summary of the elements of both descriptive and inferential statistical problems is given in the following boxes.

### Four Elements of Descriptive Statistical Problems

1. The population or sample of interest
2. One or more variables (characteristics of the population or sample units) that are to be investigated
3. Tables, graphs, or numerical summary tools
4. Identification of patterns in the data

### Five Elements of Inferential Statistical Problems

1. The population of interest
2. One or more variables (characteristics of the experimental units) that are to be investigated
3. The sample of experimental units
4. The inference about the population based on information contained in the sample
5. A measure of reliability for the inference

## Applied Exercises

1.1 *STEM experiences for girls.* Over the past several decades, the National Science Foundation (NSF) has promoted girls participation in informal science, technology, engineering or mathematics (STEM) programs. What has been the impact of these informal STEM experiences? This was the question of interest in the published study, *Cascading Influences: Long-Term Impacts of Informal STEM Experiences for Girls* (March, 2013). A sample of 159 young women who recently participated in a STEM program were recruited to complete an on-line survey. Of these, only 27% felt that participation in the STEM program increased their interest in science.

a. Identify the population of interest to the researchers.
b. Identify the sample.
c. Use the information in the study to make an inference about the relevant population.

1.2 *Corrosion prevention of buried steel structures.* Steel structures, such as piping, that are buried underground are susceptible to corrosion. Engineers have designed tests on the structures that measure the potential for corrosion. In *Materials Performance* (March 2013), two tests for steel corrosion—called "instant-off" and "instant-on" potential—were compared. The tests were applied to buried piping at a petrochemical plant in Turkey. Both the "instant-off" and "instant-on" corrosion measurements were made at each of 19 different randomly selected pipe locations. One objective of the study is to determine if one test is more desirable (i.e., can more accurately predict the potential for corrosion) than the other when applied to buried steel piping.

a. What are the experimental units for this study?
b. Describe the sample.
c. Describe the population.
d. Is this an example of descriptive or inferential statistics?

1.3 *Visual attention skills test.* Researchers at Griffin University (Australia) conducted a study to determine whether video game players have superior visual attention skills compared to non-video game players. (*Journal of Articles in Support of the Null Hypothesis,* Vol. 6, No. 1, 2009.) Each in a sample of 65 male students was classified as a video game player or a non-player. The two groups were then subjected to a series of visual attention tasks that included the "field of view" test. No differences in the performance of the two groups were found. From this analysis, the researchers inferred "a limited role for video game playing in the modification of visual attention". Thus, inferential statistics was applied to arrive at this conclusion. Identify the relevant populations and samples for this study.

**SWREUSE**

1.4 *Success/failure of software reuse.* The PROMISE Software Engineering Repository, hosted by the University of Ottawa, is a collection of publicly available data sets to serve researchers in building prediction software models. A PROMISE data set on software reuse, saved in the **SWREUSE** file, provides information on the success or failure of reusing previously developed software for each project in a sample of 24 new software development projects. (Data source: *IEEE Transactions on Software Engineering*, Vol. 28, 2002.) Of the 24 projects, 9 were judged failures and 15 were successfully implemented.

a. Identify the experimental units for this study.
b. Describe the population from which the sample is selected.
c. Use the sample information to make an inference about the population.

1.5 *Ground motion of earthquakes.* In the *Journal of Earthquake Engineering* (Nov. 2004), a team of civil and environmental engineers studied the ground motion characteristics of 15 earthquakes that occurred around the world. Three (of many) variables measured on each earthquake were the type of ground motion (short, long, or forward directive), earthquake magnitude (Richter scale) and peak ground acceleration (feet per second). One of the goals of the study was to estimate the inelastic spectra of any ground motion cycle.

a. Identify the experimental units for this study.
b. Do the data for the 15 earthquakes represent a population or a sample? Explain.

1.6 *Precooling vegetables.* Researchers have developed a new precooling method for preparing Florida vegetables for market. The system employs an air and water mixture designed to yield effective cooling with a much lower water flow than conventional hydrocooling. To compare the effectiveness of the two systems, 20 batches of green tomatoes were divided into two groups; one group was precooled with the new method, and the other with the conventional method. The water flow (in gallons) required to effectively cool each batch was recorded.

a. Identify the population, the samples, and the type of statistical inference to be made for this problem.
b. How could the sample data be used to compare the cooling effectiveness of the two systems?

**COGAS**

1.7 *Weekly carbon monoxide data.* The World Data Centre for Greenhouse Gases collects and archives data for greenhouse and related gases in the atmosphere. One such data set lists the level of carbon monoxide gas (measured in parts per billion) in the atmosphere each week at the Cold Bay, Alaska, weather station. The weekly data for the years 2000–2002 are saved in the **COGAS** file.

a. Identify the variable measured and the corresponding experimental unit.
b. If you are interested in describing only the weekly carbon monoxide values at Cold Bay station for the years 2000–2002, does the data represent a population or a sample? Explain.

1.8 *Monitoring defective items.* Checking all manufactured items coming off an assembly line for defectives would be a costly and time-consuming procedure. One effective and economical method of checking for defectives involves the selection and examination of a portion of the items by a quality control engineer. The percentage of examined items that are defective is computed and then used to estimate the percentage of all items manufactured on the line that are defective. Identify the population, the sample, and a type of statistical inference to be made for this problem.

## 1.3 Types of Data

Data can be one of two types, quantitative or qualitative. **Quantitative data** are those that represent the quantity or amount of something, measured on a numerical scale. For example, the power frequency (measured in megahertz) of a semiconductor is a quantitative variable, as is the breaking strength (measured in pounds per square inch) of steel pipe. In contrast, **qualitative** (or **categorical**) **data** possess no quantitative interpretation. They can only be classified. The set of $n$ occupations corresponding to a group of $n$ engineering graduates is a qualitative data set. The type of pigment (zinc or mica) used in an anticorrosion epoxy coating also represents qualitative data.*

*A finer breakdown of data types into nominal, ordinal, interval, and ratio data is possible. **Nominal** data are qualitative data with categories that cannot be meaningfully ordered. **Ordinal** data are also qualitative data, but a distinct ranking of the groups from high to low exists. **Interval** and **ratio** data are two different types of quantitative data. For most statistical applications (and all the methods presented in this introductory text), it is sufficient to classify data as either quantitative or qualitative.

**Definition 1.9**

**Quantitative data** are those that are recorded on a naturally occurring numerical scale, i.e., they represent the quantity or amount of something.

**Definition 1.10**

**Qualitative data** are those that cannot be measured on a natural numerical scale, i.e., they can only be classified into categories.

**Example 1.2**

Characteristics of Water Pipes

The *Journal of Performance of Constructed Facilities* reported on the performance dimensions of water distribution networks in the Philadelphia area. For one part of the study, the following variables were measured for each sampled water pipe section. Identify the data produced by each as quantitative or qualitative.

a. Pipe diameter (measured in inches)
b. Pipe material (steel or PVC)
c. Pipe location (Center City or suburbs)
d. Pipe length (measured in feet)

Solution

Both pipe diameter (in inches) and pipe length (in feet) are measured on a meaningful numerical scale; hence, these two variables produce quantitative data. Both type of pipe material and pipe location can only be classified—material is either steel or PVC; location is either Center City or the suburbs. Consequently, pipe material and pipe location are both qualitative variables.

**The proper statistical tool used to describe and analyze data will depend on the type of data.** Consequently, it is important to differentiate between quantitative and qualitative data.

## Applied Exercises

1.9 *Properties of cemented soils.* The properties of natural and cemented sandy soils in Cyprus was investigated in the *Bulletin of Engineering Geology and the Environment* (Vol. 69, 2010). For each of 20 soil specimens, the following variables were measured. Determine the type, quantitative or qualitative, of each variable.

a. Sampling method (rotary core, metal tube, or plastic tube)
b. Effective stress level (Newtons per meters squared)
c. Damping ratio (percentage)

1.10 *Satellite database.* The Union for Concerned Scientists (UCS) maintains the Satellite Database—a listing of the more than 1000 operational satellites currently in orbit around Earth. Several of the many variables stored in the database include country of operator/owner, primary use (civil, commercial, government, or military), class of orbit (low Earth, medium Earth, or geosynchronous), longitudinal position (degrees), apogee (i.e., altitude farthest from Earth's center of mass, in kilometers), launch mass (kilograms), usable electric power (watts), and expected lifetime (years). Which of the variables measured are qualitative? Which are quantitative?

1.11 *Drinking-water quality study. Disasters* (Vol. 28, 2004) published a study of the effects of a tropical cyclone on the quality of drinking water on a remote Pacific island. Water samples (size 500 milliliters) were collected approximately 4 weeks after Cyclone Ami hit the island. The following variables were recorded for each water sample. Identify each variable as quantitative or qualitative.

a. Town where sample was collected
b. Type of water supply (river intake, stream, or borehole)
c. Acidic level (pH scale, 1 to 14)
d. Turbidity level (nephalometric turbidity units = NTUs)
e. Temperature (degrees Centigrade)
f. Number of fecal coliforms per 100 milliliters
g. Free chlorine-residual (milligrams per liter)
h. Presence of hydrogen sulphide (yes or no)

1.12 *Extinct New Zealand birds.* Environmental engineers at the University of California (Riverside) are studying the patterns of extinction in the New Zealand bird population. (*Evolutionary Ecology Research*, July 2003.) The following characteristics were determined for each bird species that inhabited New Zealand at the time of the Maori colonization (i.e., prior to European Contact). Identify each variable as quantitative or qualitative.

a. Flight capability (volant or flightless)
b. Habitat type (aquatic, ground terrestrial, or aerial terrestrial)
c. Nesting site (ground, cavity within ground, tree, cavity above ground)
d. Nest density (high or low)
e. Diet (fish, vertebrates, vegetables, or invertebrates)
f. Body mass (grams)
g. Egg length (millimeters)
h. Extinct status (extinct, absent from island, present)

1.13 *CT scanning for lung cancer.* A new type of screening for lung cancer, computed tomography (CT), has been developed. Medical physicists believe CT scans are more sensitive than regular X-rays in pinpointing small tumors. The H. Lee Moffitt Cancer Center at the University of South Florida is currently conducting a clinical trial of 50,000 smokers nationwide to compare the effectiveness of CT scans with X-rays for detecting lung cancer. (*Todays' Tomorrows*, Fall 2002.) Each participating smoker is randomly assigned to one of two screening methods, CT or chest X-ray, and their progress tracked over time. In addition to the type of screening method used, the physicists recorded the age at which the scanning method first detects a tumor for each smoker.

a. Identify the experimental units of the study.
b. Identify the two variables measured for each experimental unit.
c. Identify the type (quantitative or qualitative) of the variables measured.
d. What is the inference that will ultimately be drawn from the clinical trial?

1.14 *National Bridge Inventory.* All highway bridges in the United States are inspected periodically for structural deficiency by the Federal Highway Administration (FHWA). Data from the FHWA inspections are compiled into the National Bridge Inventory (NBI). Several of the nearly 100 variables maintained by the NBI are listed below. Classify each variable as quantitative or qualitative.

a. Length of maximum span (feet)
b. Number of vehicle lanes
c. Toll bridge (yes or no)
d. Average daily traffic
e. Condition of deck (good, fair, or poor)
f. Bypass or detour length (miles)
g. Route type (interstate, U.S., state, county, or city)

## 1.4 Collecting Data: Sampling

Once you decide on the type of data—quantitative or qualitative—appropriate for the problem at hand, you'll need to collect the data. Generally, you can obtain the data in three different ways:

1. Data from a *published source*
2. Data from a *designed experiment*
3. Data from an *observational study* (e.g., a *survey*)

Sometimes, the data set of interest has already been collected for you and is available in a **published source,** such as a book, journal, newspaper, or Web site. For example, a transportation engineer may want to examine and summarize the automobile accident death rates in the 50 states of the United States. You can find this data set (as well as numerous other data sets) at your library in the *Statistical Abstract of the United States,* published annually by the U.S. government. The Internet (World Wide Web) now provides a medium by which data from published sources are readily available.*

A second, more common, method of collecting data in engineering and the sciences involves conducting a **designed experiment,** in which the researcher exerts strict control over the units (people, objects, or events) in the study. For example, an often-cited medical study investigated the potential of aspirin in preventing heart attacks. Volunteer physicians were divided into two groups—the *treatment* group and the *control* group. In the treatment group, each physician took one aspirin tablet a day for 1 year, while each physician in the control group took an aspirin-free placebo (no drug) made to look like an aspirin tablet. The researchers, not the physicians under

* With published data, we often make a distinction between the *primary source* and *secondary source*. If the publisher is the original collector of the data, the source is primary. Otherwise, the data are secondary source.

study, controlled who received the aspirin (the treatment) and who received the placebo. As you will learn in Chapter 13, a properly designed experiment allows you to extract more information from the data than is possible with an uncontrolled study.

Finally, observational studies can be employed to collect data. In an **observational study,** the researcher observes the experimental units in their natural setting and records the variable(s) of interest. For example, an industrial engineer might observe and record the level of productivity of a sample of assembly line workers. Unlike a designed experiment, an observational study is one in which the researcher makes no attempt to control any aspect of the experimental units. A common type of observational study is a survey, where the researcher samples a group of people, asks one or more questions, and records the responses.

**Definition 1.11**

A **designed experiment** is a data-collection method where the researcher exerts full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned the *treatment* and an untreated (or, *control*) group.

**Definition 1.12**

An **observational study** is a data-collection method where the experimental units sampled are observed in their natural setting. No attempt is made to control the characteristics of the experimental units sampled. (Examples include *opinion polls* and *surveys*.)

Regardless of the data-collection method employed, it is likely that the data will be a sample from some population. And if we wish to apply inferential statistics, we must obtain a *representative sample*.

**Definition 1.13**

A **representative sample** exhibits characteristics typical of those possessed by the population of interest.

For example, consider a poll conducted to estimate the percentage of all U.S. citizens who believe in global warming. The pollster would be unwise to base the estimate on survey data collected for a sample of citizens who belong to the Greenpeace organization (a group who exposes and confronts environmental abuse). Such an estimate would almost certainly be *biased* high; consequently, it would not be very reliable.

The most common way to satisfy the representative sample requirement is to select a simple random sample. A **simple random sample** ensures that every subset of fixed size in the population has the same chance of being included in the sample. If the pollster samples 1,500 of the 150 million U.S. citizens in the population so that every subset of 1,500 citizens has an equal chance of being selected, she has devised a simple random sample.

**Definition 1.14**

A **simple random sample** of $n$ experimental units is a sample selected from the population in such a way that every different sample of size $n$ has an equal chance of selection.

The procedure for selecting a simple random sample typically relies on a **random number generator.** Random number generators are available in table form, online,* and in most statistical software packages. The statistical software packages presented in this text all have easy-to-use random number generators for creating a random sample. The next two examples illustrate the procedure.

* One of many free online random number generators is available at www.randomizer.org.

**Example 1.3**

Obtaining a Simple Random Sample for Strength Testing

Suppose you want to randomly sample 5 glass-fiber strips from a lot of 100 strips for strength testing. (Note: In Chapter 3 we demonstrate that there are 75,287,520 possible samples that could be selected.) Use a random number generator to select a simple random sample of 5 glass-fiber strips.

Solution

To ensure that each of the possible samples has an equal chance of being selected (as required for simple random sampling), we will use the *random number table* provided in Table 1 of Appendix B. Random number tables are constructed in such a way that every number in the table occurs (approximately) the same number of times (i.e., each number has an equal chance of being selected). Furthermore, the occurrence of any one number in a position in the table is independent of any of the other numbers that appear in the table. Since the lot of glass-fiber strips contains 100 strips, the size of our target population is 100 and we want to sample 5 strips. Consequently, we will number the strips from 1 to 100 (i.e., number the strips 1, 2, 3, . . . , 99, 100). Then turn to Table 1 and select (arbitrarily) a starting number in the table. Proceeding from this number either across the row or down the column, remove and record a total of 5 numbers (the random sample) from the table.

**TABLE 1.1 Partial Reproduction of Table 1 in Appendix B**

| Row \ Column | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 10480 | 15011 | 01536 | 02011 | 81647 |
| 2 | 22368 | 46573 | 25595 | 85393 | 30995 |
| 3 | 24130 | 48360 | 22527 | 97265 | 76393 |
| 4 | 42167 | 93093 | 06243 | 61680 | 07856 |
| 5 | 37570 | 39975 | 81837 | 16656 | 06121 |
| 6 | 77921 | 06907 | 11008 | 42751 | 27756 |
| 7 | 99562 | 72905 | 56420 | 69994 | 98872 |
| 8 | 96301 | 91977 | 05463 | 07972 | 18876 |
| 9 | 89579 | 14342 | 63661 | 10281 | 17453 |
| 10 | 85475 | 36857 | 53342 | 53988 | 53060 |
| 11 | 28918 | 69578 | 88231 | 33276 | 70997 |
| 12 | 63553 | 40961 | 48235 | 03427 | 49626 |
| 13 | 09429 | 93969 | 52636 | 92737 | 88974 |
| 14 | 10365 | 61129 | 87529 | 85689 | 48237 |
| 15 | 07119 | 97336 | 71048 | 08178 | 77233 |
| 16 | 51085 | 12765 | 51821 | 51259 | 77452 |
| 17 | 02368 | 21382 | 52404 | 60268 | 89368 |
| 18 | 01011 | 54092 | 33362 | 94904 | 31273 |
| 19 | 52162 | 53916 | 46369 | 58586 | 23216 |
| 20 | 07056 | 97628 | 33787 | 09998 | 42698 |
| 21 | 48663 | 91245 | 85828 | 14346 | 09172 |
| 22 | 54164 | 58492 | 22421 | 74103 | 47070 |
| 23 | 32639 | 32363 | 05597 | 24200 | 13363 |
| 24 | 29334 | 27001 | 87637 | 87308 | 58731 |
| 25 | 02488 | 33062 | 28834 | 07351 | 19731 |

To illustrate, turn to a page of Table 1, say the first page. (A partial reproduction of the first page of the random number table is shown in Table 1.1.) Now, arbitrarily select a starting number, say the random number appearing in row 13, column 1. This number is 09429. Using only the first three digits (since the highest numbered strip is 100) yields the random number 94. Therefore, the first strip in the sample is strip #94. Now proceed down (an arbitrary choice) the column in this fashion, using only the first three digits of the random number and skipping any number that is greater than 100 until you obtain five random numbers. This method yields the random numbers 94, 103 (skip), 71, 510 (skip), 23, 10, 521 (skip), and 70.* (These numbers are highlighted in Table 1.1.) Consequently, our sample will include the glass-fiber strips numbered 94, 71, 23, 10, and 70.

The random number table is a convenient and easy-to-use random number generator as long as the size of the sample is not very large. For scientific studies that require a large sample, computers are used to generate the random sample. For example, suppose we require a random sample of 25 glass-fiber strips from a lot of 100,000 strips. Here, we can employ the random number generator of SAS statistical software. Figure 1.1 shows the SAS output listing 25 random numbers from a population of 100,000. The strips with these identification numbers (e.g., 2660, 25687, . . . . , 87662) would be included in the simple random sample of size 25.

**FIGURE 1.1**
SAS-generated random sample of 25 glass-fiber strips

Random sample of lot Table

| | STRIP |
|---|---|
| 1 | 2660 |
| 2 | 25687 |
| 3 | 67895 |
| 4 | 84928 |
| 5 | 89964 |
| 6 | 81903 |
| 7 | 14460 |
| 8 | 83165 |
| 9 | 18842 |
| 10 | 29611 |
| 11 | 41550 |
| 12 | 26712 |
| 13 | 40847 |
| 14 | 28990 |
| 15 | 32908 |
| 16 | 60969 |
| 17 | 11605 |
| 18 | 96324 |
| 19 | 52357 |
| 20 | 25882 |
| 21 | 77347 |
| 22 | 13609 |
| 23 | 27791 |
| 24 | 14724 |
| 25 | 87662 |

* If, in the course of recording the random numbers from the table, you select a number that has been previously selected, simply discard the duplicate and select a replacement at the end of the sequence. (This is called *sampling without replacement*.) Thus, you may have to record more random numbers from the table than the size of the sample in order to obtain the simple random sample.

The notion of random selection and randomization is also key to conducting good research with a designed experiment. The next example illustrates a basic application.

**Example 1.4**

Randomization in a designed experiment

An experiment was carried out by engineers at Georgia Tech (and published in *Human Factors*) to gauge the reaction times of people with cognitively demanding jobs (e.g., air traffic controller or radar/sonar operator) when they perform a visual search task. Volunteers were randomly divided into two groups. One group was trained to search using the "continuously consistent" method (Method A), while the other was trained using the "adjusted consistent" method (Method B). One goal was to compare the reaction times of the two groups. Assume 20 people volunteered for the study. Use a random number generator to randomly assign half of the volunteers to Method A and half to Method B.

Solution

Essentially, we want to select a random sample of 10 volunteers from the 20. The first 10 selected will be assigned to the Method A group; the remaining 10 will be assigned to the Method B group. (Alternatively, we could randomly assign each volunteer, one by one, to either Method A or B. However, this would not guarantee exactly 10 volunteers in each group.)

The MINITAB random sample procedure was employed, producing the printout shown in Figure 1.2. Numbering the volunteers from 1 to 20, we see that volunteers 8, 13, 9, 19, 16, 1, 12, 15, 18, and 14 are assigned to the group trained by Method A. The remaining volunteers are assigned to the group trained by Method B.

**FIGURE 1.2**
MINITAB worksheet with random assignment of volunteers

VISUAL-SEARCH.MTW ***

| ↓ | C1 | C2 | C3 |
|---|---|---|---|
| | Volunteer | MethodA | |
| 1 | 1 | 8 | |
| 2 | 2 | 13 | |
| 3 | 3 | 9 | |
| 4 | 4 | 19 | |
| 5 | 5 | 16 | |
| 6 | 6 | 1 | |
| 7 | 7 | 12 | |
| 8 | 8 | 15 | |
| 9 | 9 | 18 | |
| 10 | 10 | 14 | |
| 11 | 11 | | |
| 12 | 12 | | |
| 13 | 13 | | |
| 14 | 14 | | |
| 15 | 15 | | |
| 16 | 16 | | |
| 17 | 17 | | |
| 18 | 18 | | |
| 19 | 19 | | |
| 20 | 20 | | |
| 21 | | | |

In addition to simple random samples, there are more complex random sampling designs that can be employed. These include (but are not limited to) **stratified random sampling, cluster sampling**, and **systematic sampling**. Brief descriptions of each follow. (For more details on the use of these sampling methods, consult the references at the end of this chapter.)

*Stratified random sampling* is typically used when the experimental units associated with the population can be separated into two or more groups of units, called *strata*, where the characteristics of the experimental units are more similar within strata than across strata. Random samples of experimental units are obtained for each strata, then the units are combined to form the complete sample. For example, a transportation engineer interested in estimating average vehicle travel time in a city may want to stratify on a road's maximum speed limit (e.g., 25 mph, 40 mph, or 55 mph), making sure that representative samples of vehicles (in proportion to those in the target population) traveling on each of the road strata are included in the sample.

Sometimes it is more convenient and logical to sample natural groupings (*clusters*) of experimental units first, then collect data from all experimental units within each cluster. This involves the use of *cluster sampling*. For example, suppose a software engineer wants to estimate the proportion of lines of computer code with errors in 150 programs associated with a certain project. Rather than collect a simple random sample of all lines of code in the 150 programs (which would be very difficult and costly to do), the engineer will randomly sample 10 of the 150 programs (clusters), then examine all lines of code in each sampled program.

Another popular sampling method is *systematic sampling*. This method involves systematically selecting every *k*th experimental unit from a list of all experimental units. For example, a quality control engineer at a manufacturing plant may select every 10th item.

No matter what type of sampling design you employ to collect the data for your study, be careful to avoid **selection bias**. Selection bias occurs when some experimental units in the population have less chance of being included in the sample than others. This results in samples that are not representative of the population. Consider an opinion poll on whether a device to prevent cell phone use while driving should be installed in all cars. Suppose the poll employs either a telephone survey or mail survey. After collecting a random sample of phone numbers or mailing addresses, each person in the sample is contacted via telephone or the mail and a survey conducted. Unfortunately, these types of surveys often suffer from selection bias due to *nonresponse*. Some individuals may not be home when the phone rings, or others may refuse to answer the questions or mail back the questionnaire. As a consequence, no data is obtained for the nonrespondents in the sample. If the nonrespondents and respondents differ greatly on an issue, then **nonresponse bias** exits. For example, those who choose to answer the question on cell phone usage while driving may have a vested interest in the outcome of the survey—say, parents of teenagers with cell phones, or employees of a company that produces cell phones. Others with no vested interest may have an opinion on the issue but might not take the time to respond. Finally, we caution that you may encounter a biased sample that was intentional, with the sole purpose of misleading the public. Such a researcher would be guilty of **unethical statistical practice**.

**Definition 1.15**

**Selection bias** results when a subset of experimental units in the population have little or no chance of being selected for the sample.

**Definition 1.16**

**Nonresponse bias** is a type of selection bias that results when data on all experimental units in a sample are not obtained.

**Definition 1.17**

Intentionally selecting a biased sample in order to produce misleading statistics is considered **unethical statistical practice.**

We conclude this section with two examples involving actual sampling studies.

**EXAMPLE 1.5**

**Method of Data Collection—Study of a Reinforced Concrete Building**

As part of a cooperative research agreement between the United States and Japan, a full-scale reinforced concrete building was designed and tested under simulated earthquake conditions in Japan. For one part of the study (published in the *Journal of Structural Engineering*), several U.S. design engineers located on the west coast were asked to evaluate the new design. Of the 48 engineers surveyed, 75% believed the shear wall of the structure to be too lightly reinforced.

a. Identify the data-collection method.
b. Identify the target population.
c. Are the sample data representative of the population?

Solution

a. The data-collection method is a survey of 48 U.S. design engineers. Consequently, it is an observational study.
b. Presumably, the researchers are interested in the opinions of all west coast U.S. design engineers on the quality of the reinforced concrete building, not just the 48 engineers who were surveyed. Consequently, the target population is *all* west coast U.S. design engineers.
c. Because the 48 engineers surveyed make up a subset of the target population, they do form a sample. Whether or not the sample is representative of the population is unclear because the journal article provided no detailed information on how the 48 engineers were selected other than they were from the west coast of the U.S. If the engineers were randomly selected from a listing of all west coast design engineers, then the sample is likely to be representative. However, if the 48 engineers all worked at one company on the west coast (a company which may or may not be part of the cooperative research agreement with Japan), then they represent a *convenience sample*—one which may not be representative of all west coast U.S. design engineers. The survey result (75% believed the shear wall of the structure to be too lightly reinforced) may be biased high or low, depending on the affiliation of the west coast company.

**EXAMPLE 1.6**

**Method of Data Collection—Study of a Stacked Menu Displays**

One feature of a user-friendly computer interface is a stacked menu display. Each time a menu item is selected, a submenu is displayed partially over the parent menu, thus creating a series of "stacked" menus. A study (published in the *Special Interest Group on Computer Human Interaction Bulletin*) was designed to determine the effect of stacked menus on computer search time. Suppose 20 experienced on-line video game players were randomly selected from all experienced players attending a video gaming conference. The participants were then randomly assigned to one of two groups, half in the experimental group and half in the control group. Each participant was asked to search a menu-driven software package for a particular item. In the experimental group, the stacked menu format was used; in the control group, only the current menu was displayed. The search times (in minutes) of the two groups were compared.

a. Identify the data-collection method.
b. Are the sample data representative of the target population?

Solution

a. Here, the experimental units are the on-line video game players. Because the researchers controlled which group (stacked menu or current menu group) the experimental units (players) were assigned to, a designed experiment was used to collect the data.
b. The sample of 20 video game players was randomly selected from all experienced video game players who attended the conference. If the target population is all experienced video game players, then the sample is likely to be representative of the population. However, if the target population is, more broadly, all potential computer users, then the sample is likely to be biased. Since experienced on-line video

game players are more familiar with navigating menus and screens than the typical computer user, search times for these experienced users are likely to be low regardless of whether or not stacked menus are shown.

## Applied Exercises

**MTBE**

1.15 *Groundwater contamination in wells. Environmental Science & Technology* (Jan. 2005) published a study of methyl *tert*-butyl ether (MTBE) contamination in 223 New Hampshire wells. The data for the wells is saved in the **MTBE** file. Suppose you want to sample 5 of these wells and conduct a thorough analysis of the water contained in each. Use a random number generator to select a random sample of 5 wells from the 223. List the wells in your sample.

**EARTHQUAKE**

1.16 *Earthquake aftershock magnitudes.* Seismologists use the term *aftershock* to describe the smaller earthquakes that follow a main earthquake. Following a major earthquake in the Los Angeles area, the U.S. Geological Survey recorded information on 2,929 aftershocks. Data on the magnitudes (measured on the Richter scale) for the 2,929 aftershocks are saved in the **EARTHQUAKE** file. Use a random number generator to select a random sample of 30 aftershocks from the **EARTHQUAKE** file. Identify the aftershocks in your sample.

**COGAS**

1.17 *Weekly carbon monoxide data.* Refer to Exercise 1.7 (p. 6) and the World Data Centre for Greenhouse Gases collection of weekly carbon monoxide gas measurements at the Cold Bay, Alaska, weather station. The data for 590 weeks for the years 2000–2002 are saved in the **COGAS** file. Use a random number generator to select a random sample of 15 weeks from the **COGAS** file. Identify the weeks in your sample.

1.18 *CT scanning for lung cancer.* Refer to Exercise 1.13 (p. 8) and the University of South Florida clinical trial of smokers to compare the effectiveness of CT scans with X-rays for detecting lung cancer. (*Today's Tomorrows*, Fall 2002.) Recall that each participating smoker will be randomly assigned to one of two screening methods, CT or chest X-ray, and the age (in years) at which the scanning method first detects a tumor will be determined. One goal of the study is to compare the mean ages when cancer is first detected by the two screening methods. Assuming 120 smokers participate in the trial, use a random number generator to randomly assign 60 smokers to each of the two screening methods.

1.19 *Annual survey of computer crimes.* The Computer Security Institute (CSI) conducts an annual survey of computer crime at United States businesses. CSI sends survey questionnaires to computer security personnel at all U.S. corporations and government agencies. The 2010 CSI survey was sent by post or email to 5,412 firms and 351 organizations responded. Forty-one percent of the respondents admitted unauthorized use of computer systems at their firms during the year. (*CSI Computer Crime and Security Survey, 2010/2011*.)

a. Identify the population of interest to CSI.
b. Identify the data collection method used by CSI. Are there any potential biases in the method used?
c. Describe the variable measured in the CSI survey. Is it quantitative or qualitative?
d. What inference can be made from the study result?

1.20 *Corporate sustainability and firm characteristics. Corporate sustainability* refers to business practices designed around social and environmental considerations (e.g., "going green"). *Business and Society* (March 2011) published a paper on how firm size and firm type impacts sustainability behaviors. The researchers added questions on sustainability to a quarterly survey of Certified Public Accountants (CPAs). The survey was sent to approximately 23,500 senior managers at CPA firms, of which 1,293 senior managers responded. (Note: It is not clear how the 23,500 senior managers were selected.) Due to missing data (incomplete survey answers), only 992 surveys were analyzed. These data were used to infer whether larger firms are more likely to report sustainability policies than smaller firms and whether public firms are more likely to report sustainability policies than private firms.

a. Identify the population of interest to the researchers.
b. What method was used to collect the sample data?
c. Comment on the representativeness of the sample.
d. How will your answer to part c impact the validity of the inferences drawn from the study?

1.21 *Selecting archaeological dig sites.* Archaeologists plan to perform test digs at a location they believe was inhabited several thousand years ago. The site is approximately 10,000 meters long and 5,000 meters wide. They first draw rectangular grids over the area, consisting of lines every 100 meters, creating a total of $100 \cdot 50 = 5{,}000$ intersections (not counting one of the outer boundaries). The plan is to randomly sample 50 intersection points and dig at the sampled intersections. Explain how you could use a random number generator to obtain a random sample of 50 intersections. Develop at least two plans: one that numbers the intersections from 1 to 5,000 prior to selection and another that selects the row and column of each sampled intersection (from the total of 100 rows and 50 columns).

## 1.5 The Role of Statistics in Critical Thinking

Experimental research in engineering and the sciences typically involves the use of experimental data—a sample—to infer the nature of some conceptual population that characterizes a phenomenon of interest to the experimenter. This inferential process is an integral part of the scientific method. Inference based on experimental data is first used to develop a theory about some phenomenon. Then the theory is tested against additional sample data.

How does the science of statistics contribute to this process? To answer this question, we must note that inferences based on sample data will almost always be subject to error, because a sample will not provide an exact image of the population. The nature of the information provided by a sample depends on the particular sample chosen and thus will change from sample to sample. For example, suppose you want to estimate the proportion of all steel alloy failures at U.S. petrochemical plants caused by stress corrosion cracking. You investigate the cause of failure for a sample of 100 steel alloy failures and find that 47 were caused by stress corrosion cracking. Does this mean that exactly 47% of all steel alloy failures at petrochemical plants are caused by stress corrosion cracking? Of course, the answer is "no." Suppose that, unknown to you, the true percentage of steel alloy failures caused by stress corrosion cracking is 44%. One sample of 100 failures might yield 47 that were caused by cracking, whereas another sample of 100 might yield only 42. Thus, an inference based on sampling is always subject to *uncertainty*.

On the other hand, suppose one petrochemical plant experienced a steel alloy failure rate of 81%. Is this an unusually high failure rate, given the sample rate of 47%? The theory of statistics uses *probability* to measure the uncertainty associated with an inference. It enables engineers and scientists to calculate the probabilities of observing specific samples or data measurements, under specific assumptions about the population. These probabilities are used to evaluate the uncertainties associated with sample inferences; for example, we can determine whether the plant's steel alloy failure rate of 81% is unusually high by calculating the chance of observing such a high rate given the sample information.

Thus, a major contribution of statistics is that it enables engineers and scientists to make inferences—estimates and decisions about the target population—with a known measure of reliability. With this ability, an engineer can make intelligent decisions and inferences from data; that is, statistics helps engineers to *think critically* about their results.

**Definition 1.17**

**Statistical thinking** involves applying rational thought and the science of statistics to critically assess data and inferences.

## 1.6 A Guide to Statistical Methods Presented in This Text

Although we present some useful methods for exploring and describing data sets (Chapter 2), the major emphasis in this text and in modern statistics is in the area of inferential statistics. The flowchart in Figure 1.3 (p. 17) is provided as an outline of the chapters in this text and as a guide to selecting the statistical method appropriate for your particular analysis.

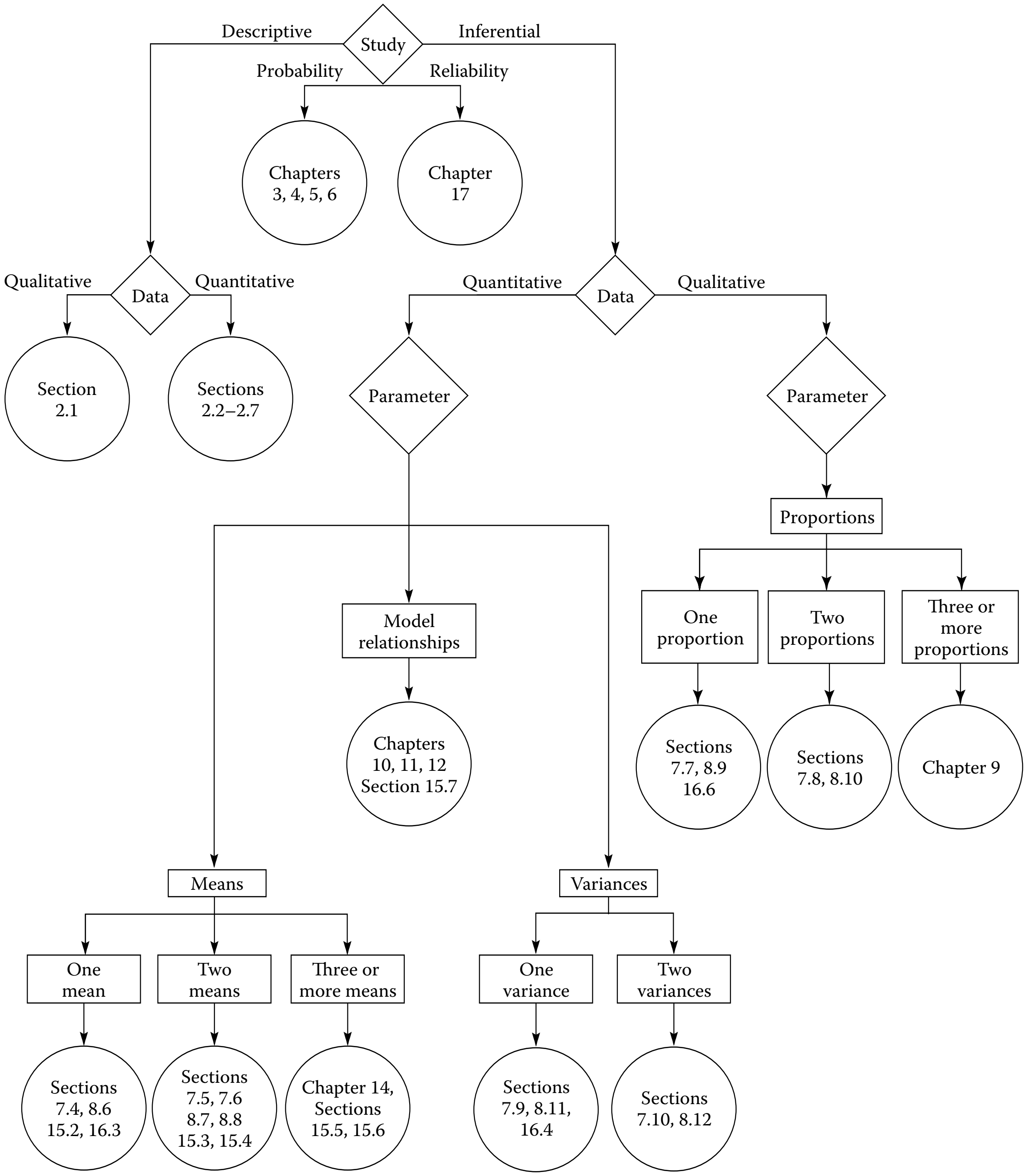



**FIGURE 1.3**
Flowchart of statistical methods described in the text

## STATISTICS IN ACTION REVISITED

### DDT Contamination of Fish in the Tennessee River — Identifying the Data Collection Method, Population, Sample, and Types of Data

We now return to the U.S. Army Corps of Engineers study of the level of DDT contamination of fish in the Tennessee River (Alabama). Recall that the engineers collected fish specimens at different locations along the Tennessee River (TR) and three tributary creeks: Flint Creek (FC), Limestone Creek (LC), and Spring Creek (SC). Consequently, each fish specimen represents the *experimental unit* for this study. Five *variables* were measured for each captured fish: location of capture, species, weight (in grams), length (in centimeters), and DDT concentration (ppm). These data are saved in the **DDT** file. Upon examining the data you will find that capture location is represented by the columns "River" and "Mile". The possible values of "River" are TR, FC, LC, and SC (as described above), while "Mile" gives the distance (in miles) from the mouth of the river or creek. Three species of fish were captured: channel catfish, largemouth bass, and smallmouth buffalofish. Both capture location and species are categorical in nature, hence they are *qualitative* variables. In contrast, weight, length, and DDT concentration are measured on numerical scales; thus, these three variables are *quantitative*.

The data collection method is actually a *designed experiment*, one involving a stratified sample. Why? The Corps of Engineers made sure to collect samples of fish at each of the river and tributary creek locations. These locations represent the different strata for the study. The MINITAB printout shown in Figure SIA1.1 shows the number of fish specimens collected at each river location. You can see that 6 fish were captured at each of the three tributary creeks, and either 6, 8, 10, or 12 fish were captured at various locations (miles upstream) along the Tennessee River, for a total of 144 fish specimens. Of course the data for the 144 captured fish represent a *sample* selected from the much larger *population* of all fish in the Tennessee River and its tributaries.

The U.S. Army Corps of Engineers used the data in the **DDT** file to compare the DDT levels of fish at different locations and among different species, and to determine if any of the quantitative variables (e.g., length and weight) are related to DDT content. In subsequent chapters, we demonstrate several of these analyses.

```
Rows: River  Columns: MILE
       1  3  5  275  280  285  290  295  300  305  310  315  320  325  330  340

FC     0  0  6    0    0    0    0    0    0    0    0    0    0    0    0    0
LC     0  6  0    0    0    0    0    0    0    0    0    0    0    0    0    0
SC     6  0  0    0    0    0    0    0    0    0    0    0    0    0    0    0
TR     0  0  0    6   12   12   12    6   12    6   12    6   12    6    8   10
All    6  6  6    6   12   12   12    6   12    6   12    6   12    6    8   10

     345  All

FC     0    6
LC     0    6
SC     0    6
TR     6  126
All    6  144

Cell Contents:    Count
```

**FIGURE SIA1.1**
MINITAB Output Showing Number of Captured Fish at Each Location

# Quick Review

## Key Terms

| | | | |
|---|---|---|---|
| Data 2 | Measure of reliability 5 | Quantitative data 6 | Selection bias 13 |
| Descriptive statistics 3 | Measurement error 4 | Random number generator 9 | Simple random sample 9 |
| Designed experiment 8 | Nonresponse bias 13 | Reliability 4 | Statistical thinking 16 |
| Experimental unit 4 | Observational study 9 | Representative sample 9 | Statistics 2 |
| Inference 3 | Population 3 | Sample 3 | Survey 8 |
| Inferential statistics 3 | Qualitative data 6 | | Variable 4 |

## Chapter Summary Notes

- Two types of statistical applications: ***descriptive*** and ***inferential***
- Fundamental elements of statistics: ***population, experimental units, variable, sample, inference, measure of reliability***
- ***Descriptive statistics*** involves summarizing and describing data sets.
- ***Inferential statistics*** involves using a sample to make inferences about a population.
- Two types of data: ***quantitative*** and ***qualitative***
- Three data collection methods: **published source, designed experiment, observational study**.
- Types of random sampling: **simple random sample, stratified random sampling, cluster sampling,** and **systematic sampling**.

## Supplementary Exercises

1.22 *Steel anticorrosion study.* Researchers at the Department of Materials Science and Engineering, National Technical University (Athens, Greece), examined the anticorrosive behavior of different epoxy coatings on steel. (*Pigment & Resin Technology*, Vol. 32, 2003.) Flat panels cut from steel sheets were randomly selected from the production line and coated with one of four different types of epoxy (S1, S2, S3, and S4). (Note: The panels were randomly assigned to an epoxy type.) After exposing the panels to water for one day, the corrosion rate (nanoamperes per square centimeter) was determined for each panel.

a. What are the experimental units for the study?
b. What data collection method was used?
c. Suppose you are interested in describing only the corrosion rates of steel panels coated with epoxy type S1. Define the target population and relevant sample.

1.23 *Reliability of a computer system.* The reliability of a computer system is measured in terms of the lifelength of a specified hardware component (e.g., the hard disk drive). To estimate the reliability of a particular system, 100 computer components are tested until they fail, and their lifelengths are recorded.

a. What is the population of interest?
b. What is the sample?
c. Are the data quantitative or qualitative?
d. How could the sample information be used to estimate the reliability of the computer system?

1.24 *Traveling turtle hatchlings.* Hundreds of sea turtle hatchlings, instinctively following the bright lights of condominiums, wandered to their deaths across a coastal highway in Florida (*Tampa Tribune*, Sept. 16, 1990). This incident led researchers to begin experimenting with special low-pressure sodium lights. One night, 60 turtle hatchlings were released on a dark beach and their direction of travel noted. The next night, the special lights were installed and the same 60 hatchlings were released. Finally, on the third night, tar paper was placed over the sodium lights. Consequently, the direction of travel was recorded for each hatchling under three experimental conditions—darkness, sodium lights, and sodium lights covered with tar paper.

a. Identify the population of interest to the researchers.
b. Identify the sample.
c. What type of data were collected, quantitative or qualitative?
d. Identify the data collection method.

1.25 *Acid neutralizer experiment.* A chemical engineer conducts an experiment to determine the amount of hydrochloric acid necessary to neutralize 2 milliliters (ml) of a newly developed cleaning solution. The chemist prepares five 2-ml portions of the solution and adds a known concentration of hydrochloric acid to each. The amount of acid necessary to achieve neutrality of the solution is determined for each of the five portions.

a. Identify the experimental units for the study.
b. Identify the variable measured.
c. Describe the population of interest to the chemical engineer.
d. Describe the sample.

1.26 *Deep hole drilling.* "Deep hole" drilling is a family of drilling processes used when the ratio of hole depth to hole diameter exceeds 10. Successful deep hole drilling depends on the satisfactory discharge of the drill chip. An experiment was conducted to investigate the performance of deep hole drilling when chip congestion exists (*Journal of Engineering for Industry*, May 1993). Some important variables in the drilling process are described here. Identify the data type for each variable.

a. Chip discharge rate (number of chips discarded per minute)
b. Drilling depth (millimeters)
c. Oil velocity (millimeters per second)
d. Type of drilling (single-edge, BTA, or ejector)
e. Quality of hole surface

1.27 *Intellectual development of engineering students.* Perry's model of intellectual development was applied to undergraduate engineering students at Penn State (*Journal of Engineering Education*, Jan. 2005). Perry scores (ranging from 1 to 5) were determined for 21 students in a first-year, project-based design course. (*Note*: A Perry score of 1 indicates the lowest level of intellectual development, and a Perry score of 5 indicates the highest level.) The average Perry score for the 21 students was 3.27.

a. Identify the experimental units for this study.
b. What is the population of interest? The sample?
c. What type of data, quantitative or qualitative, are collected?
d. Use the sample information to make an inference about the population.
e. Use a random number generator to select 3 of the 21 students for further testing.

1.28 *Type of data.* State whether each of the following data sets is quantitative or qualitative.

a. Arrival times of 16 reflected seismic waves
b. Types of computer software used in a database management system
c. Brands of calculator used by 100 engineering students on campus
d. Ash contents in pieces of coal from three different mines
e. Mileages attained by 12 automobiles powered by alcohol
f. Life-lengths of laser printers
g. Shift supervisors in charge of computer operations at an airline company
h. Accident rates at 46 machine shops

1.29 *Structurally deficient bridges.* Refer to Exercise 1.14 (p. 8). The most recent NBI data were analyzed, and the results made available at the FHWA web site (www.fhwa.dot.gov). Using the FHWA inspection ratings, each of the nearly 600,000 highway bridges in the United States was categorized as structurally deficient, functionally obsolete, or safe. About 12% of the bridges were found to be structurally deficient, and 14% were functionally obsolete.

a. What is the variable of interest to the researchers?
b. Is the variable of part a quantitative or qualitative?
c. Is the data set analyzed a population or a sample? Explain.
d. How did the researchers obtain the data for their study?
e. Use a random number generator to determine which bridges to include in a random sample of 25 bridges selected from the 600,000 bridges.

CHAPTER 2

# Descriptive Statistics

**OBJECTIVE**

*To present graphical and numerical methods for exploring, summarizing, and describing data*

**CONTENTS**



*STATISTICS IN ACTION*

Characteristics of Contaminated Fish in the Tennessee River, Alabama

## *STATISTICS IN ACTION*
## Characteristics of Contaminated Fish in the Tennessee River, Alabama

**DDT**

Recall (*Statistics in Action*, Chapter 1, p. 18) that the U.S. Army Corps of Engineers collected data on fish contaminated from the toxic discharges of a chemical plant once located on the banks of the Tennessee River in Alabama. Ecologists fear that contaminated fish migrating from the mouth of the river to a nearby reservoir and wildlife refuge could endanger other wildlife that prey on the fish.

The variables measured for each of the 144 captured fish are: species (channel catfish, largemouth bass, or smallmouth buffalofish), river/creek where captured (Tennessee River, Flint Creek, Limestone Creek, or Spring Creek), weight (in grams), length (in centimeters), and level of DDT contamination (in parts per million). The data are saved in the **DDT** file.

One goal of the study is to describe the characteristics of the captured fish. Some key questions to be answered are: Where (i.e., what river or creek) are the different species most likely to be captured? What is the typical weight and length of the fish? What is the level of DDT contamination of the fish? Does the level of contamination vary by species? These questions can be partially answered by applying the descriptive methods of this chapter. We demonstrate the application in the Statistics in Action Revisited at the end of this chapter.

Assuming you have collected a data set of interest to you, how can you make sense out of it? That is, how can you organize and summarize the data set to make it more comprehensible and meaningful? In this chapter, we look at several basic statistical tools for describing data. These involve graphs and charts that rapidly convey a visual picture of the data, and numerical measures that describe certain features of the data. The proper procedure to use depends on the type of data (quantitative or qualitative) that we want to describe.

## 2.1 Graphical and Numerical Methods for Describing Qualitative Data

Recall from Chapter 1 (see Definition 1.10) that data categorical in nature is called qualitative data. When describing qualitative observations, we define the categories in such a way that each observation can fall in one and only one category (or class). The data set is then described numerically by giving the number of observations, or the proportion of the total number of observations, that fall in each of the categories.

**Definition 2.1**
A **class** is one of the categories into which qualitative data can be classified.

**Definition 2.2**
The **category (or class) frequency** for a given category is the number of observations that fall in that category.

**Definition 2.3**
The **category (or class) relative frequency** for a given category is the proportion of the total number of observations $n$ that fall in that category, i.e.,

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

To illustrate, consider a problem of interest to researchers investigating the safety of nuclear power reactors and the hazards of using energy. The researchers discovered 62 energy-related accidents worldwide since 1979 that resulted in multiple fatalities.

**FATAL**

**TABLE 2.1 Summary Frequency Table for Cause of Energy-Related Fatal Accidents**

| Category *(Cause)* | Frequency *(Number of Accidents)* | Relative Frequency *(Proportion)* |
|---|---|---|
| Coal mine collapse | 9 | .145 |
| Dam failure | 4 | .065 |
| Gas explosion | 40 | .645 |
| Nuclear reactor | 1 | .016 |
| Oil fire | 6 | .097 |
| Other (e.g., Lightning, Power plant) | 2 | .032 |
| **Totals** | 62 | 1.000 |

*Source:* "Safety of nuclear power reactors." *World Nuclear Association*, May 2012.

Table 2.1 summarizes the researcher's findings. In this application, the qualitative variable of interest is the cause of the fatal energy-related accident. You can see from Table 2.1 that the data for the 62 accidents fall into six categories (causes). The summary table gives both the frequency and relative frequency of each cause category. Clearly, a gas explosion is the most frequent cause of an accident, occurring in 40 of the 62 accidents (or approximately 65%). The least likely cause (occuring only 1 time) is a nuclear reactor failure.

Graphical descriptions of qualitative data sets are usually achieved using bar graphs or pie charts; these figures are often constructed using statistical software. **Bar graphs** give the frequency (or relative frequency) corresponding to each category, with the height or length of the bar proportional to the category frequency (or relative frequency). **Pie charts** divide a complete circle (a pie) into slices, one corresponding to each category, with the central angle of the slice proportional to the category relative frequency. Examples of these familiar graphical methods are shown in Figures 2.1 and 2.2.

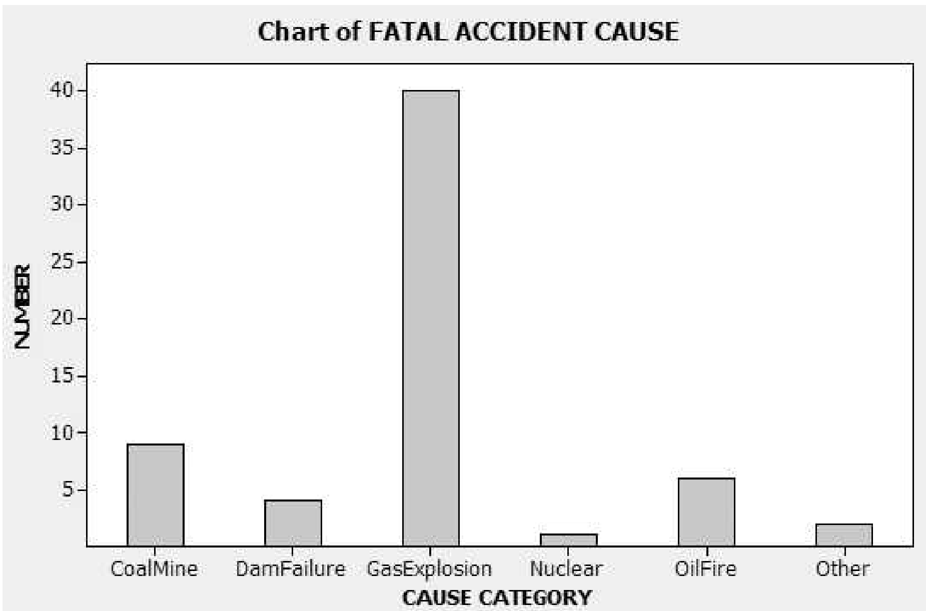



**FIGURE 2.1**
MINITAB Bar Graph for Cause of Energy-Related Fatal Accidents

**FIGURE 2.2**
MINITAB Pie Chart for Cause of Energy-Related Fatal Accidents

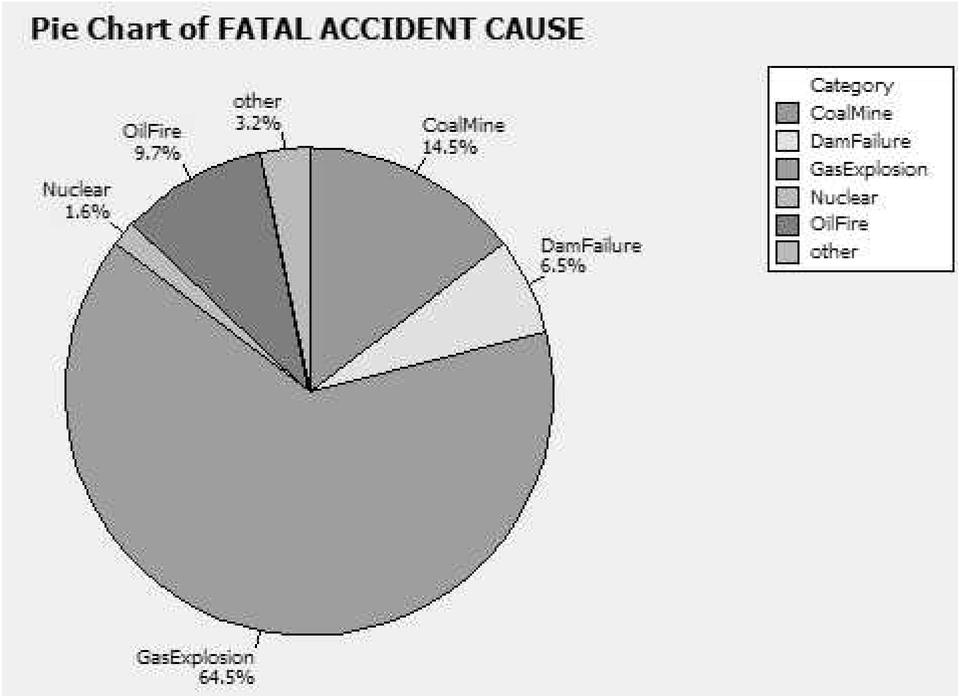


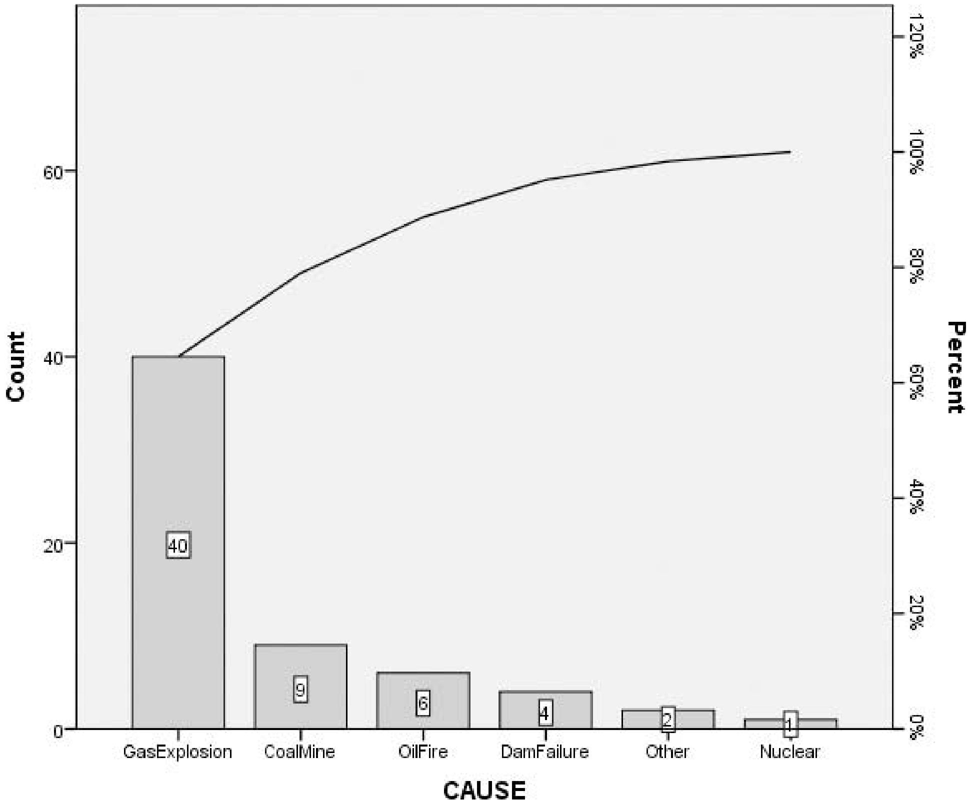

Figure 2.1 is a vertical bar graph produced by MINITAB that describes the data in Table 2.1. (Bar graphs can be vertical or horizontal.) Each bar corresponds to one of the six causes, and the height of the bar is proportional to the number of fatal accidents that fall in that cause category. The height of the vertical bar for Gas Explosion—much larger than all the other categories—highlights this as the most frequent cause of fatal accidents.

Figure 2.2 is a MINITAB pie chart showing the percentages of energy-related fatal accidents associated with the cause categories. A pie chart shows a section of the pie for each category, where the size of the pie slice is proportional to the category relative frequency (percentage). The pie chart not only gives the exact percentage of accidents for each cause but it also provides a rapid visual comparison of the relative frequencies. You can clearly see that gas explosion (64.5%) is the major cause of fatal accidents.

Vertical bar graphs like Figure 2.1 can be enhanced by arranging the bars on the graph in the form of a **Pareto diagram**. A Pareto diagram (named for the Italian economist Vilfredo Pareto) is a frequency bar graph with the bars displayed in order of height, starting with the tallest bar on the left. Pareto diagrams are popular graphical tools in process and quality control, where the heights of the bars often represent frequencies of problems (e.g., defects, accidents, breakdowns, and failures) in the production process. Because the bars are arranged in descending order of height, it is easy to identify the areas with the most severe problems.

An SPSS Pareto diagram for the energy-related accident data summarized in Table 2.1 is displayed in Figure 2.3. Since the relative frequencies associated with the six cause categories are arranged in decreasing order, it is easy to identify the cause (gas explosion) of the most accidents and the cause (nuclear reactor) of the least accidents. In addition to the bars with decreasing heights, the Pareto diagram also shows a plot of the cumulative proportion of accidents (called a “cum” line) superimposed over the bars. The cum line scale appears on the right side of the Pareto diagram in Figure 2.3.

**FIGURE 2.3**
SPSS Pareto Diagram for Cause of Energy-Related Fatal Accidents




**Example 2.1**

Graphing Qualitative Data Characteristics of Ice Meltponds

**PONDICE**

The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice meltponds in the Canadian Arctic. Environmental engineers at the University of Colorado are using these data to study how climate impacts the sea ice. Data for 504 ice meltponds located in the Barrow Strait in the Canadian Arctic are saved in the PONDICE file. One variable of interest is the type of ice observed for each pond. Ice type is classified as first-year ice, multiyear ice, or landfast ice. Construct a summary table and a horizontal bar graph to describe the ice types of the 504 meltponds. Interpret the results.

Solution

The data in the **PONDICE** file were analyzed using SAS. Figure 2.4 shows a SAS summary table for the three ice types. Of the 504 meltponds, 88 had first-year ice, 220 had multiyear ice, and 196 had landfast ice. The corresponding proportions (or relative frequencies) are $88/504 = .175$, $220/504 = .437$, and $196/504 = .389$. These proportions are shown in the "Percent" column in the table and in the accompanying SAS horizontal bar graph in Figure 2.4. The University of Colorado researchers used this information to estimate that about 17% of meltponds in the Canadian Arctic have first-year ice.

### Summary of Graphical Descriptive Methods for Qualitative Data

**Bar Graph:** The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.

**Pie Chart:** The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the class relative frequency.

**Pareto Diagram:** A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged by height in descending order from left to right.

The FREQ Procedure

| ICETYPE | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| First-year | 88 | 17.46 | 88 | 17.46 |
| Landfast | 196 | 38.89 | 284 | 56.35 |
| Multi-year | 220 | 43.65 | 504 | 100.00 |

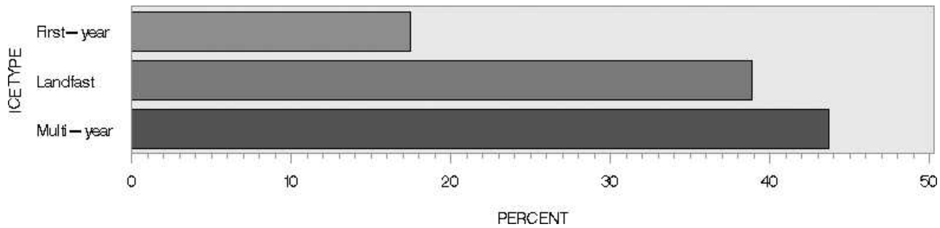
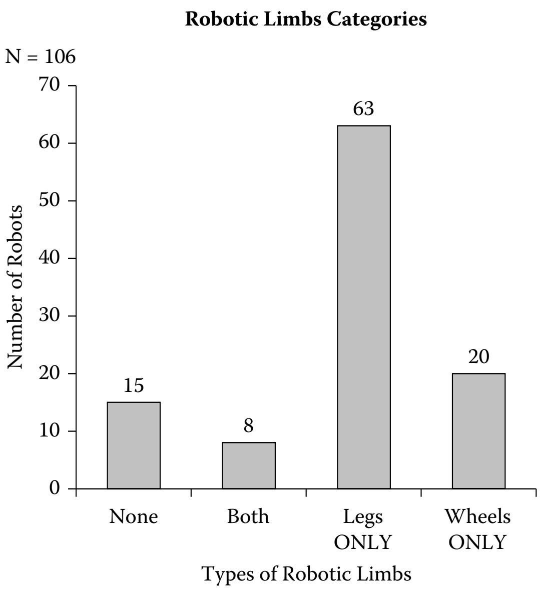



**FIGURE 2.4**
SAS analysis of ice types for meltponds

## Applied Exercises

2.1 *Do social robots walk or roll?* According to the United Nations, social robots now outnumber industrial robots worldwide. A social (or service) robot is designed to entertain, educate, and care for human users. In a paper published by the *International Conference on Social Robotics* (Vol. 6414, 2010), design engineers investigated the trend in the design of social robots. Using a random sample of 106 social robots obtained through a web search, the engineers found that 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels. This information is portrayed in the accompanying figure.

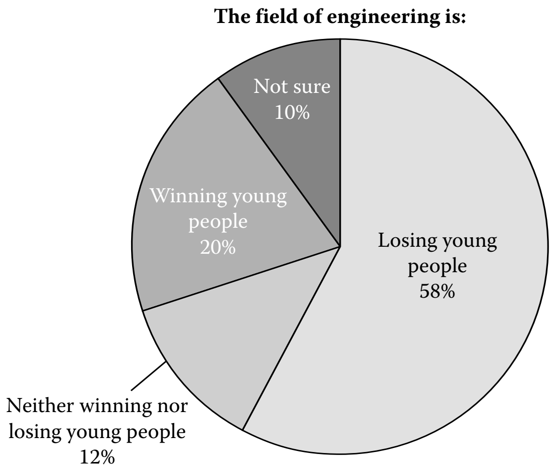
a. What type of graph is used to describe the data?
b. Identify the variable measured for each of the 106 robot designs.
c. Use graph to identify the social robot design that is currently used the most.
d. Compute class relative frequencies for the different categories shown in the graph.
e. Use the results, part d, to construct a Pareto diagram for the data




Graph for Exercise 2.1

2.2 *Americans' view of engineering.* Duke University's Pratt School of Engineering commissioned a survey on Americans' attitudes toward engineering. A telephone survey was conducted among a representative national sample of 808 American adults in January, 2009. One of the survey questions asked, "Do you believe the field of engineering is winning or losing young people?" The results are summarized in the pie chart on pg. 27.

a. What variable is described in the pie chart? What are the categories (classes)?
b. Explain what the "20%" represents in the chart.
c. Convert the pie chart into a Pareto diagram.
d. Based on the graphs, what is the majority opinion of American adults responding to the survey question?



*Source:* Jan 2009–Hart Research for Pratt School/Duke University

2.3 *STEM experiences for girls.* The National Science Foundation (NSF) sponsored a study on girls participation in informal science, technology, engineering or mathematics (STEM) programs (see Exercise 1.1). The results of the study were published in *Cascading Influences: Long-Term Impacts of Informal STEM Experiences for Girls* (March 2013). The researchers sampled 174 young women who recently participated in a STEM program. They used a pie chart to describe the geographic location (urban, suburban, or rural) of the STEM programs attended. Of the 174 STEM participants, 107 were in urban areas, 57 in suburban areas, and 10 in rural areas. Use this information to construct the pie chart. Interpret the results.

2.4 *Microsoft program security issues.* The dominance of Microsoft in the computer software market has led to numerous malicious attacks (e.g., worms, viruses) on its programs. To help its users combat these problems, Microsoft periodically issues a Security Bulletin that reports the software affected by the vulnerability. In *Computers & Security* (July 2013), researchers focused on reported security issues with three Microsoft products: Office, Windows, and Explorer. In a sample of 50 security bulletins issued in 2012, 32 reported a security issue with Windows, 6 with Explorer, and 12 with Office. The researchers also categorized the security bulletins according to the expected repercussion of the vulnerability. Categories were Denial of service, Information disclosure, Remote code execution, Spoofing, and Privilege elevation. Suppose that of the 50 bulletins sampled, the following numbers of bulletins were classified into each respective category: 6, 8, 22, 3, 11.

a. Construct a pie chart to describe the Microsoft products with security issues. Which product had the lowest proportion of security issues in 2012?
b. Construct a Pareto diagram to describe the expected repercussions from security issues. Based on the graph, what repercussion would you advise Microsoft to focus on?

2.5 *Beach erosional hotspots.* Beaches that exhibit high erosion rates relative to the surrounding beach are defined as *erosional hotspots*. The U.S. Army Corps of Engineers conducted a study of beach hotspots using an online questionnaire. Information on six beach hotspots was collected. Some of the data are listed in the table.

a. Identify each variable recorded as quantitative or qualitative.
b. Form a pie chart for the beach condition of the six hotspots.
c. Form a pie chart for the nearshore bar condition of the six hotspots.
d. Comment on the reliability of using the pie charts to make inferences about all beach hotspots in the country.

| Beach Hotspot | Beach Condition | Nearshore Bar Condition | Long-Term Erosion Rate (miles/year) |
|---|---|---|---|
| Miami Beach, FL | No dunes/flat | Single, shore parallel | 4 |
| Coney Island, NY | No dunes/flat | Other | 13 |
| Surfside, CA | Bluff/scarp | Single, shore parallel | 35 |
| Monmouth Beach, NJ | Single dune | Planar | Not estimated |
| Ocean City, NJ | Single dune | Other | Not estimated |
| Spring Lake, NJ | Not observed | Planar | 14 |

*Source:* "Identification and characterization of erosional hotspots." William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March 18, 2002.

2.6 *Management system failures.* The U.S. Chemical Safety and Hazard Investigation Board (CSB) is responsible for determining the root cause of industrial accidents. Since its creation in 1998, the CSB has identified 83 incidents that were caused by management system failures. (*Process Safety Progress*, Dec. 2004.) The accompanying table gives a breakdown of the root causes of these 83 incidents. Construct a Pareto diagram for the data and interpret the graph.

| Management System Cause Category | Number of Incidents |
|---|---|
| Engineering & Design | 27 |
| Procedures & Practices | 24 |
| Management & Oversight | 22 |
| Training & Communication | 10 |
| **Total** | 83 |

*Source:* Blair, A. S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board." *Process Safety Progress*, Vol. 23, No. 4, Dec. 2004 (Table 1).

2.7 *Satellites in orbit.* According to the Union of Concerned Scientists (www.ucsusa.org), as of November 2012, there were 502 low Earth orbit (LEO) and 432 geosynchronous orbit (GEO) satellites in space. Each satellite is owned by an entity in either the government, military, commercial, or civil sector. A breakdown of the number of satellites in orbit for each sector is displayed in the accompanying table. Use this information to construct a pair of graphs that compare the ownership sectors of LEO and GEO satellites in orbit. What observations do you have about the data?

| LEO Satellites | | GEO Satellites | |
|---|---|---|---|
| Government | — 229 | Government | — 59 |
| Military | — 109 | Military | — 91 |
| Commercial | — 118 | Commercial | — 281 |
| Civil | — 46 | Civil | — 1 |
| **Total** | — 502 | **Total** | — 432 |

2.8 *Railway track allocation.* One of the problems faced by transportation engineers is the assignment of tracks to trains at a busy railroad station. Overused and/or underused tracks cause increases in maintenance costs and inefficient allocation of resources. The *Journal of Transportation Engineering* (May, 2013) investigated the optimization of track allocation at a Chinese railway station with 11 tracks. Using a new algorithm designed to minimize waiting time and bottlenecks, engineers assigned tracks to 53 trains in a single day as shown in the accompanying table. Construct a Pareto diagram for the data. Use the diagram to help the engineers determine if the allocation of tracks to trains is evenly distributed, and, if not, which tracks are underutilized and overutilized.

| Track Assigned | Number of Trains |
|---|---|
| Track #1 | 3 |
| Track #2 | 4 |
| Track #3 | 4 |
| Track #4 | 4 |
| Track #5 | 7 |
| Track #6 | 5 |
| Track #7 | 5 |
| Track #8 | 7 |
| Track #9 | 4 |
| Track #10 | 5 |
| Track #11 | 5 |
| **Total** | 53 |

*Source:* Wu, J., et al. "Track allocation optimization at a railway station: Mean-Variance model and case study", *Journal of Transportation Engineering,* Vol. 39, No. 5, May 2013 (extracted from Table 4).

2.9 *Benford's Law of Numbers.* According to *Benford's Law*, certain digits (1, 2, 3, . . . , 9) are more likely to occur as the first significant digit in a randomly selected number than other digits. For example, the law predicts that the number 1 is the most likely to occur (30% of the time) as the first digit. In a study reported in the *American Scientist* (July–Aug. 1998) to test Benford's Law, 743 first-year college students were asked to write down a six-digit number at random. The first significant digit of each number was recorded and its distribution summarized in the following table.

**DIGITS**

| First Digit | Number of Occurrences |
|---|---|
| 1 | 109 |
| 2 | 75 |
| 3 | 77 |
| 4 | 99 |
| 5 | 72 |
| 6 | 117 |
| 7 | 89 |
| 8 | 62 |
| 9 | 43 |
| **Total** | 743 |

*Source:* Hill, T.P. "The first digit phenomenon." *American Scientist*, Vol. 86, No. 4, July–Aug. 1998, p. 363 (Figure 5).

a. Describe the first digit of the "random guess" data with a Pareto diagram.
b. Does the graph support Benford's Law? Explain.

**SWDEFECTS**

2.10 *Software defects.* The PROMISE Software Engineering Repository is a collection of data sets available to serve researchers in building predictive software models. One such data set, saved in the **SWDEFECTS** file, contains information on 498 modules of software code. Each module was analyzed for defects and classified as "true" if it contained defective code and "false" if not. Access the data file and produce a pie chart for the defect variable. Use the pie chart to make a statement about the likelihood of defective software code.

**NZBIRDS**

2.11 *Extinct New Zealand birds.* Refer to the *Evolutionary Ecology Research* (July, 2003) study of the patterns of extinction in the New Zealand bird population, Exercise 1.10 (p. 6). Data on flight capability (volant or flightless), habitat (aquatic, ground terrestrial, or aerial terrestrial), nesting site (ground, cavity within ground, tree, cavity above ground), nest density (high or low), diet (fish, vertebrates, vegetables, or invertebrates), body mass (grams), egg length (millimeters), and extinct status (extinct, absent from island, present) for 132 bird species at the time of the Maori colonization of New Zealand are saved in the

**NZBIRDS** file. Use a graphical method to investigate the theory that extinct status is related to flight capability, habitat, and nest density.

2.12 *Groundwater contamination in wells.* In New Hampshire, about half the counties mandate the use of reformulated gasoline. This has led to an increase in the contamination of groundwater with methyl *tert*-butyl ether (MTBE). *Environmental Science & Technology* (Jan. 2005) reported on the factors related to MTBE contamination in public and private New Hampshire wells. Data were collected for a sample of 223 wells. These data are saved in the MTBE file. Three of the variables are qualitative in nature: well class (public or private), aquifer (bedrock or unconsolidated), and detectable level of MTBE (below limit or detect). (*Note:* A detectable level of MTBE occurs if the MTBE value exceeds .2 micrograms per liter.) The data for 10 selected wells are shown in the accompanying table. Use graphical methods to describe each of the three qualitative variables for all 223 wells.

**MTBE** (10 selected observations from 223)

| Well Class | Aquifier | Detect MTBE |
|---|---|---|
| Private | Bedrock | Below limit |
| Private | Bedrock | Below limit |
| Public | Unconsolidated | Detect |
| Public | Unconsolidated | Below limit |
| Public | Unconsolidated | Below limit |
| Public | Unconsolidated | Below limit |
| Public | Unconsolidated | Detect |
| Public | Unconsolidated | Below limit |
| Public | Unconsolidated | Below limit |
| Public | Bedrock | Detect |
| Public | Bedrock | Detect |

*Source:* Ayotte, J.D., Argue, D.M., and McGarry, F.J., "Methyl *tert*-butyl ether occurrence and related factors in public and private wells in southeast New Hampshire." *Environmental Science & Technology*, Vol. 39, No. 1, Jan. 2005.

## 2.2 Graphical Methods for Describing Quantitative Data

Recall from Section 1.3 that quantitative data sets consist of data that are recorded on a meaningful numerical scale. For describing, summarizing, and detecting patterns in such data, we can use three graphical methods: *dot plots, stem-and-leaf displays*, and *histograms*. Since most statistical software packages can be used to construct these displays, we'll focus here on their interpretation rather than their construction.

For example, the Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. Suppose that the 100 measurements in Table 2.2 represent the results (miles per gallon) of such tests on a certain new car model. How can we summarize the information in this rather large sample?

A visual inspection of the data indicates some obvious facts. For example, most of the mileages are in the 30s, with a smaller fraction in the 40s. But it is difficult to provide much additional information on the 100 mileage ratings without resorting to some method of summarizing the data. One such method is a dot plot.

**EPAGAS**

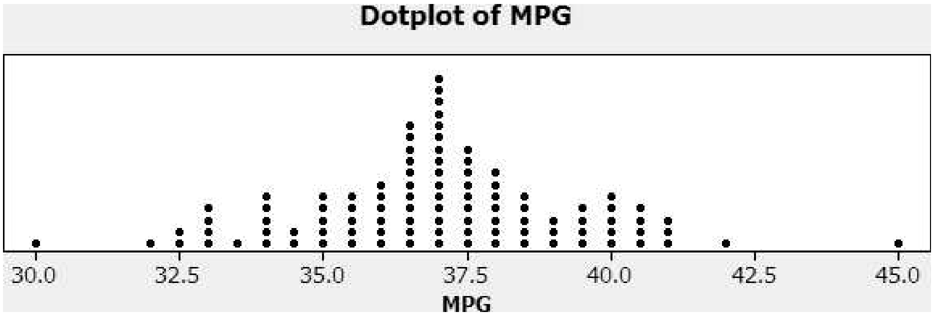
**TABLE 2.2 EPA Mileage Ratings on 100 Cars**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 36.3 | 41.0 | 36.9 | 37.1 | 44.9 | 36.8 | 30.0 | 37.2 | 42.1 | 36.7 |
| 32.7 | 37.3 | 41.2 | 36.6 | 32.9 | 36.5 | 33.2 | 37.4 | 37.5 | 33.6 |
| 40.5 | 36.5 | 37.6 | 33.9 | 40.2 | 36.4 | 37.7 | 37.7 | 40.0 | 34.2 |
| 36.2 | 37.9 | 36.0 | 37.9 | 35.9 | 38.2 | 38.3 | 35.7 | 35.6 | 35.1 |
| 38.5 | 39.0 | 35.5 | 34.8 | 38.6 | 39.4 | 35.3 | 34.4 | 38.8 | 39.7 |
| 36.3 | 36.8 | 32.5 | 36.4 | 40.5 | 36.6 | 36.1 | 38.2 | 38.4 | 39.3 |
| 41.0 | 31.8 | 37.3 | 33.1 | 37.0 | 37.6 | 37.0 | 38.7 | 39.0 | 35.8 |
| 37.0 | 37.2 | 40.7 | 37.4 | 37.1 | 37.8 | 35.9 | 35.6 | 36.7 | 34.5 |
| 37.1 | 40.3 | 36.7 | 37.0 | 33.9 | 40.1 | 38.0 | 35.2 | 34.8 | 39.5 |
| 39.9 | 36.9 | 32.9 | 33.8 | 39.8 | 34.0 | 36.8 | 35.0 | 38.1 | 36.9 |

## Dot Plots

A MINITAB **dot plot** for the 100 EPA mileage ratings is shown in Figure 2.5. The horizontal axis of Figure 2.5 is a scale for the quantitative variable in miles per gallon (mpg). The rounded (to the nearest half gallon) numerical value of each measurement in the data set is located on the horizontal scale by a dot. When data values repeat, the dots are placed above one another, forming a pile at that particular numerical location. As you can see, this dot plot verifies that almost all of the mileage ratings are in the 30s, with most falling between 35 and 40 miles per gallon.

**FIGURE 2.5**
MINITAB dot plot for 100 EPA mileage ratings




## Stem-and-Leaf Display

Another graphical representation of these same data, a MINITAB **stem-and-leaf display**, is shown in Figure 2.6. In this display the *stem* is the portion of the measurement (mpg) to the left of the decimal point, and the remaining portion to the right of the decimal point is the *leaf*.

In Figure 2.6, the stems for the data set are listed in the second column from the smallest (30) to the largest (44). Then the leaf for each observation is listed to the right in the row of the display corresponding to the observation's stem.* For example, the

**FIGURE 2.6**
MINITAB stem-and-leaf display for 100 mileage ratings

**Stem-and-Leaf Display: MPG**

```
Stem-and-leaf of MPG  N  = 100
Leaf Unit = 0.10


  1    30  0
  2    31  8
  6    32  5799
 12    33  126899
 18    34  024588
 29    35  01235667899
 49    36  01233445566777888999
(21)   37  000011122334456677899
 30    38  0122345678
 20    39  00345789
 12    40  0123557
  5    41  002
  2    42  1
  1    43
  1    44  9
```

* The first column of the MINITAB stem-and-leaf display represents the cumulative number of measurements from the class interval to the nearest extreme class interval.

leaf 3 of the first observation (36.3) in Table 2.2 appears in the row corresponding to the stem 36. Similarly, the leaf 7 for the second observation (32.7) in Table 2.2 appears in the row corresponding to the stem 32, and the leaf 5 for the third observation (40.5) appears in the row corresponding to the stem 40. (The stems and leaves for these first three observations are highlighted in Figure 2.6.) Typically, the leaves in each row are ordered as shown in the MINITAB stem-and-leaf display.

The stem-and-leaf display presents another compact picture of the data set. You can see at a glance that the 100 mileage readings were distributed between 30.0 and 44.9, with most of them falling in stem rows 35 to 39. The six leaves in stem row 34 indicate that six of the 100 readings were at least 34.0 but less than 35.0. Similarly, the eleven leaves in stem row 35 indicate that eleven of the 100 readings were at least 35.0 but less than 36.0. Only five cars had readings equal to 41 or larger, and only one was as low as 30.

### Steps to Follow in Constructing a Stem-and-Leaf Display

*Step 1* Divide each observation in the data set into two parts, the **stem** and the **leaf**. For example, the stem and leaf of the mileage 31.8 are 31 and 8, respectively:

| Stem | Leaf |
|---|---|
| 31 | 8 |

*Step 2* List the stems in order in a column, starting with the smallest stem and ending with the largest.

*Step 3* Proceed through the data set, placing the leaf for each observation in the appropriate stem row. Arbitrarily, you may want to arrange the leaves in each row in ascending order.

## Histograms

An SPSS **histogram** for these 100 EPA mileage readings is shown in Figure 2.7. The horizontal axis of Figure 2.7, which gives the miles per gallon for a given automobile, is divided into **class intervals** commencing with the interval from 30–31 and proceeding in intervals of equal size to 44–45 mpg. The vertical axis gives the number (or *frequency*) of the 100 readings that fall in each interval. It appears that about 21 of the 100 cars, or 21%, obtained a mileage between 37 and 38 mpg. This class interval contains the highest frequency, and the intervals tend to contain a smaller number of the measurements as the mileages get smaller or larger.

Histograms can be used to display either the frequency or relative frequency of the measurements falling into the class intervals. The class intervals, frequencies, and relative frequencies for the EPA car mileage data are shown in the summary table, Table 2.3.*

By summing the relative frequencies in the intervals 35–36, 36–37, 37–38 and 38–39, you can see that 65% of the mileages are between 35.0 and 39.0. Similarly, only 2% of the cars obtained a mileage rating over 42.0. Many other summary statements can be made by further study of the histogram and accompanying summary table. Note that the sum of all class frequencies will always equal the sample size, $n$.

*SPSS, like many software packages, will classify an observation that falls on the borderline of a class interval into the next highest interval. For example, the gas mileage of 37.0, which falls on the border between the class intervals 36–37 and 37–38, is classified into the 37–38 class. The frequencies in Table 2.3 reflect this convention.

**FIGURE 2.7**
SPSS Histogram for EPA Gas Mileage Ratings

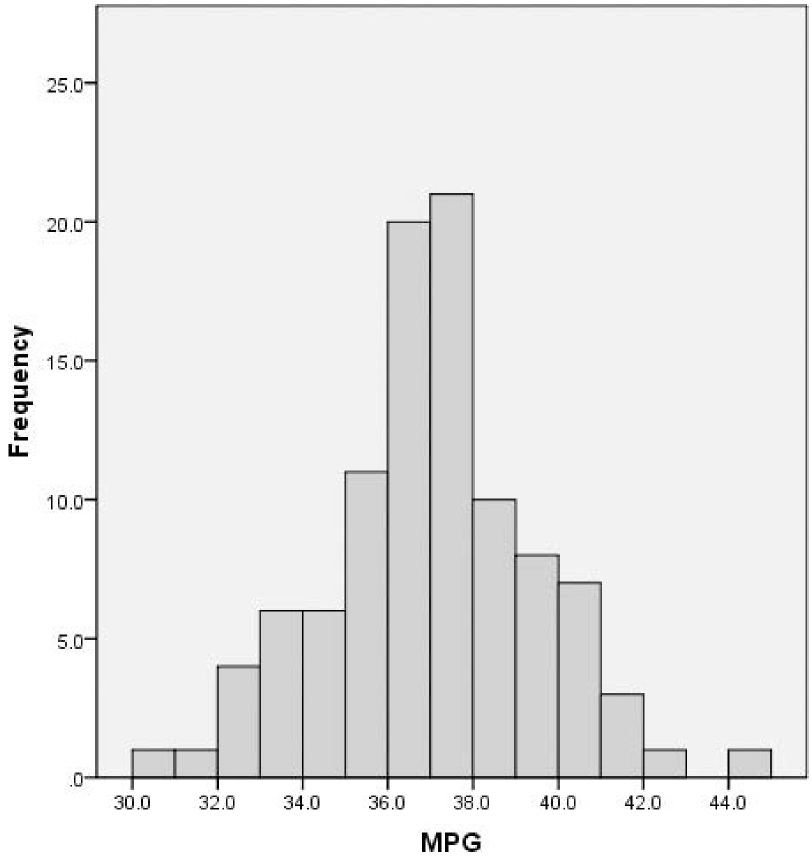



**TABLE 2.3 Class Intervals, Frequencies, and Relative Frequencies for the Car Mileage Data**

| Class Interval | Frequency | Relative Frequency |
|---|---|---|
| 30–31 | 1 | .01 |
| 31–32 | 1 | .01 |
| 32–33 | 4 | .04 |
| 33–34 | 6 | .06 |
| 34–35 | 6 | .06 |
| 35–36 | 11 | .11 |
| 36–37 | 20 | .20 |
| 37–38 | 21 | .21 |
| 38–39 | 10 | .10 |
| 39–40 | 8 | .08 |
| 40–41 | 7 | .07 |
| 41–42 | 3 | .03 |
| 42–43 | 1 | .01 |
| 43–44 | 0 | .00 |
| 44–45 | 1 | .01 |
| **Totals** | 100 | 1.00 |

Some recommendations for selecting the number of intervals in a histogram for smaller data sets are given in the following box.

### Determining the Number of Classes in a Histogram

| Number of Observations in Data Set | Number of Classes |
|---|---|
| Less than 25 | 5–6 |
| 25–50 | 7–10 |
| More than 50 | 11–15 |

Although histograms provide good visual descriptions of data sets—particularly very large ones—they do not let us identify individual measurements. In contrast, each of the original measurements is visible to some extent in a dot plot and clearly visible in a stem-and-leaf display. The stem-and-leaf display arranges the data in ascending order, so it's easy to locate the individual measurements. For example, in Figure 2.6 we can easily see that two of the gas mileage measurements are equal to 36.3, but can't see that fact by inspecting the histogram in Figure 2.7. However, stem-and-leaf displays can become unwieldy for very large data sets. A very large number of stems and leaves causes the vertical and horizontal dimensions of the display to become cumbersome, diminishing the usefulness of the visual display.

### Steps to Follow in Constructing a Histogram

*Step 1* Calculate the range of the data:

$$\text{Range} = \text{Largest observation} - \text{Smallest observation}$$

*Step 2* Divide the range into between 5 and 15 classes of equal width. The number of classes is arbitrary, but you will obtain a better graphical description if you use a small number of classes for a small amount of data and a larger number of classes for larger data sets (see the rule of thumb in the previous box). The lowest (or first) class boundary should be located below the smallest measurement, and the class width should be chosen so that no observation can fall on a class boundary.

*Step 3* For each class, count the number of observations that fall in that class. This number is called the **class frequency**.

*Step 4* Calculate each **class relative frequency:**

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of measurements}}$$

*Step 5* The **histogram** is essentially a bar graph in which the categories are classes. In a **frequency histogram**, the heights of the bars are determined by the class frequency. Similarly, in a **relative frequency histogram**, the heights of the bars are determined by the class relative frequency.

**Example 2.2**

Graphing a Quantitative Variable-Iron Content

**IRONORE**

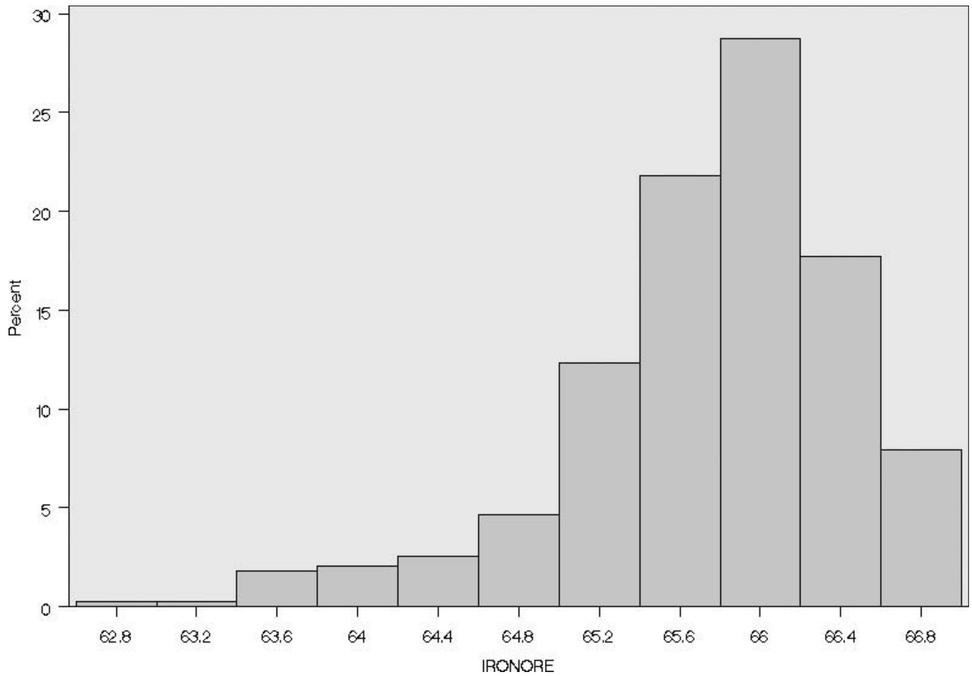
The **IRONORE** file contains data on the percentage iron content for 390 iron-ore specimens collected in Japan. Figure 2.8 is a relative frequency histogram for the 390 iron-ore measurements produced using SAS.

a. Interpret the graph.
b. Visually estimate the fraction of iron-ore measurements that lie between 64.6 and 65.8.

Solution

a. Note that the classes are marked off in intervals of .4 along the horizontal axis of the SAS histogram in Figure 2.8, with the midpoint (rather than the lower and upper boundaries) of each interval shown. The histogram shows that the percentage iron-ore measurements tend to pile up near 66; that is, the class from 65.8 to 66.2 has the greatest relative frequency.
b. The bars that fall in the interval from 64.6 to 65.8 are shaded in Figure 2.8. This shaded portion represents approximately 40% of the total area of the bars for the complete distribution. Thus, about 40% of the 390 iron-ore measurements lie between 64.6 and 65.8.

**FIGURE 2.8**
SAS histogram for iron-ore data




### Interpreting a Relative Frequency Distribution

The percentage of the total number of measurements falling within a particular interval is proportional to the area of the bar that is constructed above the interval. For example, if 30% of the area under the distribution lies over a particular interval, then 30% of the observations fall in that interval.

Most statistical software packages can be used to generate histograms, stem-and-leaf displays, and dot plots. All three are useful tools for graphically describing data sets. We recommend that you generate and compare the displays whenever you can. You'll find that histograms are generally more useful for very large data sets, while stem-and-leaf displays and dot plots provide useful detail for smaller data sets.

### Summary of Graphical Descriptive Methods for Quantitative Data

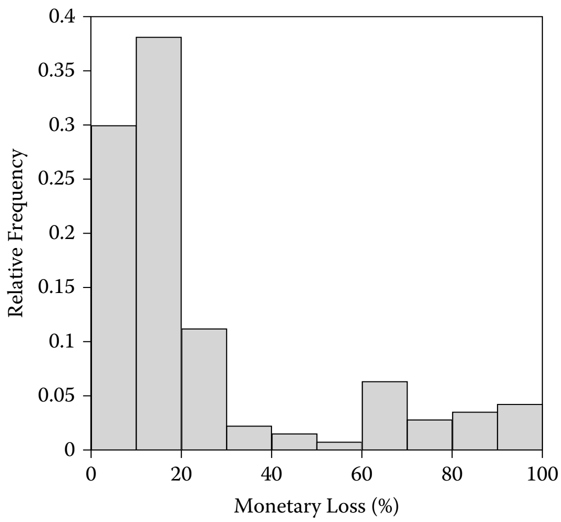
**Dot Plot:** The numerical value of each quantitative measurement in the data set is represented by a dot on a horizontal scale. When data values repeat, the dots are placed above one another vertically.

**Stem-and-Leaf Display:** The numerical value of the quantitative variable is partitioned into a "stem" and a "leaf." The possible stems are listed in order in a column. The leaf for each quantitative measurement in the data set is placed in the corresponding stem row. Leaves for observations with the same stem value are listed in increasing order horizontally.

**Histogram:** The possible numerical values of the quantitative variable are partitioned into class intervals, where each interval has the same width. These intervals form the scale of the horizontal axis. The frequency or relative frequency of observations in each class interval is determined. A vertical bar is placed over each class interval with height equal to either the class frequency or class relative frequency.

## Applied Exercises

2.13 *Annual survey of computer crimes.* Refer to the 2010 CSI *Computer Crime and Security Survey,* Exercise 1.19 (p. 15). Recall that 351 organizations responded to the survey on unauthorized use of computer systems. One of the survey questions asked respondents to indicate the percentage of monetary losses attributable to malicious actions by individuals within the organization (i.e., malicious insider actions). The following histogram summarizes the data for the 144 firms who experienced some monetary loss due to malicious insider actions.




a. Which measurement class contains the highest proportion of respondents?
b. What is the approximate proportion of the 144 organizations that reported a percentage monetary loss from malicious insider actions less than 20%?
c. What is the approximate proportion of the 144 organizations that reported a percentage monetary loss from malicious insider actions greater than 60%?
d. About how many of the 144 organizations reported a percentage monetary loss from malicious insider actions between 20% and 30%?

2.14 *Cheek teeth of extinct primates.* The characteristics of cheek teeth (e.g., molars) can provide anthropologists with information on the dietary habits of extinct mammals. The cheek teeth of an extinct primate species was the subject of research reported in the *American Journal of Physical Anthropology* (Vol. 142, 2010). A total of 18 cheek teeth extracted from skulls discovered in western Wyoming were analyzed. Each tooth was classified according to degree of wear (unworn, slight, light-moderate, moderate, moderate-heavy, or heavy). In addition, the researchers recorded the dentary depth of molars (in millimeters) for each tooth. These depth measurements are listed in the table on page 36.

a. Summarize the data graphically with a dot plot.
b. Summarize the data graphically with a stem-and-leaf display.
c. Is there a particular molar depth that occurs more frequently in the sample? If so, identify the value.

CHEEKTEETH

**Data for Exercise 2.14**

| | |
|---|---|
| 18.12 | 16.55 |
| 19.48 | 15.70 |
| 19.36 | 17.83 |
| 15.94 | 13.25 |
| 15.83 | 16.12 |
| 19.70 | 18.13 |
| 15.76 | 14.02 |
| 17.00 | 14.04 |
| 13.96 | 16.20 |

*Source:* Boyer, D.M., Evans, A.R., and Jernvall, J. "Evidence of Dietary Differentiation Among Late Paleocene-Early Eocene Plesiadapids (Mammalia, Primates)", *American Journal of Physical Anthropology*, Vol. 142, 2010. (Table A3.)

2.15 *Radioactive lichen.* Lichen has a high absorbance capacity for radiation fallout from nuclear accidents. Since lichen is a major food source for Alaskan caribou, and caribou are, in turn, a major food source for many Alaskan villagers, it is important to monitor the level of radioactivity in lichen. Researchers at the University of Alaska, Fairbanks, collected data on nine lichen specimens at various locations for this purpose. The amount of the radioactive element, cesium-137, was measured (in microcuries per milliliter) for each specimen. The data values, converted to logarithms, are given in the table. (Note, the closer the value is to zero, the greater the amount of cesium in the specimen.)

LICHEN

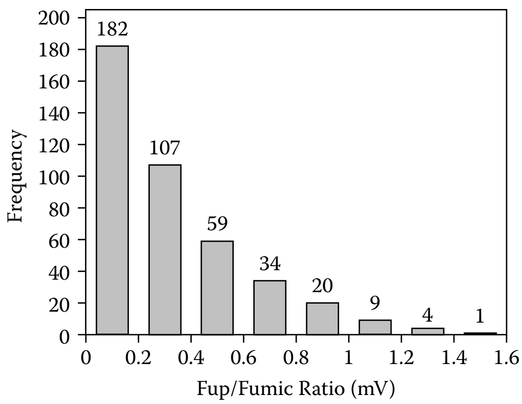
| Location | | | |
|---|---|---|---|
| Bethel | –5.50 | –5.00 | |
| Eagle Summit | –4.15 | –4.85 | |
| Moose Pass | –6.05 | | |
| Turnagain Pass | –5.00 | | |
| Wickersham Dome | –4.10 | –4.50 | –4.60 |

*Source:* Lichen Radionuclide Baseline Research Project, 2003.

a. Construct a dot plot for the nine measurements.
b. Construct a stem-and-leaf display for the nine measurements.
c. Construct a histogram plot for the nine measurements.
d. Which of the three graphs, parts a–c, is more informative about where most of the measurements lie?
e. What proportion of the measurements has a radioactivity level of –5.00 or lower?

2.16 *Stability of compounds in new drugs.* Testing the metabolic stability of compounds used in drugs is the cornerstone of new drug discovery. Two important values computed from the testing phase are the fraction of compound unbound to plasma (*fup*) and the fraction of compound unbound to microsomes (*fumic*). A key formula for assessing stability assumes that the fup/fumic ratio is 1. Pharmacologists at Pfizer Global Research and Development investigated this phenomenon and reported the results in *ACS Medicinal Chemistry Letters* (Vol. 1, 2010). The fup/fumic ratio was determined for each of 416 drugs in the Pfizer database. A graph describing the fup/fumic ratios is shown below.

a. What type of graph is displayed?
b. What is the quantitative variable summarized in the graph?
c. Determine the proportion of fup/fumic ratios that fall above 1.
d. Determine the proportion of fup/fumic ratios that fall below .4.




2.17 *Sound waves from a basketball.* An experiment was conducted to characterize sound waves in a spherical cavity. (*American Journal of Physics,* June , 2010.) A fully inflated

BBALL

| Resonance | Frequency | Resonance | Frequency |
|---|---|---|---|
| 1 | 979 | 13 | 4334 |
| 2 | 1572 | 14 | 4631 |
| 3 | 2113 | 15 | 4711 |
| 4 | 2122 | 16 | 4993 |
| 5 | 2659 | 17 | 5130 |
| 6 | 2795 | 18 | 5210 |
| 7 | 3181 | 19 | 5214 |
| 8 | 3431 | 20 | 5633 |
| 9 | 3638 | 21 | 5779 |
| 10 | 3694 | 22 | 5836 |
| 11 | 4038 | 23 | 6259 |
| 12 | 4203 | 24 | 6339 |

*Source:* Russell, D.A. "Basketballs as spherical acoustic cavities", *American Journal of Physics,* Vol. 48, No. 6, June 2010. (Table I.)

basketball, hanging from rubber bands, was struck with a metal rod, producing a series of metallic sounding pings. Of particular interest were the frequencies of sound waves resulting from the first 24 resonances (echoes). A mathematical formula, well known in physics, was used to compute the theoretical frequencies. These frequencies (measured in hertz) are listed in the table on page 36. Use a graphical method to describe the distribution of sound frequencies for the first 24 resonances.

2.18 *Crude oil biodegradation.* In order to protect their valuable resources, oil companies spend millions of dollars researching ways to prevent biodegradation of crude oil. The *Journal of Petroleum Geology* (April, 2010) published a study of the environmental factors associated with biodegradation in crude oil reservoirs. Sixteen water specimens were randomly selected from various locations in a reservoir on the floor of a mine. Two of the variables measured were (1) the amount of dioxide (milligrams/liter) present in the water specimen and (2) whether or not oil was present in the water specimen. These data are listed in the accompanying table. Construct a stem-and-leaf display for the dioxide data. Locate the dioxide levels associated with water specimens that contain oil. Highlight these data points on the stem-and-leaf display. Is there a tendency for crude oil to be present in water with lower levels of dioxide?

**BIODEG**

| Dioxide Amount | Crude Oil Present |
|---|---|
| 3.3 | No |
| 0.5 | Yes |
| 1.3 | Yes |
| 0.4 | Yes |
| 0.1 | No |
| 4.0 | No |
| 0.3 | No |
| 0.2 | Yes |
| 2.4 | No |
| 2.4 | No |
| 1.4 | No |
| 0.5 | Yes |
| 0.2 | Yes |
| 4.0 | No |
| 4.0 | No |
| 4.0 | No |

*Source:* Permanyer, A., et al. "Crude oil biodegradation and environmental factors at the Riutort oil shale mine, SE Pyrenees", *Journal of Petroleum Geology,* Vol. 33, No. 2, April 2010 (Table 1).

2.19 *Sanitation inspection of cruise ships.* To minimize the potential for gastrointestinal disease outbreaks, all passenger cruise ships arriving at U.S. ports are subject to unannounced sanitation inspections. Ships are rated on a 100-point scale by the Centers for Disease Control and Prevention. A score of 86 or higher indicates that the ship is providing an accepted standard of sanitation. The sanitation scores for 186 cruise ships are saved in the **SHIPSANIT** file. The first five and last five observations in the data set are listed in the accompanying table.

**SHIPSANIT** (selected observations)

| Ship Name | Sanitation Score |
|---|---|
| *Adonia* | 96 |
| *Adventure of the Seas* | 93 |
| *AIDAAura* | 86 |
| *AID Abella* | 95 |
| *AID Aluna* | 93 |
| . | . |
| . | . |
| *Voyager of the Seas* | 96 |
| *Vspbeta* | 100 |
| *Westerdam* | 98 |
| *Zaaddam* | 100 |
| *Zuiderdam* | 96 |

*Source:* National Center for Environmental Health, Centers for Disease Control and Prevention, August 5, 2013.

a. Generate both a stem-and-leaf display and histogram of the data.
b. Use the graphs to estimate the proportion of ships that have an accepted sanitation standard. Which graph did you use?
c. Locate the inspection score of 69 (*MS Columbus 2*) on the graph. Which graph did you use?

2.20 *Surface roughness of pipe.* Oil field pipes are internally coated in order to prevent corrosion. Engineers at the University of Louisiana, Lafayette, investigated the influence that coating may have on the surface roughness of oil field pipes (*Anti-corrosion Methods and Materials*, Vol. 50, 2003). A scanning probe instrument was used to measure the surface roughness of 20 sample sections of coated interior pipe. The data (in micrometers) is provided in the table. Describe the sample data with an appropriate graph.

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F. and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

**MTBE**

2.21 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of the factors related to MTBE contamination in 223 New Hampshire wells, Exercise 2.12 (p. 29). The data are saved in the **MTBE** file. Two of the many quantitative variables measured for each well are the pH level (standard units) and the MTBE level (micrograms per liter).

a. Construct a histogram for the pH levels of the sampled wells. From the histogram, estimate the proportion of wells with pH values less than 7.0.

b. For those wells with detectable levels of MTBE, construct a histogram for the MTBE values. From the histogram, estimate the proportion of contaminated wells with MTBE values that exceed 5 micrograms per liter.

2.22 *Estimating the age of glacial drifts.* Tills are glacial drifts consisting of a mixture of clay, sand, gravel, and boulders. Engineers from the University of Washington's Department of Earth and Space Sciences studied the chemical makeup of buried tills in order to estimate the age of the glacial drifts in Wisconsin. (*American Journal of Science*, Jan. 2005.) The ratio of the elements aluminum (Al) and beryllium (Be) in sediment is related to the duration of burial. The Al/Be ratios for a sample of 26 buried till specimens are given in the table. With the aid of a graph, estimate the proportion of till specimens with an Al/Be ratio that exceeds 4.5.

**TILLRATIO**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.75 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 |
| 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 | 2.73 | 2.95 | 2.25 | 2.73 | 2.55 | 3.06 |

*Source:* Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

2.23 *Mineral flotation in water study.* A high concentration of calcium and gypsum in water can impact the water quality and limit mineral flotation. In *Minerals Engineering* (Vol. 46-47, 2013), chemical and materials engineers published a study of the impact of calcium and gypsum on the flotation properties of silica in water. Solutions of deionized water were prepared both with and without calcium/gypsum, and the level of flotation of silica in the solution was measured using a variable called *zeta potential* (measured in millivolts, mV). Assume that 50 specimens for each type of liquid solution were prepared and tested for zeta potential. The data (simulated, based on information provided in the journal article) are provided in the table and saved in the **SILICA** data file. Create side-by-side graphs to compare the zeta potential distributions for the two types of solutions. How does the addition of calcium/gypsum to the solution impact water quality (measured by zeta potential of silica)?

**SILICA**

Without calcium/gypsum

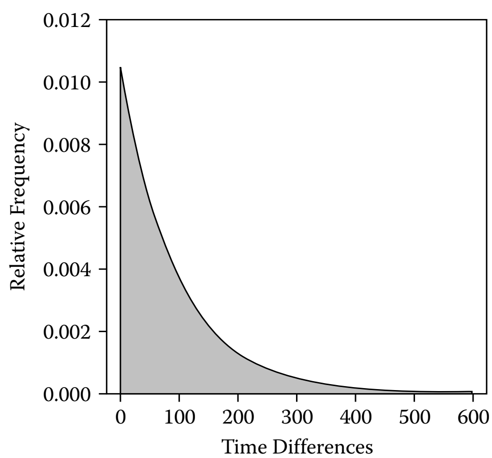
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −47.1 | −53.0 | −50.8 | −54.4 | −57.4 | −49.2 | −51.5 | −50.2 | −46.4 | −49.7 |
| −53.8 | −53.8 | −53.5 | −52.2 | −49.9 | −51.8 | −53.7 | −54.8 | −54.5 | −53.3 |
| −50.6 | −52.9 | −51.2 | −54.5 | −49.7 | −50.2 | −53.2 | −52.9 | −52.8 | −52.1 |
| −50.2 | −50.8 | −56.1 | −51.0 | −55.6 | −50.3 | −57.6 | −50.1 | −54.2 | −50.7 |
| −55.7 | −55.0 | −47.4 | −47.5 | −52.8 | −50.6 | −55.6 | −53.2 | −52.3 | −45.7 |

With calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −9.2 | −11.6 | −10.6 | −8.0 | −10.9 | −10.0 | −11.0 | −10.7 | −13.1 | −11.5 |
| −11.3 | −9.9 | −11.8 | −12.6 | −8.9 | −13.1 | −10.7 | −12.1 | −11.2 | −10.9 |
| −9.1 | −12.1 | −6.8 | −11.5 | −10.4 | −11.5 | −12.1 | −11.3 | −10.7 | −12.4 |
| −11.5 | −11.0 | −7.1 | −12.4 | −11.4 | −9.9 | −8.6 | −13.6 | −10.1 | −11.3 |
| −13.0 | −11.9 | −8.6 | −11.3 | −13.0 | −12.2 | −11.3 | −10.5 | −8.8 | −13.4 |

**PHISHING**

2.24 *Phishing attacks to email accounts.* Phishing is the term used to describe an attempt to extract personal/financial information (e.g., PIN numbers, credit card information, bank account numbers) from unsuspecting people through fraudulent email. An article in *Chance* (Summer, 2007), demonstrates how statistics can help identify phishing attempts and make e-commerce safer. Data from an actual phishing attack against an organization were used to determine whether the attack may have been an "inside job" that originated within the company. The company set up a publicized email account — called a "fraud box" — which enabled employees to notify them if they suspected an email phishing attack. The interarrival times, i.e., the time differences (in seconds), for 267 fraud box email notifications were recorded. *Chance* showed that if there is minimal or no collaboration or collusion from within the company, the interarrival times would have a frequency distribution similar to the one shown in the accompanying figure. The 267 interarrival times are saved in the **PHISHING** file. Construct a frequency histogram for the interarrival times. Give your opinion on whether the phishing attack against the organization was an "inside job".



## 2.3 Numerical Methods for Describing Quantitative Data

**Numerical descriptive measures** are numbers computed from a data set to help us create a mental image of its relative frequency histogram. The measures that we will present fall into three categories: (1) those that help to locate the *center* of the relative frequency distribution, (2) those that measure its *spread* around the center, and (3) those that describe the *relative position* of an observation within the data set. These categories are called, respectively, **measures of central tendency, measures of variation**, and **measures of relative standing**. In the definitions that follow, we will denote the *variable* observed to create a data set by the symbol $y$ and the $n$ measurements of a data set by $y_1, y_2, \ldots, y_n$.

Numerical descriptive measures computed from sample data are often called **statistics**. In contrast, numerical descriptive measures of the population are called **parameters**. Their values are typically unknown and are usually represented by Greek symbols. For example, we will see that the average value of the population is represented by the Greek letter $\mu$. Although we *could* calculate the value of this parameter if we actually had access to the entire population, we generally wish to avoid doing so, for economic or other reasons. Thus, as you will subsequently see, we will *sample* the population and then use the sample statistic to infer, or make decisions about, the value of the population parameter of interest.

**Definition 2.4**

A **statistic** is a numerical descriptive measure computed from sample data.

**Definition 2.5**

A **parameter** is a numerical descriptive measure of a population.

## 2.4 Measures of Central Tendency

The three most common measures of central tendency are the **arithmetic mean**, the **median**, and the **mode**. Of the three, the arithmetic mean (or **mean**, as it is commonly called) is used most frequently in practice.

**Definition 2.6**

The **arithmetic mean** of a set of $n$ measurements, $y_1, y_2, \ldots, y_n$, is the average of the measurements:

$$\frac{\sum_{i=1}^{n} y_i}{n}$$

Typically, the symbol $\bar{y}$ is used to represent the **sample mean** (i.e., the mean of a sample of $n$ measurements), whereas the Greek letter $\mu$ represents the **population mean**.

To illustrate, we will calculate the mean for the set of $n = 5$ sample measurements: 4, 6, 1, 2, 3. Substitution into the formula for $\bar{y}$ yields

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{4 + 6 + 1 + 2 + 3}{5} = 3.2$$

**Definition 2.7**

The **median** of a set of $n$ measurements, $y_1, y_2, \ldots, y_n$, is the middle number when the measurements are arranged in ascending (or descending) order, i.e., the value of $y$ located so that half the area under the relative frequency histogram lies to its left and half the area lies to its right. We will use the symbol $m$ to represent the *sample median* and the symbol $\tau$ to represent the *population median*.

If the number of measurements in a data set is odd, the median is the measurement that falls in the middle when the measurements are arranged in increasing order. For example, the median of the $n = 5$ sample measurements of Example 2.3 is $m = 3$. If the number of measurements is even, the median is defined to be the mean of the two middle measurements when the measurements are arranged in increasing order. For example, the median of the $n = 6$ measurements, 1, 4, 5, 8, 10, 11, is

$$m = \frac{5 + 8}{2} = 6.5$$

### Calculating the Median of Small Sample Data Sets

Let $y_{(i)}$ denote the $i$th value of $y$ when the sample of $n$ measurements is arranged in ascending order. Then the sample median is calculated as follows:

$$m = \begin{cases} y_{[(n+1)/2]} & \text{if } n \text{ is odd} \\ \dfrac{y_{(n/2)} + y_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

**Definition 2.8**

The **mode** of a set of $n$ measurements, $y_1, y_2, \ldots, y_n$, is the value of $y$ that occurs with the greatest frequency.

If the outline of a relative frequency histogram were cut from a piece of plywood, it would be perfectly balanced over the point that locates its mean, as illustrated in Figure 2.9a. As noted in Definition 2.6, half the area under the relative frequency distribution will lie to the left of the median, and half will lie to the right,

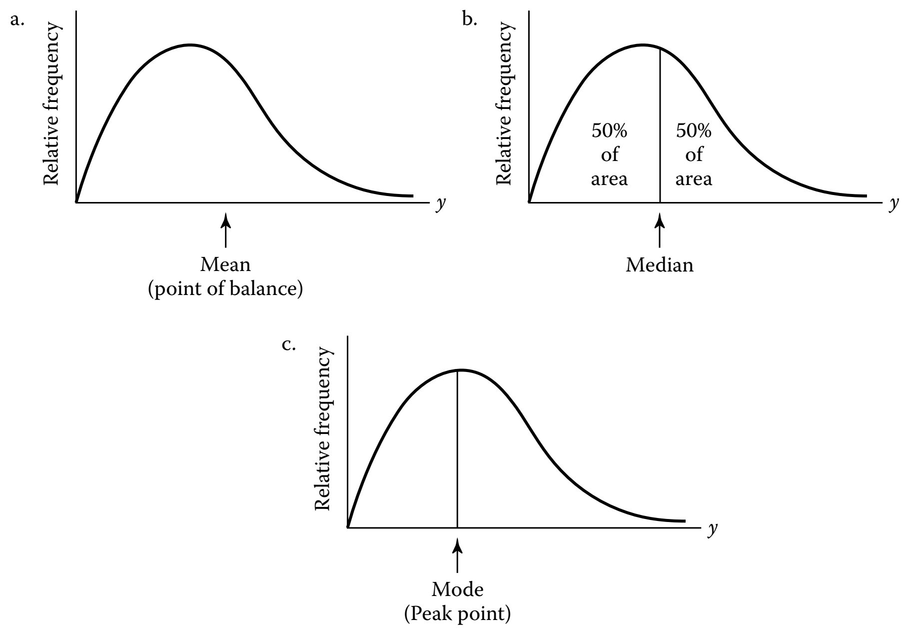



**FIGURE 2.9**
Interpretations of the mean, median, and mode for a relative frequency distribution

as shown in Figure 2.9b. The mode will locate the point at which the greatest frequency occurs, i.e., the peak of the relative frequency distribution, as shown in Figure 2.9c.

Although the mean is often the preferred measure of central tendency, it is sensitive to very large or very small observations. Consequently, the mean will shift toward the direction of **skewness** (i.e., the tail of the distribution) and may be misleading in some situations. For example, if a data set consists of the first-year starting salaries of civil engineering graduates, the high starting salaries of a few graduates will influence the mean more than the median. For this reason, the median is sometimes called a *resistant* measure of central tendency, since it, unlike the mean, is resistant to the influence of extreme observations. For data sets that are extremely skewed, (e.g., the starting salaries of civil engineering graduates), the median would better represent the "center" of the distribution data.

Rarely is the mode the preferred measure of central tendency. The mode is preferred over the mean or median only if the relative frequency of occurrence of $y$ is of interest. For example, a supplier of carpenter's materials would be interested in the modal length (in inches) of nails he sells.

In summary, the best measure of central tendency for a data set depends on the type of descriptive information you want. Most of the inferential statistical methods discussed in this text are based, theoretically, on **mound-shaped distributions** of data with little or no skewness. For these situations, the mean and the median will be, for all practical purposes, the same. Since the mean has nicer mathematical properties than the median, it is the preferred measure of central tendency for these inferential techniques.

**Example 2.3**

Comparing the Mean, Median, and Mode — Earthquake Aftershocks

**EARTHQUAKE**

**Problem:** Seismologists use the "aftershock" to describe the smaller earthquakes that follow a main earthquake. Following the Northridge earthquake in 1994, the Los Angeles area experienced 2,929 aftershocks in a three-week period. The magnitudes (measured on the Richter scale) of these aftershocks as well as their inter-arrival times (in minutes) were recorded by the U.S. Geological Survey. (The data are saved in the **EARTHQUAKE** file.) Find and interpret the mean, median, and mode for both of these variables. Which measure of central tendency is better for describing the magnitude distribution? The distribution of inter-arrival times?

Solution

Measures of central tendency for the two variables, magnitude and inter-arrival time, were produced using MINITAB. The means, medians, and modes are displayed in Figure 2.10.

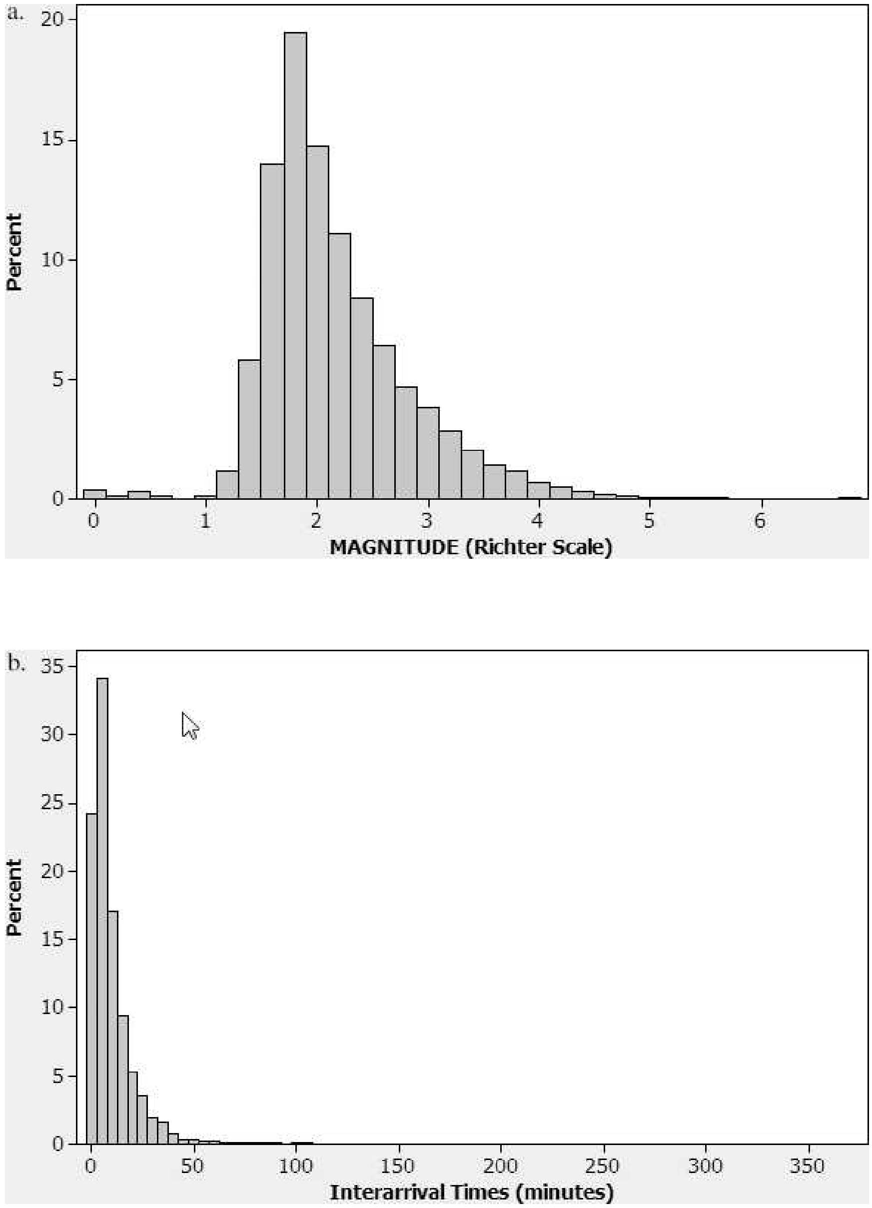
For magnitude, the mean, median, and mode are 2.12, 2.00, and 1.8, respectively, on the Richter scale. The average magnitude is 2.12; half the magnitudes fall below 2.0; and, the most commonly occurring magnitude is 1.8. These values are nearly identical, with the mean slightly larger than the median. This implies a slight rightward skewness in the data, which is shown graphically in the MINITAB histogram for magnitude displayed in Figure 2.11a. Because the distribution is nearly symmetric, any of the three measures would be adequate for describing the "center" of the earthquake aftershock magnitude distribution.

**FIGURE 2.10**
MINITAB Descriptive Statistics for Earthquake Data

**Descriptive Statistics: MAGNITUDE, INT-TIME**

| Variable | N | Mean | Median | Mode | N for Mode |
|---|---|---|---|---|---|
| MAGNITUDE | 2929 | 2.1197 | 2.0000 | 1.8 | 298 |
| INT-TIME | 2928 | 9.771 | 6.000 | 2 | 354 |

**FIGURE 2.11**
MINITAB Histograms for Magnitude and Inter-Arrival Times of Aftershocks




The mean, median, and mode of the inter-arrival times of the aftershocks are 9.77, 6.0, and 2.0 minutes, respectively. On average, the aftershocks arrive 9.77 minutes apart; half the aftershocks have inter-arrival times below 6.0 minutes; and, the most commonly occurring inter-arrival time is 2.0 minutes. Note that the mean is much larger than the median, implying that the distribution of inter-arrival times is highly skewed to the right. This extreme rightward skewness is shown graphically in the histogram, Figure 2.11b. The skewness is due to several exceptionally large inter-arrival times. Consequently, we would probably want to use the median of 6.0 minutes as the "typical" inter-arrival time for the aftershocks. You can see that the mode of 2.0 minutes is not very descriptive of the "center" of the inter-arrival time distribution.

## Applied Exercises

2.25 *Measures of central tendency.* Find the mean, median, and mode for each of the following data sets.

a. 4, 3, 10, 8, 5
b. 9, 6, 12, 4, 4, 2, 5, 6

2.26 *Highest paid engineers.* According to *Electronic Design's* 2012 Engineering Salary Survey, the mean base salary of a software engineering manager is $126,417—the highest mean among all types of engineers. In contrast, a manufacturing/production engineer has a mean base salary of $92,360. Assume these values are accurate and represent population means. Determine whether the following statements are true or false.

a. All software engineering managers earn a base salary of $126,417.
b. Half of all manufacturing/production engineers earn a base salary less than $92,360.
c. A randomly selected software engineering manager will always earn more in base salary than a randomly selected manufacturing/production engineer.

2.27 *Cheek teeth of extinct primates.* Refer to the *American Journal of Physical Anthropology* (Vol. 142, 2010) study of the characteristics of cheek teeth (e.g., molars) in an extinct primate species, Exercise 2.14 (p. 35). The data on dentary depth of molars (in millimeters) for 18 cheek teeth extracted from skulls are reproduced below.

a. Find and interpret the mean of the data set. If the largest depth measurement in the sample were doubled, how would the mean change? Would it increase or decrease?
b. Find and interpret the median of the data set. If the largest depth measurement in the sample were doubled, how would the median change? Would it increase or decrease?
c. Note that there is no single measurement that occurs more than once. How does this fact impact the mode?

**CHEEKTEETH**

| | |
|---|---|
| 18.12 | 16.55 |
| 19.48 | 15.70 |
| 19.36 | 17.83 |
| 15.94 | 13.25 |
| 15.83 | 16.12 |
| 19.70 | 18.13 |
| 15.76 | 14.02 |
| 17.00 | 14.04 |
| 13.96 | 16.20 |

*Source:* Boyer, D.M., Evans, A.R., and Jernvall, J. "Evidence of Dietary Differentiation Among Late Paleocene-Early Eocene Plesiadapids (Mammalia, Primates)", *American Journal of Physical Anthropology,* Vol. 142, 2010. (Table A3.)

2.28 *Radioactive lichen.* Refer to the University of Alaska study to monitor the level of radioactivity in lichen, Exercise 2.15 (p. 36). The amount of the radioactive element cesium-137 (measured in microcuries per milliliter) for each of nine lichen specimens is repeated in the table.

**LICHEN**

| Location | | | |
|---|---|---|---|
| Bethel | –5.50 | –5.00 | |
| Eagle Summit | –4.15 | –4.85 | |
| Moose Pass | –6.05 | | |
| Turnagain Pass | –5.00 | | |
| Wickersham Dome | –4.10 | –4.50 | –4.60 |

*Source:* Lichen Radionuclide Baseline Research Project, 2003.

a. Find the mean, median, and mode of the radioactivity levels.
b. Interpret the value of each measure of central tendency, part **a.**

2.29 *Characteristics of a rock fall.* In *Environmental Geology* (Vol. 58, 2009) computer simulation was employed to estimate how far a block from a collapsing rock wall will bounce—called *rebound length*—down a soil slope. Based on the depth, location, and angle of block-soil impact marks left on the slope from an actual rock fall, the following 13 rebound lengths (meters) were estimated. Compute the mean and median of the rebound lengths and interpret these values.

**ROCKFALL**

| | | | | | | |
|---|---|---|---|---|---|---|
| 10.94 | 13.71 | 11.38 | 7.26 | 17.83 | 11.92 | |
| 11.87 | 5.44 | 13.35 | 4.90 | 5.85 | 5.10 | 6.77 |

*Source:* Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy", *Environmental Geology,* Vol. 58, 2009. (Table 2.)

2.30 *Ammonia in car exhaust.* Three-way catalytic converters have been installed in new vehicles in order to reduce pollutants from motor vehicle exhaust emissions. However, these converters unintentionally increase the level of ammonia in the air. *Environmental Science & Technology* (Sept. 1, 2000) published a study on the ammonia levels near the exit ramp of a San Francisco highway tunnel. The data in the table represent daily ammonia concentrations (parts per million) on eight randomly selected days during afternoon drive-time in the summer of a recent year.

**AMMONIA**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.53 | 1.50 | 1.37 | 1.51 | 1.55 | 1.42 | 1.41 | 1.48 |

a. Find the mean daily ammonia level in air in the tunnel.
b. Find the median ammonia level.
c. Interpret the values obtained in parts **a** and **b.**

2.31 *Crude oil biodegradation.* Refer to the *Journal of Petroleum Geology* (April, 2010) study of the environmental factors associated with biodegradation in crude oil reservoirs, Exercise 2.18 (p. 37). Recall that amount of dioxide (milligrams/liter) and presence/absence of crude oil was determined for each of 16 water specimens collected from a mine reservoir. The data are repeated in the accompanying table.

a. Find the mean dioxide level of the 16 water specimens. Interpret this value.
b. Find the median dioxide level of the 16 water specimens. Interpret this value.
c. Find the mode of the 16 dioxide levels. Interpret this value.
d. Find the median dioxide level of the 10 water specimens with no crude oil present.
e. Find the median dioxide level of the 6 water specimens with crude oil present.
f. Compare the results, parts d and e. Make a statement about the association between dioxide level and presence/absence of crude oil.

**BIODEG**

| Dioxide Amount | Crude Oil Present |
|---|---|
| 3.3 | No |
| 0.5 | Yes |
| 1.3 | Yes |
| 0.4 | Yes |
| 0.1 | No |
| 4.0 | No |
| 0.3 | No |
| 0.2 | Yes |
| 2.4 | No |
| 2.4 | No |
| 1.4 | No |
| 0.5 | Yes |
| 0.2 | Yes |
| 4.0 | No |
| 4.0 | No |
| 4.0 | No |

*Source:* Permanyer, A., et al. "Crude oil biodegradation and environmental factors at the Riutort oil shale mine, SE Pyrenees", *Journal of Petroleum Geology,* Vol. 33, No. 2, April 2010 (Table 1).

**SHIPSANIT**

2.32 *Sanitation inspection of cruise ships.* Refer to the Centers for Disease Control study of sanitation levels for 186 international cruise ships, Exercise 2.19 (p. 37). (Recall that sanitation scores ranged from 0 to 100.) Find and interpret numerical descriptive measures of central tendency for the sanitation levels.

**SANDSTONE**

2.33 *Permeability of sandstone during weathering.* Natural stone, such as sandstone, is a popular building construction material. An experiment was carried out in order to better understand the decay properties of sandstone when exposed to the weather. (*Geographical Analysis,* Vol. 42, 2010.) Blocks of sandstone were cut into 300 equal-sized slices and the slices randomly divided into three groups of 100 slices each. Slices in group A were not exposed to any type of weathering; slices in group B were repeatedly sprayed with a 10% salt solution (to simulate wetting by driven rain) under temperate conditions; and, slices in group C were soaked in a 10% salt solution and then dried (to simulate blocks of sandstone exposed during a wet winter and dried during a hot summer). All sandstone slices were then tested for permeability, measured in milliDarcies (mD). These permeability values measure pressure decay as a function of time. The data for the study (simulated) are saved in the **SANDSTONE** file. Measures of central tendency for the permeability measurements of each sandstone group are displayed in the MINITAB printout below.

a. Interpret the mean and median of the permeability measurements for Group A sandstone slices.
b. Interpret the mean and median of the permeability measurements for Group B sandstone slices.
c. Interpret the mean and median of the permeability measurements for Group C sandstone slices.
d. Interpret the mode of the permeability measurements for Group C sandstone slices.
e. The lower the permeability value, the slower the pressure decay in the sandstone over time. Which type of weathering (type B or type C) appears to result in faster decay?

MINITAB Output for Exercise 2.33

**Descriptive Statistics: PermA, PermB, PermC**

```
                                                                N for
Variable    N    Mean  Median                         Mode   Mode
PermA     100   73.62   70.45        59.9, 60, 60.1, 60.4      2
PermB     100  128.54  139.30  146.4, 146.6, 147.9, 148.3      3
PermC     100   83.07   78.65                         70.9      3

The data contain at least five mode values. Only the smallest four are shown.
```

**SILICA**

2.34 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercise 2.23 (p. 38). The *zeta potential* (mV) was determined for each of 50 liquid solutions prepared without calcium/gypsum and for 50 liquid solutions prepared with calcium/gypsum. (These data are saved in the **SILICA** file.)

a. Find the mean, median, and mode for the zeta potential measurements of the liquid solutions prepared without calcium/gypsum. Interpret these values.

b. Find the mean, median, and mode for the zeta potential measurements of the liquid solutions prepared with calcium/gypsum. Interpret these values.

c. In Exercise 2.23, you used graphs to compare the zeta potential distributions for the two types of solutions. Now use the measures of central tendency to make the comparison. How does the addition of calcium/gypsum to the solution impact water quality (measured by zeta potential of silica)?

2.35 *Contact lenses for myopia.* Myopia (i.e., nearsightedness) is a visual condition that affects over 100 million Americans. Two treatments that may slow myopia progression is the use of (1) corneal reshaping contact lenses and (2) bifocal soft contact lenses. In *Optometry and Vision Science* (Jan., 2013), university optometry professors compared the two methods for treating myopia. A sample of 14 myopia patients participated in the study. Each patient was fitted with a contact lens of each type for the right eye, and the peripheral refraction was measured for each type of lens. The differences (bifocal soft minus corneal reshaping) are shown in the following table. (These data, simulated based on information provided in the journal article, are saved in the **MYOPIA** file.)

a. Find measures of central tendency for the difference measurements and interpret their values.

b. Note that the data contains one unusually large (negative) difference relative to the other difference measurements. Find this difference. (In Section 2.7, we call this value an **outlier**.)

c. The large negative difference of -8.11 is actually a typographical error. The actual difference for this patient is -0.11. Rerun the analysis, part a, using the corrected difference. Which measure of central tendency is most affected by the correcting of the outlier?

**MYOPIA**

| | | | | | | |
|---|---|---|---|---|---|---|
| −0.15 | −8.11 | −0.79 | −0.80 | −0.81 | −0.39 | −0.68 |
| −1.13 | −0.32 | −0.01 | −0.63 | −0.05 | −0.41 | −1.11 |

2.36 *Active nuclear power plants.* The U.S. Energy Information Administration monitors all nuclear power plants operating in the United States. The table lists the number of active nuclear power plants operating in each of a sample of 20 states.

a. Find the mean, median, and mode of this data set.

b. Eliminate the largest value from the data set and repeat part **a**. What effect does dropping this measurement have on the measures of central tendency found in part **a**?

c. Arrange the 20 values in the table from lowest to highest. Next, eliminate the lowest two values and the highest two values from the data set and find the mean of the remaining data values. The result is called a *10% trimmed mean*, since it is calculated after removing the highest 10% and the lowest 10% of the data values. What advantages does a trimmed mean have over the regular arithmetic mean?

**NUCLEAR**

| State | Number of Power Plants |
|---|---|
| Alabama | 5 |
| Arizona | 3 |
| California | 4 |
| Florida | 5 |
| Georgia | 4 |
| Illinois | 11 |
| Kansas | 1 |
| Louisiana | 2 |
| Massachusetts | 1 |
| Mississippi | 1 |
| New Hampshire | 1 |
| New York | 6 |
| North Carolina | 5 |
| Ohio | 3 |
| Pennsylvania | 9 |
| South Carolina | 7 |
| Tennessee | 3 |
| Texas | 4 |
| Vermont | 1 |
| Wisconsin | 3 |

*Source: Statistical Abstract of the United States*, 2012 (Table 942). U.S. Energy Information Administration, *Electric Power Annual*.

## 2.5 Measures of Variation

Measures of central tendency provide only a partial description of a quantitative data set. The description is incomplete without a **measure of variability**, or **spread** of the data. The most commonly used measures of data variation are the **range**, the **variance**, and the **standard deviation**.

**Definition 2.9**

The **range** is equal to the difference between the largest and the smallest measurements in a data set:

$$\text{Range} = \text{Largest measurement} - \text{Smallest measurement}$$

**Definition 2.10**

The **variance** of a **sample** of $n$ measurements, $y_1, y_2, \ldots, y_n$, is defined to be

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}$$

The **population variance** is defined to be

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n}$$

for a finite population with $n$ measurements.

**Definition 2.11**

The **standard deviation** of a **sample** of $n$ measurements is equal to the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The **population standard deviation** is $\sigma = \sqrt{\sigma^2}$.

**Example 2.4**

Computing measures of variation

Find the range, variance and standard deviation for the $n = 5$ sample observations: 1, 3, 2, 2, 4.

Solution

The range is simply the difference between the largest (4) and smallest (1) measurement, i.e.,

$$\text{Range} = 4 - 1 = 3$$

To obtain the variance and standard deviation we must first calculate $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} y_i^2$:

$$\sum_{i=1}^{n} y_i = 1 + 3 + 2 + 2 + 4 = 12 \qquad \sum_{i=1}^{n} y_i^2 = (1)^2 + (3)^2 + (2)^2 + (2)^2 + (4)^2 = 34$$

Then the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{34 - \frac{(12)^2}{5}}{4} = 1.3$$

and the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.1402$$

It is possible that two different data sets could possess the same range but differ greatly in the amount of variation in the data. Consequently, the range is a relatively insensitive measure of data variation. It is used primarily in industrial quality control where the inferential procedures are based on small samples (i.e., small values of $n$). The variance has theoretical significance but is difficult to interpret since the units of measurement on the variable $y$ of interest are squared (e.g., feet$^2$, ppm$^2$, etc.). The units of measurement on the standard deviation, however, are the same as the units on $y$ (e.g., feet, ppm). When combined with the mean of the data set, the standard deviation is easily interpreted.

Two useful rules for interpreting the standard deviation are the **Empirical Rule** and **Chebyshev's Rule**.

### The Empirical Rule

If a data set has an approximately mound-shaped, symmetric distribution, then the following rules of thumb may be used to describe the data set (see Figure 2.12a):

1. Approximately 68% of the measurements will lie within 1 standard deviation of their mean (i.e., within the interval $\bar{y} \pm s$ for samples and $\mu \pm \sigma$ for populations).
2. Approximately 95% of the measurements will lie within 2 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 2s$ for samples and $\mu \pm 2\sigma$ for populations).
3. Almost all the measurements will lie within 3 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 3s$ for samples and $\mu \pm 3\sigma$ for populations).

### Chebyshev's Rule

**Chebyshev's Rule** applies to any data set, regardless of the shape of the frequency distribution of the data (see Figure 2.12b).

a. It is possible that very few of the measurements will fall within 1 standard deviation of the mean, i.e., within the interval $(\bar{y} \pm s)$ for samples and $(\mu \pm \sigma)$ for populations.

b. At least $\frac{3}{4}$ of the measurements will fall within 2 standard deviations of the mean, i.e., within the interval $(\bar{y} \pm 2s)$ for samples and $(\mu \pm 2\sigma)$ for populations.

c. At least $\frac{8}{9}$ of the measurements will fall within 3 standard deviations of the mean, i.e., within the interval $(\bar{y} \pm 3s)$ for samples and $(\mu \pm 3\sigma)$ for populations.

d. Generally, for any number $k$ greater than 1, at least $(1 - 1/k^2)$ of the measurements will fall within $k$ standard deviations of the mean, i.e., within the interval $(\bar{y} \pm ks)$ for samples and $(\mu \pm k\sigma)$ for populations.

**FIGURE 2.12a**
Empirical Rule

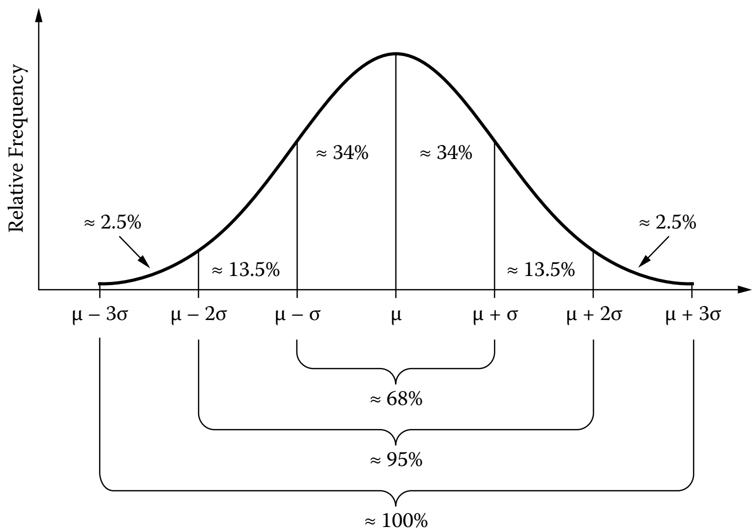
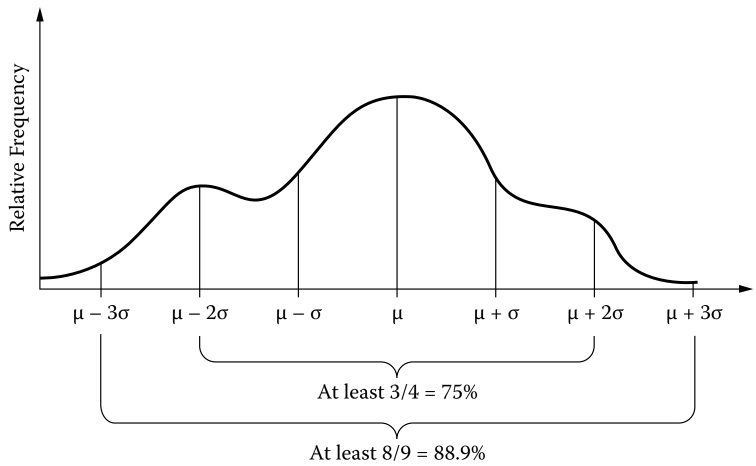


**FIGURE 2.12b**
Chebyshey's Rule




The Empirical Rule is the result of the practical experience of researchers in many fields who have observed many different types of real-life data sets. Chebyshev's Rule is derived from a theorem proved by the Russian mathematician Pafnuty L. Chebyshev (1821–1894). Both rules, described in the boxes, give the percentage of measurements in a data set that fall in the interval $\bar{y} \pm ks$, where $k$ is any integer.

## Example 2.5

Applying Rules for Describing the Distribution of Iron Ore Contents

**IRONORE**

Refer to Example 2.2 (p. 34) and the data on percent iron content of iron-ore specimens. Use a rule of thumb to describe the distribution of iron content measurements. In particular, estimate the number of the 390 iron-ore specimens that have iron content measurements that fall within 2 standard deviations of the mean.

Solution

We used SAS to obtain the mean and standard deviation of the iron contents. From the SAS printout, Figure 2.13, the sample mean is $\bar{y} = 65.74\%$ and the standard deviation is $s = .69\%$. Using these values, we form the intervals $\bar{y} \pm s$, $\bar{y} \pm 2s$, and $\bar{y} \pm 3s$. Applying both the Empirical Rule and Chebyshev's Rule, we can estimate the proportions of the 390 iron content measurements to fall within the intervals. These proportions are given in Table 2.4.

**FIGURE 2.13**
SAS Descriptive Statistics for Iron Ore Contents

Summary Statistics

Results

The MEANS Procedure

| Analysis Variable : IRONORE | | | | | |
|---|---|---|---|---|---|
| Mean | Std Dev | Variance | Minimum | Maximum | N |
| 65.7424359 | 0.6936098 | 0.4810946 | 62.7700000 | 66.8600000 | 390 |

**TABLE 2.4 Applying Rules of Thumb to the 390 Iron Content Measurements**

| $k$ | $\bar{y} \pm ks$ | Expected Proportion Using Empirical Rule | Expected Proportion Using Chebyshev's Rule | Actual Proportion |
|---|---|---|---|---|
| 1 | (65.05, 66.43) | $\approx$ .68 | at least 0 | .744 |
| 2 | (64.36, 67.12) | $\approx$ .95 | at least .75 | .947 |
| 3 | (63.67, 67.81) | $\approx$ 1.00 | at least .889 | .980 |

You can see that, for each of the three intervals, the actual proportion (obtained using SAS) of the $n = 390$ iron-ore specimens that have iron measurements in the interval is very close to that approximated by the Empirical Rule. Such a result is expected since the relative frequency histogram of the 390 measurements (shown in Figure 2.8, p. 34) is mound-shaped and nearly symmetric. Although it can be applied to any data set, Chebyshev's Rule tends to be conservative, providing a lower bound on the percentage of measurements that fall in the interval. Consequently, our best estimate of the percentage of iron content measurements that fall within 2 standard deviations of the mean is obtained using the Empirical Rule—namely, approximately 95%.

Since many data sets encountered in engineering and the sciences are approximately mound-shaped, scientists often apply the Empirical Rule to estimate a range where most of the measurements fall. The interval $\bar{y} \pm 2s$ is typically selected since it captures about 95% of the data.

## Applied Exercises

2.37 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study on the current trend in the design of social robots, Exercise 2.1 (p. 26). Recall that in a random sample of social robots obtained through a web search, 28 were built with wheels. The number of wheels on each of the 28 robots is listed in the accompanying table.

a. Generate a histogram for the sample data set. Is the distribution of number of wheels mound-shaped and symmetric?
b. Find the mean and standard deviation for the sample data set.
c. Form the interval, $\bar{y} \pm 2s$.
d. According to Chebychev's Rule, what proportion of sample observations will fall within the interval, part c?
e. According to the Empirical Rule, what proportion of sample observations will fall within the interval, part c?
f. Determine the actual proportion of sample observations that fall within the interval, part c. Even though the histogram, part a, is not perfectly symmetric, does the Empirical Rule provide a good estimate of the proportion?

**ROBOTS**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 3 | 3 | 3 | 6 | 4 | 2 | 2 | 2 | 1 | 3 | 3 | 3 |
| 3 | 4 | 4 | 3 | 2 | 8 | 2 | 2 | 3 | 4 | 3 | 3 | 4 | 2 |

*Source:* Chew, S., et al. "Do social robots walk or roll?", *International Conference on Social Robotics,* Vol. 6414, 2010 (adapted from Figure 2).

2.38 *Highest paid engineers.* Recall (from Exercise 2.26) that the mean base salary of a software engineering manager is \$126,417 (*Electronic Design's* 2012 Engineering Salary Survey). Assume the distribution of base salaries for all software engineers is mound-shaped with a variance of 225,000,000. Sketch the distribution, showing the intervals $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$ on the graph. Estimate the proportion of software engineering managers with base salaries in each interval.

2.39 *Ammonia in car exhaust.* Refer to the *Environmental Science & Technology* (Sept. 1, 2000) study on the ammonia levels near the exit ramp of a San Francisco highway tunnel, Exercise 2.30 (p. 43). The data (in parts per million) for 8 days during afternoon drive-time are reproduced in the table.

**AMMONIA**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.53 | 1.50 | 1.37 | 1.51 | 1.55 | 1.42 | 1.41 | 1.48 |

a. Find the range of the ammonia levels.
b. Find the variance of the ammonia levels.
c. Find the standard deviation of the ammonia levels.
d. Suppose the standard deviation of the daily ammonia levels during morning drive-time at the exit ramp is 1.45 ppm. Which time, morning or afternoon drive-time, has more variable ammonia levels?

**SILICA**

2.40 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercise 2.23 and 2.34 (p. 38, 45). Recall that one flotation property is zeta potential (measured in mV units). Zeta potential was determined for each of 50 liquid solutions prepared without calcium/gypsum and for 50 liquid solutions prepared with calcium/gypsum.

a. Find the standard deviation for the zeta potential measurements of the liquid solutions prepared without calcium/gypsum. Give an interval that contains most (about 95%) of the zeta potential measurements in this data set.
b. Find the standard deviation for the zeta potential measurements of the liquid solutions prepared with calcium/gypsum. Give an interval that contains most (about 95%) of the zeta potential measurements in this data set.
c. Use the intervals, parts a and b, to make a statement about whether the addition of calcium/gypsum to the liquid solution impacts the flotation property of silica.

**SANDSTONE**

2.41 *Permeability of sandstone during weathering.* Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercise 2.33 (p.). Recall that slices of sandstone blocks were tested for permeability under three conditions: no exposure to any type of weathering (A), repeatedly sprayed with a 10% salt solution (B), and soaked in a 10% salt solution and dried (C). Measures of variation for the permeability measurements (mV) of each sandstone group are displayed in the accompanying MINITAB printout.

**Descriptive Statistics: PermA, PermB, PermC**

| Variable | N | StDev | Variance | Minimum | Maximum | Range |
|---|---|---|---|---|---|---|
| PermA | 100 | 14.48 | 209.53 | 55.20 | 122.40 | 67.20 |
| PermB | 100 | 21.97 | 482.75 | 50.40 | 150.00 | 99.60 |
| PermC | 100 | 20.05 | 401.94 | 52.20 | 129.00 | 76.80 |

a. Find the range of the permeability measurements for Group A sandstone slices. Verify its value using the minimum and maximum values shown on the printout.
b. Find the standard deviation of the permeability measurements for Group A sandstone slices. Verify its value using the variance shown on the printout.
c. Combine the mean (from Exercise 2.33) and standard deviation to make a statement about where most of the permeability measurements for Group A sandstone slices will fall. Which rule (and why) did you use to make this inference?
d. Repeat parts a–c for Group B sandstone slices.
e. Repeat parts a–c for Group C sandstone slices.
f. Based on all your analyses, which type of weathering (type B or type C) appears to result in faster decay (i.e., higher permeability measurements)?

**NUCLEAR**

2.42 *Active nuclear power plants.* Refer to Exercise 2.36 (p. 45) and the U.S. Energy Information Administration's data on the number of nuclear power plants operating in each of 20 states. The data are saved in the **NUCLEAR** file.

a. Find the range, variance, and standard deviation of this data set.
b. Eliminate the largest value from the data set and repeat part **a**. What effect does dropping this measurement have on the measures of variation found in part **a**?
c. Eliminate the smallest and largest value from the data set and repeat part **a**. What effect does dropping both of these measurements have on the measures of variation found in part **a**?

2.43 *Shopping vehicle and judgment.* Engineers who design shopping vehicles (e.g., carts) for retail stores take into account issues such as maneuverability, shopping behavior, child safety, and maintenance cost. Interestingly, while shopping at the grocery store you may be more likely to buy a vice product (e.g., a candy bar) when pushing a shopping cart than when carrying a shopping basket. This possibility was explored in the *Journal of Marketing Research* (Dec., 2011). The researchers believe that when your arm is flexed (as when carrying a basket) you are more likely to choose a vice product than when your arm is extended (as when pushing a cart). To test this theory in a laboratory setting, the researchers recruited 22 consumers and had each push their hand against a table while they were asked a serious of shopping questions. Half of the consumers were told to put their arm in a flex position (similar to a shopping basket) and the other half were told to put their arm in an extended position (similar to a shopping cart). Participants were offered several choices between a vice and a virtue (e.g., a movie ticket vs. a shopping coupon, pay later with a larger amount vs. pay now) and a choice score (on a scale of 0 to 100) was determined for each. (Higher scores indicate a greater preference for vice options.) The average choice score for consumers with a flexed arm was 59, while the average for consumers with an extended arm was 43.

a. Suppose the standard deviations of the choice scores for the flexed arm and extended arm conditions are 4 and 2, respectively. Does this information support the researchers' theory? Explain.
b. Suppose the standard deviations of the choice scores for the flexed arm and extended arm conditions are 10 and 15, respectively. Does this information support the researchers' theory? Explain.

**Descriptive Statistics: MLOC**

| Variable | defect | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|
| MLOC | false | 449 | 26.17 | 37.64 | 1.10 | 8.00 | 15.00 | 29.00 | 423.00 |
| | true | 49 | 61.51 | 67.49 | 1.00 | 21.00 | 41.00 | 85.50 | 411.00 |

MINITAB output for Exercise 2.44

**SWDEFECTS**

2.44 *Software defects.* Refer to Exercise 2.10 (p. 28) and the PROMISE Software Engineering Repository data set that contains information on 498 modules of software code. One possible predictor of whether a module of code contains defects is the number of lines of code. The MINITAB printout above shows summary statistics for number of lines of code for modules that contain defects and modules that do not. Use the means and standard deviations to compare the distributions of lines of code for defective ("true") and nondefective ("false") modules.

**MTBE**

2.45 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of the MTBE contamination in New Hampshire wells, Exercise 2.12 (p. 29). Consider only the data for those wells with detectable levels of MTBE. The MINITAB printout below gives summary statistics for MTBE levels (micrograms per liter) of public and private wells.

a. Find an interval that will contain most (about 95%) of the MTBE values for private New Hampshire wells.
b. Find an interval that will contain most (about 95%) of the MTBE values for public New Hampshire wells.

2.46 *Monitoring impedance to leg movements.* In an experiment to monitor the impedance to leg movement, Korean engineers attached electrodes to the ankles and knees of volunteers. Of interest was the signal-to-noise ratio (SNR) of impedance changes, where the signal is the magnitude of the leg movement and noise is the impedance change resulting from interferences such as knee flexes and hip extensions. For a particular ankle–knee electrode pair, a sample of 10 volunteers had SNR values with a mean of 19.5 and a standard deviation of 4.7. (*IEICE Transactions on Information & Systems*, Jan. 2005.) Assuming the distribution of SNR values in the population is mound-shaped and symmetric, give an interval that contains about 95% of all SNR values in the population. Would you expect to observe an SNR value of 30?

2.47 *Bearing strength of concrete FRP strips.* Fiber-reinforced polymer (FRP) composite materials are the standard for strengthening, retrofitting, and repairing concrete structures. Typically, FRP strips are fastened to the concrete with epoxy adhesive. Engineers at the University of Wisconsin–Madison have developed a new method of fastening the FRP strips using mechanical anchors. (*Composites Fabrication Magazine*, Sept. 2004.) To evaluate the new fastening method, 10 specimens of pultruded FRP strips mechanically fastened to highway bridges were tested for bearing strength. The strength measurements (recorded in mega Pascal units, Mpa) are shown in the table. Use the sample data to give an interval that is likely to contain the bearing strength of a pultruded FRP strip.

**FRP**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 240.9 | 248.8 | 215.7 | 233.6 | 231.4 | 230.9 | 225.3 | 247.3 | 235.5 | 238.0 |

*Source:* Data are simulated from summary information provided in *Composites Fabrication Magazine*, Sept. 2004, p. 32 (Table 1).

2.48 *Velocity of Winchester bullets.* The *American Rifleman* reported on the velocity of ammunition fired from the FEG P9R pistol, a 9-mm gun manufactured in Hungary. Field tests revealed that Winchester bullets fired from the

**Descriptive Statistics: MTBE-Level (WellClass = Private)**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| MTBE-Level | 22 | 1.000 | 0.950 | 0.240 | 0.330 | 0.520 | 1.390 | 3.860 |

**Descriptive Statistics: MTBE-Level (WellClass = Public)**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| MTBE-Level | 48 | 4.56 | 10.39 | 0.230 | 0.353 | 1.04 | 3.07 | 48.10 |

MINITAB output for Exercise 2.45

pistol had a mean velocity (at 15 feet) of 936 feet per second and a standard deviation of 10 feet per second. Tests were also conducted with Uzi and Black Hills ammunition.

a. Describe the velocity distribution of Winchester bullets fired from the FEG P9R pistol.

b. A bullet, brand unknown, is fired from the FEG P9R pistol. Suppose the velocity (at 15 feet) of the bullet is 1000 feet per second. Is the bullet likely to be manufactured by Winchester? Explain.

## 2.6 Measures of Relative Standing

We've seen that numerical measures of central tendency and variation help describe the distribution of a quantitative data set. In addition, you may want to describe the location of an observation *relative* to the other values in the distribution. Two **measures of the relative standing** of an observation are **percentiles** and **$z$-scores**.

**Definition 2.12**

The **100$p$th percentile** of a data set is a value of $y$ located so that 100$p$% of the area under the relative frequency distribution for the data lies to the left of the 100$p$th percentile and $100(1-p)$% of the area lies to its right. (*Note:* $0 \leq p \leq 1$.)

For example, if your grade in an industrial engineering class was located at the 84th percentile, then 84% of the grades were lower than your grade and 16% were higher.

The median is the 50th percentile. The 25th percentile, the median, and the 75th percentile are called the **lower quartile**, the **midquartile**, and the **upper quartile**, respectively, for a data set.

**Definition 2.13**

The **lower quartile**, $Q_L$, for a data set is the 25th percentile.

**Definition 2.14**

The **midquartile** (or median), $m$, for a data set is the 50th percentile.

**Definition 2.15**

The **upper quartile**, $Q_U$, for a data set is the 75th percentile.

For large data sets (e.g., populations), quartiles are found by locating the corresponding areas under the curve (relative frequency distribution). However, when the data set of interest is small, it may be impossible to find a measurement in the data set that exceeds, say, *exactly* 25% of the remaining measurements. Consequently, the 25th percentile (or lower quartile) for the data set is not well defined. The following box contains a few rules for finding quartiles and other percentiles with small data sets.

### Finding Quartiles (and Percentiles) with Small Data Sets

*Step 1* Rank the measurements in the data set in increasing order of magnitude. Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ represent the ranked measurements.

*Step 2* Calculate the quantity $\ell = \frac{1}{4}(n+1)$ and round to the nearest integer. The measurement with this rank, denoted $y_{(\ell)}$, represents the *lower quartile* or 25th percentile. [*Note:* If $\ell = \frac{1}{4}(n+1)$ falls halfway between two integers, round up.]

*Step 3* Calculate the quantity $u = \frac{3}{4}(n+1)$ and round to the nearest integer. The measurement with this rank, denoted $y_{(u)}$, represents the *upper quartile* or 75th percentile. [*Note:* If $u = \frac{3}{4}(n+1)$ falls halfway between two integers, round down.]

**General** To find the $p$th percentile, calculate the quantity $i = p(n + 1)/100$ and round to the nearest integer. The measurement with this rank, denoted $y_{(i)}$, is the $p$th percentile.

## Example 2.6

Finding Quartiles—Ingot Freckling

Freckles are defects that sometimes form during the solidification of alloy ingots. A freckle index has been developed to measure the level of freckling on the ingot. A team of engineers conducted several experiments to measure the freckle index of a certain type of superalloy (*Journal of Metallurgy*, Sept. 2004). The data for $n = 18$ alloy tests is shown in Table 2.5. Create a stem-and-leaf display for the data and use it to find the lower quartile for the 18 freckle indexes.

Solution

The data of Table 2.5 are saved in the **FRECKLE** file. A MINITAB stem-and-leaf display for the data is shown in Figure 2.14. We'll use this graph to help find the lower quartile for the data set.

From the box, the lower quartile $Q_L$ is the observation $y_{(\ell)}$ when the data are arranged in increasing order, where $\ell = \frac{1}{4}(n + 1)$. Since $n = 18$, $\ell = \frac{1}{4}(19) = 4.75$. Rounding up, we obtain $\ell = 5$. Thus, the lower quartile, $Q_L$, will be the fifth observation when the data are arranged in order from smallest to largest, i.e., $Q_L = y_{(5)}$. For small data sets, a stem-and-leaf display is useful for finding quartiles and percentiles. You can see that the fifth observation is the fifth leaf in stem row 0. This value corresponds to a freckle index of 4.1. Thus, for this small data set, $Q_L = 4.1$.

FRECKLE

**TABLE 2.5 Freckle Indexes for 18 Superalloys**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 30.1 | 22.0 | 14.6 | 16.4 | 12.0 | 2.4 | 22.2 | 10.0 | 15.1 |
| 12.6 | 6.8 | 4.1 | 2.5 | 1.4 | 33.4 | 16.8 | 8.1 | 3.2 |

*Source:* Yang, W. H., et al., "A freckle criterion for the solidification of superalloys with a tilted solidification front," *Journal of Metallurgy*, Vol. 56, No. 9, Sept. 2004 (Table IV).

**FIGURE 2.14**
MINITAB stem-and-leaf display for freckle index of alloys

```
Stem-and-leaf of F-INDEX  N  = 18
Leaf Unit = 1.0


  5   0  12234
  7   0  68
 (4)  1  0224
  7   1  566
  4   2  22
  2   2
  2   3  03
```

Another useful measure of relative standing is a ***z*-score**. By definition, a $z$-score describes the location of an observation $y$ relative to the mean in units of the standard deviation. Negative $z$-scores indicate that the observation lies to the left of the mean; positive $z$-scores indicate that the observation lies to the right of the mean. Also, we know from the Empirical Rule that most of the observations in a data set will be less than 2 standard deviations from the mean (i.e., will have $z$-scores less than 2 in absolute value) and almost all will be within 3 standard deviations of the mean (i.e., will have $z$-scores less than 3 in absolute value).

**Definition 2.16**

The ***z*-score** for a value *y* of a data set is the distance that *y* lies above or below the mean, measured in units of the standard deviation:

$$\text{Sample } z\text{-score:} \quad z = \frac{y - \bar{y}}{s}$$

$$\text{Population } z\text{-score:} \quad z = \frac{y - \mu}{\sigma}$$

**Example 2.7**

Finding z-scores—Iron ore contents

Refer to Example 2.5 and the data on percentage iron content for 390 iron-ore specimens. Find and interpret the *z*-score for the measurement of 66.56%.

**Solution**

Recall that the mean and standard deviation of the sample data (shown in Figure 2.10) are $\bar{y} = 65.74$ and $s = .69$. Substituting $y = 66.56$ into the formula for $z$, we obtain

$$z = (y - \bar{y})/s = (66.56 - 65.74)/.69 = 1.19$$

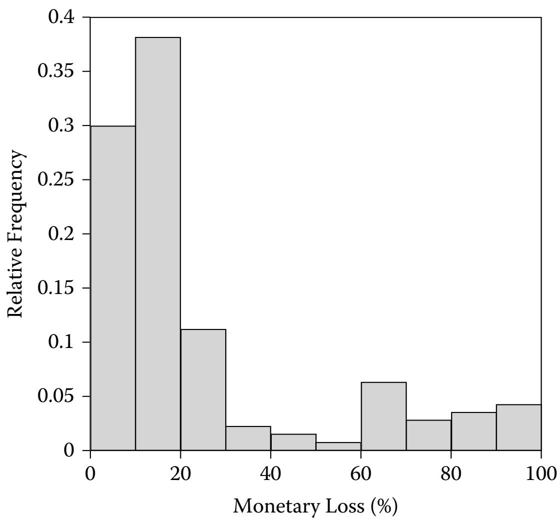
Since the *z*-score is positive, we conclude that the iron content value of 66.56% lies a distance of 1.19 standard deviations above (to the right of) the sample mean of 65.74%.

## Applied Exercises

2.49 *Annual survey of computer crimes.* Refer to the 2010 *CSI Computer Crime and Security Survey,* Exercise 2.13 (p. 35). Recall that the percentage of monetary losses attributable to malicious insider actions was recorded for 144 firms. The histogram for the data is reproduced below.

a. Based on the histogram, what (approximate) monetary loss value represents the 30th percentile?

b. Based on the histogram, what (approximate) monetary loss value represents the 95th percentile?




2.50 *Stability of compounds in new drugs.* Refer to the Pfizer Global Research and Development study (reported in *ACS Medicinal Chemistry Letters,* Vol. 1, 2010) of the metabolic stability of drugs, Exercise 2.16 (p. 36). Recall that the stability of each of 416 drugs was measured by the fup/fumic ratio.

a. In Exercise 2.16 you determined the proportion of fup/fumic ratios that fall above 1. Use this proportion to determine the percentile rank of 1.

b. In Exercise 2.16 you determined the proportion of fup/fumic ratios that fall below .4. Use this proportion to determine the percentile rank of .4.

2.51 *Highest paid engineers.* Recall (from Exercise 2.26) that the mean base salary of a software engineering manager is \$126,417 (*Electronic Design's* 2012 Engineering Salary Survey). Assume (as in Exercise 2.38) that the distribution of base salaries for all software engineers is mound-shaped and symmetric with a standard deviation of \$15,000. Use your understanding of the Empirical Rule to find:

a. the 84th percentile.

b. the 2.5th percentile.

c. the z-score for a salary of \$100,000.

2.52 *Phosphorous standards in the Everglades.* A key pollutant of the Florida Everglades is total phosphorous (TP). *Chance* (Summer 2003) reported on a study to establish standards for TP water quality in the Everglades. The Florida Department of Environmental Protection (DEP) collected data on TP concentrations at 28 Everglades sites. The 75th percentile of the TP distribution was found to be 10 micrograms per liter. The DEP recommended this value be used as a TP standard for the Everglades; i.e., any site with a TP reading exceeding 10 micrograms per liter would be considered unsafe. Interpret this 75th percentile value. Give a reason why it was selected as a TP standard by the DEP.

2.53 *Voltage sags and swells.* The power quality of a transformer is measured by the quality of the voltage. Two causes of poor power quality are "sags" and "swells". A sag is an unusual dip and a swell is an unusual increase in the voltage level of a transformer. The power quality of transformers built in Turkey was investigated in *Electrical Engineering* (Vol. 95, 2013). For a sample of 103 transformers built for heavy industry, the mean number of sags per week was 353 and the mean number of swells per week was 184. Assume the standard deviation of the sag distribution is 30 sags per week and the standard deviation of the swell distribution is 25 swells per week. Suppose one of the transformers is randomly selected and found to have 400 sags and 100 swells in a week.

a. Find the z-score for the number of sags for this transformer. Interpret this value.
b. Find the z-score for the number of swells for this transformer. Interpret this value.

**NZBIRDS**

2.54 *Extinct New Zealand birds.* Refer to the *Evolutionary Ecology Research* (July 2003) study of the patterns of extinction in the New Zealand bird population, Exercise 2.11 (p. 28). Consider the data on the egg length (measured in millimeters) for the 116 bird species saved in the **NZBIRDS** file.

a. Find the 10th percentile for the egg length distribution and interpret its value.
b. The *Moas, P. australis* bird species has an egg length of 205 millimeters. Find the $z$-score for this species of bird and interpret its value.

**SHIPSANIT**

2.55 *Sanitation inspection of cruise ships.* Refer to the sanitation levels of cruise ships, Exercise 2.19 (p. 37), saved in the **SHIPSANIT** file.

a. Give a measure of relative standing for the *Nautilus Explorer*'s score of 74. Interpret the result.
b. Give a measure of relative standing for the *Rotterdam*'s score of 86. Interpret the result.

2.56 *Lead in drinking water.* The US. Environmental Protection Agency (EPA) sets a limit on the amount of lead permitted in drinking water. The EPA *Action Level* for lead is .015 milligrams per liter (mg/L) of water. Under EPA guidelines, if 90% of a water system's study samples have a lead concentration less than .015 mg/L, the water is considered safe for drinking. I (co-author Sincich) received a report on a study of lead levels in the drinking water of homes in my subdivision. The 90th percentile of the study sample had a lead concentration of .00372 mg/L. Are water customers in my subdivision at risk of drinking water with unhealthy lead levels? Explain.

**SILICA**

2.57 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercises 2.23, 2.34 and 2.40 (p. 50). Recall that zeta potential (mV) was determined for each of 50 liquid solutions prepared without calcium/gypsum and for 50 liquid solutions prepared with calcium/gypsum.

a. For solutions prepared without calcium/gypsum, find the z-score for a zeta potential measurement of $-9.0$.
b. For solutions prepared with calcium/gypsum, find the z-score for a zeta potential measurement of $-9.0$.
c. Based on the results, parts a and b, which solution is more likely to have a zeta potential measurement of $-9.0$? Explain.

## 2.7 Methods for Detecting Outliers

Sometimes inconsistent observations are included in a data set. For example, when we discuss starting salaries for college graduates with bachelor's degrees, we generally think of traditional college graduates—those near 22 years of age with 4 years of college education. But suppose one of the graduates is a 34-year-old PhD chemical engineer who has returned to the university to obtain a bachelor's degree in metallurgy. Clearly, the starting salary for this graduate could be much larger than the other starting salaries because of the graduate's additional education and experience, and we probably would not want to include it in the data set. Such an errant observation, which lies outside the range of the data values that we want to describe, is called an **outlier**.

**Definition 2.17**

An observation $y$ that is unusually large or small relative to the other values in a data set is called an **outlier**. Outliers typically are attributable to one of the following causes:

1. The measurement is observed, recorded, or entered into the computer incorrectly.
2. The measurement comes from a different population.
3. The measurement is correct, but represents a rare (chance) event.

The most obvious method for determining whether an observation is an outlier is to calculate its $z$-score (Section 2.6).

**Example 2.8**

Deleting Outliers—Energy-Related Fatalities

FATAL

Refer to the sample data on 62 energy-related accidents worldwide since 1979 that resulted in multiple fatalities. (The data are saved in the FATAL file.) In addition to the cause of the fatal energy-related accident, the data set also contains information on the number of fatalities for each accident. The first observation in the data set is a dam failure accident that occurred in India in 1979, killing 2500 people. Is this observation an outlier?

Solution

Descriptive statistics on the number of fatalities for the 62 energy-related accidents are displayed in the MINITAB printout, Figure 2.15. The mean and standard deviation, highlighted on the printout, are $\bar{y} = 208.3$ and $s = 344.6$. Consequently, the $z$-score for the observation with a number of fatalities of $y = 2500$ is

$$z = (y - \bar{y})/s = (2500 - 208.3)/344.6 = 6.65$$

The Empirical Rule states that almost all the observations in a data set will have $z$-scores less than 3 in absolute value, while Chebyshev's Rule guarantees that at most $\frac{1}{9}$ (or, 11%) will have $z$-scores greater than 3 in absolute value. Since a $z$-score as large as 6.65 is rare, the measurement $y = 2500$ is called an *outlier*. Although this value was correctly recorded, the 1979 accident was attributed to heavy flooding in India, causing one of the first hydroelectric dams in the country to collapse.

**Descriptive Statistics: Fatalities**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Fatalities | 62 | 208.3 | 344.6 | 11.0 | 69.5 | 106.5 | 185.8 | 2500.0 | 116.3 |

**FIGURE 2.15**
MINITAB Descriptive Statistics for Number of Energy-Related Fatal Accidents

Another procedure for detecting outliers is to construct a **box plot** of the sample data. With this method, we construct intervals similar to the $\bar{y} \pm 2s$ and $\bar{y} \pm 3s$ intervals of the Empirical Rule; however, the intervals are based on a quantity called the **interquartile range** instead of the standard deviation $s$.

**Definition 2.18**
The **interquartile range**, IQR, is the distance between the upper and lower quartiles:

$$\text{IQR} = Q_U - Q_L$$

The intervals $[Q_L - 1.5(\text{IQR}), Q_U + 1.5(\text{IQR})]$ and $[Q_L - 3(\text{IQR}), Q_U + 3(\text{IQR})]$ are the key to detecting outliers with a box plot.

The elements of a box plot are listed in the next box. A box plot is relatively easy to construct for small data sets because the quartiles and interquartile range can be quickly determined. However, since almost all statistical software includes box plot routines, we'll use the computer to construct a box plot.

### Elements of a Box Plot (See Figure 2.16)

1. A rectangle (the **box**) is drawn with the ends (the **hinges**) drawn at the lower and upper quartiles ($Q_L$ and $Q_U$). The median of the data is shown in the box, usually by a line.
2. The points at distances 1.5(IQR) from each hinge mark the **inner fences** of the data set. Lines (the **whiskers**) are drawn from each hinge to the most extreme measurement inside the inner fence.

$$\text{Lower inner fence} = Q_L - 1.5(\text{IQR})$$

$$\text{Upper inner fence} = Q_U + 1.5(\text{IQR})$$

3. A second pair of fences, the **outer fences**, appear at a distance of 3 interquartile ranges, 3(IQR), from the hinges. One symbol (e.g., "*") is used to represent measurements falling between the inner and outer fences, and another (e.g., "0") is used to represent measurements beyond the outer fences. Thus, outer fences are not shown unless one or more measurements lie beyond them.

$$\text{Lower outer fence} = Q_L - 3(\text{IQR})$$

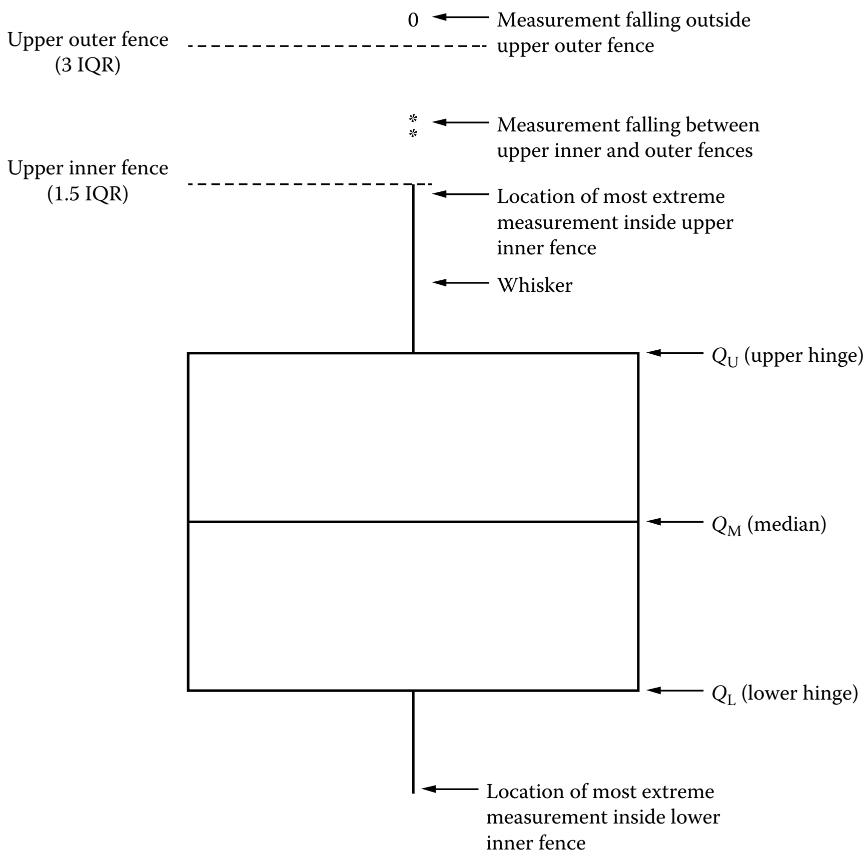
$$\text{Upper outer fence} = Q_U + 3(\text{IQR})$$

4. The symbols used to represent the median and the extreme data points (those beyond the fences) will vary depending on the software you use to construct the box plot. (You may use your own symbols if you are constructing a box plot by hand.) You should consult the program's documentation to determine exactly which symbols are used.

## Aids to the Interpretation of Box Plots

1. Examine the length of the box. The IQR is a measure of the sample's variability and is especially useful for the comparison of two samples.
2. Visually compare the lengths of the whiskers. If one is clearly longer, the distribution of the data is probably skewed in the direction of the longer whisker.

**FIGURE 2.16**
Key Elements of a Box Plot



3. Analyze any measurements that lie beyond the fences. Fewer than 5% should fall beyond the inner fences, even for very skewed distributions. Measurements beyond the outer fences are probably outliers, with one of the following explanations:
   a. The measurement is incorrect. It may have been observed, recorded, or entered into the computer incorrectly.
   b. The measurement belongs to a population different from the population that the rest of the sample was drawn from.
   c. The measurement is correct *and* from the same population as the rest. Generally, we accept this explanation only after carefully ruling out all others.

**Example 2.9**

Constructing a Box Plot—Energy-Related Fatalities

**FATAL**

Refer to Example 2.8 (p. 56) and the data on number of fatalities for the 62 energy-related accidents saved in the FATAL file. Construct a box plot for the data and use it to identify any outliers.

**Solution**

We used MINITAB to form a box plot for the fatalities data. The box plot is shown in Figure 2.17. Recall that description statistics for the data are shown in Figure 2.15.

From Figure 2.15, the lower and upper quartiles are $Q_L = 69.5$ and $Q_U = 185.8$, respectively. These values form the edges (hinges) of the box in Figure 2.17. (The median, $m = 106.5$, is shown inside the box with a horizontal line.) The interquartile range, $IQR = Q_U - Q_L = 185.8 - 69.5 = 116.3$, is used to form the fences and whiskers of the box plot. Several highly suspect outliers (identified by asterisks) are shown on Figure 2.17. There appear to be several outliers with values of around 500 fatalities, one with about 1000 fatalities, and one with 2500 fatalities. (*Note:* The largest outlier is the observation identified in Example 2.8.)

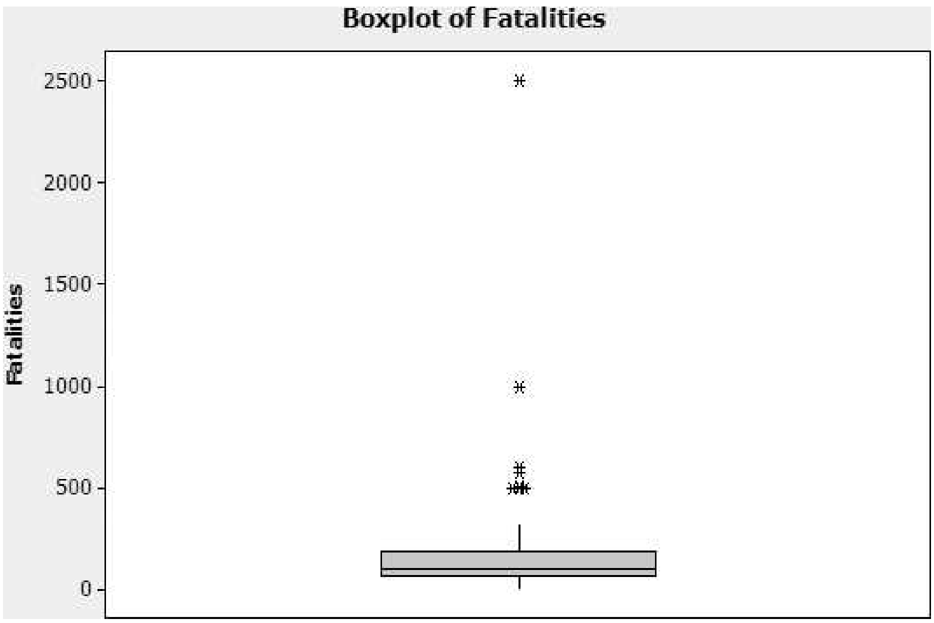



**FIGURE 2.17**
MINITAB Boxplot for Number of Energy-Related Fatal Accidents

The $z$-score and box plot methods both establish rule-of-thumb limits outside of which a $y$ value is deemed to be an outlier. Usually, the two methods produce similar results. However, the presence of one or more outliers in a data set can inflate the value of $s$ used to calculate the $z$-score. Consequently, it will be less likely that an errant observation would have a $z$-score larger than 3 in absolute value. In contrast, the values of the quartiles used to calculate the fences for a box plot are not affected by the presence of outliers.

### Rules of Thumb for Detecting Outliers*

| | Suspect Outliers | Highly Suspect Outliers |
|---|---|---|
| Box Plots: | Data points between inner and outer fences | Data points beyond outer fences |
| $z$-Scores: | $2 \leq \lvert z \rvert \leq 3$ | $\lvert z \rvert > 3$ |

* The $z$-score and box plot methods both establish rule-of-thumb limits outside of which a measurement is deemed to be an outlier. Usually, the two methods produce similar results. However, the presence of one or more outliers in a data set can inflate the computed value of $z$. Consequently, it will be less likely that an errant observation would have a $z$-score larger than 3 in absolute value. In contrast, the values of the quartiles used to calculate the intervals for a box plot are not affected by the presence of outliers.

## Applied Exercises

2.58 *Highest paid engineers.* Recall (from Exercise 2.26) that the mean base salary of a software engineering manager is \$126,417 (*Electronic Design's* 2012 Engineering Salary Survey). Assume (as in Exercises 2.38 and 2.51) that the distribution of base salaries for all software engineers is mound-shaped and symmetric with a standard deviation of \$15,000. Suppose a software engineering manager claims his salary is \$180,000. Is this claim believable? Explain.

2.59 *Barium content of clinkers.* Paving bricks—called clinkers—were examined for trace elements in order to determine the origin (e.g., factory) of the clinker. (*Advances in Cement Research*, Jan. 2004.) The barium content (mg/kg) for each in a sample of 200 clinkers was measured, yielding the following summary statistics: $Q_L = 115$, $m = 170$, and $Q_U = 260$.

a. Interpret the value of the median, $m$.
b. Interpret the value of the lower quartile, $Q_L$.
c. Interpret the value of the upper quartile, $Q_U$.
d. Find the interquartile range, IQR.
e. Find the endpoints of the inner fence in a box plot for barium content.
f. The researchers found no clinkers with a barium content beyond the boundaries of the inner fences. What does this imply?

2.60 *Characteristics of a rock fall.* Refer to the *Environmental Geology* (Vol. 58, 2009) study of how far a block from a collapsing rock wall will bounce, Exercise 2.29 (p. 43). The computer simulated rebound lengths (meters) for 13 block-soil impact marks left on a slope from an actual rock fall are reproduced in the next table. Do you detect any outliers in the data? Explain.

**ROCKFALL**

| | | | | | | |
|---|---|---|---|---|---|---|
| 10.94 | 13.71 | 11.38 | 7.26 | 17.83 | 11.92 | 11.87 |
| 5.44 | 13.35 | 4.90 | 5.85 | 5.10 | 6.77 | |

*Source:* Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy", *Environmental Geology,* Vol. 58, 2009. (Table 2.)

2.61 *Voltage sags and swells.* Refer to the *Electrical Engineering* (Vol. 95, 2013) study of power quality (measured by "sags" and "swells") in Turkish transformers, Exercise 2.53 (p. 54). For a sample of 103 transformers built for heavy industry, the mean and standard deviation of the number of sags per week was 353 and 30, respectively; also, the mean and standard deviation of the number of swells per week was 184 and 25, respectively. Consider a transformer that has 400 sags and 100 swells in a week.

a. Would you consider 400 sags per week unusual, statistically? Explain.
b. Would you consider 100 swells per week unusual, statistically? Explain.

**SHIPSANIT**

2.62 *Sanitation inspection of cruise ships.* Refer to the data on sanitation levels of cruise ships, Exercise 2.17 (p. 36).

a. Use the box plot method to detect any outliers in the data.
b. Use the $z$-score method to detect any outliers in the data.
c. Do the two methods agree? If not, explain why.

2.63 *Zinc phosphide in sugarcane.* A chemical company produces a substance composed of 98% cracked corn particles and 2% zinc phosphide for use in controlling rat populations in sugarcane fields. Production must be carefully controlled to maintain the 2% zinc phosphide because too much zinc phosphide will cause damage to the sugarcane and too little will be ineffective in controlling the rat population. Records from past production indicate that the distribution of the actual percentage of zinc phosphide present in the substance is approximately mound-shaped, with a mean of 2.0% and a standard deviation of .08%. Suppose one batch chosen randomly actually contains 1.80% zinc phosphide. Does this indicate that there is too little zinc phosphide in today's production? Explain your reasoning.

2.64 *Sensor motion of a robot.* Researchers at Carnegie Mellon University developed an algorithm for estimating the sensor motion of a robotic arm by mounting a camera with inertia sensors on the arm. (*The International Journal of Robotics Research*, Dec. 2004.) Two variables of interest were the error of estimating arm rotation (measured in radians) and the error of estimating arm translation (measured in centimeters). Data for 11 experiments are listed in the table. In each experiment, the perturbation of camera intrinsics and projections were varied.

**SENSOR**

| Trial | Perturbed Intrinsics | Perturbed Projections | Rotation Error (radians) | Translation Error (cm) |
|---|---|---|---|---|
| 1 | No | No | .0000034 | .0000033 |
| 2 | Yes | No | .032 | 1.0 |
| 3 | Yes | No | .030 | 1.3 |
| 4 | Yes | No | .094 | 3.0 |
| 5 | Yes | No | .046 | 1.5 |
| 6 | Yes | No | .028 | 1.3 |
| 7 | No | Yes | .27 | 22.9 |
| 8 | No | Yes | .19 | 21.0 |
| 9 | No | Yes | .42 | 34.4 |
| 10 | No | Yes | .57 | 29.8 |
| 11 | No | Yes | .32 | 17.7 |

*Source:* Strelow, D., and Singh, S., "Motion estimation form image and inertial measurements." *The International Journal of Robotics Research*, Vol. 23, No. 12, Dec. 2004 (Table 4).

a. Find $\bar{y}$ and $s$ for translation errors in trials with perturbed intrinsics but no perturbed projections.
b. Find $\bar{y}$ and $s$ for translation errors in trials with perturbed projections but no perturbed intrinsics.
c. A trial resulted in a translation error of 4.5 cm. Is this value an outlier for trials with perturbed intrinsics but no perturbed projections? For trials with perturbed projections but no perturbed intrinsics? What type of camera perturbation most likely occurred for this trial?

**SANDSTONE**

2.65 *Permeability of sandstone during weathering.* Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercises 2.33 and 2.41 (p. 50). Recall that slices of sandstone blocks were tested for permeability under three conditions: no exposure to any type of weathering (A), repeatedly sprayed with a 10% salt solution (B), and soaked in a 10% salt solution and dried (C).

a. Identify any outliers in the permeability measurements for Group A sandstone slices.
b. Identify any outliers in the permeability measurements for Group B sandstone slices.
c. Identify any outliers in the permeability measurements for Group C sandstone slices.
d. If you remove the outliers detected in parts a-c, how will descriptive statistics like the mean, median, and standard deviation be effected? If you are unsure of your answer, carry out the analysis.

## 2.8 Distorting the Truth with Descriptive Statistics

A picture may be "worth a thousand words," but pictures can also color messages or distort them. In fact, the pictures in statistics—histograms, bar charts, and other graphical descriptions—are susceptible to distortion, so we have to examine each of them with care. In this section, we begin by mentioning a few of the pitfalls to watch for when interpreting a chart or a graph. Then we discuss how numerical descriptive statistics can be used to distort the truth.

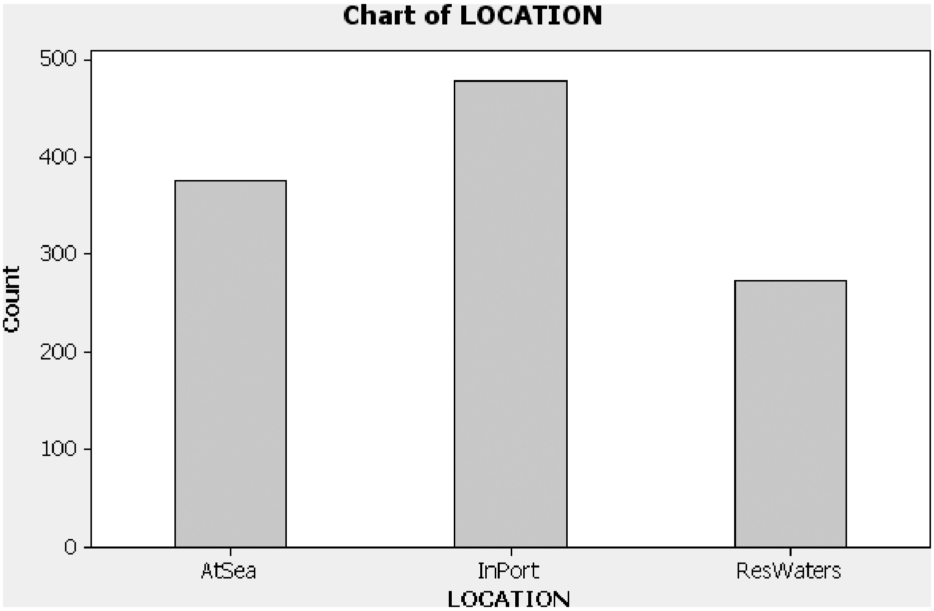
COLLISION

**TABLE 2.6 Collisions of Marine Vessels by Location**

| Location | Number of Ships |
|---|---|
| At Sea | 376 |
| Restricted Waters | 273 |
| In Port | 478 |
| **Total** | 1,127 |

One common way to change the impression conveyed by a graph is to change the scale on the vertical axis, the horizontal axis, or both. For example, consider the data on collisions of large marine vessels operating in European waters over a 5-year period summarized in Table 2.6. Figure 2.18 is a MINITAB bar graph showing the frequency of collisions for each of the three locations. The graph shows that "in port" collisions occur more often than collisions "at sea" or collisions in "restricted waters."

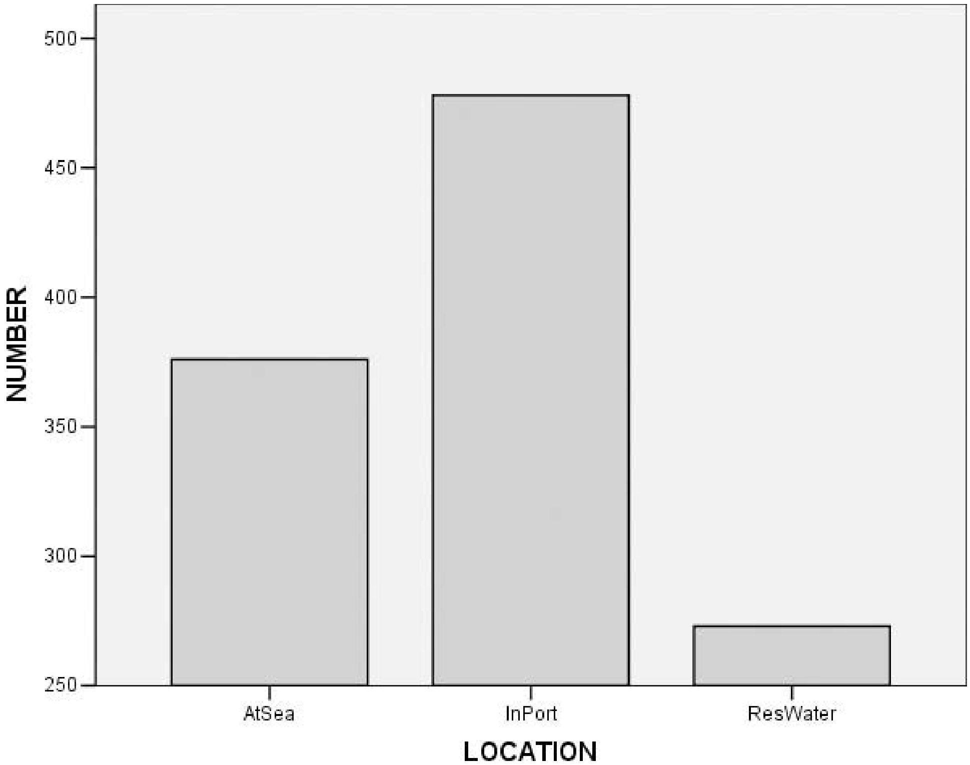
**FIGURE 2.18**
MINITAB bar graph of vessel collisions by location




Suppose you want to use the same data to exaggerate the difference between the number of "in port" collisions and the number of collisions in "restricted waters." One way to do this is to increase the distance between successive units on the vertical axis—that is, *stretch* the vertical axis by graphing only a few units per inch. A telltale sign of stretching is a long vertical axis, but this is often hidden by starting the vertical axis at some point above the origin, 0. Such a graph is shown in the SPSS printout, Figure 2.19. By starting the bar chart at 250 collisions (instead of 0), it appears that the frequency of "in port" collisions is many times larger than the frequency of collisions in "restricted waters."

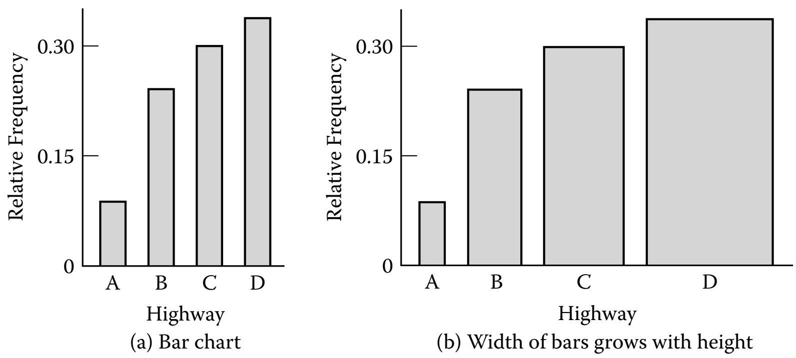
The changes in categories indicated by a bar graph can also be emphasized or deemphasized by stretching or shrinking the vertical axis. Another method of achieving visual distortion with bar graphs is by making the width of the bars proportional to height. For example, look at the bar chart in Figure 2.20a, which depicts the percentage of the total number of motor vehicle deaths in a year that occurred on each of four major highways. Now suppose we make both the width and the height grow as the percentage of fatal accidents grows. This change is shown in Figure 2.20b. The reader may tend to equate the *area* of the bars with the percentage of deaths occurring at each highway. But in fact, the true relative frequency of fatal accidents is proportional only to the *height* of the bars.

**FIGURE 2.19**
SPSS bar graph of vessel collisions by location—adjusted vertical axis




Although we've discussed only a few of the ways that graphs can be used to convey misleading pictures of phenomena, the lesson is clear. Look at all graphical descriptions of data with a critical eye. Particularly, check the axes and the size of the units on each axis. Ignore the visual changes and concentrate on the actual numerical changes indicated by the graph or chart.

**FIGURE 2.20**
Relative frequency of fatal motor vehicle accidents on each of four major highways




The information in a data set can also be distorted by using numerical descriptive measures. Consider the data on 62 energy-related accidents analyzed in Examples 2.8 and 2.9 (and saved in the **FATAL** file). Suppose you want a single number that best describes the "typical" number of fatalities that occur in such an accident. One choice is the mean number of fatalities. In Example 2.8 we found the mean to be $\bar{y} = 208.3$ fatalities. However, if you examine the data in the **FATAL** file, you will find that 48 of the 62 accidents (or 77%) had fatalities below the mean. In other words, the value of 208.3 is not very "typical" of the accidents in the data set. This is because (as we discussed in Section 2.4) the mean is inflated by the extreme values in a data set. Recall (Example 2.9) that one accident had 2500 fatalities and another had

**FIGURE 2.21**
MINITAB Descriptive Statistics for Number of Energy-Related Fatal Accidents

**Descriptive Statistics: Fatalities (All Data)**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| Fatalities | 62 | 208.3 | 344.6 | 11.0 | 69.5 | 106.5 | 185.8 | 2500.0 |

**Descriptive Statistics: Fatalities (1 Outlier Deleted)**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| Fatalities | 61 | 170.8 | 178.2 | 11.0 | 69.0 | 105.0 | 179.5 | 1000.0 |

**Descriptive Statistics: Fatalities (2 Outliers Deleted)**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| Fatalities | 60 | 156.9 | 142.9 | 11.0 | 68.5 | 103.0 | 176.3 | 600.0 |

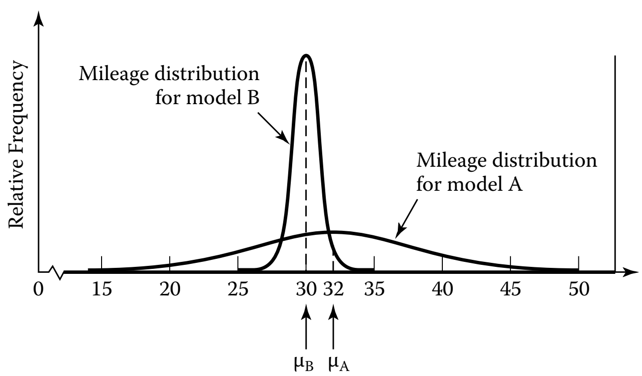
1000 fatalities. These two very atypical values inflate the mean. Figure 2.21 shows how the value of the mean changes as these outliers are removed from the data set. When the most deadly accident (2500 fatalities) is deleted, the mean drops to $\bar{y} = 170.8$. When both outliers are deleted, the mean drops to $\bar{y} = 156.9$.

A better measure of central tendency for the number of fatalities in the 62 energy-related accidents is the median. In Example 2.9 we found the median to be $m = 106.5$ fatalities. We know, by definition, that half the accidents have a fatality value below 106.5 and half above. Consequently, the median is more "typical" of the values in the data set. As you can see from Figure 2.21, the median is $m = 105$ when the largest outlier is deleted, and is $m = 103$ when both outliers are deleted. Thus, the median does not dramatically change as the largest observations in the data set are removed.

Another distortion of information in a sample occurs when *only* a measure of central tendency is reported. Both a measure of central tendency and a measure of variability are needed to obtain an accurate mental image of a data set. For example, suppose the Environmental Protection Agency (EPA) wants to rank two car models based on their estimated (mean) EPA city mileage ratings. Assume that model A has a mean EPA mileage rating of 32 miles per gallon and that model B has a mean EPA mileage rating of 30 miles per gallon. Based on the mean, the EPA should rank model A ahead of model B.

However, the EPA did not take into account the variability associated with the mileage ratings. As an extreme example, suppose that the standard deviation for model A is 5 miles per gallon, whereas that for model B is only 1 mile per gallon. If the mileages form a mound-shaped distribution, they might appear as shown in Figure 2.22.

**FIGURE 2.22**
Mileage distributions for two car models

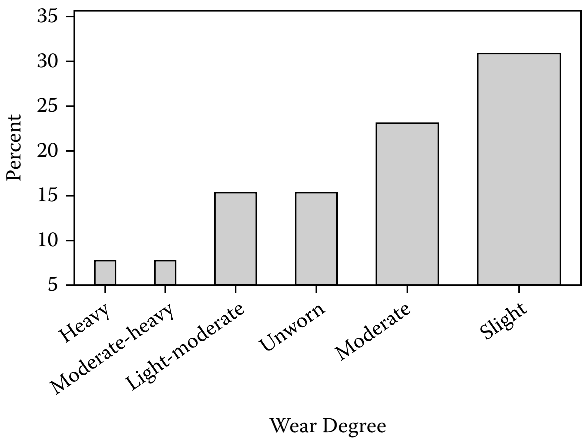


Note that the larger amount of variability associated with model A implies that a model A car is more likely to have a low mileage rating than a model B car. If the ranking is based on selecting the model with the lowest chance of a low mileage rating, model B will be ranked ahead of model A.

## Applied Exercises

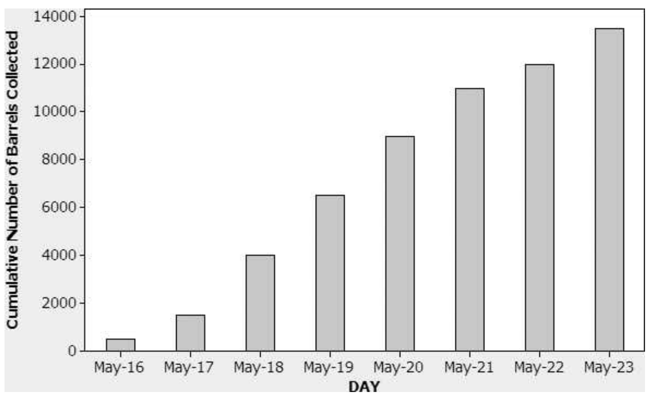
2.66 *Cheek teeth of extinct primates.* Refer to the *American Journal of Physical Anthropology* (Vol. 142, 2010) study of the dietary habits of extinct primates, Exercise 2.14 (p. 35). Recall that cheek teeth were extracted from skulls discovered in western Wyoming and analyzed for wear (unworn, slight, light-moderate, moderate, moderate-heavy, or heavy). A summary of the 13 teeth that could be classified is shown in the accompanying table. Consider the bar graph shown below. Identify two ways in which the bar graph might mislead the viewer by overemphasizing the importance of one of the performance measures.

| Wear Category | Number of teeth | Proportion |
|---|---|---|
| Heavy | 1 | .077 |
| Moderate-heavy | 1 | .077 |
| Moderate | 3 | .231 |
| Light-moderate | 2 | .154 |
| Slight | 4 | .308 |
| Unworn | 2 | .154 |




BPOIL

2.67 *BP oil leak.* In the summer of 2010, an explosion on the Deepwater Horizon oil drilling rig caused a leak in one of British Petroleum (BP) Oil Company's wells in the Gulf of Mexico. Crude oil rushed unabated for three straight months into the Gulf until BP could fix the leak. During the disaster, BP used suction tubes to capture some of the gushing oil. In May of 2010, a BP representative presented a graphic on the daily number of 42-gallon barrels (bbl) of oil collected by the suctioning process in an effort to demonstrate the daily improvement in the process. A MINITAB graphic similar to the one used by BP is shown below.




a. Note that the vertical axis represents the "cumulative" number of barrels collected per day. This is calculated by adding the amounts of the previous days' oil collection to the current day's oil collection. Explain why this graph is misleading.
b. Estimates of the actual number of barrels of oil collected per day for each of the 8 days are listed in the accompanying table. Construct a graph for this data that accurately depicts BP's progress in its daily collection of oil. What conclusions can you draw from the graph?

| Day | Number of Barrels (bbl) |
|---|---|
| May-16 | 500 |
| May-17 | 1,000 |
| May-18 | 3,000 |
| May-19 | 2,500 |
| May-20 | 2,500 |
| May-21 | 2,000 |
| May-22 | 1,000 |
| May-23 | 1,500 |

**PHISHING**

2.68 *Phishing attacks to email accounts.* Recall (Exercise 2.24, p. 38) that *phishing* is the term used to describe an attempt to extract personal/financial information from unsuspecting people through fraudulent email. Data from an actual phishing attack against an organization are provided in the **PHISHING** file. The company set up a publicized email account—called a "fraud box"—which enabled employees to notify them if they suspected an email phishing attack. The data represent interarrival times, i.e., the difference (in seconds) between the time of the actual phishing attack and the time when an employee notified the company of the attack, for 267 fraud box email notifications. The greater the "typical" interarrival time, the more likely the phishing attack was an "inside job" that originated within the company. Suppose the Technical Support Group at the company, investigating the phishing attack, will classify the attack as an inside job if the typical interarrival time is 80 or more seconds. Descriptive statistics for the interarrival times are shown in the MINITAB printout below. Based on the high mean value of 95.52 seconds, a technical support manager claims the phishing attack is an inside job. Do you agree?

**Descriptive Statistics: INTTIME**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| INTTIME | 267 | 95.52 | 91.54 | 1.86 | 30.59 | 70.88 | 133.34 | 513.52 |

## *STATISTICS IN ACTION REVISITED*

### Characteristics of Contaminated Fish in the Tennessee River, Alabama

**DDT**

We now return to the U.S. Army Corps of Engineers study of fish contaminated from the toxic discharges of a chemical plant once located on the banks of the Tennessee River in Alabama. The study data are saved in the **DDT** file.

The key questions to be answered are: Where (i.e., what river or creek) are the different species most likely to be captured? What is the typical weight and length of the fish? What is the level of DDT contamination of the fish? Does the level of contamination vary by species? These questions can be partially answered by applying the descriptive methods of this chapter. Of course, the method used will depend on the type (quantitative or qualitative) of the variable analyzed.

Consider, first, the qualitative variable species. A bar graph for species, produced using SAS, is shown in Figure SIA2.1. You can see that the majority (about 67%) of the captured fish were channel catfish, another 25% were smallmouth buffalofish, and about 8% were largemouth bass. To determine where these species of fish were captured, we examine the MINITAB pie charts in Figure SIA2.2. One pie chart is produced for each of the four river/creek locations. The charts show that the only species captured in the tributary creeks (LC, SC, or FC) was channel catfish. Since these creeks are closest to the reservoir and wildlife reserve, ecologists focused their investigation on wildlife that prey on channel catfish.

To examine the quantitative variables length, weight, and DDT level, we produced descriptive statistics for each variable by species. These statistics are shown in the SAS printout, Figure SIA2.3. Histograms of these variables for channel catfish are shown in the MINITAB printout, Figure SIA2.4. The histograms reveal mound-shaped, nearly symmetric distributions for the lengths and weights of channel catfish. Thus, we can apply the Empirical Rule to describe these distributions.

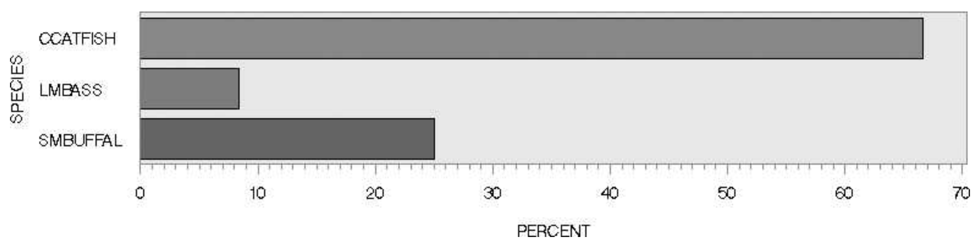



**FIGURE SIA2.1**
SAS horizontal bar graph for species of fish

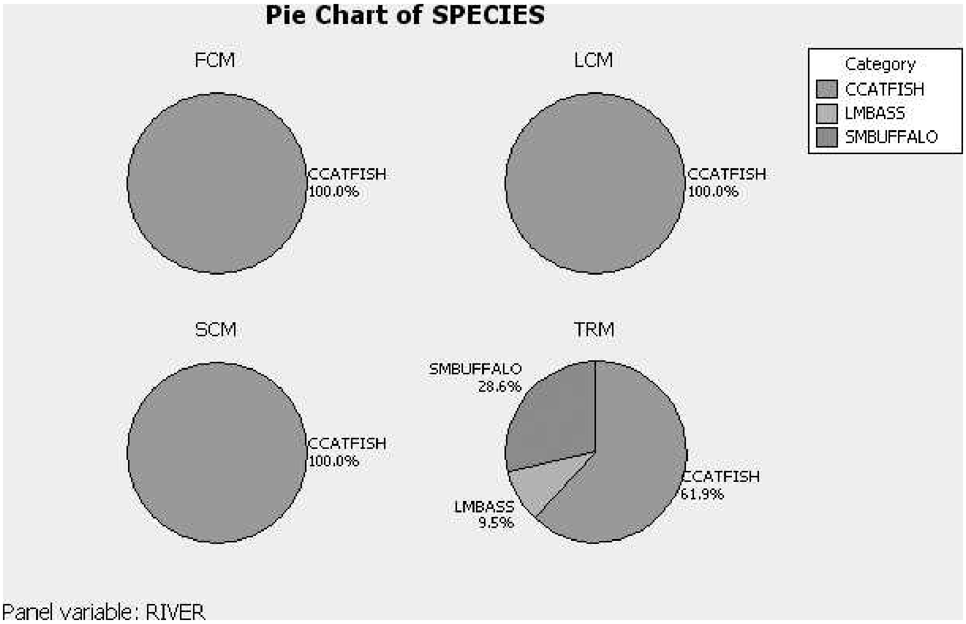



**FIGURE SIA2.2**
MINITAB pie charts of species by river

For channel catfish lengths, $\bar{y} = 44.73$ and $s = 4.58$. Therefore, about 95% of the channel catfish lengths fall in the interval $44.73 \pm 2(4.58)$, i.e., between 35.57 and 53.89 centimeters. For channel catfish weights, $\bar{y} = 987.3$ and $s = 262.7$. This implies that about 95% of the channel catfish weights fall in the interval $987.3 \pm 2(262.7)$, i.e., between 461.9 and 1512.7 grams.

The histogram at the bottom of Figure SIA2.4 shows that channel catfish DDT levels are highly skewed to the right. The skewness appears to be caused by a few extremely large DDT values. The SAS printout, Figure SIA2.3, shows that the largest (maximum) DDT level is 1100 ppm. Is this value an outlier? For channel catfish DDT levels, $\bar{y} = 33.3$ and $s = 119.5$. Thus, the $z$-score for this large DDT value is $z = (1100 - 33.3)/119.5 = 8.93$. Since it is extremely unlikely to find an observation in a data set that is almost 9 standard deviations from the mean, the DDT value is considered a highly suspect outlier. Some research by the U.S. Army Corps of Engineers revealed that this DDT value was correctly measured and recorded but that the fish was one of the few found at the exact location where the manufacturing plant was discharging its toxic waste materials into the water. Consequently, the researchers removed this observation from the data set and reanalyzed the DDT levels of channel catfish.

The MINITAB printout, Figure SIA2.5, gives summary statistics for channel catfish DDT levels when the outlier is deleted. Now, $\bar{y} = 22.1$ and $s = 46.8$. According to Chebyshev's Rule, at least 75% of the DDT levels for channel catfish will lie in the interval $22.1 \pm 2(46.8)$, i.e., between 0 and 115.7 ppm. Also, the SAS histogram for the reduced data set is shown in Figure SIA2.6. The histogram reveals that a large percentage of the DDT levels are above 5 ppm—the maximum level deemed safe by the Environmental Protection Agency. This provided further evidence for the ecologists to focus on wildlife that prey on channel catfish. ●

The MEANS Procedure

| SPECIES | N Obs | Variable | Mean | Std Dev | Minimum | Maximum | Median |
|---|---|---|---|---|---|---|---|
| CCATFISH | 96 | LENGTH | 44.7291667 | 4.5808047 | 29.5000000 | 52.0000000 | 45.0000000 |
| | | WEIGHT | 987.2916667 | 262.6813926 | 353.0000000 | 1770.00 | 983.5000000 |
| | | DDT | 33.2993750 | 119.4746053 | 0.7400000 | 1100.00 | 9.5500000 |
| LMBASS | 12 | LENGTH | 26.5416667 | 4.4795410 | 17.5000000 | 36.0000000 | 26.2500000 |
| | | WEIGHT | 629.0000000 | 324.7564122 | 173.0000000 | 1433.00 | 538.0000000 |
| | | DDT | 1.3800000 | 2.0426142 | 0.1100000 | 7.4000000 | 0.5300000 |
| SMBUFFAL | 36 | LENGTH | 43.1250000 | 5.4134423 | 32.5000000 | 52.0000000 | 46.0000000 |
| | | WEIGHT | 1356.42 | 436.7343995 | 520.0000000 | 2302.00 | 1440.50 |
| | | DDT | 8.1616667 | 11.2832608 | 0.2500000 | 48.0000000 | 4.6500000 |

**FIGURE SIA2.3**
SAS descriptive statistics by species of fish

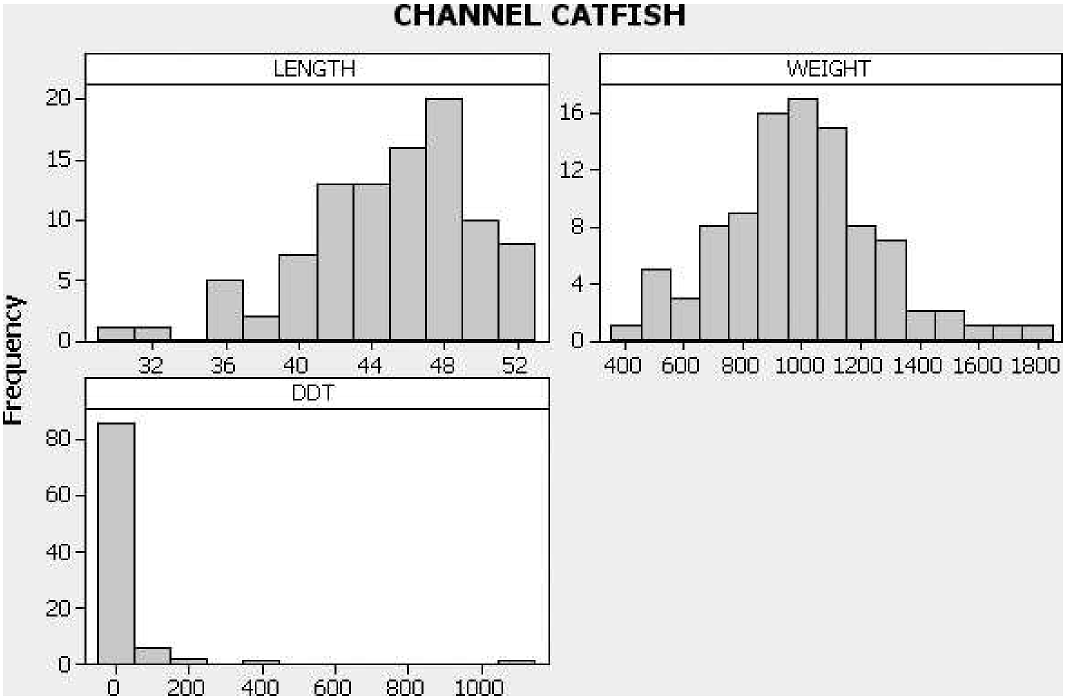



**FIGURE SIA2.4**
MINITAB histograms for channel catfish

**FIGURE SIA2.5**
MINITAB summary statistics for DDT levels of channel catfish, outlier deleted

**Descriptive Statistics: DDT**

| Variable | N | Mean | StDev | Minimum | Median | Maximum |
|---|---|---|---|---|---|---|
| DDT | 95 | 22.07 | 46.84 | 0.740 | 9.40 | 360.00 |

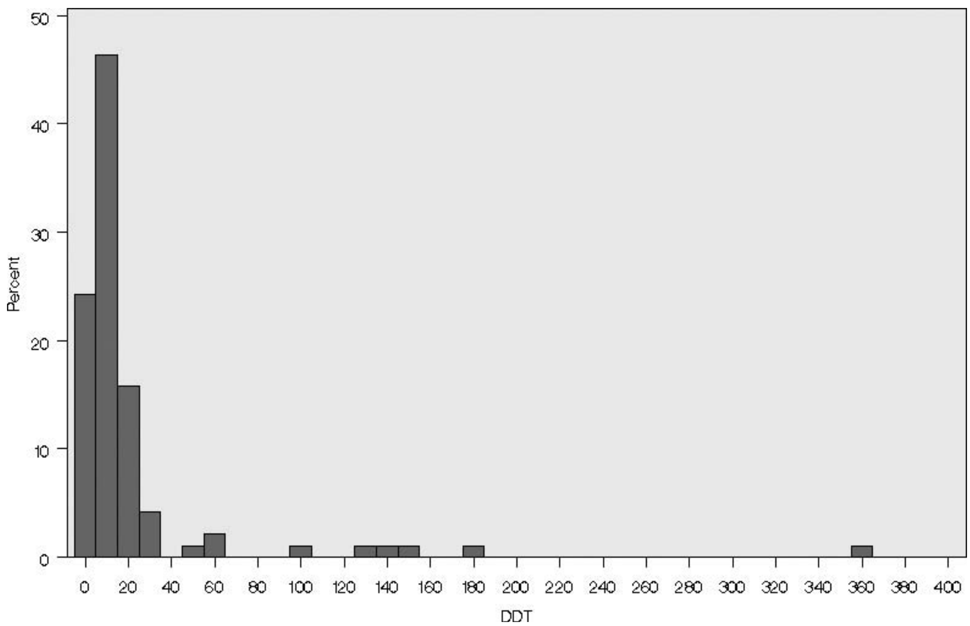



**FIGURE SIA2.6**
SAS histogram for DDT levels of channel catfish, outlier deleted

# Quick Review

## Key Terms



## Key Formulas

| | |
|---|---|
| $\dfrac{\text{Category frequency}}{n}$ | Category relative frequency 23 |
| $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$ | Sample mean 39 |
| $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \dfrac{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$ | Sample variance 46 |
| $s = \sqrt{s^2}$ | Sample standard deviation 46 |
| $z = \dfrac{y - \bar{y}}{s}$ | Sample $z$-score 52 |
| $z = \dfrac{y - \mu}{\sigma}$ | Population $z$-score 52 |
| $\text{IQR} = Q_U - Q_L$ | Interquartile range 56 |
| $Q_L - 1.5(\text{IQR})$ | Lower inner fence 56 |
| $Q_U + 1.5(\text{IQR})$ | Upper inner fence 56 |
| $Q_L - 3(\text{IQR})$ | Lower outer fence 57 |
| $Q_U + 3(\text{IQR})$ | Upper outer fence 57 |

## Chapter Summary Notes

- Graphical methods for qualitative data: **pie chart, bar graph**, and **Pareto diagram**
- Graphical methods for quantitative data: **dot plot, stem-and-leaf display**, and **histogram**
- Numerical measures of central tendency: **mean, median**, and **mode**

- Numerical measures of variation: **range, variance**, and **standard deviation**
- Sample numerical descriptive measures are called **statistics**.
- Population numerical descriptive measures are called **parameters**.
- Rules for determining the percentage of measurements in the interval (mean) ± 2 (std. dev.): **Chebyshev's Rule** (at least 75%) and **Empirical Rule** (approximately 95%)
- Measures of relative standing: **percentile score** and ***z*-score**
- Methods for detecting outliers: **box plots** and ***z*-scores**

## Supplementary Exercises

2.69 *Fate of scrapped tires.* According to the Rubber Manufacturers Association, there are approximately 300 million tires scrapped each year in the U.S. The summary table below describes the fate of these scrapped tires.

a. Identify the variable measured for each scrapped tire.
b. What are the classes (categories)?
c. Calculate the class relative frequencies.
d. Use the results, part c, to form a pie chart for the data.
e. Use the results, part c, to form a Pareto diagram for the data. Interpret the graph.

| Fate of Tires | Number (millions) |
|---|---|
| Burned for fuel | 155 |
| Recycled into new products | 96 |
| Exported | 7 |
| Land disposed | 42 |
| **Totals** | 300 |

*Source:* Rubber Manufacturers Association, May 2009 report.

2.70 *Microbial fuel cells.* A promising new technology for generating electricity uses microbial fuel cells (MFCs)—a product of natural human wastewaters. Over the past several years, research in employing MFCs for this purpose has dramatically increased. The graph below, extracted from the *Biochemical Engineering Journal* (Vol. 73, 2013), summarizes the areas of investigation for a sample of 54 recently published research articles on MFCs. (Note: Each of the 54 articles was classified according to a particular area of investigation, e.g., microbial metabolism, biocathodes, patents, etc.)

a. Identify the qualitative variable measured for each research article.
b. What type of graph is portrayed?
c. Convert the graph into a Pareto diagram. Then, use the diagram to identify the investigation area with the largest proportion of MFC research articles.

2.71 *Unsafe Florida roads.* In Florida, civil engineers are designing roads with the latest safety-oriented construction methods in response to the fact that more people in Florida are killed by bad roads than by guns. One year, a total of 135 traffic accidents that occurred was attributed to poorly

Graph for Exercise 2.70

constructed roads. A breakdown of the poor road conditions that caused the accidents is shown in the following table. Construct and interpret a Pareto diagram for the data.

**BADROADS**

| Poor Road Condition | Number of Fatalities |
|---|---|
| Obstructions without warning | 7 |
| Road repairs/Under construction | 39 |
| Loose surface material | 13 |
| Soft or low shoulders | 20 |
| Holes, ruts, etc. | 8 |
| Standing water | 25 |
| Worn road surface | 6 |
| Other | 17 |
| **Total** | 135 |

*Source:* Florida Department of Highway Safety and Motor Vehicles.

2.72 *Process voltage readings.* A Harris Corporation/University of Florida study was undertaken to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the old and new locations and voltage readings on the process were obtained. A "good process" was considered to be one with voltage readings of at least 9.2 volts (with larger readings being better than smaller readings). The table contains voltage readings for 30 production runs at each location.

**VOLTAGE**

| Old Location | | | New Location | | |
|---|---|---|---|---|---|
| 9.98 | 10.12 | 9.84 | 9.19 | 10.01 | 8.82 |
| 10.26 | 10.05 | 10.15 | 9.63 | 8.82 | 8.65 |
| 10.05 | 9.80 | 10.02 | 10.10 | 9.43 | 8.51 |
| 10.29 | 10.15 | 9.80 | 9.70 | 10.03 | 9.14 |
| 10.03 | 10.00 | 9.73 | 10.09 | 9.85 | 9.75 |
| 8.05 | 9.87 | 10.01 | 9.60 | 9.27 | 8.78 |
| 10.55 | 9.55 | 9.98 | 10.05 | 8.83 | 9.35 |
| 10.26 | 9.95 | 8.72 | 10.12 | 9.39 | 9.54 |
| 9.97 | 9.70 | 8.80 | 9.49 | 9.48 | 9.36 |
| 9.87 | 8.72 | 9.84 | 9.37 | 9.64 | 8.68 |

*Source:* Harris Corporation, Melbourne, FL.

a. Construct a relative frequency histogram for the voltage readings of the old process.
b. Construct a stem-and-leaf display for the voltage readings of the old process. Which of the two graphs in parts **a** and **b** is more informative about where most of the voltage readings lie?
c. Construct a relative frequency histogram for the voltage readings of the new process.
d. Compare the two graphs in parts **a** and **c**. (You may want to draw the two histograms on the same graph.) Does it appear that the manufacturing process can be established locally (i.e., is the new process as good as or better than the old)?
e. Find and interpret the mean, median, and mode for each of the voltage readings data sets. Which is the preferred measure of central tendency? Explain.
f. Calculate the $z$-score for a voltage reading of 10.50 at the old location.
g. Calculate the $z$-score for a voltage reading of 10.50 at the new location.
h. Based on the results of parts **f** and **g**, at which location is a voltage reading of 10.50 more likely to occur? Explain.
i. Construct a box plot for the data at the old location. Do you detect any outliers?
j. Use the method of $z$-scores to detect outliers at the old location.
k. Construct a box plot for the data at the new location. Do you detect any outliers?
l. Use the method of $z$-scores to detect outliers at the new location.
m. Compare the distributions of voltage readings at the two locations by placing the box plots, parts **i** and **k**, side by side vertically.

2.73 *Surface roughness of pipe.* Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated oil field pipes, Exercise 2.20 (p. 37). The data (in micrometers) are repeated in the table. Give an interval that will likely contain about 95% of all coated pipe roughness measurements.

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T., "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

**CRASH**

2.74 *Crash tests on new cars.* Each year, the National Highway Traffic Safety Administration (NHTSA) crash tests new car models to determine how well they protect the driver and front-seat passenger in a head-on collision. The NHTSA has developed a "star" scoring system for the frontal crash test, with results ranging from one star (*) to five stars (*****). The more stars in the rating, the better the level of crash protection in a head-on collision. The NHTSA crash test results for 98 cars in a recent model year are stored in the data file named **CRASH**. The driver-side star ratings for the 98 cars are summarized

**TILLRATIO**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.75 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 |
| 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 | 2.73 | 2.95 | 2.25 | 2.73 | 2.55 | 3.06 |

*Source:* Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

in the following MINITAB printout. Use the information in the printout to form a pie chart. Interpret the graph.

**Tally for Discrete Variables: DRIVSTAR**

| DRIVSTAR | Count | Percent |
|---|---|---|
| 2 | 4 | 4.08 |
| 3 | 17 | 17.35 |
| 4 | 59 | 60.20 |
| 5 | 18 | 18.37 |
| N= | 98 | |

2.75 *Crash tests on new cars.* Refer to Exercise 2.74 and the NHTSA crash test data. One quantitative variable recorded by the NHTSA is driver's severity of head injury (measured on a scale from 0 to 1500). The mean and standard deviation for the 98 driver head-injury ratings in the **CRASH** file are displayed in the MINITAB printout at the bottom of the page. Use these values to find the $z$-score for a driver head-injury rating of 408. Interpret the result.

2.76 *Chemical makeup of glacial drifts.* Refer to the *American Journal of Science* (Jan., 2005) study of the chemical makeup of buried glacial drifts (tills), Exercise 2.22 (p. 38). The data on the Al/Be ratios for a sample of 26 buried till specimens are repeated in the table at the top of the page.

a. Compute and interpret three numerical descriptive measures of central tendency for the Al/Be ratios.
b. Compute and interpret three numerical descriptive measures of variation for the Al/Be ratios.
c. Construct a box plot for the data. Do you detect any outliers?

2.77 *Red dye in gasoline.* Dyes are used in coloration products, such as textiles, paper, leather, and foodstuffs, and are required by law to be in gasoline to indicate the presence of lead. To monitor environmental contamination, analytical methods must be developed to identify and quantify these dyes. In one study, thermospray high-performance liquid chromatography/mass spectrometry was used to characterize dyes in wastewater and gasoline. The next table gives the relative abundance (relative frequency of occurrence) of commercial Diazo Red dye components in gasoline. Describe the relative abundance of red dye compounds with a bar graph. Interpret the graph.

**REDDYE**

| Red Dye Compound | Relative Abundance |
|---|---|
| H | .021 |
| $CH_3$ | .210 |
| $C_2H_5$ | .354 |
| $C_3H_7$ | .072 |
| $C_7H_{15}$ | .054 |
| $C_8H_{17}$ | .127 |
| $C_9H_{19}$ | .118 |
| $C_{10}H_{21}$ | .025 |
| Others | .019 |

2.78 *Deep-hole drilling.* Refer to the *Journal of Engineering for Industry* (May 1993) study of deep hole drilling described in Exercise 1.26 (p. 20). An analysis of drill chip congestion was performed using data generated via computer simulation. The simulated distribution of the length (in millimeters) of 50 drill chips is displayed in a frequency histogram, shown on the top of the next page.

a. Convert the frequency histogram into a relative frequency histogram.
b. Based on the graph in part **a**, would you expect to observe a drill chip with a length of at least 190 mm? Explain.

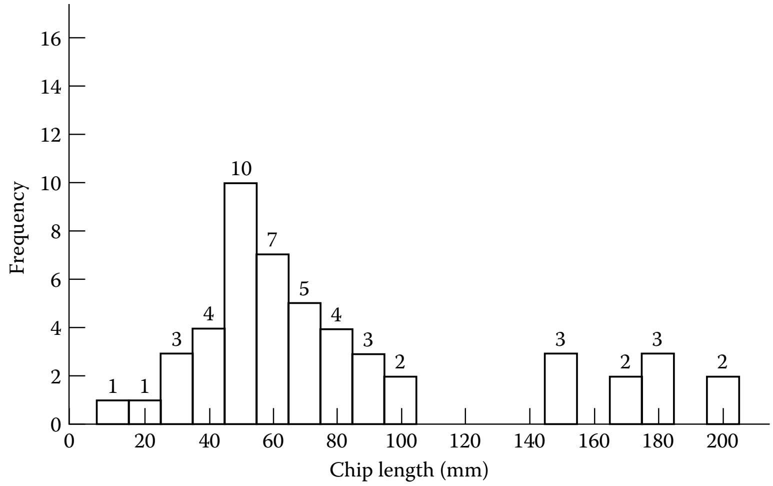
2.79 *Lumpy iron ore.* Sixty-six bulk specimens of Chilean lumpy iron ore (95% particle size, 150 millimeters) were randomly sampled from a 35,325-long-ton shipload of ore, and the percentage of iron in each ore specimen was determined. The data are shown in the table on p. 72.

a. Describe the population from which the sample was selected.
b. Give one possible objective of this sampling procedure.
c. Construct a relative frequency histogram for the data.
d. Calculate $\bar{y}$ and $s$.

MINITAB output for Exercise 2.75

**Descriptive Statistics: DRIVHEAD**

| Variable | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| DRIVHEAD | 98 | 603.7 | 185.4 | 216.0 | 475.0 | 605.0 | 724.3 | 1240.0 |




Frequency histogram for Exercise 2.78

*Source:* Chin, Jih-Hua, et al. "The computer simulation and experimental analysis of chip monitoring for deep hole drilling." *Journal of Engineering for Industry, Transactions of the ASME*, Vol. 115, May 1993, p. 187 (Figure 12).

**LUMPYORE**

| | | |
|---|---|---|
| 62.66 | 61.82 | 62.24 |
| 62.87 | 63.01 | 63.43 |
| 63.22 | 63.01 | 62.87 |
| 63.01 | 62.80 | 63.64 |
| 62.10 | 62.80 | 63.92 |
| 63.43 | 63.01 | 63.71 |
| 63.22 | 62.10 | 63.64 |
| 63.57 | 63.29 | 64.06 |
| 61.75 | 63.37 | 62.73 |
| 63.15 | 61.75 | 62.52 |
| 63.08 | 63.29 | 62.10 |
| 63.22 | 62.38 | 63.29 |
| 63.22 | 62.59 | 63.01 |
| 63.08 | 63.92 | 63.36 |
| 62.87 | 63.29 | 63.08 |
| 61.68 | 63.57 | 62.03 |
| 62.45 | 62.80 | 64.34 |
| 62.10 | 62.31 | 64.06 |
| 62.87 | 63.01 | 62.87 |
| 62.87 | 62.94 | 63.50 |
| 62.94 | 63.08 | 63.78 |
| 62.38 | 63.43 | 62.10 |

e. Find the percentage of the total number ($n = 66$) of observations that lie in the interval $\bar{y} \pm 2s$. Does this percentage agree with the Empirical Rule?

f. Find the 25th, 50th, 75th, and 90th percentiles for the data set. Interpret these values.

2.80 *Mongolian desert ants.* The *Journal of Biogeography* (Dec. 2003) published an article on the first comprehensive study of ants in Mongolia (Central Asia). Botanists placed seed baits at 11 study sites and observed the ant species attracted to each site. Some of the data recorded at each study site are provided in the table at the top of p. 73.

a. Find the mean, median, and mode for the number of ant species discovered at the 11 sites. Interpret each of these values.

b. Which measure of central tendency would you recommend to describe the center of the number of ant species distribution? Explain.

c. Find the mean, median, and mode for the total plant cover percentage at the 5 Dry Steppe sites only.

d. Find the mean, median, and mode for the total plant cover percentage at the 6 Gobi Desert sites only.

e. Based on the results, parts **c** and **d**, does the center of the total plant cover percentage distribution appear to be different at the two regions?

2.81 *Unplanned nuclear scrams.* *Scram* is the term used by nuclear engineers to describe a rapid emergency shutdown

Data for Exercise 2.80

**GOBIANTS**

| Site | Region | Annual Rainfall (mm) | Max. Daily Temp. (°C) | Total Plant Cover (%) | Number of Ant Species | Species Diversity Index |
|---|---|---|---|---|---|---|
| 1 | Dry Steppe | 196 | 5.7 | 40 | 3 | .89 |
| 2 | Dry Steppe | 196 | 5.7 | 52 | 3 | .83 |
| 3 | Dry Steppe | 179 | 7.0 | 40 | 52 | 1.31 |
| 4 | Dry Steppe | 197 | 8.0 | 43 | 7 | 1.48 |
| 5 | Dry Steppe | 149 | 8.5 | 27 | 5 | .97 |
| 6 | Gobi Desert | 112 | 10.7 | 30 | 49 | .46 |
| 7 | Gobi Desert | 125 | 11.4 | 16 | 5 | 1.23 |
| 8 | Gobi Desert | 99 | 10.9 | 30 | 4 | . |
| 9 | Gobi Desert | 125 | 11.4 | 56 | 4 | .76 |
| 10 | Gobi Desert | 84 | 11.4 | 22 | 5 | 1.26 |
| 11 | Gobi Desert | 115 | 11.4 | 14 | 4 | .69 |

*Source:* Pfeiffer, M., et al., "Community organization and species richness of ants in Mongolia along an ecological gradient from steppe to Gobi Desert." *Journal of Biogeography*, Vol. 30, No. 12, Dec. 2003 (Tables 1 and 2).

of a nuclear reactor. The nuclear industry has made a concerted effort to significantly reduce the number of unplanned scrams. The accompanying table gives the number of scrams at each of 56 U.S. nuclear reactor units in a recent year. Would you expect to observe a nuclear reactor in the future with 11 unplanned scrams? Explain.

**SCRAMS**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | 1 | 4 | 2 | 10 | 6 | 5 | 2 | 0 | 3 | 1 | 5 |
| 4 | 2 | 7 | 12 | 0 | 3 | 8 | 2 | 0 | 9 | 3 | 3 | 4 | 7 |
| 2 | 4 | 5 | 3 | 2 | 7 | 13 | 4 | 2 | 3 | 3 | 7 | 0 | 9 |
| 4 | 3 | 5 | 2 | 7 | 8 | 5 | 2 | 4 | 3 | 4 | 0 | 1 | 7 |

2.82 *Work measurement data.* Industrial engineers periodically conduct "work measurement" analyses to determine the time required to produce a single unit of output. At a large processing plant, the number of total worker-hours required per day to perform a certain task was recorded for 50 days. The data are shown in the next table.

a. Compute the mean, median, and mode of the data set.
b. Find the range, variance, and standard deviation of the data set.
c. Construct the intervals $\bar{y} \pm s$, $\bar{y} \pm 2s$, and $\bar{y} \pm 3s$. Count the number of observations that fall within each interval and find the corresponding proportions. Compare the results to the Empirical Rule. Do you detect any outliers?
d. Construct a box plot for the data. Do you detect any outliers?
e. Find the 70th percentile for the data on total daily worker-hours. Interpret its value.

**WORKHRS**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 128 | 119 | 95 | 97 | 124 | 128 | 142 | 98 | 108 | 120 |
| 113 | 109 | 124 | 132 | 97 | 138 | 133 | 136 | 120 | 112 |
| 146 | 128 | 103 | 135 | 114 | 109 | 100 | 111 | 131 | 113 |
| 124 | 131 | 133 | 131 | 88 | 118 | 116 | 98 | 112 | 138 |
| 100 | 112 | 111 | 150 | 117 | 122 | 97 | 116 | 92 | 122 |

2.83 *Oil spill impact on seabirds.* The *Journal of Agricultural, Biological, and Environmental Statistics* (Sept. 2000) published a study on the impact of the *Exxon Valdez* tanker oil spill on the seabird population in Prince William Sound, Alaska. A subset of the data analyzed is stored in the EVOS file. Data were collected on 96 shoreline locations (called transects) of constant width but variable length. For each transect, the number of seabirds found is recorded as well as the length (in kilometers) of the transect and whether or not the transect was in an oiled area. (The first five and last five observations in the **EVOS** file are listed in the table on page 74.)

**Descriptive Statistics: Density**

```
Variable  Oil   N   Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Density   no   36   3.27   6.70    0.000  0.000   0.890   3.87    36.23
          yes  60  3.495  5.968   0.0000  0.000   0.700  5.233   32.836
```

Data for Exercise 2.83

**EVOS**

(Selected observations)

| Transect | Seabirds | Length | Oil |
|---|---|---|---|
| 1 | 0 | 4.06 | No |
| 2 | 0 | 6.51 | No |
| 3 | 54 | 6.76 | No |
| 4 | 0 | 4.26 | No |
| 5 | 14 | 3.59 | No |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 92 | 7 | 3.40 | Yes |
| 93 | 4 | 6.67 | Yes |
| 94 | 0 | 3.29 | Yes |
| 95 | 0 | 6.22 | Yes |
| 96 | 27 | 8.94 | Yes |

*Source:* McDonald, T. L., Erickson, W. P. and McDonald, L. L., "Analysis of count data from before-after control-impact studies." *Journal of Agricultural, Biological, and Environmental Statistics*, Vol 5, No. 3, Sept. 2000, pp.277–278 (Table A.1).

a. Identify the variables measured as quantitative or qualitative.
b. Identify the experimental unit.
c. Use a pie chart to describe the percentage of transects in oiled and unoiled areas.
d. Use a graphical method to examine the relationship between observed number of seabirds and transect length.
e. Observed seabird density is defined as the observed count divided by the length of the transect. MINITAB descriptive statistics for seabird densities in unoiled and oiled transects are displayed in the printout shown at the bottom of page 73. Assess whether the distribution of seabird densities differs for transects in oiled and unoiled areas.
f. For unoiled transects, give an interval of values that is likely to contain at least 75% of the seabird densities.
g. For oiled transects, give an interval of values that is likely to contain at least 75% of the seabird densities.
h. Which type of transect, an oiled or unoiled one, is more likely to have a seabird density of 16? Explain.

2.84 *Speed of light from galaxies.* Astronomers theorize that cold dark matter (CDM) caused the formation of galaxies. The theoretical CDM model requires an estimate of the velocity of light emitted from the galaxy. *The Astronomical Journal* (July, 1995) published a study of galaxy velocities. One galaxy, named A1775, is thought to be a *double cluster*; that is, two clusters of galaxies in close proximity. Fifty-one velocity observations (in kilometers per second, km/s) from cluster A1775 are listed in the table.

**GALAXY2**

| | | | | | | |
|---|---|---|---|---|---|---|
| 22922 | 20210 | 21911 | 19225 | 18792 | 21993 | 23059 |
| 20785 | 22781 | 23303 | 22192 | 19462 | 19057 | 23017 |
| 20186 | 23292 | 19408 | 24909 | 19866 | 22891 | 23121 |
| 19673 | 23261 | 22796 | 22355 | 19807 | 23432 | 22625 |
| 22744 | 22426 | 19111 | 18933 | 22417 | 19595 | 23408 |
| 22809 | 19619 | 22738 | 18499 | 19130 | 23220 | 22647 |
| 22718 | 22779 | 19026 | 22513 | 19740 | 22682 | 19179 |
| 19404 | 22193 | | | | | |

*Source:* Oegerle, W. R., Hill, J. M., and Fitchett, M. J., "Observations of high dispersion clusters of galaxies: Constraints on cold dark matter." *The Astronomical Journal*, Vol. 110, No. 1, July 1995, p. 34 (Table 1). p. 37 (Figure 1).

a. Use a graphical method to describe the velocity distribution of galaxy cluster A1775.
b. Examine the graph, part **a**. Is there evidence to support the double cluster theory? Explain.
c. Calculate numerical descriptive measures (e.g., mean and standard deviation) for galaxy velocities in cluster A1775. Depending on your answer to part **b**, you may need to calculate two sets of numerical descriptive measures, one for each of the clusters (say, A1775A and A1775B) within the double cluster.
d. Suppose you observe a galaxy velocity of 20,000 km/s. Is this galaxy likely to belong to cluster A1775A or A1775B? Explain.

**OILSPILL**

2.85 *Hull failures of oil tankers.* Owing to several major ocean oil spills by tank vessels, Congress passed the 1990 Oil Pollution Act, which requires all tankers to be designed with thicker hulls. Further improvements in the structural design of a tank vessel have been proposed since then, each with the objective of reducing the likelihood of an oil spill and decreasing the amount of outflow in the event of a hull puncture. To aid in this development, *Marine Technology* (Jan. 1995) reported on the spillage amount (in thousands of metric tons) and cause of puncture for 50 recent major oil spills from tankers and carriers. [*Note:* Cause of puncture is classified as either collision (C), fire/explosion (FE), hull failure (HF), or grounding (G).] The data are saved in the **OILSPILL** file.

a. Use a graphical method to describe the cause of oil spillage for the 50 tankers. Does the graph suggest that any one cause is more likely to occur than any other? How is this information of value to the design engineers?
b. Find and interpret descriptive statistics for the 50 spillage amounts. Use this information to form an interval that can be used to predict the spillage amount of the next major oil spill.

2.86 *Manual materials handling.* Engineers have a team for unaided human acts of lifting, lowering, pushing, pulling, carrying, or holding and releasing an object—*manual materials handling activities (MMHA).* M. M. Ayoub, et al. (1980) have attempted to develop strength and capacity guidelines for MMHA. The authors point out that a clear distinction between strength and capacity must be made: "Strength implies what a person can do in a single attempt, whereas capacity implies what a person can do for an extended period of time. Lifting strength, for example, determines the amount that can be lifted at frequent intervals." The accompanying table presents a portion of the recommendations of Ayoub, et al. for the lifting capacities of males and females. It gives the means and standard deviations of the maximum weight (in kilograms) of a box 30 centimeters wide that can be safely lifted from the floor to knuckle height at two different lift rates—1 lift per minute and 4 lifts per minute.

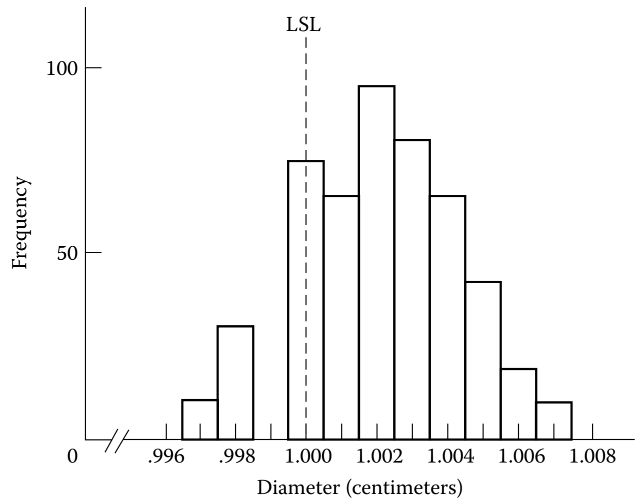
| Gender | Lifts/Minute | Mean | Standard Deviation |
|---|---|---|---|
| Male | 1 | 30.25 | 8.56 |
| | 4 | 23.83 | 6.70 |
| Female | 1 | 19.79 | 3.11 |
| | 4 | 15.82 | 3.23 |

*Source*: Ayoub, M. M., Mital, A., Bakken, G. M., Asfour, S. S., and Bethea, N. J., "Development of strength capacity norms for manual materials handling activities: The state of the art." *Human Factors*, June 1980, Vol. 22, pp. 271–283. Copyright 1980 by the Human Factors Society, Inc. and reproduced by permission.

a. Roughly sketch the relative frequency distribution of maximum recommended weight of lift for each of the four gender/lifts-per-minute combinations. The Empirical Rule will help you do this.
b. Construct the interval $\bar{y} \pm 2s$ for each of the four data sets and give the approximate proportion of measurements that fall within the interval.
c. Assuming the MMHA recommendations of Ayoub et al. are reasonable, would you expect that an average male could safely lift a box (30 centimeters wide) weighing 25 kilograms from the floor to knuckle height at a rate of 4 lifts per minute? An average female? Explain.

2.87 *Steel rod quality.* In his essay "Making Things Right," W. Edwards Deming considered the role of statistics in the quality control of industrial products.* In one example, Deming examined the quality control process for a manufacturer of steel rods. Rods produced with diameters smaller than 1 centimeter fit too loosely in their bearings and ultimately must be rejected (thrown out). To determine whether the diameter setting of the machine that produces the rods is correct, 500 rods are selected from the day's production and their diameters are recorded. The distribution of the 500 diameters for one day's production is shown in the figure below. Note that the symbol LSL in the figure represents the 1-centimeter lower specification limit of the steel rod diameters.

a. What type of data, quantitative or qualitative, does the figure portray?
b. What type of graphical method is being used to describe the data?
c. Use the figure to estimate the proportion of rods with diameters between 1.0025 and 1.0045 centimeters.
d. There has been speculation that some of the inspectors are unaware of the trouble that an undersized rod diameter would cause later in the manufacturing process. Consequently, these inspectors may be passing rods with diameters that were barely below the lower specification limit and recording them in the interval centered at 1.000 centimeter. According to the figure, is there any evidence to support this claim? Explain.




*From Tanur, J., et al., eds. *Statistics: A Guide to the Unknown*. San Francisco: Holden-Day, 1978. pp. 279–81.

CHAPTER 3

# Probability

**OBJECTIVE**

*To present an introduction to the theory of probability and to suggest the role that probability will play in statistical inference*

**CONTENTS**

## *STATISTICS IN ACTION:*
## Assessing Predictors of Software Defects in NASA Spacecraft Instrument Code

Software engineers are responsible for testing and evaluating computer software code. Generally, the more rigorous the evaluation, the higher the costs. Given finite budgets, software engineers usually focus on code that is believed to be the most critical. Consequently, this leaves portions of software code – called "blind spots" – that may contain undetected defects. For example, the *Journal of Systems and Software* (Feb., 2003) reported on faulty ground software with NASA deep-space satellites. NASA engineers had focused on the more critical flight software code; however, the faulty ground software was not collecting data from the flight software correctly, leading to critical problems with the satellites.

This issue of "blind spots" in software code evaluation was recently addressed by professors Tim Menzies and Justin DiStefano of the Department of Computer Science & Electrical Engineering at West Virginia University.* The researchers also developed some guidelines for assessing different methods of detecting software defects.** The methods were applied to multiple data sets, one of which is the focus of this *Statistics in Action* application. The data, saved in the **SWDEFECTS** file, is publicly available at the PROMISE Software Engineering Repository hosted by the School of Information Technology and Engineering, University of Ottawa. The data contains 498 modules of software code written in "C" language for a NASA spacecraft instrument.

For each module, the software code was evaluated, line-by-line, for defects and classified as "true" (i.e., module has defective code) or "false" (i.e., module has correct code). Because line-by-line code checking is very time consuming and expensive, the researchers considered some simple, easy-to-apply, algorithms for predicting whether or not a module has defects. For example, a simple algorithm is to count the lines of code in the module; any module with, say, 100 or more lines of code is predicted to have a defect. A list of several of the prediction methods considered is provided in Table SIA3.1. The **SWDEFECTS** file contains a variable that corresponds to each method. When the method predicts a defect, the corresponding variable's value is "yes". Otherwise, it is "no".

**SWDEFECTS**

**TABLE SIA3.1 Software Defect Prediction Algorithms**

| Method | Defects Algorithm | Definitions |
|---|---|---|
| Lines of code | LOC $>$ 50 | LOC $=$ lines of code |
| Cyclomatic complexity | $v(g) > 10$ | $v(g) =$ number of linearly independent paths |
| Essential complexity | $ev(g) \geq 14.5$ | $ev(g) =$ number of subflow graphs with D-structured primes |
| Design complexity | $iv(g) \geq 9.2$ | $iv(g) =$ cyclomatic complexity of module's reduced flow graph |

Software engineers evaluate these defect prediction algorithms by computing several probability measures, called *accuracy, detection rate, false alarm rate*, and *precision*. In the Statistics in Action at the end of this chapter, we demonstrate how to compute these probabilities.

*Menzies, T. & DiStefano, J. "How good is your blind spot sampling policy?", *8th IEEE International Symposium on High Assurance Software Engineering*, March 2004.
**Menzies, T., DiStefano, J., Orrego, A., & Chapman, R. "Assessing predictors of software defects", *Proceedings, Workshop on Predictive Software Models*, Chicago, 2004.

## 3.1 The Role of Probability in Statistics

If you play poker, a popular gambling game, you know that whether you win in any one game is an outcome that is very uncertain. Similarly, investing in an oil exploration company is a venture whose success is subject to uncertainty. (In fact, some would argue that investing is a form of educated gambling—one in which knowledge, experience, and good judgment can improve the odds of winning.)

Much like playing poker and investing, making inferences based on sample data is also subject to uncertainty. A sample rarely tells a perfectly accurate story about the population from which it was selected. There is always a margin of error (as the pollsters tell us) when sample data are used to estimate the proportion of people in favor of a particular political candidate or some consumer product. Similarly, there is always uncertainty about how far the sample estimate of the mean diameter of molded rubber expansion joints selected off an assembly line will depart from the true population mean. Consequently, a measure of the amount of uncertainty associated with an estimate (which we called *the reliability of an inference* in Chapter 1) plays a major role in statistical inference.

How do we measure the uncertainty associated with events? Anyone who has observed a daily newscast can answer that question. The answer is probability. For example, it may be reported that the probability of rain on a given day is 20%. Such a statement acknowledges that it is uncertain whether it will rain on the given day and indicates that the forecaster measures the likelihood of its occurrence as 20%.

Probability also plays an important role in decision making. To illustrate, suppose you have an opportunity to invest in an oil exploration company. Past records show that for 10 out of 10 previous oil drillings (a sample of the company's experiences), all 10 resulted in dry wells. What do you conclude? Do you think the chances are better than 50–50 that the company will hit a producing well? Should you invest in this company? We think your answer to these questions will be an emphatic "no." If the company's exploratory prowess is sufficient to hit a producing well 50% of the time, a record of 10 dry wells out of 10 drilled is an event that is just too improbable. Do you agree?

In this chapter, we will examine the meaning of probability and develop some properties of probability that will be useful in our study of statistics.

## 3.2 Events, Sample Spaces, and Probability

We will begin the discussion of probability with simple examples that are easily described, thus eliminating any discussion that could be distracting. With the aid of simple examples, important definitions are introduced and the notion of probability is more easily developed.

Suppose a coin is tossed once and the up face of the coin is recorded. This is an **observation**, or **measurement**. Any process of obtaining or generating an observation is called an **experiment**. Our definition of experiment is broader than that used in the physical sciences, where you would picture test tubes, microscopes, etc. Other, more practical examples of statistical experiments are recording whether a customer prefers one of two brands of smart phones, recording a voter's opinion on an important environmental issue, measuring the amount of dissolved oxygen in a polluted river, observing the breaking strength of reinforced steel, counting the number of errors in software code, and observing the fraction of insects killed by a new insecticide. This list of statistical experiments could be continued, but the point is that our definition of an experiment is very broad.

**Definition 3.1**
An **experiment** is the process of obtaining an observation or taking a measurement.

Consider another simple experiment consisting of tossing a die and observing the number on the face of the die. The six basic possible outcomes to this experiment are

1. Observe a 1
2. Observe a 2
3. Observe a 3
4. Observe a 4
5. Observe a 5
6. Observe a 6

Note that if this experiment is conducted once, ***you can observe one and only one of these six basic outcomes***. The distinguishing feature of these outcomes is that these possibilities *cannot be decomposed* into any other outcomes. These very basic possible outcomes to an experiment are called **simple events**.

**Definition 3.2**
A **simple event** is a basic outcome of an experiment; it cannot be decomposed into simpler outcomes.

**Example 3.1**

Listing Simple Events for a Coin-Tossing Experiment

Two coins are tossed and the up faces of both coins are recorded. List all the simple events for this experiment.

Solution

Even for a seemingly trivial experiment, we must be careful when listing the simple events. At first glance the basic outcomes seem to be Observe two heads, Observe two tails, Observe one head and one tail. However, further reflection reveals that the last of these, Observe one head and one tail, can be decomposed into Head on coin 1, Tail on coin 2 and Tail on coin 1, Head on coin 2.* Thus, the simple events are as follows:

1. Observe $HH$
2. Observe $HT$
3. Observe $TH$
4. Observe $TT$

(where $H$ in the first position means "Head on coin 1," $H$ in the second position means "Head on coin 2," etc.).

We will often wish to refer to the collection of all the simple events of an experiment. This collection will be called the **sample space** of the experiment. For example, there are six simple events in the sample space associated with the die-tossing experiment. The sample spaces for the experiments discussed thus far are shown in Table 3.1.

**Definition 3.3**
The **sample space** of an experiment is the collection of all its simple events.

Just as graphs are useful in describing sets of data, a pictorial method for presenting the sample space and its simple events will often be useful. Figure 3.1 shows such a representation for each of the experiments in Table 3.1. In each case, the sample space is shown as a closed figure, labeled $S$, containing a set of points, called **sample**

*Even if the coins are identical in appearance, there are, in fact, two distinct coins. Thus, the designation of one coin as "coin 1" and the other as "coin 2" is legitimate in any case.

**TABLE 3.1 Experiments and Their Sample Spaces**

*Experiment*: Toss a coin and observe the up face.

*Sample space*:
1. Observe a head
2. Observe a tail

This sample space can be represented in set notation as a set containing two simple events

$S$: $\{H, T\}$

where $H$ represents the simple event Observe a head and $T$ represents the simple event Observe a tail.

---

*Experiment*: Toss a die and observe the up face.

*Sample space*:
1. Observe a 1
2. Observe a 2
3. Observe a 3
4. Observe a 4
5. Observe a 5
6. Observe a 6

This sample space can be represented in set notation as a set of six simple events

$S$: $\{1, 2, 3, 4, 5, 6\}$

---

*Experiment*: Toss two coins and observe the up face on each.

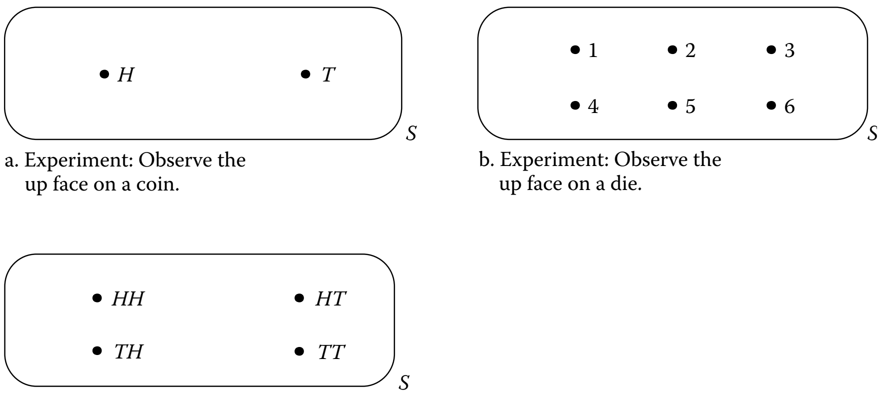
*Sample space*:
1. Observe $HH$
2. Observe $HT$
3. Observe $TH$
4. Observe $TT$

This sample space can be represented in set notation as a set of four simple events

$S$: $\{HH, HT, TH, TT\}$

---

**FIGURE 3.1**
Venn diagrams for the three experiments from Table 3.1




a. Experiment: Observe the up face on a coin.

b. Experiment: Observe the up face on a die.

c. Experiment: Observe the up faces on two coins.

**points**, with each point representing one simple event. Note that the number of sample points in a sample space $S$ is equal to the number of simple events associated with the respective experiment: two for the coin toss, six for the die toss, and four for the two-coin toss. These graphical representations are called **Venn diagrams**.

Now that we have defined simple events as the basic outcomes of the experiment and the sample space as the collection of all the simple events, we are prepared to discuss the probabilities of simple events. You have undoubtedly used the term *probability* and have some intuitive idea about its meaning. Probability is generally used synonymously with "chance," "odds," and similar concepts. We will begin our

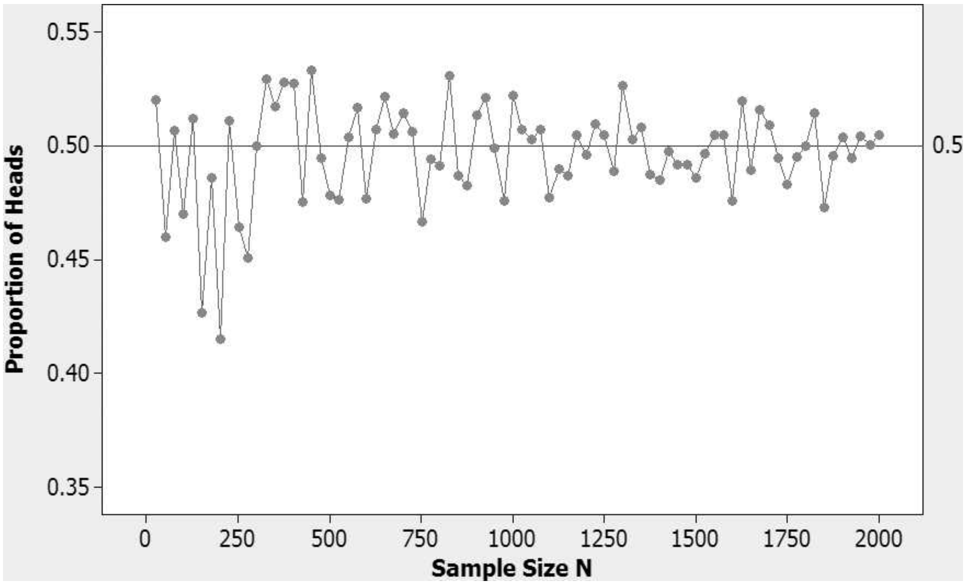



**FIGURE 3.2**
MINITAB Output Showing Proportion of Heads in N Tosses of a Coin

discussion of probability using these informal concepts. For example, if a fair coin is tossed, we might reason that both the simple events, Observe a head and Observe a tail, have the same chance of occurring. Thus, we might state that "the probability of observing a head is 50% or $\frac{1}{2}$," or "the odds of seeing a head are 50–50."

What do we mean when we say that the probability of a head is $\frac{1}{2}$? We mean that, in a very long series of tosses, approximately half would result in a head. Therefore, the number $\frac{1}{2}$ measures the likelihood of observing a head on a single toss.

Stating that the probability of observing a head is $\frac{1}{2}$ does *not* mean that exactly half of a number of tosses will result in heads. For example, we do not expect to observe exactly one head in two tosses of a coin or exactly five heads in ten tosses of a coin. Rather, we would expect the proportion of heads to vary in a random manner and to approach closer and closer the probability of a head, $\frac{1}{2}$, as the number of tosses increases. This property can be seen in the graph in Figure 3.2.

The MINITAB printout in Figure 3.2 shows the proportion of heads observed after $n = 25, 50, 75, 100, 125, \ldots, 1950, 1975,$ and 2000 simulated repetitions of a coin-tossing experiment. The number of tosses is marked along the horizontal axis of the graph, and the corresponding proportions of heads are plotted on the vertical axis above the values of $n$. We have connected the points to emphasize that the proportion of heads moves closer and closer to .5 as $n$ gets larger (as you move to the right on the graph).

**Definition 3.4**

The **probability** of a event (simple or otherwise) is a number that measures the likelihood that the event will occur when the experiment is performed. The probability can be approximated by the proportion of times that the event is observed when the experiment is repeated a very large number of times.* For a simple event $E$, we denote the probability of $E$ as $P(E)$.

---

*The result derives from an axiom in probability theory called the *Law of Large Numbers*. Phrased informally, the law states that the relative frequency of the number of times that an outcome occurs when an experiment is replicated over and over again (i.e., a large number of times) approaches the true (or theoretical) probability of the outcome.

Although we usually think of the probability of an event as the proportion of times the event occurs in a very long series of trials, some experiments can never be repeated. For example, if you invest in an oil-drilling venture, the probability that your venture will succeed has some unknown value that you will never be able to evaluate by repetitive experiments. The probability of this event occurring is a number that has some value, but it is unknown to us. The best that we could do, in estimating its value, would be to attempt to determine the proportion of similar ventures that succeeded and take this as an approximation to the desired probability. In spite of the fact that we may not be able to conduct repetitive experiments, the relative frequency definition for probability appeals to our intuition.

No matter how you assign probabilities to the simple events of an experiment, the probabilities assigned must obey the two rules (or axioms) given in the box.

### Rules for Assigning Probabilities to Simple Events

Let $E_1, E_2, \ldots, E_k$ be the simple events in a sample space.

1. All simple event probabilities *must* lie between 0 and 1:

   $$0 \leq P(E_i) \leq 1 \quad \text{for } i = 1, 2, \ldots, k$$

2. The sum of the probabilities of all the simple events within a sample space must be equal to 1:

   $$\sum_{i=1}^{k} P(E_i) = 1$$

Sometimes we are interested in the occurrence of any one of a collection of simple events. For example, in the die-tossing experiment of Table 3.1, we may be interested in observing an odd number on the die. This will occur if any one of the following three simple events occurs:

1. Observe a 1
2. Observe a 3
3. Observe a 5

In fact, the event Observe an odd number is clearly defined if we specify the collection of simple events that imply its occurrence. Such specific collections of simple events are called **events**.

**Definition 3.5**

An **event** is a specific collection of sample points (or simple events).

The probability of an event is computed by summing the probabilities of the simple events that comprise it. This rule agrees with the relative frequency concept of probability, as Example 3.2 illustrates.

### The Probability of an Event

The **probability of an event** $A$ is equal to the sum of the probabilities of the sample points in event $A$.

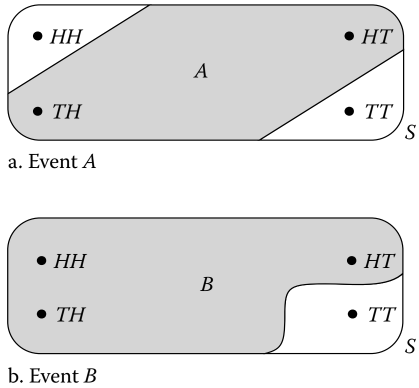
**Example 3.2**

Summing Probabilities of Simple Events

Consider the experiment of tossing two coins. If the coins are balanced, then the correct probabilities associated with the simple events are as follows:

| *Simple Event* | *Probability* | *Simple Event* | *Probability* |
|---|---|---|---|
| $HH$ | $\frac{1}{4}$ | $TH$ | $\frac{1}{4}$ |
| $HT$ | $\frac{1}{4}$ | $TT$ | $\frac{1}{4}$ |

**FIGURE 3.3**
Coin-tossing experiment showing events $A$ and $B$ as collections of simple events




Define the following events:

$A$: {Observe exactly one head}
$B$: {Observe at least one head}

Calculate the probability of $A$ and the probability of $B$.

**Solution** Note that each of the four simple events has the same probability since we expect each to occur with approximately equal relative frequency $\left(\frac{1}{4}\right)$ if the coin-tossing experiment were repeated a large number of times. Since the event $A$: {Observe exactly one head} will occur if either of the two simple events $HT$ or $TH$ occurs (see Figure 3.3), then approximately $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ of the large number of experiments will result in event $A$. This additivity of the relative frequencies of simple events is consistent with our rule for finding $P(A)$:

$$\begin{aligned} P(A) &= P(HT) + P(TH) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

Applying this rule to find $P(B)$, we note that event $B$ contains the simple events $HH$, $HT$, and $TH$—that is, $B$ will occur if any one of these three simple events occurs. Therefore,

$$\begin{aligned} P(B) &= P(HH) + P(HT) + P(TH) \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4} \end{aligned}$$

We can now summarize the steps for calculating the probability of any event:*

### Steps for Calculating Probabilities of Events

1. Define the experiment, i.e., describe the process used to make an observation and the type of observation that will be recorded.
2. Define and list the simple events.
3. Assign probabilities to the simple events.
4. Determine the collection of simple events contained in the event of interest.
5. Sum the simple event probabilities to get the event probability.

*A thorough treatment of this topic can be found in Feller (1968).

**Example 3.3**

**Probabilities of Simple Events—Quality Control Application**

A quality control engineer must decide whether an assembly line that produces manufactured items is "out of control"—that is, producing defective items at a higher rate than usual. At this stage of our study, we do not have the tools to solve this problem, but we can say that one of the important factors affecting the solution is the proportion of defectives manufactured by the line. To illustrate, what is the probability that an item manufactured by the line will be defective? What is the probability that the next two items produced by the line will be defective? What is the probability for the general case of $k$ items? Explain how you might solve this problem.

Solution

*Step 1* Define the experiment. The experiment corresponding to the inspection of a single item is identical in underlying structure to the coin-tossing experiment illustrated in Figure 3.1a. An item, either a nondefective (call this a head) or a defective (call this a tail), is observed and its operating status is recorded.

*Experiment:* Observe the operating status of a single manufactured item.

*Step 2* List the simple events. There are only two possible outcomes of the experiment. These simple events are

*Simple events:*
1. $N$: {Item is nondefective}
2. $D$: {Item is defective}

*Step 3* Assign probabilities to the simple events. The difference between this problem and the coin-tossing problem becomes apparent when we attempt to assign probabilities to the two simple events. What probability should we assign to the simple event $D$? Some people might say .5, as for the coin-tossing experiment, but you can see that finding $P(D)$, the probability of simple event $D$, is not so easy. Suppose that when the assembly line is in control, 10% of the items produced will be defective. Then, at first glance, it would appear that $P(D)$ is .10. But this may not be correct, because the line may be out of control, producing defectives at a higher rate. So, the important point to note is that this is a case where equal probabilities are not assigned to the simple events. How can we find these probabilities? A good procedure might be to monitor the assembly line for a period of time, and record the number of defective and nondefective items produced. Then the proportions of the two types of items could be used to approximate the probabilities of the two simple events.

We could then continue with steps 4 and 5 to calculate any probability of interest for this experiment with two simple events.

The experiment, assessing the operating status of two items, is identical to the experiment of Example 3.2, tossing two coins, except that the probabilities of the simple events are not the same. We will learn how to find the probabilities of the simple events for this experiment, or for the general case of $k$ items, in Section 3.6.

**Example 3.4**

**Probability of an Event—Debugging Software Code**

A software engineer must select three program modules from among five that need to be checked for defective code. If, unknown to the engineer, the modules vary in the effort required to debug them, what is the probability that

a. The engineer selects the two modules that require the least amount of effort?
b. The engineer selects the three modules that require the most effort?

Solution

*Step 1* The experiment consists of selecting three modules from among the five that require debugging.

*Step 2* We will denote the available modules by the symbols $M_1, M_2, \ldots, M_5$, where $M_1$ is the module that requires the least effort and $M_5$ requires the most effort. The notation $M_i\,M_j$ will denote the selection of modules $M_i$ and $M_j$. For example, $M_1\,M_3$ denotes the selection of modules $M_1$ and $M_3$. Then the 10 simple events associated with the experiment are as follows:

| *Simple Event* | *Probability* | *Simple Event* | *Probability* |
|---|---|---|---|
| $M_1\,M_2\,M_3$ | $\frac{1}{10}$ | $M_1\,M_4\,M_5$ | $\frac{1}{10}$ |
| $M_1\,M_2\,M_4$ | $\frac{1}{10}$ | $M_2\,M_3\,M_4$ | $\frac{1}{10}$ |
| $M_1\,M_2\,M_5$ | $\frac{1}{10}$ | $M_2\,M_3\,M_5$ | $\frac{1}{10}$ |
| $M_1\,M_3\,M_4$ | $\frac{1}{10}$ | $M_2\,M_4\,M_5$ | $\frac{1}{10}$ |
| $M_1\,M_3\,M_5$ | $\frac{1}{10}$ | $M_3\,M_4\,M_5$ | $\frac{1}{10}$ |

*Step 3* If we assume that the selection of any set of three modules is as likely as any other, then the probability of each of the 10 simple events is $\frac{1}{10}$.

*Step 4* Define the events $A$ and $B$ as follows:

$A$: {The engineer selects the two modules that require the least amount of effort}

$B$: {The engineer selects the three modules that require the most effort}

Event $A$ will occur for any simple events in which modules $M_1$ and $M_2$ are selected—namely, the three simple events $M_1\,M_2\,M_3$, $M_1\,M_2\,M_4$ and $M_1\,M_2\,M_5$. Similarly, the event $B$ is made up of the single event $M_3\,M_4\,M_5$.

*Step 5* We now sum the probabilities of the simple events in $A$ and $B$ to obtain

$$
\begin{aligned}
P(A) &= P(M_1M_2M_3) + P(M_1M_2M_4) + P(M_1M_2M_5) \\
&= \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{3}{10}
\end{aligned}
$$

and

$$P(B) = P(M_3M_4M_5) = \frac{1}{10}$$

[*Note*: For the experiments discussed thus far, listing the simple events has been easy. For more complex experiments, the number of simple events may be so large that listing them is impractical. In solving probability problems for experiments with many simple events, we use the same principles as for experiments with few simple events. The only difference is that we need **counting rules** for determining the number of simple events without actually enumerating all of them. In Section 3.8, we present several of the more useful counting rules.]

## Applied Exercises

3.1 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study of the trend in the design of social robots, Exercise 2.1 (p. 26). Recall that in a random sample of 106 social (or service) robots designed to entertain, educate, and care for human users, 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels. One of the 106 social robots is randomly selected and the design (e.g., wheels only) is noted.

a. List the simple events for this study.
b. Assign reasonable probabilities to the simple events.
c. What is the probability that the selected robot is designed with wheels?
d. What is the probability that the selected robot is designed with legs?

3.2 *STEM experiences for girls.* Refer to the 2013 National Science Foundation (NSF) study on girls participation in informal science, technology, engineering or mathematics (STEM) programs, Exercise 2.3 (p. 27). Recall that the researchers sampled 174 young women who recently participated in a STEM program. Of the 174 STEM participants, 107 were in urban areas, 57 in suburban areas, and 10 in rural areas. If one of the participants is selected at random, what is the probability that she is from an urban area? Not a rural area?

3.3 *Rare underwater sounds.* Acoustical engineers conducted a study of rare underwater sounds in a specific region of the Pacific Ocean, such as humpback whale screams, dolphin whistles, and sounds from passing ships (*Acoustical Physics*, Vol. 56, 2010). During the month of September (non-rainy season), research revealed the following probabilities of rare sounds: P(whale scream) = .03, P(ship sound) = .14, and P(rain) = 0. If a sound is picked up by the acoustical equipment placed in this region of the Pacific Ocean, is it more likely to be a whale scream or a sound from a passing ship? Explain.

3.4 *Is a product "green"?* A "green" product (e.g., a product built from recycled materials) is one that has minimal impact on the environment and human health. How do consumers determine if a product is "green"? The 2011 ImagePower Green Brands Survey asked this question to over 9,000 international consumers. The results are shown in the next table.

a. What method is an international consumer most likely to use to identify a green product?
b. Find the probability that an international consumer identifies a green product by a certification mark on the product label or by the product packaging.
c. Find the probability that an international consumer identifies a green product by reading about the product or from information at the brand's website.
d. Find the probability that an international consumer does not use advertisements to identify a green product.

| Reason for saying a product is green | Percentage of consumers |
|---|---|
| Certification mark on label | 45 |
| Packaging | 15 |
| Reading information about the product | 12 |
| Advertisement | 6 |
| Brand website | 4 |
| Other | 18 |
| **TOTAL** | 100 |

*Source:* 2011 ImagePower Green Brands Survey

3.5 *Toxic chemical incidents.* *Process Safety Progress* (Sept. 2004) reported on an emergency response system for incidents involving toxic chemicals in Taiwan. The system has logged over 250 incidents since being implemented in 2000. The accompanying table gives a breakdown of the locations where these toxic chemical incidents occurred. Consider the location of a toxic chemical incident in Taiwan.

a. List the simple events for this experiment.
b. Assign reasonable probabilities to the simple events.
c. What is the probability that the incident occurs in a school laboratory?

| Location | Percent of Incidents |
|---|---|
| *School laboratory* | 6% |
| *In transit* | 26% |
| *Chemical plant* | 21% |
| *Nonchemical plant* | 35% |
| *Other* | 12% |
| **Total** | 100% |

*Source:* Chen, J. R., et al. "Emergency response of toxic chemicals in Taiwan: The system and case studies." *Process Safety Progress*, Vol. 23, No. 3, Sept. 2004 (Figure 5a).

3.6 *Beach erosional hot spots.* Refer to the U.S. Army Corps of Engineers study of beaches with high erosion rates (i.e., beach *hot spots*). Exercise 2.5 (p. 27). The data for six beach hot spots are reproduced in the table on page 87.

a. Suppose you record the nearshore bar condition of each beach hot spot. Give the sample space for this experiment.
b. Find the probabilities of the simple events in the sample space, part **a**.
c. What is the probability that a beach hot spot has either a planar or single shore parallel-nearshore bar condition?
d. Now, suppose you record the beach condition of each beach hot spot. Give the sample space for this experiment.

e. Find the probabilities of the simple events in the sample space, part **d**.
f. What is the probability that the condition of the beach at a hot spot is not flat?

| Beach Hot Spot | Beach Condition | Nearshore Bar Condition | Long-Term Erosion Rate (miles/year) |
|---|---|---|---|
| Miami Beach, FL | No dunes/flat | Single, shore parallel | 4 |
| Coney Island, NY | No dunes/flat | Other | 13 |
| Surfside, CA | Bluff/scarp | Single, shore parallel | 35 |
| Monmouth Beach, NJ | Single dune | Planar | Not estimated |
| Ocean City, NJ | Single dune | Other | Not estimated |
| Spring Lake, NJ | Not observed | Planar | 14 |

*Source:* "Identification and characterization of erosional hotspots." William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March, 18, 2002.

**MTBE**

3.7 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in New Hampshire wells, Exercise 2.12 (p. 29). Data collected for a sample of 223 wells are saved in the **MTBE** file. Recall that each well was classified according to well class (public or private), aquifer (bedrock or unconsolidated), and detectable level of MTBE (below limit or detect).

a. Consider an experiment in which the well class, aquifer, and detectable MTBE level of a well are observed. List the simple events for this experiment. (*Hint*: One simple event is Private/Bedrock/BelowLimit.)
b. Use statistical software to find the number of the 223 wells in each simple event outcome. Then, use this information to compute probabilities for the simple events.
c. Find and interpret the probability that a well has a detectable level of MTBE.

3.8 *USDA chicken inspection.* The United States Department of Agriculture (USDA) reports that, under its standard inspection system, one in every 100 slaughtered chickens passes inspection with fecal contamination.

a. If a slaughtered chicken is selected, what is the probability that it passes inspection with fecal contamination?
b. The probability of part **a** was based on a USDA study that found that 306 of 32,075 chicken carcasses passed inspection with fecal contamination. Do you agree with the USDA's statement about the likelihood of a slaughtered chicken passing inspection with fecal contamination?

3.9 *Fungi in beech forest trees.* Beechwood forests in East Central Europe are being threatened by dynamic changes in land ownership and economic upheaval. The current status of the beech tree species in this area was evaluated by Hungarian university professors in *Applied Ecology and Environmental Research* (Vol. 1, 2003). Of 188 beech trees surveyed, 49 of the trees had been damaged by fungi. Depending on the species of fungi, damage will occur either on the trunk, branches, or leaves of the tree. In the damaged trees, the trunk was affected 85% of the time, the leaves 10% of the time, and the branches 5% of the time.

a. Give a reasonable estimate of the probability of a beech tree in East Central Europe being damaged by fungi.
b. A fungi-damaged beech tree is selected and the area (trunk, leaf, or branch) affected is observed. List the sample points for this experiment and assign a reasonable probability to each sample point.

3.10 *Predicting when a Florida hurricane occurs.* Since the early 1900's, the state of Florida has exceeded $450 million in damages due to destructive hurricanes. Consequently, the value of insured property against windstorm damage in Florida is the highest in the nation. Researchers at Florida State University conducted a comprehensive analysis of damages caused by Florida hurricanes and published the results in *Southeastern Geographer* (Summer, 2009). Part of their analysis included estimating the likelihood that a hurricane develops from a tropical storm based on the sequence number of the tropical storm within a season. The researchers discovered that of the 67 Florida hurricanes since 1900, 11 developed from the 5th tropical storm of the season (the sequence with the highest frequency). Also, only 5 hurricanes developed from a tropical storm with a sequence number of 12 or greater.

a. Estimate the probability that a Florida hurricane develops from the 5th tropical storm of the season.
b. Estimate the probability that a Florida hurricane develops before the 12th tropical storm of the season.

3.11 *Using game simulation to teach a POM course.* In *Engineering Management Research* (May, 2012), a simulation game approach was proposed to teach concepts in a course on Production and Operations Management (POM). The proposed game simulation was for color television production. The products are two color television models, A and B. Each model comes in two colors, red and black. Also, the quantity ordered for each model can be 1, 2, or 3 televisions. The choice of model, color, and quantity is specified on a purchase order card.

a. For this simulation, list how many different purchase order cards are possible. (These are the simple events for the experiment.)
b. Suppose, from past history, that black color TVs are in higher demand than red TVs. For planning purposes, should the engineer managing the production process assign equal probabilities to the simple events, part a? Why or why not?

## 3.3 Compound Events

An event can often be viewed as a composition of two or more other events. Such events are called **compound events**; they can be formed (composed) in two ways.

**Definition 3.6**

The **union** of two events $A$ and $B$ is the event that occurs if either $A$ or $B$, or both, occur on a single performance of the experiment. We will denote the union of events $A$ and $B$ by the symbol $A \cup B$.

$$A \cup B = A \text{ or } B$$

**Definition 3.7**

The **intersection** of two events $A$ and $B$ is the event that occurs if both $A$ and $B$ occur on a single performance of the experiment. We will write $A \cap B$ for the intersection of events $A$ and $B$.

$$A \cap B = A \text{ and } B$$

**Example 3.5**

Finding Probabilities of Unions and Intersections: CO Poisoning

The *American Journal of Public Health* (July 1995) published a study on unintentional carbon monoxide (CO) poisoning of Colorado residents. The source of exposure was determined for 1,000 cases of CO poisoning that occurred during a recent six-year period. In addition, each case was classified as fatal or nonfatal. The proportion of the cases occurring in each of 10 source/fatal categories is shown in Table 3.2. Define the following events:

A: {CO poisoning case is caused by fire}
B: {CO poisoning case is fatal}

a. Describe the simple events for this experiment. Assign probabilities to these simple events.
b. Describe $A \cup B$.
c. Describe $A \cap B$.
d. Calculate $P(A \cup B)$ and $P(A \cap B)$.

Solution

a. The simple events for this experiment are the different combinations of exposure source and fatality status. For example, one simple event is {Fire related/Fatal}; another is {Fire related/Nonfatal}; a third is {Auto exhaust/Fatal}. You can see from Table 3.2 that there are a total of $5 \times 2 = 10$ simple events. Since the probability of an event is the likelihood that the event will occur in a long series of observations, the probability of each simple event can be approximated by the proportion of times the simple event occurs in the 1,000 cases. These proportions are listed in Table 3.2. If you sum these 10 probabilities, you will find they sum to 1.

**TABLE 3.2 CO Exposure by Source and Fatal Status**

| Source of Exposure | Fatal | Nonfatal |
|---|---|---|
| Fire related | .07 | .06 |
| Auto exhaust | .07 | .19 |
| Furnace | .02 | .37 |
| Appliance/motor | .02 | .175 |
| Other | .005 | .020 |

*Adapted from*: Cook, M., Simon, P., and Hoffman, R. "Unintentional carbon monoxide poisoning in Colorado." *American Journal of Public Health*, Vol. 85, No. 7, July 1995 (Table 1).

b. The union of $A$ and $B$ is the event that occurs if we observe a CO poisoning case caused by fire exposure (event $A$), or a case where a fatality occurs (event $B$), or both. Consequently, the simple events in the event $A \cup B$ are those for which $A$ occurs, $B$ occurs, or both $A$ and $B$ occur, i.e.,

$$A \cup B = \{\text{Fire/Fatal, Fire/Nonfatal, Auto/Fatal, Furnace/Fatal, Appliance/Fatal, Other/Fatal}\}$$

This union is illustrated in the Venn diagram, Figure 3.4.

c. The intersection of $A$ and $B$ is the event that occurs if we observe both a CO poisoning case caused by fire exposure (event $A$) and a case where a fatality occurs (event $B$). In Figure 3.4, you can see that the only simple event with both of these characteristics is

$$A \cap B = \{\text{Fire/Fatal}\}$$

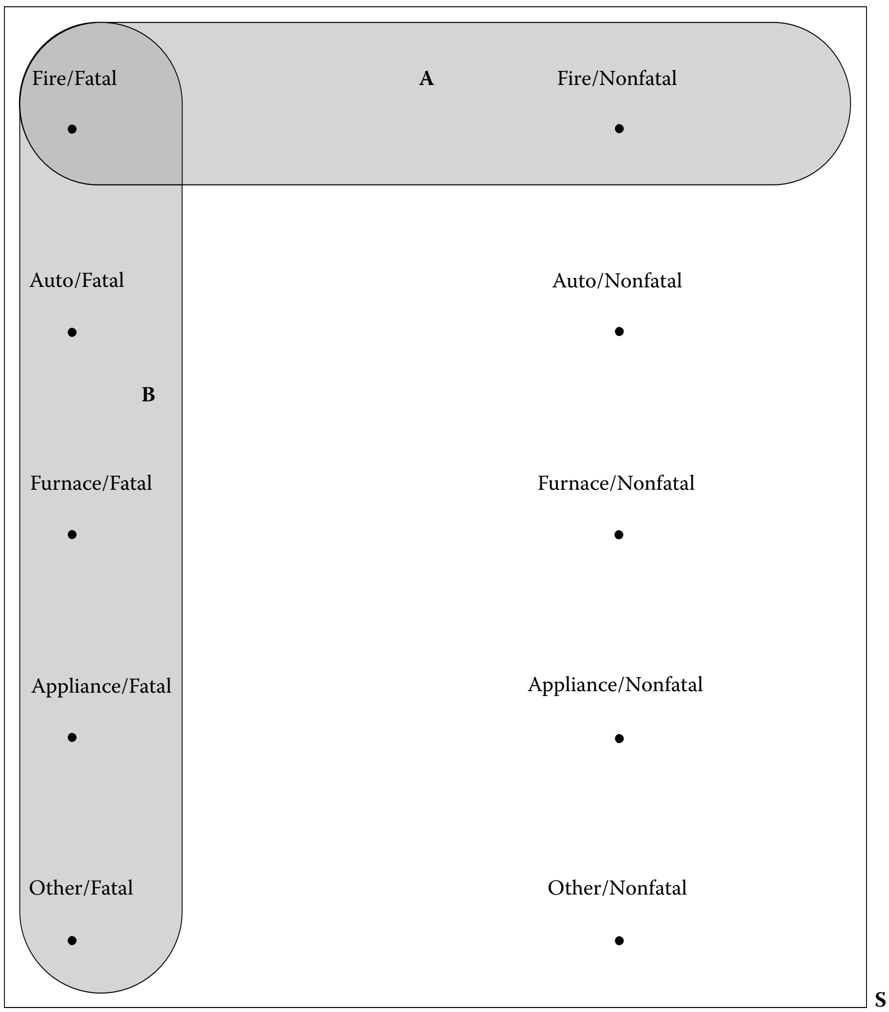



**FIGURE 3.4**
Venn diagram of $A \cup B$, Example 3.5

d. Recalling that the probability of an event is the sum of the probabilities of the simple events of which the event is composed, we have

$$\begin{aligned} P(A \cup B) &= P(\text{Fire/Fatal}) + P(\text{Fire/Nonfatal}) + P(\text{Auto/Fatal}) \\ &\quad + P(\text{Furnace/Fatal}) + P(\text{Appliance/Fatal}) + P(\text{Other/Fatal}) \\ &= .07 + .06 + .07 + .02 + .02 + .005 = .245 \end{aligned}$$

and

$$P(A \cap B) = P(\text{Fire/Fatal}) = .07$$

Unions and intersections also can be defined for more than two events. For example, the event $A \cup B \cup C$ represents the union of three events, $A$, $B$, and $C$. This event, which includes the set of simple events in $A$, $B$, or $C$, will occur if any one or more of the events $A$, $B$, or $C$ occurs. Similarly, the intersection $A \cap B \cap C$ is the event that all three of the events $A$, $B$, and $C$ occur simultaneously. Therefore, $A \cap B \cap C$ is the set of simple events that are in all three of the events $A$, $B$, and $C$.

**Example 3.6**

Probabilities of Unions and Intersections: Die-Tossing Experiment

Consider the die-tossing experiment with equally likely simple events {1, 2, 3, 4, 5, 6}. Define the events $A$, $B$, and $C$ as follows:

A: {Toss an even number} = {2, 4, 6}
B: {Toss a number less than or equal to 3} = {1, 2, 3}
C: {Toss a number greater than 1} = {2, 3, 4, 5, 6}

Find

a. $P(A \cup B \cup C)$
b. $P(A \cap B \cap C)$

Solution

a. Event $C$ contains the simple events corresponding to tossing a 2, 3, 4, 5, or 6; event $B$ contains the simple events 1, 2, and 3; and, event $A$ contains the simple events 2, 4, or 6. Therefore, the event that $A$, $B$, or $C$ occurs contains all six simple events in $S$, i.e., those corresponding to tossing a 1, 2, 3, 4, 5, or 6. Consequently, $P(A \cup B \cup C) = P(S) = 1$.
b. You can see that you will observe all of the events, $A$, $B$, and $C$, only if you observe a 2. Therefore, the intersection $A \cap B \cap C$ contains the single simple event Toss a 2 and $P(A \cap B \cap C) = P(2) = \frac{1}{6}$.

## 3.4 Complementary Events

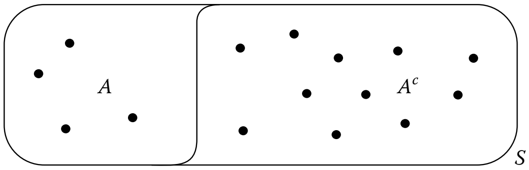
A very useful concept in the calculation of event probabilities is the notion of **complementary events**.

**Definition 3.8**

The **complement*** of an event $A$ is the event that $A$ does not occur, i.e., the event consisting of all simple events that are not in event $A$. We will denote the complement of $A$ by $A^c$. Note that $A \cup A^c = S$, the sample space.

*Some texts use the symbol $A'$ to denote the complement of an event $A$.

**FIGURE 3.5**
Venn diagram of complementary events




An event $A$ is a collection of simple events, and the simple events included in $A^c$ are those that are not in $A$. Figure 3.5 demonstrates this. You will note from the figure that all simple events in $S$ are included in *either* $A$ or $A^c$, and that *no* simple event is in both $A$ and $A^c$. This leads us to conclude that the probabilities of an event and its complement must sum to 1.

### Complementary Relationship

***The sum of the probabilities of complementary events equals 1***. That is,

$$P(A) + P(A^c) = 1$$

In many probability problems, it will be easier to calculate the probability of the complement of the event of interest rather than the event itself. Then, since

$$P(A) + P(A^c) = 1$$

we can calculate $P(A)$ by using the relationship

$$P(A) = 1 - P(A^c)$$

**Example 3.7**

Finding the Probability of a Complementary Event: Coin-Tossing Experiment

Consider the experiment of tossing two fair coins. Calculate the probability of event

$A$: {Observe at least one head}

by using the complementary relationship.

Solution

We know that the event $A$: {Observe at least one head} consists of the simple events

$A$: $\{HH, HT, TH\}$

The complement of $A$ is defined as the event that occurs when $A$ does not occur. Therefore,

$A^c$: {Observe no heads} = $\{TT\}$

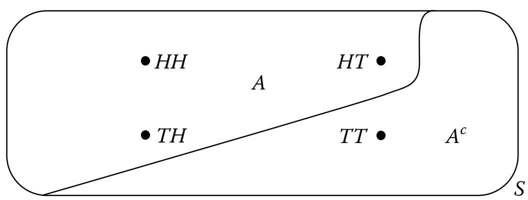
This complementary relationship is shown in Figure 3.6. Assuming the coins are balanced, we have

$$P(A^c) = P(TT) = \frac{1}{4}$$

and

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{4} = \frac{3}{4}$$

**FIGURE 3.6**
Complementary events in the toss of two coins



**Example 3.8**

**Finding the Probability of a Defective Item**

Refer to Example 3.3 (p. 84). Assume 10 items are selected for inspection from the assembly line. Also, assume the process is out of control, with a defective (*D*) item just as likely to occur as a nondefective (*N*) item. Find the probability of the event

$A$: {Observe at least one defective}

**Solution**

We will solve this problem by following the five steps for calculating probabilities of events (see Section 3.2).

*Step 1* Define the experiment. The experiment is to record the results (*D* or *N*) of the 10 items.

*Step 2* List the simple events. A simple event consists of a particular sequence of 10 defectives and nondefectives. Thus, one simple event is

*DDNNNDNDNN*

which denotes defective for the first item, defective for the second item, nondefective for the third item, etc. Others would be *DNDDDNNNNN* and *NDDNDNDNND*. There is obviously a very large number of simple events—too many to list. It can be shown (see Section 3.8) that there are $2^{10} = 1{,}024$ simple events for this experiment.

*Step 3* Assign probabilities. Since defectives and nondefectives occur at the same rate in the out-of-control process, each sequence of *N*s and *D*s has the same chance of occurring and therefore all the simple events are equally likely. Then

$$P(\text{Each simple event}) = \frac{1}{1{,}024}$$

*Step 4* Determine the simple events in event $A$. A simple event is in $A$ if at least one $D$ appears in the sequence of 10 items. However, if we consider the complement of $A$, we find that

$A^c$: {No *D*s are observed in 10 items}

Thus, $A^c$ contains only the simple event

$A^c$: {*NNNNNNNNNN*}

and therefore

$$P(A^c) = \frac{1}{1{,}024}$$

*Step 5* Since we know the probability of the complement of $A$, we use the relationship for complementary events:

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{1{,}024} = \frac{1{,}023}{1{,}024} = .999$$

That is, we are virtually certain of observing at least one defective in a sequence of 10 items produced from the out-of-control process.

## Applied Exercises

3.12 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study of the trend in the design of social robots, Exercise 3.1 (p. 86). Recall that in a random sample of 106 social robots, 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels. Use the rule of complements to find the probability that a randomly selected social robot is designed with either legs or wheels.

3.13 *Toxic chemical incidents.* Refer to the *Process Safety Progress* (Sept. 2004) study of toxic chemical incidents in Taiwan, Exercise 3.5 (p. 86).

a. Find the probability that the incident occurs in either a chemical or nonchemical plant.

b. Find the probability that the incident did not occur in a school laboratory.

3.14 *Beach erosional hot spots.* Refer to the U.S. Army Corps of Engineers study of six beaches with high erosion rates (i.e., beach *hot spots*). Exercise 3.6 (p. 86). Use the rule of complements to find the probability that a beach hot spot is not flat. Compare your answer to Exercise 3.6f.

3.15 *Cell phone handoff behavior.* A “handoff” is a term used in wireless communications to describe the process of a cell phone moving from a coverage area of one base station to another. Each base station has multiple channels (called color codes) that allow it to communicate with the cell phone. The *Journal of Engineering*, *Computing and Architecture* (Vol. 3., 2009) published a study of cell phone handoff behavior. During a sample driving trip which involved crossing from one base station to another, the different color codes accessed by the cell phone were monitored and recorded. The table below shows the number of times each color code was accessed for two identical driving trips, each using a different cell phone model. (Note: The table is similar to the one published in the article.) Suppose you randomly select one point during the combined driving trips.

| | Color Code | | | | |
|---|---|---|---|---|---|
| | **0** | **5** | **b** | **c** | **TOTAL** |
| Model 1 | 20 | 35 | 40 | 0 | 95 |
| Model 2 | 15 | 50 | 6 | 4 | 75 |
| **TOTAL** | 35 | 85 | 46 | 4 | 170 |

a. What is the probability that the cell phone is using color code 5?
b. What is the probability that the cell phone is using color code 5 or color code 0?
c. What is the probability that the cell phone used is Model 2 and the color code is 0?

**MTBE**

3.16 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in 223 New Hampshire wells, Exercise 3.7 (p. 87). Each well is classified according to well class (public or private), aquifer (bedrock or unconsolidated), and detectable level of MTBE (below limit or detect). Consider one of these 223 wells.

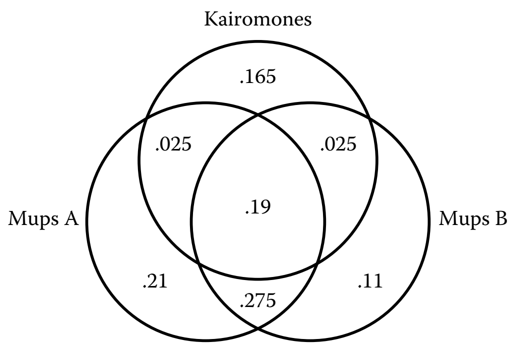
a. What is the probability that the well has a bedrock aquifer and a detected level of MTBE?
b. What is the probability that the well has a bedrock aquifer or a detected level of MTBE?

3.17 *USDA chicken inspection.* Refer to the USDA report on slaughtered chickens that pass inspection with fecal contamination, Exercise 3.8 (p. 87). Consider a sample of five chickens that have all passed the standard USDA inspection. Each chicken carcass will be classified as either passing inspection with fecal contamination or passing inspection without fecal contamination.

a. List the simple events for the sample.
b. Assuming the simple events, part **a**, are equally likely, find the probability that at least one of the five chickens passes inspection with fecal contamination.
c. Explain why the assumption of part **a** is unreasonable for this sample. (*Hint*: Refer to the results reported in Exercise 3.7.)

3.18 *Chemical signals of mice.* The ability of a mouse to recognize the odor of a potential predator (e.g., a cat) is essential to the mouse’s survival. The chemical makeup of these odors — called kairomones — was the subject of a study published in *Cell* (May 14, 2010). Typically, the source of these odors is major urinary proteins (Mups). Cells collected from lab mice were exposed to Mups from rodent species A, Mups from rodent species B, and kairomones (from a cat). The accompanying Venn diagram shows the proportion of cells that chemically responded to each of the three odors. (Note: A cell may respond to more than a single odor.)

a. What is the probability that a lab mouse responds to all three source odors?
b. What is the probability that a lab mouse responds to the kairomone?
c. What is the probability that a lab mouse responds to Mups A and Mups B, but not the kairomone?




3.19 *Inactive oil and gas structures.* U.S. federal regulations require that operating companies clear all inactive offshore oil and gas structures within one year after production ceases. Researchers at the Louisiana State University Center for Energy Studies gathered data on both active and inactive oil and gas structures in the Gulf of Mexico (*Oil & Gas Journal*, Jan. 3, 2005). They discovered that the gulf had 2,175 active and 1,225 idle (inactive) structures at the end of 2003. The table on pg. 94 breaks down these structures by type (caisson, well protector, or fixed platform). Consider the structure type and active status of one of these oil/gas structures.

| | Caisson | Well Protector | Fixed Platform | Totals |
|---|---|---|---|---|
| *Active* | 503 | 225 | 1,447 | 2,175 |
| *Inactive* | 598 | 177 | 450 | 1,225 |

*Source:* Kaiser, M., and Mesyanzhinov, D. "Study tabulates idle Gulf of Mexico structures." *Oil & Gas Journal*, Vol. 103, No. 1, Jan. 3, 2005 (Table 2).

a. List the simple events for this experiment.
b. Assign reasonable probabilities to the simple events.
c. Find the probability that the structure is active.
d. Find the probability that the structure is a well protector.
e. Find the probability that the structure is an inactive caisson.
f. Find the probability that the structure is either inactive or a fixed platform.
g. Find the probability that the structure is not a caisson.

3.20 *Outcomes in roulette.* One game that is popular in many American casinos is roulette. Roulette is played by spinning a ball on a circular wheel that has been divided into 38 arcs of equal length; these bear the numbers 00, 0, 1, 2, . . . , 35, 36. The number on the arc at which the ball comes to rest is the outcome of one play of the game. The numbers are also colored in the following manner:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Red*: | 1 | 3 | 5 | 7 | 9 | 12 | 14 | 16 | 18 |
| | 19 | 21 | 23 | 25 | 27 | 30 | 32 | 34 | 36 |
| *Black*: | 2 | 4 | 6 | 8 | 10 | 11 | 13 | 15 | 17 |
| | 20 | 22 | 24 | 26 | 28 | 29 | 31 | 33 | 35 |
| *Green*: | 00 | 0 | | | | | | | |

Players may place bets on the table in a variety of ways, including bets on odd, even, red, black, low (1–18), and high (19–36) outcomes. Consider the following events (00 and 0 are considered neither odd nor even):

$A$: {Outcome is an odd number}
$B$: {Outcome is a black number}
$C$: {Outcome is a high number}

Calculate the probabilities of the following events:
a. $A \cup B$ b. $A \cap C$ c. $B \cup C$ d. $B^c$ e. $A \cap B \cap C$

3.21 *Probability of an oil gusher.* An oil-drilling venture involves the drilling of six wildcat oil wells in different parts of the country. Suppose that each drilling will produce either a dry well or an oil gusher. Assuming that the simple events for this experiment are equally likely, find the probability that at least one oil gusher will be discovered.

3.22 *Encoding variability in software.* At the 2012 *Gulf Petrochemicals and Chemicals Association (GPCA) Forum*, Oregon State University software engineers presented a paper on modeling and implementing variation in computer software. The researchers employed the compositional choice calculus (CCC) – a formal language for representing, generating, and organizing variation in tree-structured artifacts. The CCC language was compared to two other coding languages – the annotative choice calculus (ACC) and the computational feature algebra (CFA). Their research revealed the following: any type of expression (e.g., plain expressions, dimension declarations, or lambda abstractions) found in either ACC or CFA can be found in CCC; plain expressions exist in both ACC and CFA; dimension declarations exist in ACC, but not CFA; lambda abstractions exist in CFA, but not ACC. Based on this information, draw a Venn diagram that illustrates the relationships among the three languages. (Hint: An expression represents a sample point in the Venn diagram.)

## 3.5 Conditional Probability

The event probabilities we have discussed thus far give the relative frequencies of the occurrences of the events when the experiment is repeated a very large number of times. They are called **unconditional probabilities** because no special conditions are assumed other than those that define the experiment.

Sometimes we may wish to alter our estimate of the probability of an event when we have additional knowledge that might affect its outcome. This revised probability is called the **conditional probability** of the event. For example, we have shown that the probability of observing an even number (event $A$) on a toss of a fair die is $\frac{1}{2}$. However, suppose you are given the information that on a particular throw of the die the result was a number less than or equal to 3 (event $B$). Would you still believe that the probability of observing an even number on that throw of the die is equal to $\frac{1}{2}$? If you reason that making the assumption that $B$ has occurred reduces the sample space from six simple events to three simple events (namely, those contained in event $B$), the reduced sample space is as shown in Figure 3.7.

Since the only even number of the three numbers in the reduced sample space of event $B$ is the number 2 and since the die is fair, we conclude that the probability that $A$ occurs **given that $B$ occurs** is one in three, or $\frac{1}{3}$. We will use the symbol $P(A \mid B)$ to

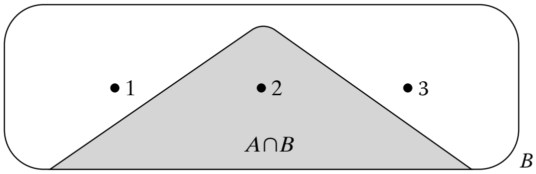



**FIGURE 3.7**
Reduced sample space for the die-tossing experiment, given that event $B$ has occurred

represent the probability of event $A$ given that event $B$ occurs. For the die-tossing example, we write

$$P(A \mid B) = \frac{1}{3}$$

To get the probability of event $A$ given that event $B$ occurs, we proceed as follows: We divide the probability of the part of $A$ that falls within the reduced sample space of event $B$, namely, $P(A \cap B)$, by the total probability of the reduced sample space, namely, $P(B)$. Thus, for the die-tossing example where event $A$: {Observe an even number} and event $B$: {Observe a number less than or equal to 3}, we find

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(2)}{P(1) + P(2) + P(3)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

This formula for $P(A \mid B)$ is true in general.

### Formula for Conditional Probability

To find the **conditional probability that event $A$ occurs** given that event **$B$ occurs**, divide the probability that *both* $A$ and $B$ occur by the probability that $B$ occurs, that is,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{where we assume that } P(B) \neq 0$$

### Example 3.9

Conditional Probability in a Process Control Study

Consider the following problem in process control. Suppose you are interested in the probability that a manufactured product (e.g., a small mechanical part) shipped to a buyer conforms to the buyer's specifications. Lots containing a large number of parts must pass inspection before they are accepted for shipment. [Assume that not all parts in a lot are inspected. For example, if the mean product characteristic (e.g., diameter) of a sample of parts selected from the lot falls within certain limits, the entire lot is accepted even though there may be one or more individual parts that fall outside specifications.] Let $I$ represent the event that a lot passes inspection and let $B$ represent the event that an individual part in a lot conforms to the buyer's specifications. Thus, $I \cap B$ is the simple event that the individual part is both shipped to the buyer (this happens when the lot containing the part passes inspection) and conforms to specifications, $I \cap B^c$ is the simple event that the individual part is shipped to the buyer but does not conform to specifications, etc. Assume that the probabilities associated with the four simple events are as shown in the accompanying table. Find the probability that an individual part conforms to the buyer's specifications given that it is shipped to the buyer.

| *Simple Event* | *Probability* |
|---|---|
| $I \cap B$ | .80 |
| $I \cap B^c$ | .02 |
| $I^c \cap B$ | .15 |
| $I^c \cap B^c$ | .03 |

Solution

If one part is selected from a lot of manufactured parts, what is the probability that the buyer will accept the part? To be accepted, the part must first be shipped to the buyer (i.e., the lot containing the part must pass inspection) *and* then the part must meet the buyer's specifications, so this *unconditional* probability is $P(I \cap B) = .80$.

In contrast, suppose you *know* that the selected part is from a lot that passes inspection. Now you are interested in the probability that the part conforms to specifications *given* that the part is shipped to the buyer, i.e., you want to determine the *conditional* probability $P(B \mid I)$. From the definition of conditional probability,

$$P(B \mid I) = \frac{P(I \cap B)}{P(I)}$$

where the event

$I$: {Part is shipped to the buyer}

contains the two simple events

$I \cap B$: {Part is shipped to buyer and conforms to specifications}

and

$I \cap B^c$: {Part is shipped to buyer but fails to meet specifications}

Recalling that the probability of an event is equal to the sum of the probabilities of its simple events, we obtain

$$P(I) = P(I \cap B) + P(I \cap B^c)$$
$$= .80 + .02 = .82$$

Then the conditional probability that a part conforms to specifications, given the part is shipped to the buyer, is

$$P(B \mid I) = \frac{P(I \cap B)}{P(I)} = \frac{.80}{.82} = .976$$

As we would expect, the probability that the part conforms to specifications, given that the part is shipped to the buyer, is higher than the unconditional probability that a part will be acceptable to the buyer.

## Example 3.10

Conditional Probability Associated with Consumer Complaints

The investigation of consumer product complaints by the Federal Trade Commission (FTC) has generated much interest by manufacturers in the quality of their products. A manufacturer of food processors conducted an analysis of a large number of consumer complaints and found that they fell into the six categories shown in Table 3.3. If a consumer complaint is received, what is the probability that the cause of the complaint was product appearance given that the complaint originated during the guarantee period?

**TABLE 3.3 Distribution of Product Complaints**

| | Reason for Complaint | | | |
|---|---|---|---|---|
| | *Electrical* | *Mechanical* | *Appearance* | Totals |
| During guarantee period | 18% | 13% | 32% | 63% |
| After guarantee period | 12% | 22% | 3% | 37% |
| **Totals** | 30% | 35% | 35% | 100% |

Solution Let $A$ represent the event that the cause of a particular complaint was product appearance, and let $B$ represent the event that the complaint occurred during the guarantee period. Checking Table 3.3, you can see that $(18 + 13 + 32)\% = 63\%$ of the complaints occurred during the guarantee time. Hence, $P(B) = .63$. The percentage of complaints that were caused by appearance and occurred during the guarantee time (the event $A \cap B$) is 32%. Therefore, $P(A \cap B) = .32$.

Using these probability values, we can calculate the conditional probability $P(A \mid B)$ that the cause of a complaint is appearance given that the complaint occurred during the guarantee time:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{.32}{.63} = .51$$

Consequently, you can see that slightly more than half the complaints that occurred during the guarantee time were due to scratches, dents, or other imperfections in the surface of the food processors.

## Applied Exercises

3.23 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study of the trend in the design of social robots, Exercises 3.1 and 3.12 (p. 92). Recall that in a random sample of 106 social robots, 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels. If a social robot is designed with wheels, what is the probability that the robot also has legs?

3.24 *Speeding linked to fatal car crashes.* According to the National Highway Traffic and Safety Administration's National Center for Statistics and Analysis (NCSA), "Speeding is one of the most prevalent factors contributing to fatal traffic crashes" (*NHTSA Technical Report,* Aug 2005). The probability that speeding is a cause of a fatal crash is .3. Furthermore, the probability that speeding and missing a curve are causes of a fatal crash is .12. Given that speeding is a cause of a fatal crash, what is the probability that the crash occurred on a curve?

3.25 *Detecting traces of TNT.* University of Florida researchers in the Department of Materials Science and Engineering have invented a technique to rapidly detect traces of TNT. (*Today*, Spring 2005.) The method, which involves shining a photoluminescence spectroscope (i.e., a laser light) on a potentially contaminated object, provides instantaneous results and gives no false positives. In this application, a false positive would occur if the laser light detects traces of TNT when, in fact, no TNT is actually present on the object. Let $A$ be the event that the laser light detects traces of TNT. Let $B$ be the event that the object contains no traces of TNT. The probability of a false positive is 0. Write this probability in terms of $A$ and $B$ using symbols such as $\cup$, $\cap$, and $\mid$.

**MTBE**

3.26 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in 223 New Hampshire wells, Exercise 3.7 (p. 87). Each well is classified according to well class (public or private), aquifer (bedrock or unconsolidated), and detectable level of MTBE (below limit or detect). Consider one of these 223 wells.

a. If the well class is a public well, what is the probability that the well has a bedrock aquifer?
b. Given that the well has a bedrock aquifer, what is the probability that it has a detected level of MTBE?

3.27 *Inactive oil and gas structures.* Refer to the *Oil & Gas Journal* (Jan. 3, 2005) study of active and inactive oil and gas structures in the Gulf of Mexico, Exercise 3.19 (p. 93). The table summarizing the results of the study is reproduced here. Consider the structure type and active status of one of these oil/gas structures.

| | Caisson | Well Protector | Fixed Platform | Totals |
|---|---|---|---|---|
| *Active* | 503 | 225 | 1,447 | 2,175 |
| *Inactive* | 598 | 177 | 450 | 1,225 |

*Source:* Kaiser, M., and Mesyanzhinov, D. "Study tabulates idle Gulf of Mexico structures." *Oil & Gas Journal*, Vol 103, No. 1, Jan. 3, 2005 (Table 2).

a. Given that the structure is a fixed platform, what is the probability that the structure is active?
b. Given that the structure is inactive, what is the probability that the structure is a well protector?

**NZBIRDS**

3.28 *Extinct New Zealand birds.* Refer to the *Evolutionary Ecology Research* (July 2003) study of the patterns of extinction in the New Zealand bird population, Exercise 2.11 (p. 28). Consider the data on extinct status (extinct, absent from island, present) for the 132 bird species saved in the **NZBIRDS** file. The data is summarized in the MINITAB printout reproduced here. Suppose you randomly select 10 of the 132 bird species (without replacement) and record the extinct status of each.

a. What is the probability that the first species you select is extinct? (*Note*: Extinct = Yes on the MINITAB printout.)
b. Suppose the first 9 species you select are all extinct. What is the probability that the 10th species you select is extinct?

**Tally for Discrete Variables: Extinct**

```
Extinct  Count  Percent
 Absent     16    12.12
     No     78    59.09
    Yes     38    28.79
     N=    132
```

3.29 *Cell phone handoff behavior.* Refer to the *Journal of Engineering, Computing and Architecture* (Vol. 3., 2009) study of cell phone handoff behavior, Exercise 3.15 (p. 93). Recall that the different color codes accessed by two different model cell phones during a combined driving trip were monitored and recorded. The results (number of times each color code was accessed) are reproduced in the table below. Suppose you randomly select one point during the combined driving trips.

| | Color Code | | | | |
|---|---|---|---|---|---|
| | 0 | 5 | b | c | TOTAL |
| Model 1 | 20 | 35 | 40 | 0 | 85 |
| Model 2 | 15 | 50 | 6 | 4 | 75 |
| **TOTAL** | 35 | 85 | 46 | 4 | 160 |

a. Given the cell phone used is Model 2, what is the probability that the phone is using color code 5?
b. Given that the cell phone is using color code 5 or color code 0, what is probability that the phone used is Model 1?

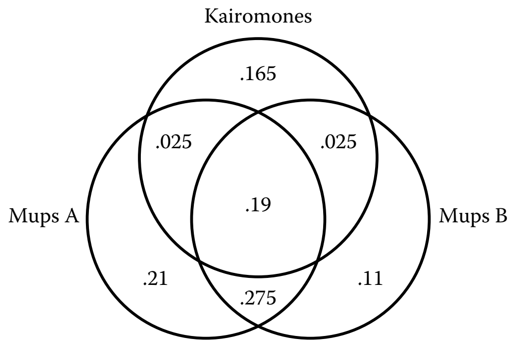
3.30 *The game of roulette.* Refer to the game of roulette and the events described in Exercise 3.20 (p. 94). Find

a. $P(A \mid B)$
b. $P(B \mid C)$
c. $P(C \mid A)$

3.31 *Data-communications system.* The probability that a data-communications system will have high selectivity is .72, the probability that it will have high fidelity is .59, and the probability that it will have both is .33. Find the probability that a system with high fidelity will also have high selectivity.

3.32 *Nuclear safety risk.* The United States Nuclear Regulatory Commission assesses the safety risks associated with nuclear power plants. The commission estimates that the probability of a nuclear reactor core meltdown is 1 in 100,000. The probability of a nuclear reactor core meltdown and less than one latent cancer fatality occurring (per year) is estimated at .00000005. Use this information to estimate the probability that at least one latent cancer fatality (per year) will occur, given a core meltdown of a nuclear reactor.

3.33 *Chemical signals of mice.* Refer to the *Cell* (May 14, 2010) study of the chemical signals of mice, Exercise 3.18 (p. 93). Recall that lab mice were exposed to odors (Mups) from rodent species A, odors from rodent species B, and kairomones (from a cat). The Venn diagram showing the proportion of cells that chemically responded to each of the three odors is reproduced below. Given that a lab mouse responds to the kairomone, how likely is it to also respond to Mups A? Mups B?




3.34 *Forest fragmentation.* Ecologists classify the cause of forest fragmentation as either anthropogenic (i.e., due to human development activities such as road construction or logging) or natural in origin (e.g., due to wetlands or wildfire). *Conservation Ecology* (Dec. 2003) published an article on the causes of fragmentation for 54 South American forests. The researchers used advanced high-resolution satellite imagery to develop fragmentation indices for each forest. A 9 $\times$ 9 grid was superimposed over an aerial photo of the forest, and each square (pixel) of the grid was classified as forest (F), anthropogenic land-use (A), or natural land-cover (N). An example of one such grid is shown here. The edges of the grid (where an "edge" is an imaginary line that separates any two adjacent pixels) are classified as A-A, N-A, F-A, F-N, N-N, or F-F edges.

| | | |
|---|---|---|
| A | A | N |
| N | F | F |
| N | F | F |

a. Refer to the grid shown. Note that there are 12 edges inside the grid. Classify each edge as A-A, N-A, F-A, F-N, N-N, or F-F.

b. The researchers calculated the fragmentation index by considering only the F-edges in the grid. Count the number of F-edges. (These edges represent the sample space for the experiment.)

c. Given that an F-edge is selected, find the probability that it is an F-A edge. (This probability is proportional to the anthropogenic fragmentation index calculated by the researchers.)

d. Given that an F-edge is selected, find the probability that it is an F-N edge. (This probability is proportional to the natural fragmentation index calculated by the researchers.)

## 3.6 Probability Rules for Unions and Intersections

Since unions and intersections of events are themselves events, we can always calculate their probabilities by adding the probabilities of the simple events that compose them. However, when the probabilities of certain events are known, it is easier to use one or both of two rules to calculate the probability of unions and intersections. How and why these rules work will be illustrated by example.

### Example 3.11

Probability of a Union: Tossing a Die

A loaded (unbalanced) die is tossed and the up face is observed. The following two events are defined:

$A$: {Observe an even number}

$B$: {Observe a number less than 3}

Suppose that $P(A) = .4$, $P(B) = .2$, and $P(A \cap B) = .1$. Find $P(A \cup B)$.
(*Note*: Assuming that we would know these probabilities in a practical situation is not very realistic, but the example will illustrate a point.)

Solution

By studying the Venn diagram in Figure 3.8, we can obtain information that will help us find $P(A \cup B)$. We can see that

$$P(A \cup B) = P(1) + P(2) + P(4) + P(6)$$

Also, we know that

$$\begin{aligned} P(A) &= P(2) + P(4) + P(6) = .4 \\ P(B) &= P(1) + P(2) = .2 \\ P(A \cap B) &= P(2) = .1 \end{aligned}$$

If we add the probabilities of the simple events that comprise events $A$ and $B$, we find

$$\begin{aligned} P(A) + P(B) &= \overbrace{P(2) + P(4) + P(6)}^{P(A)} + \overbrace{P(1) + P(2)}^{P(B)} \\ &= \overbrace{P(1) + P(2) + P(4) + P(6)}^{P(A \cup B)} + \overbrace{P(2)}^{P(A \cap B)} \end{aligned}$$

Thus, by subtraction, we have

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= .4 + .2 - .1 = .5 \end{aligned}$$

**FIGURE 3.8**
Venn diagram for die toss

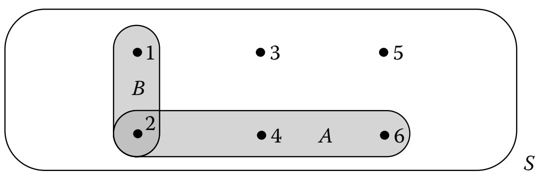
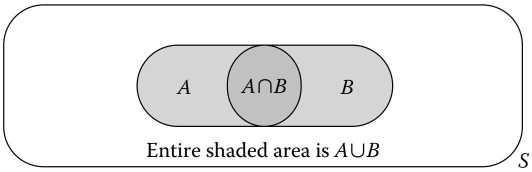


**FIGURE 3.9**
Venn diagram of union




Examining the Venn diagram in Figure 3.9, you can see that the method used in Example 3.11 may be generalized to find the union of two events for any experiment. The probability of the union of two events, $A$ and $B$, can always be obtained by summing $P(A)$ and $P(B)$ and subtracting $P(A \cap B)$. Note that we must subtract $P(A \cap B)$ because the simple event probabilities in $(A \cap B)$ have been included twice—once in $P(A)$ and once in $P(B)$.

The formula for calculating the probability of the union of two events, often called the **additive rule of probability**, is given in the box.

### Additive Rule of Probability

The probability of the union of events $A$ and $B$ is the sum of the probabilities of events $A$ and $B$ minus the probability of the intersection of events $A$ and $B$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### Example 3.12

Probability of a Union: Industrial Accidents

Records at an industrial plant show that 12% of all injured workers are admitted to a hospital for treatment, 16% are back on the job the next day, and 2% are both admitted to a hospital for treatment and back on the job the next day. If a worker is injured, what is the probability that the worker will be either admitted to a hospital for treatment, or back on the job the next day, or both?

Solution

Consider the following events:

$A$: {An injured worker is admitted to the hospital for treatment}
$B$: {An injured worker returns to the job the next day}

Then, from the information given in the statement of the example, we know that

$$P(A) = .12 \qquad P(B) = .16$$

and the probability of the event that an injured worker receives hospital treatment and returns to the job the next day is

$$P(A \cap B) = .02$$

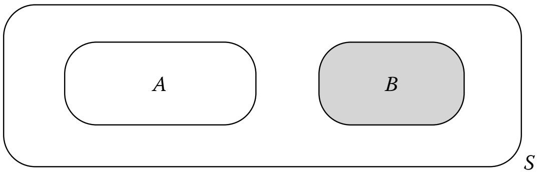
The event that an injured worker is admitted to the hospital, or returns to the job the next day, or both, is the union, $A \cup B$. The probability of $A \cup B$ is given by the additive rule of probability:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= .12 + .16 - .02 = .26 \end{aligned}$$

Thus, 26% of all injured workers either are admitted to the hospital, or return to the job the next day, or both.

A very special relationship exists between events $A$ and $B$ when $A \cap B$ contains no simple events. In this case, we call the events $A$ and $B$ **mutually exclusive** events.

**FIGURE 3.10**
Venn diagram of mutually exclusive events




**Definition 3.9**
Events $A$ and $B$ are **mutually exclusive** if $A \cap B$ contains no simple events.

Figure 3.10 shows a Venn diagram of two mutually exclusive events. The events $A$ and $B$ have no simple events in common, i.e., $A$ and $B$ cannot occur simultaneously, and $P(A \cap B) = 0$. Thus, we have the important relationship shown in the next box.

### Additive Rule for Mutually Exclusive Events

**If two events $A$ and $B$ are mutually exclusive**, the probability of the union of $A$ and $B$ equals the sum of the probabilities of $A$ and $B$:

$$P(A \cup B) = P(A) + P(B)$$

### Example 3.13

Probability of a Union: Coin Tossing Experiment

Consider the experiment of tossing two balanced coins. Find the probability of observing at least one head.

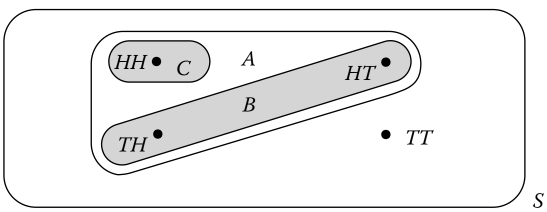
**Solution** Define the events

$A$: {Observe at least one head}
$B$: {Observe exactly one head}
$C$: {Observe exactly two heads}

Note that $A = B \cup C$ and that $B \cap C$ contains no simple events (see Figure 3.11). Thus, $B$ and $C$ are mutually exclusive, so that

$$P(A) = P(B \cup C) = P(B) + P(C)$$
$$= \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

**FIGURE 3.11**
Venn diagram for coin-toss experiment




Although Example 3.13 is very simple, the concept of writing events with verbal descriptions that include the phrases "at least" or "at most" as unions of mutually exclusive events is a very useful one. This enables us to find the probability of the event by adding the probabilities of the mutually exclusive events.

The second rule of probability, which will help us find the probability of the intersection of two events, is illustrated by Example 3.14.

**Example 3.14**

Probability of an Intersection: Sealed Bid Monitoring

A state Department of Transportation engineer is responsible for monitoring the sealed bids submitted by construction companies competing to build a new road. Of interest is the event that a construction company called FMPaving will rig the bidding process in order to improve its chances of winning the right to build the road at a noncompetitive price. This event is the intersection of the following two events:

$A$: {FMPaving bids on the job}
$B$: {FMPaving rigs the bid}

Based on information obtained from the state's computerized bid-monitoring system, the engineer estimates that the probability is .90 that FMPaving will bid on the job and that the probability is .25 that the bid will be rigged given that FMPaving bids on the job. That is,

$$P(A) = .90 \quad \text{and} \quad P(B \mid A) = .25$$

Based on the information provided, what is the probability that FMPaving will submit a rigged bid for building the new road? That is, find $P(A \cap B)$.

Solution

As you will see, we have already developed a formula for finding the probability of an intersection of two events. Recall that the conditional probability of $B$ given $A$ is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Multiplying both sides of this equation by $P(A)$, we obtain a formula for the probability of the intersection of events $A$ and $B$. This is often called the **multiplicative rule of probability** and is given by

$$P(A \cap B) = P(A)\, P(B \mid A)$$

Thus,

$$P(A \cap B) = (.90)(.25) = .225$$

The probability that FMPaving will submit a rigged bid is .225.

We formally state the multiplicative rule in the next box.

### Multiplicative Rule of Probability

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

**Example 3.15**

Independent Events: Coin Tossing Experiment

Consider the experiment of tossing a fair coin twice and recording the up face on each toss. The following events are defined:

$A$: {First toss is a head}
$B$: {Second toss is a head}

Does *knowing* that event $A$ has occurred affect the probability that $B$ will occur?

Solution

Intuitively the answer should be no, since what occurs on the first toss should in no way affect what occurs on the second toss. Let us check our intuition. Recall the sample space for this experiment:

1. Observe $HH$
2. Observe $HT$
3. Observe $TH$
4. Observe $TT$

Each of these simple events has a probability of $\frac{1}{4}$. Thus,

$$P(B) = P(HH) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \quad \text{and} \quad P(A) = P(HH) + P(HT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Now, what is $P(B \mid A)$?

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(HH)}{P(A)}$$

$$= \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

We can now see that $P(B) = \frac{1}{2}$ and $P(B \mid A) = \frac{1}{2}$. Knowing that the first toss resulted in a head does not affect the probability that the second toss will be a head. The probability is $\frac{1}{2}$ whether or not we know the result of the first toss. When this occurs, we say that the two events $A$ and $B$ are **independent**.

**Definition 3.10**

Events $A$ and $B$ are **independent** if the occurrence of $B$ does not alter the probability that $A$ has occurred, i.e., events $A$ and $B$ are independent if

$$P(A \mid B) = P(A)$$

When events $A$ and $B$ are **independent**, it will also be true that

$$P(B \mid A) = P(B)$$

Events that are not independent are said to be **dependent**.

**Example 3.16**

Independent Events: Die Tossing Experiment

Consider the experiment of tossing a fair die and define the following events:

$A$: {Observe an even number}

$B$: {Observe a number less than or equal to 4}

Are events $A$ and $B$ independent?

Solution

The Venn diagram for this experiment is shown in Figure 3.12. We first calculate

$$P(A) = P(2) + P(4) + P(6) = \frac{1}{2}$$

$$P(B) = P(1) + P(2) + P(3) + P(4) = \frac{4}{6} = \frac{2}{3}$$

$$P(A \cap B) = P(2) + P(4) = \frac{2}{6} = \frac{1}{3}$$

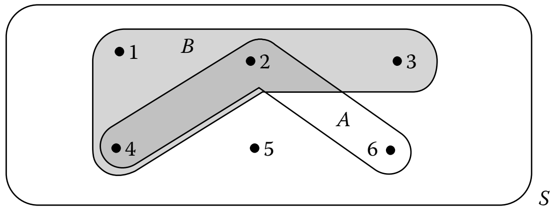
Now assuming $B$ has occurred, the conditional probability of $A$ given $B$ is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2} = P(A)$$

Thus, assuming that event $B$ occurs does not alter the probability of observing an even number—it remains $\frac{1}{2}$. Therefore, the events $A$ and $B$ are independent. Note that if we calculate the conditional probability of $B$ given $A$, our conclusion is the same:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} = P(B)$$

**FIGURE 3.12**
Venn diagram for Example 3.16




**Example 3.17**

**Dependent Events: Consumer Complaints**

Refer to the consumer product complaint study in Example 3.10 (p. 96). The percentages of complaints of various types in the pre- and postguarantee periods are shown in Table 3.3. Define the following events:

$A$: {Cause of complaint is product appearance}
$B$: {Complaint occurred during the guarantee term}

Are $A$ and $B$ independent events?

**Solution** Events $A$ and $B$ are independent if $P(A \mid B) = P(A)$. We calculated $P(A \mid B)$ in Example 3.10 to be .51, and from Table 3.2 we can see that

$$P(A) = .32 + .03 = .35$$

Therefore, $P(A \mid B)$ is not equal to $P(A)$, and $A$ and $B$ are not independent events.

We will make three final points about independence. The first is that the property of independence, unlike the mutually exclusive property, cannot be shown on or gleaned from a Venn diagram. In general, the best way to check for independence is by performing the calculations of the probabilities in the definition.

The second point concerns the relationship between the mutually exclusive and independence properties. Suppose that events $A$ and $B$ are mutually exclusive, as shown in Figure 3.10. Are these events independent or dependent? That is, does the assumption that $B$ occurs alter the probability of the occurrence of $A$? It certainly does, because if we assume that $B$ has occurred, it is impossible for $A$ to have occurred simultaneously. Thus, **mutually exclusive events are dependent events**.*

The third point is that the probability of the intersection of independent events is very easy to calculate. Referring to the formula for calculating the probability of an intersection, we find

$$P(A \cap B) = P(B)P(A \mid B)$$

Thus, since $P(A \mid B) = P(A)$ when $A$ and $B$ are independent, we have the useful rule stated in the following box.

### Multiplicative Rule for Independent Events

**If events $A$ and $B$ are independent**, the probability of the intersection of $A$ and $B$ equals the product of the probabilities of $A$ and $B$, i.e.,

$$P(A \cap B) = P(A)P(B)$$

In the die-tossing experiment, we showed in Example 3.16 that the events $A$: {Observe an even number} and $B$: {Observe a number less than or equal to 4} are independent if the die is fair. Thus,

$$P(A \cap B) = P(A)P(B) = \left(\frac{1}{2}\right)\left(\frac{2}{3}\right) = \frac{1}{3}$$

*The result holds unless one of the events has zero probability.

This agrees with the result

$$P(A \cap B) = P(2) + P(4) = \frac{2}{6} = \frac{1}{3}$$

that we obtained in the example.

**Example 3.18**

Intersection of Independent Events: Quality Control

In Example 3.3 (p. 84), a quality control engineer considered the problem of determining whether an assembly line is out of control. In the example, we discussed the problem of finding the probability that one, two, or, in general, $k$ items arriving off the assembly line are defective. We are now ready to find the probability that both of two items arriving in succession off the line are defective. Suppose that the line is out of control and that 20% of the items being produced are defective.

a. If two items arrive in succession off the line, what is the probability that they are both defective?
b. If $k$ items arrive in succession off the line, what is the probability that all are defective?

Solution

a. Let $D_1$ be the event that item 1 is defective, and let $D_2$ be a similar event for item 2. The event that *both* items will be defective is the intersection $D_1 \cap D_2$. Then, since it is not unreasonable to assume that the operating conditions of the items would be independent of one another, the probability that both will be defective is

$$\begin{aligned} P(D_1 \cap D_2) &= P(D_1)P(D_2) \\ &= (.2)(.2) = (.2)^2 = .04 \end{aligned}$$

b. Let $D_i$ represent the event that the $i$th item arriving in succession off the line is defective. Then the event that all three of three items arriving in succession will be defective is the intersection of the event $D_1 \cap D_2$ (from part **a**) with the event $D_3$. Assuming independence of the events $D_1$, $D_2$, and $D_3$, we have

$$\begin{aligned} P(D_1 \cap D_2 \cap D_3) &= P(D_1 \cap D_2)P(D_3) \\ &= (.2)^2(.2) = (.2)^3 = .008 \end{aligned}$$

Noting the pattern, you can see that the probability that all $k$ out of $k$ arriving items are defective is the probability of $D_1 \cap D_2 \cap \cdots \cap D_k$, or

$$P(D_1 \cap D_2 \cap \cdots \cap D_k) = (.2)^k \quad \text{for } k = 1, 2, 3, \ldots$$

## Applied Exercises

3.35 *Firefighters use of gas detection devices.* Two deadly gases that can be present in fire smoke are hydrogen cyanide and carbon monoxide. *Fire Engineering* (March, 2013) reported the results of a survey of 244 firefighters conducted by the Fire Smoke Coalition. The purpose of the survey was to assess the base level of knowledge of firefighters regarding the use of gas detection devices at the scene of a fire. The survey revealed the following: Eighty percent of firefighters had no standard operating procedures (SOP) for detecting/monitoring hydrogen cyanide in fire smoke; 49% had no SOP for detecting/monitoring carbon monoxide in fire smoke. Assume that 94% of firefighters had no SOP for detecting either hydrogen cyanide or carbon monoxide in fire smoke. What is the probability that a firefighter has no SOP for detecting hydrogen cyanide and no SOP for detecting carbon monoxide in fire smoke?

3.36 *Fingerprint expertise.* Software engineers are working on developing a fully automated fingerprint identification algorithm. Currently, expert examiners are required to identify the person who left the fingerprint. A study published in *Psychological Science* (August, 2011) tested the accuracy of experts and novices in identifying fingerprints. Participants were presented pairs of fingerprints and asked to judge whether the prints in each pair matched. The pairs were presented under three different conditions: prints from the same individual (*match condition*), non-matching but similar prints (*similar distracter condition*), and non-matching and very dissimilar prints (*non-similar distracter condition*). The percentages of correct decisions made by the two groups under each of the three conditions are listed in the table.

| Condition | Fingerprint experts | Novices |
|---|---|---|
| Match | 92.12% | 74.55% |
| Similar Distracter | 99.32% | 44.82% |
| Non-similar Distracter | 100.00% | 77.03% |

*Source:* Tangen, J.M., et al. "Identifying fingerprint expertise", *Psychological Science,* Vol. 22, No. 8, August, 2011(Figure 1).

a. Given a pair of matched prints, what is the probability that an expert will fail to identify the match?
b. Given a pair of matched prints, what is the probability that a novice will fail to identify the match?
c. Assume the study included 10 participants, 5 experts and 5 novices. Suppose that a pair of matched prints are presented to a randomly selected study participant and the participant fails to identify the match. Is the participant more likely to be an expert or a novice?

3.37 *Monitoring quality of power equipment. Mechanical Engineering* (Feb. 2005) reported on the need for wireless networks to monitor the quality of industrial equipment. For example, consider Eaton Corp., a company that develops distribution products. Eaton estimates that 90% of the electrical switching devices it sells can monitor the quality of the power running through the device. Eaton further estimates that of the buyers of electrical switching devices capable of monitoring quality, 90% do not wire the equipment up for that purpose. Use this information to estimate the probability that an Eaton electrical switching device is capable of monitoring power quality and is wired up for that purpose.

3.38 *Selling electric power.* Researchers for the University of Maryland Department of Civil and Environmental Engineering used stochastic dynamic programming to determine optimal load estimates for electric power (*Journal of Energy Engineering*, Apr. 2004). One objective was to determine the probability that a supplier of electric power would reach or exceed a specific net profit goal for varied load estimates. All load estimates in the study yielded a probability of .95. Consider two different suppliers of electric power (Supplier A and Supplier B) acting independently.

a. What is the probability that both suppliers reach their net profit goal?
b. What is the probability that neither supplier reaches its net profit goal?
c. What is the probability that either Supplier A or Supplier B reaches its net profit goal?

3.39 *Intrusion detection systems.* A computer intrusion detection system (IDS) is designed to provide an alarm whenever an intrusion (e.g., unauthorized access) is being attempted into a computer system. A probabilistic evaluation of a system with two independently operating intrusion detection systems (a double IDS) was published in the *Journal of Research of the National Institute of Standards and Technology* (Nov.–Dec. 2003). Consider a double IDS with system A and system B. If there is an intruder, system A sounds an alarm with probability .9 and system B sounds an alarm with probability .95. If there is no intruder, the probability that system A sounds an alarm (i.e., a false alarm) is .2 and the probability that system B sounds an alarm is .1.

a. Express the four probabilities given using symbols.
b. If there is an intruder, what is the probability that both systems sound an alarm?
c. If there is no intruder, what is the probability that both systems sound an alarm?
d. Given an intruder, what is the probability that at least one of the systems sounds an alarm?

3.40 *Ambulance response time. Geographical Analysis* (Jan., 2010) presented a study of Emergency Medical Services (EMS) ability to meet the demand for an ambulance. In one example, the researchers presented the following scenario. An ambulance station has one vehicle and two demand locations, A and B. The probability that the ambulance can travel to a location in under eight minutes is .58 for location A and .42 for location B. The probability that the ambulance is busy at any point in time is .3.

a. Find the probability that EMS can meet demand for an ambulance at location A.
b. Find the probability that EMS can meet demand for an ambulance at location B.

3.41 *Confidence of feedback information for improving quality.* In the semiconductor manufacturing industry, companies strive to improve product quality. One key to improved quality is having confidence in the feedback generated by production equipment. A study of the confidence level of feedback information was published in *Engineering Applications of Artificial Intelligence* (Vol. 26, 2013). At any point in time during the production process, a report can be generated. The report is classified as either "OK" or "not OK". As an example, the researchers provided the following probabilities: The probability of an "OK" report at one time period $(t + 1)$, given an "OK" report in the previous time period $(t)$, is .20. Also, the probability of an "OK" report at one time period $(t + 1)$, given a "not OK" report in the previous time period $(t)$, is .55. Use this information to find the probability of an "OK" report in two consecutive time periods, $t + 1$ and $t + 2$, given an "OK" report in time period $t$.

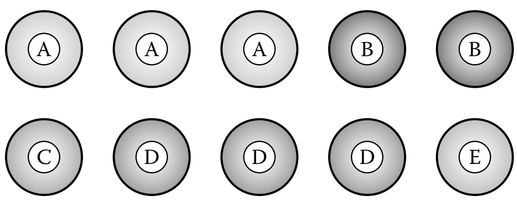
3.42 *Reflection of nuclear particles.* The transport of neutral particles in an evacuated duct is an important aspect of nuclear fusion reactor design. In one experiment, particles entering through the duct ends streamed unimpeded until they collided with the inner duct wall. Upon colliding, they were either scattered (reflected) or absorbed by the wall (*Nuclear Science and Engineering*, May 1986). The reflection probability (i.e., the probability that a particle is reflected off the wall) for one type of duct was found to be .16.

a. If two particles are released into the duct, find the probability that both will be reflected.
b. If five particles are released into the duct, find the probability that all five will be absorbed.
c. What assumption about the simple events in parts **a** and **b** is required to calculate the probabilities?

3.43 *Cornmaize seeds.* The genetic origin and properties of maize (modern-day corn) was investigated in *Economic Botany* (Jan.–Mar. 1995). Seeds from maize ears carry either single spikelets or paired spikelets, but not both. Progeny tests on approximately 600 maize ears revealed the following information. Forty percent of all seeds carry single spikelets, and 60% carry paired spikelets. A seed with single spikelets will produce maize ears with single spikelets 29% of the time and paired spikelets 71% of the time. A seed with paired spikelets will produce maize ears with single spikelets 26% of the time and paired spikelets 74% of the time.

a. Find the probability that a randomly selected maize ear seed carries a single spikelet and produces ears with single spikelets.
b. Find the probability that a randomly selected maize ear seed produces ears with paired spikelets.

3.44 *Cloning credit or debit cards.* Wireless identity theft is a technique of stealing an individual's personal information from radio-frequency-enabled cards (e.g., credit or debit cards). Upon capturing this data, thieves are able to program their own cards to respond in an identical fashion via *cloning*. A method for detecting cloning attacks in radio-frequency identification (RFID) applications was explored in *IEEE Transactions on Information Forensics and Security* (March, 2013). The method was illustrated using a simple ball drawing game. Consider a group of 10 balls, 5 representing genuine RFID cards and 5 representing clones of one or more of these cards. A labeling system was used to distinguish among the different genuine cards. Since there are 5 genuine cards, 5 letters—A, B, C, D, and E—were used. Balls labeled with the same letter represent either the genuine card or a clone of the card. Suppose the 10 balls are labeled as follows: 3 A's, 2 B's, 1 C, 3 D's, 1 E. (See figure below.) Note that the singleton C and E balls must represent the genuine cards (i.e., there are no clones of these cards). If two balls of the same letter are drawn (without replacement) from the 10 balls, then a cloning attack is detected. For this example, find the probability of detecting a cloning attack.




3.45 *Encryption systems with erroneous ciphertexts.* In cryptography, ciphertext is encrypted or encoded text that is unreadable by a human or computer without the proper algorithm to decrypt it into plaintext. The impact of erroneous ciphertexts on the performance of an encryption system was investigated in *IEEE Transactions on Information Forensics and Security* (April, 2013). For one data encryption system, the probability of receiving an erroneous ciphertext is assumed to be $\beta$, where $0 < \beta < 1$. The researchers showed that if an erroneous ciphertext occurs, the probability of an error in restoring plaintext using the decryption system is .5. When no error occurs in the received ciphertext, the probability of an error in restoring plaintext using the decryption system is $\alpha\beta$, where $0 < \alpha < 1$. Use this information to give an expression for the probability of an error in restoring plaintext using the decryption system.

3.46 *Truck accident study. Transportation Quarterly* (Jan. 1993) identified several studies on truck accidents that utilized misleading or inappropriate probability analysis. Consider the following excerpt from the article.* Can you find the flaw in the probability argument?

> *For example, consider a situation where only two vehicle types are present in the traffic mix: trucks at 20% and cars at 80% of the total. If only two vehicle accidents are considered, the probability of occurrence for all events in the sample space would be as follows:*

$$\begin{aligned}
P\text{ (truck impacting truck)} &= P(\text{TT}) = .20 \times .20 = .04 \\
P\text{ (truck impacting car)} &= P(\text{TC}) = .20 \times .80 = .16 \\
P\text{ (car impacting truck)} &= P(\text{CT}) = .80 \times .20 = .16 \\
P\text{ (car impacting car)} &= P(\text{CC}) = .80 \times .80 = \underline{.64} \\
&\phantom{=} 1.00
\end{aligned}$$

> *Many analysts [used the calculation* $P(\text{TT}) + P(\text{TC}) + P(\text{CT}) = .36$ *to conclude] that trucks, constituting 20% of the traffic, are involved in 36% of all two-vehicle accidents. Hence cars, which constitute 80% of the traffic, are involved in* 100% − 36% = 64% *of the accidents.*

## Optional Applied Exercises

3.47 Writing in *Environmental Science & Technology* (May 1986), Joseph Fiksel researched the problem of compensating victims of chronic diseases (such as cancer and birth defects) who are exposed to hazardous and toxic substances. The key to compensation, as far as the U.S. judicial system is concerned, is the *probability of causation* (i.e., the likelihood, for a person developing the disease, that the cause is due to exposure to the hazardous substance). Usually, the

*Bowman, B. L., and Hummzer, J. "Data validity barriers to determining magnitude of large truck accident problem." *Transportation Quarterly*, Vol. 47, No. 1, Jan. 1993, p.40.

probability of causation must be greater than .50 for the court to award compensation. Fiksel gave examples of how to calculate the probability of causation for several different scenarios.†

a. "*Ordinary causation,*" as defined by Fiksel, "describes a situation in which the presence of a single factor, such as asbestos insulation, is believed to cause an effect, such as mesothelioma." For this situation, define the following events:

   $D$: {Effect (disease) occurs}
   $A$: {Factor A present}

   Under ordinary causation, if $A$ occurs, then $D$ must occur. However, $D$ can also occur when factor A is not present. The probability of causation for factor A, then, is the conditional probability $P(A \mid D)$. Show that the probability of causation for factor A is

$$P(A \mid D) = \frac{P_1 - P_0}{P_1}$$

   where $P_0$ is the probability that the effect occurs when factor A is *not* present, and $P_1$ is the probability that the effect occurs when factor A is either present or not. (*Note*: To epidemiologists, $P_1$ is often called the *overall risk rate* for the disease and $P_1 - P_0$ is the *additional risk* attributable to the presence of factor A.) (*Hint*: The simple events for this experiment are $\{D^c \cap A^c, D \cap A, \text{ and } D \cap A^c\}$. Write $P_0$ and $P_1$ in terms of the simple events.)

b. Fiksel defined *simultaneous exclusive causation* as "a situation in which two or more causal factors are present but the resulting effect is caused by *one and only one* of these factors." Consider two factors, A and B, and let $B$ be the event that factor B is present. In this situation, if either $A$ or $B$ occurs, then $D$ must occur. However, both $A$ and $B$ cannot occur simultaneously (i.e., $A$ and $B$ are mutually exclusive). Assuming that $D$ cannot occur when neither $A$ nor $B$ occurs, show that the probabilities of causation for factors A and B are, respectively,

$$P(A \mid D) = \frac{P_1}{P_1 + P_2} \text{ and } P(B \mid D) = \frac{P_2}{P_1 + P_2}$$

   where $P_1$ is the probability that the effect occurs when factor A is present and $P_2$ is the probability that the effect occurs when factor B is present. (*Hint*: The simple events for this experiment are $D \cap A \cap B^c$, and $D \cap A^c \cap B\}$.)

c. *Simultaneous joint causation*, wrote Fiksel, describes a more realistic situation "in which several factors can contribute in varying degrees to the occurrence of an effect. For example, a cigarette smoker who is exposed to radiation and chemical carcinogens in the workplace may develop a lung tumor. Whether the tumor was caused wholly by one factor or by a combination of factors is, at present, impossible to determine." For this case, consider two factors, A and B, which affect $D$ independently. Also, assume that if either $A$ or $B$, or both occur, then $D$ must occur; $D$ cannot occur if neither $A$ nor $B$ occurs. Show that the probabilities of causation for factors A and B are, respectively,

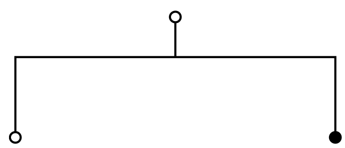
$$P(A \mid D) = \frac{P_1}{P_1 + P_2 - P_1P_2}$$

   and

$$P(B \mid D) = \frac{P_2}{P_1 + P_2 - P_1P_2}$$

   where $P_1$ is the probability that the effect occurs when factor A is present and $P_2$ is the probability that the effect occurs when factor B is present. (*Hint*: The simple events for this experiment are $\{D^c \cap A^c \cap B^c, D \cap A \cap B^c, D \cap A^c \cap B, \text{ and } D \cap A \cap B\}$.)

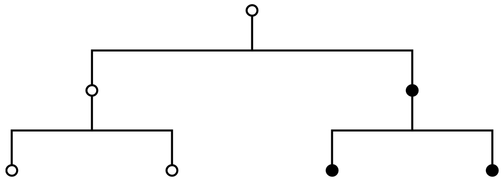
3.48 *Random mutation of cells. Chance* (Spring, 2010) presented an article on the random mutation hypothesis developed by microbiologists. Under this hypothesis, when a wild-type organic cell (e.g., a bacteria cell) divides, there is a chance that one of the two "daughter" cells is a mutant. When a mutant cell divides, both offspring will be mutant. The schematic below shows a possible pedigree from a single cell that has divided. Note that one "daughter" cell is mutant (black) and one is a normal cell.



a. Consider a single, normal cell that divides into two offspring. List the different possible pedigrees.
b. Assume that a "daughter" cell is equally likely to be mutant or normal. What is the probability that a single, normal cell that divides into two offspring will result in at least one mutant cell?
c. Now assume that the probability of a mutant "daughter" cell is .2. What is the probability that a single, normal cell that divides into two offspring will result in at least one mutant cell?

†Fiksel, J. "Victim compensation: Understanding the problem of indeterminate causation." *Environmental Science and Technology*, May 1986. 

d. The schematic below shows a possible 2nd-generation pedigree from a single cell that has divided. Note that the 1st generation mutant cell automatically produces two mutant cells in the 2nd-generation. List the different possible 2nd-generation pedigrees. (Hint: Use your answer to part a.)



e. Assume that a "daughter" cell is equally likely to be mutant or normal. What is the probability that a single, normal cell that divides into two offspring will result in at least one mutant cell after the 2nd generation?

## 3.7 Bayes' Rule (*Optional*)

An early attempt to employ probability in making inferences is the basis for a branch of statistical methodology known as **Bayesian statistical methods**. The logic employed by the English philosopher, the Reverend Thomas Bayes in the mid-1700's, involves converting an unknown conditional probability, $P(B \mid A)$, to one involving a known conditional probability, $P(A \mid B)$. The method is illustrated in the next example.

### Example 3.19

Applying Bayes' Logic—Intruder Detection System

An unmanned monitoring system uses high-tech video equipment and microprocessors to detect intruders. A prototype system has been developed and is in use outdoors at a weapons munitions plant. The system is designed to detect intruders with a probability of .90. However, the design engineers expect this probability to vary with weather condition. The system automatically records the weather condition each time an intruder is detected. Based on a series of controlled tests, in which an intruder was released at the plant under various weather conditions, the following information is available: Given the intruder was, in fact, detected by the system, the weather was clear 75% of the time, cloudy 20% of the time, and raining 5% of the time. When the system failed to detect the intruder, 60% of the days were clear, 30% cloudy, and 10% rainy. Use this information to find the probability of detecting an intruder, given rainy weather conditions. (Assume that an intruder has been released at the plant.)

Solution

Define $D$ to be the event that the intruder is detected by the system. Then $D^c$ is the event that the system failed to detect the intruder. Our goal is to calculate the conditional probability, $P(D \mid \text{Rainy})$. From the statement of the problem, the following information is available:

$$P(D) = .90 \qquad P(D^c) = .10$$
$$P(\text{Clear} \mid D) = .75 \qquad P(\text{Clear} \mid D^c) = .60$$
$$P(\text{Cloudy} \mid D) = .20 \qquad P(\text{Cloudy} \mid D^c) = .30$$
$$P(\text{Rainy} \mid D) = .05 \qquad P(\text{Rainy} \mid D^c) = .10$$

Then

$$P(\text{Rainy} \cap D) = P(D)P(\text{Rainy} \mid D) = (.90)(.05) = .045$$

and

$$P(\text{Rainy} \cap D^c) = P(D^c)P(\text{Rainy} \mid D^c) = (.10)(.10) = .01$$

The event Rainy is the union of two mutually exclusive events, $(\text{Rainy} \cap D)$ and $(\text{Rainy} \cap D^c)$. Thus,

$$P(\text{Rainy}) = P(\text{Rainy} \cap D) + P(\text{Rainy} \cap D^c) = .045 + .01 = .055$$

We now apply the formula for conditional probability to obtain:

$$P(D \mid \text{Rainy}) = \frac{P(\text{Rainy} \cap D)}{P(\text{Rainy})} = \frac{P(\text{Rainy} \cap D)}{P(\text{Rainy} \cap D) + P(\text{Rainy} \cap D^c)}$$

$$= \frac{.045}{.055} = .818$$

Therefore, under rainy weather conditions, the prototype system can detect the intruder with a probability of .818—a value lower than the designed probability of .90.

The technique utilized in Example 3.19, called **Bayes' method**, can be applied when an observed event $E$ occurs with any one of $k$ mutually exclusive and exhaustive states of nature (or events), $A_1, A_2, \ldots, A_k$. The formula for finding the appropriate conditional probabilities is given in the box.

### Bayes' Rule

Given $k$ mutually exclusive and exhaustive states of nature (events), $A_1, A_2, \ldots, A_k$, and an observed event $E$, then $P(A_i \mid E)$, for $i = 1, 2, \ldots, k$, is

$$P(A_i \mid E) = \frac{P(A_i \cap E)}{P(E)}$$

$$= \frac{P(A_i)P(E \mid A_i)}{P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) + \cdots + P(A_k)P(E \mid A_k)}$$

In applying Bayes' rule to Example 3.19, the observed event $E$ is $\{\text{Rainy}\}$ and the mutually exclusive and exhaustive states of nature are $D$ (intruder detected) and $D^c$ (intruder not detected). Hence, the formula

$$P(D \mid \text{Rainy}) = \frac{P(D)P(\text{Rainy} \mid D)}{P(D)P(\text{Rainy} \mid D) + P(D^c)P(\text{Rainy} \mid D^c)}$$

$$= \frac{(.90)(.05)}{(.90)(.05) + (.10)(.10)} = .818$$

[*Note*: In Exercise 3.98, you use Bayes' rule to find $P(D \mid \text{Clear})$ and $P(D \mid \text{Cloudy})$.]

## Applied Exercises

3.49 *Electric wheelchair control.* Electric wheelchairs are difficult to maneuver for many disabled people. In a paper presented at the *1st International Workshop on Advances in Service Robotics* (March, 2003), researchers applied Bayes' Rule to evaluate an "intelligent" robotic controller that aims to capture the intent of a wheelchair user and aid in navigation. Consider the following scenario. From a certain location in a room, a wheelchair user will either (1) turn sharply to the left and navigate through a door, (2) proceed straight to the other side of the room, or (3) turn slightly right and stop at a table. Denote these three events as D (for door), S (Straight), and T (for table). Based on previous trips, $P(D) = .5$, $P(S) = .2$, and $P(T) = .3$. The wheelchair is installed with a robot-controlled joystick. When the user intends to go through the door, he points the joystick straight 30% of the time; when the user intends to go straight, he points the joystick straight 40% of the time; and, when the user intends to go to the table, he points the joystick straight 5% of the time. If the wheelchair user points the joystick straight, where is his most likely destination?

3.50 *Nondestructive evaluation.* Nondestructive evaluation (NDE) describes methods that quantitatively characterize materials, tissues, and structures by noninvasive means such as X-ray computed tomography, ultrasonics, and acoustic emission. Recently, NDE was used to detect defects in steel castings. (*JOM*, May 2005.) Assume that the probability that NDE detects a "hit" (i.e., predicts a defect in a steel casting) when, in fact, a defect exists is .97. (This is often called the probability of detection.) Also assume that the probability that NDE detects a "hit" when, in fact, no defect exists is .005. (This is called the probability of a false call.) Past experience has shown that a defect occurs once in every 100 steel castings. If NDE detects a "hit" for a particular steel casting, what is the probability that an actual defect exists?

3.51 *Fingerprint expertise.* Refer to the *Psychological Science* (August, 2011) study comparing the accuracy of experts and novices in identifying fingerprints, Exercise 3.36 (p. 105). The percentages of correct decisions made by the two groups under each of the three conditions are reproduced in the table. Assume the study included 10 participants, 5 experts and 5 novices. Suppose that a pair of matched prints are presented to a randomly selected study participant and the participant fails to identify the match. Is the participant more likely to be an expert or a novice?

| Condition | Fingerprint experts | Novices |
|---|---|---|
| Match | 92.12% | 74.55% |
| Similar Distracter | 99.32% | 44.82% |
| Non-similar Distracter | 100% | 77.03% |

*Source:* Tangen, J.M., et al. "Identifying fingerprint expertise", *Psychological Science,* Vol. 22, No. 8, August, 2011(Figure 1).

3.52 *Drug testing in athletes.* Due to inaccuracies in drug testing procedures (e.g., false positives and false negatives), in the medical field the results of a drug test represent only one factor in a physician's diagnosis. Yet when Olympic athletes are tested for illegal drug use (i.e., doping), the results of a single test are used to ban the athlete from competition. In *Chance* (Spring 2004), University of Texas biostatisticians D. A. Berry and L. Chastain demonstrated the application of Bayes' Rule for making inferences about testosterone abuse among Olympic athletes. They used the following example. In a population of 1,000 athletes, suppose 100 are illegally using testosterone. Of the users, suppose 50 would test positive for testosterone. Of the nonusers, suppose 9 would test positive.

a. Given the athlete is a user, find the probability that a drug test for testosterone will yield a positive result. This probability represents the *sensitivity* of the drug test.)
b. Given the athlete is a nonuser, find the probability that a drug test for testosterone will yield a negative result. (This probability represents the *specificity* of the drug test.)
c. If an athlete tests positive for testosterone, use Bayes' Rule to find the probability that the athlete is really doping. (This probability represents the *positive predictive value* of the drug test.)

3.53 *Errors in estimating job costs.* A construction company employs three sales engineers. Engineers 1, 2, and 3 estimate the costs of 30%, 20%, and 50%, respectively, of all jobs bid by the company. For $i = 1,2,3$, define $E_i$ to be the event that a job is estimated by engineer $i$. The following probabilities describe the rates at which the engineers make serious errors in estimating costs:

$$P(\text{error} \mid E_1) = .01,\ P(\text{error} \mid E_2) = .03, \text{ and}$$
$$P(\text{error} \mid E_3) = .02$$

a. If a particular bid results in a serious error in estimating job cost, what is the probability that the error was made by engineer 1?
b. If a particular bid results in a serious error in estimating job cost, what is the probability that the error was made by engineer 2?
c. If a particular bid results in a serious error in estimating job cost, what is the probability that the error was made by engineer 3?
d. Based on the probabilities, parts a–c, which engineer is most likely responsible for making the serious error?

3.54 *Intrusion detection systems.* Refer to the *Journal of Research of the National Institute of Standards and Technology* (Nov.–Dec. 2003) study of a double intrusion detection system with independent systems, Exercise 3.39 (p. 106). Recall that if there is an intruder, system A sounds an alarm with probability .9 and system B sounds an alarm with probability .95. If there is no intruder, system A sounds an alarm with probability .2 and system B sounds an alarm with probability .1. Now, assume that the probability of an intruder is .4. If both systems sound an alarm, what is the probability that an intruder is detected?

3.55 *Mining for dolomite.* Dolomite is a valuable mineral that is found in sedimentary rock. During mining operations, dolomite is often confused with shale. The radioactivity features of rock can aid miners in distinguishing between dolomite and shale rock zones. For example, if the Gamma ray reading of a rock zone exceeds 60 API units, the area is considered to be mostly shale (and is not mined); if the Gamma ray reading of a rock zone is less than 60 API units, the area is considered to be abundant in dolomite (and is mined). Data on 771 core samples in a rock quarry collected by the Kansas Geological Survey revealed the following: 476 of the samples are dolomite and 295 of the samples are shale. Of the 476 dolomite core samples, 34 had a Gamma ray reading greater than 60. Of the 295 shale core samples, 280 had a Gamma ray reading greater than 60. Suppose you obtain a Gamma ray reading greater than 60 at a certain depth of the rock quarry. Should this area be mined?

3.56 *Purchasing microchips.* An important component of your desktop or laptop personal computer (PC) is a microchip. The table gives the proportions of microchips that a certain PC manufacturer purchases from seven suppliers.

| Supplier | Proportion |
|---|---|
| $S_1$ | .15 |
| $S_2$ | .05 |
| $S_3$ | .10 |
| $S_4$ | .20 |
| $S_5$ | .12 |
| $S_6$ | .20 |
| $S_7$ | .18 |

a. It is known that the proportions of defective microchips produced by the seven suppliers are .001, .0003, .0007, .006, .0002, .0002, and .001, respectively. If a single PC microchip failure is observed, which supplier is most likely responsible?

b. Suppose the seven suppliers produce defective microchips at the same rate, .0005. If a single PC microchip failure is observed, which supplier is most likely responsible?

## Optional Applied Exercise

3.57 *Forensic analysis of JFK assassination bullets.* Following the assassination of President John F. Kennedy (JFK) in 1963, the House Select Committee on Assassinations (HSCA) conducted an official government investigation. The HSCA concluded that although there was a probable conspiracy, involving at least one additional shooter other than Lee Harvey Oswald, the additional shooter missed all limousine occupants. A recent analysis of assassination bullet fragments, reported in the *Annals of Applied Statistics* (Vol. 1, 2007), contradicted these findings, concluding that evidence used to rule out a second assassin by the HSCA is fundamentally flawed. It is well documented that at least two different bullets were the source of bullet fragments used in the assassination. Let E = {bullet evidence used by the HSCA}, T = {two bullets used in the assassination}, and $T^C$ = {more than two bullets used in the assassination}. Given the evidence (E), which is more likely to have occurred—two bullets used (T) or more than two bullets used ($T^C$)?

a. The researchers demonstrated that the ratio, $P(T \mid E)/P(T^C \mid E)$, is less than 1. Explain why this result supports the theory of more than two bullets used in the assassination of JFK.

b. To obtain the result, part a, the researchers first showed that

$$P(T \mid E)/P(T^C \mid E) = [P(E \mid T)\, P(T)]/[P(E \mid T^C)\, P(T^C)]$$

Demonstrate this equality using Bayes' Theorem.

## 3.8 Some Counting Rules

In Section 3.2 we pointed out that experiments sometimes have so many simple events that it is impractical to list them all. However, many of these experiments possess simple events with identical characteristics. If you can develop a **counting rule** to count the number of simple events for such an experiment, it can be used to aid in the solution of the problems.

**Example 3.20**

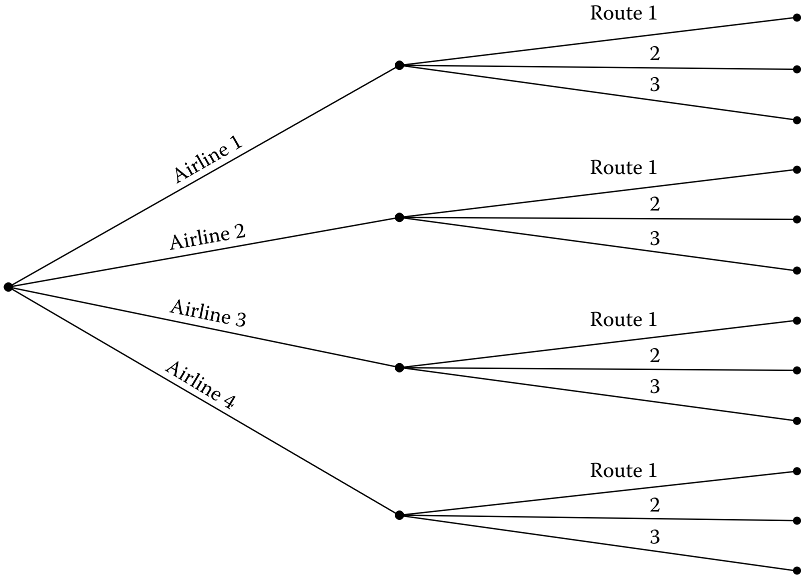
Multiplicative Rule: Routing Problem

A product (e.g., hardware for a networked computer system) can be shipped by four different airlines, and each airline can ship via three different routes. How many distinct ways exist to ship the product?

**Solution** A pictorial representation of the different ways to ship the product will aid in counting them. This representation, called a **decision tree**, is shown in Figure 3.13. At the starting point (stage 1), there are four choices—the different airlines—to begin the journey. Once we have chosen an airline (stage 2), there are three choices—the different routes—to complete the shipment and reach the final destination. Thus, the decision tree clearly shows that there are $(4)(3) = 12$ distinct ways to ship the product.

The method of solving Example 3.20 can be generalized to any number of stages with sets of different elements. The framework is provided by the **multiplicative rule**.

**FIGURE 3.13**
Decision tree for shipping problem




### THEOREM 3.1

**The Multiplicative Rule** You have $k$ sets of elements—$n_1$ in the first set, $n_2$ in the second set, . . . , and $n_k$ in the $k$th set. Suppose you want to form a sample of $k$ elements by taking one element from each of the $k$ sets. The number of different samples that can be formed is the product

$$n_1 n_2 n_3 \ldots\ldots n_k$$

**Outline of Proof of Theorem 3.1** The proof of Theorem 3.1 can be obtained most easily by examining Table 3.4. Each of the pairs that can be formed from two sets of elements—$a_1, a_2, \ldots, a_{n1}$ and $b_1, b_2, \ldots, b_{n2}$—corresponds to a cell of Table 3.4.

Since the table contains $n_1$ rows and $n_2$ columns, there will be $n_1 n_2$ pairs corresponding to each of the $n_1 n_2$ cells of the table. To extend the proof to the case in which $k = 3$, note that the number of triplets that can be formed from three sets of elements—$a_1, a_2, \ldots, a_{n1}$; $b_1, b_2, \ldots, b_{n2}$; and $c_1, c_2, \ldots, c_{n3}$—is equal to the number of pairs that can be formed by associating one of the $a_\mathrm{i} b_\mathrm{j}$ pairs with one of the $c$ elements. Since there are $(n_1, n_2)$ of the $a_\mathrm{i} b_\mathrm{j}$ pairs and $n_3$ of the $c$ elements, we can form $(n_1 n_2) n_3 = n_1 n_2 n_3$ triplets consisting of one $a$ element, one $b$ element, and one $c$ element. The proof of the multiplicative rule for any number, say, $k$, of sets is obtained by mathematical induction. We leave this proof as an exercise for the student.

**TABLE 3.4 Pairings of $a_1, a_2, \ldots, a_{n1}$, and $b_1, b_2, \ldots, b_{n2}$**

| | $b_1$ | $b_2$ | $b_3$ | . . . | $b_{n2}$ |
|---|---|---|---|---|---|
| $a_1$ | $a_1 b_1$ | $a_1 b_2$ | $a_1 b_3$ | . . . | $a_1 b_{n2}$ |
| $a_2$ | $a_2 b_1$ | . . . | . . . | . . . | . . . |
| $a_3$ | $a_3 b_1$ | . . . | . . . | . . . | . . . |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a_{n1}$ | $a_{n1} b_1$ | . . . | . . . | . . . | $a_{n1} b_{n2}$ |

**Example 3.21**

Multiplicative Rule: Candidate Selection Problem

There are 20 candidates for three different mechanical engineer positions, $E_1$, $E_2$, and $E_3$. How many different ways could you fill the positions?

Solution

This example consists of the following $k = 3$ sets of elements:

Set 1: Candidates available to fill position $E_1$
Set 2: Candidates remaining (after filling $E_1$) that are available to fill $E_2$
Set 3: Candidates remaining (after filling $E_1$ and $E_2$) that are available to fill $E_3$

The numbers of elements in the sets are $n_1 = 20$, $n_2 = 19$, $n_3 = 18$. Therefore, the number of different ways of filling the three positions is

$$n_1 n_2 n_3 = (20)(19)(18) = 6{,}840$$

**Example 3.22**

Multiplicative Rule: Assembly Line Inspection

Consider an experiment that consists of selecting 10 items from an assembly line for inspection, with each item classified as a defective (*D*) or nondefective (*N*). (Recall Example 3.8.) Show that there are $2^{10} = 1{,}024$ simple events for this experiment.

Solution

There are $k = 10$ sets of elements for this experiment. Each set contains two elements, a defective and a nondefective. Thus, there are

$$(2)(2)(2)(2)(2)(2)(2)(2)(2)(2) = 2^{10} = 1{,}024$$

different outcomes (simple events) of this experiment.

**Example 3.23**

Permutation Rule: Assignment Problem

Suppose there are five different space flights scheduled, each requiring one astronaut. Assuming that no astronaut can go on more than one space flight, in how many different ways can 5 of the country's top 100 astronauts be assigned to the 5 space flights?

Solution

We can solve this problem by using the multiplicative rule. The entire set of 100 astronauts is available for the first flight, and after the selection of one astronaut for that flight, 99 are available for the second flight, etc. Thus, the total number of different ways of choosing five astronauts for the five space flights is

$$n_1 n_2 n_3 n_4 n_5 = (100)(99)(98)(97)(96) = 9{,}034{,}502{,}400$$

The *arrangement of elements in a distinct order* is called a **permutation**. Thus, from Example 3.23, we see that there are more than 9 billion different *permutations* of 5 elements (astronauts) drawn from a set of 100 elements!

**THEOREM 3.2**

Permutations Rule Given a single set of $N$ distinctly different elements, you wish to select $n$ elements from the $N$ and arrange them within $n$ positions in a distinct order. The number of different permutations of the $N$ elements taken $n$ at a time is denoted by $P_n^N$ and is equal to

$$P_n^N = N(N-1)(N-2)\cdot\cdots\cdot(N-n+1) = \frac{N!}{(N-n)!}$$

where $n! = n(n-1)(n-2)\cdot\cdots\cdot(3)(2)(1)$ and is called ***n* factorial**. (Thus, for example, $5! = 5\cdot 4\cdot 3\cdot 2\cdot 1 = 120$.) The quantity $0!$ is defined to be equal to 1.

**Proof of Theorem 3.2** The proof of Theorem 3.2 is a generalization of the solution to Example 3.23. There are $N$ ways of filling the first position. After it is filled, there are $N - 1$ ways of filling the second, $N - 2$ ways of filling the third, $\ldots$, and $(N - n + 1)$ ways of filling the $n$th position. We apply the multiplicative rule to obtain

$$P_n^N = (N)(N-1)(N-2)\cdot\cdots\cdot(N-n+1) = \frac{N!}{(N-n)!}$$

**Example 3.24**

Permutation Rule: Transportation Engineering

Consider the following transportation engineering problem: You want to drive, in sequence, from a starting point to each of five cities, and you want to compare the distances and average speeds of the different routings. How many different routings would have to be compared?

Solution

Denote the cities as $C_1, C_2, \ldots, C_5$. Then a route moving from the starting point to $C_2$ to $C_1$ to $C_3$ to $C_4$ to $C_5$ would be represented as $C_2C_1C_3C_4C_5$. The total number of routings would equal the number of ways you could rearrange the $N = 5$ cities in $n = 5$ positions. This number is

$$P_n^N = P_5^5 = \frac{5!}{(5-5)!} = \frac{5!}{0!} = \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{1} = 120$$

(Recall that $0! = 1$.)

**Example 3.25**

Partitions Rule: Assignment Problem

There are four system analysts, and you must assign three to job 1 and one to job 2. In how many different ways can you make this assignment?

Solution

To begin, suppose that each system analyst is to be assigned to a distinct job. Then, using the multiplicative rule, we obtain $(4)(3)(2)(1) = 24$ ways of assigning the system analysts to four distinct jobs. The 24 ways are listed in four groups in Table 3.5 (where ABCD indicates that system analyst A was assigned the first job; system analyst B, the second; etc.).

Now, suppose the first three positions represent job 1 and the last position represents job 2. We can now see that all the listings in group 1 represent the same outcome of the experiment of interest. That is, system analysts A, B, and C are assigned to job 1 and system analyst D is assigned to job 2. Similarly, group 2 listings are equivalent, as are group 3 and group 4 listings. Thus, there are only four different assignments of four system analysts to the two jobs. These are shown in Table 3.6.

**TABLE 3.5 Ways to Assign System Analysts to Four Distinct Jobs**

| Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|
| ABCD | ABDC | ACDB | BCDA |
| ACBD | ADBC | ADCB | BDCA |
| BACD | BADC | CADB | CBDA |
| BCAD | BDAC | CDAB | CDBA |
| CABD | DABC | DACB | DBCA |
| CBAD | DBAC | DCAB | DCBA |

**TABLE 3.6 Ways to Assign Three System Analysts to Job 1 and One System Analyst to Job 2**

| Job 1 | Job 2 |
|---|---|
| ABC | D |
| ABD | C |
| ACD | B |
| BCD | A |

To generalize the result obtained in Example 3.25, we point out that the final result can be found by

$$\frac{(4)(3)(2)(1)}{(3)(2)(1)(1)} = 4$$

The (4)(3)(2)(1) is the number of different ways (*permutations*) the system analysts could be assigned four distinct jobs. The division by (3)(2)(1) is to remove the duplicated permutations resulting from the fact that three system analysts are assigned the same jobs. And the division by (1) is associated with the system analyst assigned to job 2.

**THEOREM 3.3**

**Partitions Rule** There exists a single set of $N$ distinctly different elements and you want to partition them into $k$ sets, the first set containing $n_1$ elements, the second containing $n_2$ elements, . . . , and the $k$th set containing $n_k$ elements. The number of different partitions is

$$\frac{N!}{n_1!n_2! \cdot \cdots \cdot n_k!} \quad \text{where } n_1 + n_2 + n_3 + \cdots + n_k = N$$

**Proof of Theorem 3.3** Let $A$ equal the number of ways that you can partition $N$ distinctly different elements into $k$ sets. We want to show that

$$A = \frac{N!}{n_1!n_2! \cdot \cdots \cdot n_k!}$$

We will find $A$ by writing an expression for arranging $N$ distinctly different elements in $N$ positions. By Theorem 3.2, the number of ways this can be done is

$$P_N^N = \frac{N!}{(N-N)!} = \frac{N!}{0!} = N!$$

But, by Theorem 3.1, $P_N^N$ is also equal to the product

$$P_N^N = N! = (A)(n_1!)(n_2!)\cdots(n_k!)$$

where $A$ is the number of ways of partitioning $N$ elements into $k$ groups of $n_1, n_2, \ldots, n_k$ elements, respectively; $n_1!$ is the number of ways of arranging the $n_1$ elements in group 1; $n_2!$ is the number of ways of arranging the $n_2$ elements in group 2; . . . ; and $n_k!$ is the number of ways of arranging the $n_k$ elements in group $k$. We obtain the desired result by solving for $A$:

$$A = \frac{N!}{n_1!n_2! \cdot \cdots \cdot n_k!}$$

**Example 3.26**

**Partitions Rule: Another Assignment Problem**

You have 12 system analysts and you want to assign three to job 1, four to job 2, and five to job 3. In how many different ways can you make this assignment?

**Solution**

For this example, $k = 3$ (corresponding to the $k = 3$ different jobs), $N = 12$, $n_1 = 3$, $n_2 = 4$, and $n_3 = 5$. Then the number of different ways to assign the system analysts to the jobs is

$$\frac{N!}{n_1!n_2!n_3!} = \frac{12!}{3!4!5!} = \frac{12 \cdot 11 \cdot 10 \cdot \cdots \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(4 \cdot 3 \cdot 2 \cdot 1)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = 27{,}720$$

**Example 3.27**

Combinations Rule: Sampling Problem

How many samples of 4 tin-lead solder joints can be selected from a lot of 25 tin-lead solder joints available for strength tests?

Solution

For this example, $k = 2$ (corresponding to the $n_1 = 4$ solder joints you *do* choose and the $n_2 = 21$ solder joints you do *not* choose) and $N = 25$. Then, the number of different ways to choose the 4 solder joints from 25 is

$$\frac{N!}{n_1!n_2!} = \frac{25!}{(4!)(21!)} = \frac{25 \cdot 24 \cdot 23 \cdot \cdots \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1)(21 \cdot 20 \cdot \cdots \cdot 2 \cdot 1)} = 12{,}650$$

The special application of the partitions rule illustrated by Example 3.27—partitioning a set of $N$ elements into $k = 2$ groups (the elements that appear in a sample and those that do not)—is very common. Therefore, we give a different name to the rule for counting the number of different ways of partitioning a set of elements into two parts—the **combinations rule**.

**THEOREM 3.4**

**The Combinations Rule** A sample of $n$ elements is to be chosen from a set of $N$ elements. Then the number of different samples of $n$ elements that can be selected from $N$ is denoted by $\binom{N}{n}$ and is equal to

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Note that the order in which the $n$ elements are drawn is not important.

**Proof of Theorem 3.4** The proof of Theorem 3.4 follows directly from Theorem 3.3. Selecting a sample of $n$ elements from a set of $N$ elements is equivalent to partitioning the $N$ elements into $k = 2$ groups—the $n$ that are selected for the sample and the remaining $(N - n)$ that are not selected. Therefore, by applying Theorem 3.3 we obtain

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

**Example 3.28**

Combinations Rule: Group Selection Problem

Five sales engineers will be hired from a group of 100 applicants. In how many ways (*combinations*) can groups of 5 sales engineers be selected?

Solution

This is equivalent to sampling $n = 5$ elements from a set of $N = 100$ elements. Thus, the number of ways is the number of possible combinations of 5 applicants selected from 100, or

$$\binom{100}{5} = \frac{100!}{(5!)(95!)} = \frac{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \cdot 95 \cdot 94 \cdot \cdots \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)(95 \cdot 94 \cdot \cdots \cdot 2 \cdot 1)}$$

$$= \frac{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 75{,}287{,}520$$

Compare this result with that of Example 3.23, where we found that the number of permutations of 5 elements drawn from 100 was more than 9 billion. ***Because the order of the elements does not affect combinations, there are fewer combinations than permutations***.

When working a probability problem, you should carefully examine the experiment to determine whether you can use one or more of the rules we have discussed in this section. A summary of these rules is shown below. We will illustrate in Examples 3.29 and 3.30 how these rules can help solve a probability problem.

### Summary of Counting Rules

1. *Multiplicative rule:* If you are drawing one element from each of $k$ sets of elements, with the sizes of the sets $n_1, n_2, \cdots, n_k$, the number of different results is

$$n_1 n_2 n_3 \cdot \cdots \cdot n_k$$

2. *Permutations rule:* If you are drawing $n$ elements from a set of $N$ elements and arranging the $n$ elements in a distinct order, the number of different results is

$$P_n^N = \frac{N!}{(N-n)!}$$

3. *Partitions rule:* If you are partitioning the elements of a set of $N$ elements into $k$ groups consisting of $n_1, n_2, \ldots, n_k$ elements $(n_1 + n_2 + \cdots + n_k = N)$, the number of different results is

$$\frac{N!}{n_1! n_2! \cdot \cdots \cdot n_k!}$$

4. *Combinations rule:* If you are drawing $n$ elements from a set of $N$ elements without regard to the order of the $n$ elements, the number of different results is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

(*Note*: The combinations rule is a special case of the partitions rule when $k = 2$.)

**Example 3.29**

**Counting Rule Application: Ranking LCD Monitors**

A computer rating service is commissioned to rank the top 3 brands of flat-screen LCD monitors. A total of 10 brands is to be included in the study.

a. In how many different ways can the computer rating service arrive at the final ranking?

b. If the rating service can distinguish no difference among the brands and therefore arrives at the final ranking by chance, what is the probability that company Z's brand is ranked first? In the top 3?

**Solution**

a. Since the rating service is drawing 3 elements (brands) from a set of 10 elements and arranging the 3 elements in a distinct order, we use the permutations rule to find the number of different results:

$$P_3^{10} = \frac{10!}{(10-3)!} = 10 \cdot 9 \cdot 8 = 720$$

b. The steps for calculating the probability of interest are as follows:

*Step 1* The experiment is to select and rank 3 brands of flat-screen LCD monitors from 10 brands.

*Step 2* There are too many simple events to list. However, we know from part **a** that there are 720 different outcomes (i.e., simple events) of this experiment.

*Step 3* If we assume the rating service determines the rankings by chance, each of the 720 simple events should have an equal probability of occurrence. Thus,

$$P(\text{Each simple event}) = \frac{1}{720}$$

*Step 4* One event of interest to company Z is that its brand receives top ranking. We will call this event $A$. The list of simple events that result in the occurrence of event $A$ is long, but the *number* of simple events contained in event $A$ is determined by breaking event $A$ into two parts:

| Company Z brand | 9 other brands |
|---|---|
| ↓ | ↙ ↘ |
| Rank 1 | Rank 2 Rank 3 |
| Only one possibility | $P_2^9 = \frac{9!}{(9-2)!} = 72$ possibilities |

Event $A$

$$1 \cdot 72 = 72 \text{ simple events}$$

Thus, event $A$ can occur in 72 different ways.

Now define $B$ as the event that company Z's brand is ranked in the top three. Since event $B$ specifies only that brand Z appear in the top three, we repeat the calculations above, fixing brand Z in position 2 and then in position 3. We conclude that the number of simple events contained in event $B$ is $3(72) = 216$.

*Step 5* The final step is to calculate the probabilities of events $A$ and $B$. Since the 720 simple events are equally likely to occur, we find

$$P(A) = \frac{\text{Number of simple events in } A}{\text{Total number of simple events}} = \frac{72}{720} = \frac{1}{10}$$

Similarly,

$$P(B) = \frac{216}{720} = \frac{3}{10}$$

**Example 3.30**

Counting Rule Application: Selecting LCD Monitors

Refer to Example 3.29. Suppose the computer rating service is to choose the top 3 flat-screen LCD monitors from the group of 10, but is *not to rank the three*.

a. In how many different ways can the rating service choose the 3 to be designated as top-of-the-line flat-screen LCD monitors?

b. Assuming that the rating service makes its choice by chance and that company X has 2 brands in the group of 10, what is the probability that exactly 1 of the company X brands is selected in the top 3? At least 1?

**Solution**

a. The rating service is selecting 3 elements (brands) from a set of 10 elements *without regard to order*, so we can apply the combinations rule to determine the number of different results:

$$\binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$$

b. We will follow the five-step procedure.

*Step 1* The experiment is to select (but *not rank*) 3 brands from 10.

*Step 2* There are 120 simple events for this experiment.

*Step 3* Since the selection is made by chance, each simple event is equally likely:

$$P(\text{Each simple event}) = \frac{1}{120}$$

*Step 4* Define events $A$ and $B$ as follows:

$A$: {Exactly one company X brand is selected}
$B$: {At least one company X brand is selected}

Since each of the simple events is equally likely to occur, we need to know only the number of simple events in $A$ and $B$ to determine their probabilities.

For event $A$ to occur, exactly 1 company X brand must be selected, along with 2 of the remaining 8 brands. We thus break $A$ into two parts:

2 company X brands $\downarrow$ Select 1

$$\binom{2}{1} = 2 \text{ different possibilities}$$

8 other brands $\downarrow$ Select 2

$$\binom{8}{2} = \frac{8!}{2!(8-2)!} = 28 \text{ different possibilities}$$

Event $A$

$$2 \cdot 28 = 56 \text{ simple events}$$

Note that the one company X brand can be selected in 2 ways, whereas the two other brands can be selected in 28 ways (we use the combinations rule because the order of selection is not important). Then, we use the multiplicative rule to combine one of the 2 ways to select a company X brand with one of the 28 ways to select two other brands, yielding a total of 56 simple events for event $A$.

The simple events in event $B$ would include all simple events containing either one or two company X brands. We already know that the number containing exactly one company X brand is 56, the number of elements in event $A$. The number containing exactly two company X brands is equal to the product of the number of ways of selecting two company X brands out of a possible 2 and the number of ways of selecting the third brand from the remaining 8:

2 company X brands $\downarrow$ Select 2

$$\binom{2}{2} = 1 \text{ possibility}$$

8 other brands $\downarrow$ Select 1

$$\binom{8}{1} = 8 \text{ different possibilities}$$

$$(1)(8) = 8 \text{ simple events}$$

Then the number of simple events that imply the selection of either one or two company X brands is

$$\binom{\text{Number containing}}{\text{one X brand}} + \binom{\text{Number containing}}{\text{two X brands}}$$

or

$$56 + 8 = 64$$

*Step 5* Since all the simple events are equally likely, we have

$$P(A) = \frac{\text{Number of simple events in } A}{\text{Total number of simple events}} = \frac{56}{120} = \frac{7}{15}$$

and

$$P(B) = \frac{\text{Number of simple events in } B}{\text{Total number of simple events}} = \frac{64}{120} = \frac{8}{15}$$

Learning how to decide whether a particular counting rule applies to an experiment takes patience and practice. If you want to develop this skill, use the rules to solve the following exercises and some of the supplementary exercises given at the end of this chapter.

## Applied Exercises

3.58 *Using game simulation to teach a POM course.* Refer to the *Engineering Management Research* (May, 2012) study of using a simulation game to a POM course, Exercise 3.11 (p. 87). Recall that a purchase order card in the simulated game consists of a choice of one of two color television brands (A or B), one of two colors (red or black), and the quantity ordered (1, 2, or 3 TVs). Use a counting rule to determine the number of different purchase order cards that are possible. Does your answer agree with the list you produced in Exercise 3.11?

3.59 *Selecting a maintenance support system.* ARTHUR is the Norwegian Army's high-tech radar system designed to identify and track "unfriendly" artillery grenades, calculate where enemy positions are, and direct counterattacks on the enemy. In the *Journal of Quality in Maintenance Engineering* (Vol. 9, 2003), researchers used an analytic hierarchy process to help build a preferred maintenance organization for ARTHUR. The process requires the builder to select alternatives in three different stages (called echelons). In the first echelon, the builder must choose one of two mobile units (regular soldiers or soldiers with engineering training). In the second echelon, the builder chooses one of three heavy mobile units (units in the Norwegian Army, units from Supplier 2, or shared units). Finally, in the third echelon the builder chooses one of three maintenance workshops (Norwegian Army, Supplier 1, or Supplier 2).

a. How many maintenance organization alternatives exist when choices are made in the three echelons?
b. The researchers determined that only four of the alternatives in part **a** are feasible alternatives for ARTHUR. If one of the alternatives is randomly selected, what is the probability that it is a feasible alternative?

3.60 *Monitoring impedance to leg movements.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) study of impedance to leg movement, Exercise 2.46 (p. 51). Recall that Korean engineers attached electrodes to the ankles and knees of volunteers and measured the voltage readings between pairs of electrodes. These readings were used to determine the signal-to-noise ratio (SNR) of impedance changes such as knee flexes and hip extensions.

a. Six voltage electrodes were attached to key parts of the ankle. How many electrode pairs on the ankle are possible?
b. Ten voltage electrodes were attached to key parts of the knee. How many electrode pairs on the knee are possible?
c. Determine the number of possible electrode pairs, where one electrode is attached to the knee and one is attached to the ankle.

3.61 *Mathematical theory of partitions.* Mathematicians at the University of Florida solved a 30-year-old math problem using the theory of partitions. (*Explore*, Fall 2000.) In math terminology, a partition is a representation of an integer as a sum of positive integers. (For example, the number 3 has three possible partitions: 3, 2 + 1 and 1 + 1 + 1.) The researchers solved the problem by using "colored partitions" of a number, where the colors correspond to the four suits—red hearts, red diamonds, black spades, and black clubs—in a standard 52-card bridge deck. Consider forming colored partitions of an integer.

a. How many colored partitions of the number 3 are possible? (*Hint*: One partition is 3♥; another is 2♦ + 1♣.)
b. How many colored partitions of the number 5 are possible?

3.62 *Modeling the behavior of granular media. Granular media* are substances made up of many distinct grains—including sand, rice, ball bearings, and flour. The properties of these materials were theoretically modeled in *Engineering Computations: International Journal for Computer-Aided Engineering and Software* (Vol. 30, No. 2, 2013). The model assumes there is a system of $N$ non-interacting granular particles. The particles are grouped according to energy level. Assume there are $r$ energy levels, with Ni particles at energy level $i$, $i = 1, 2, 3, \ldots, r$. Consequently, $N = N_1 + N_2 + \ldots + N_r$. A microset is defined as a possible grouping of the particles among the energy levels. For example, suppose $N = 7$ and $r = 3$. Then one possible microset is $N_1 = 1, N_2 = 2$, and $N_3 = 4$. That is, there is one particle at energy level 1, two particles at energy level 2, and four particles at energy level 3. Determine the number of different microsets possible when $N = 7$ and $r = 3$.

3.63 *Chemical catalyst study.* A study was conducted by Union Carbide to identify the optimal catalyst preparation conditions in the conversion of monoethanolamine (MEA) to ethylenediamine (EDA), a substance used commercially in soaps.* The initial experimental plan was chosen to screen four metals (Fe, Co, Ni, and Cu) and four catalyst support classes (low acidity, high acidity, porous, and high surface area).

a. How many metal–support combinations are possible for this experiment?
b. All four catalyst supports are tested in random order with one of the metals. How many different orderings of the four supports are possible with each metal?

3.64 *Brain-wave study.* Can man communicate with a machine through brain-wave processing? This question was the topic of research reported in *IEEE Engineering in Medicine and Biology Magazine* (Mar. 1990). Volunteers were wired to both a computer and an electroencephalogram (EEG) monitor. Each subject performed five tasks under two conditions—eyes opened and eyes closed.

a. Determine the number of experimental conditions under which each subject was tested.
b. List the conditions of part **a**.
c. Two measurements were recorded for each subject—one after 2 seconds of artifact-free EEG and one after only .25 second of artifact-free EEG. What is the total number of measurements obtained for each subject?

3.65 *Concrete building evaluation.* A full-scale reinforced concrete building was designed and tested under simulated earthquake loading conditions (*Journal of Structural Engineering*, Jan. 1986). After completion of the experiments, several design engineers were administered a questionnaire in which they were asked to evaluate two building parameters (size and reinforcement) for each of three parts (shear wall, columns, and girders). For each parameter–part combination, the design engineers were asked to choose one of the following three responses: too heavy, about right, and too light.

a. How many different responses are possible on the questionnaire?
b. Suppose the design engineers are also asked to select the three parameter–part combinations with the overall highest ratings and rank them from 1 to 3. How many different rankings are possible?

3.66 *Alarm code combinations.* A security alarm system is activated and deactivated by correctly entering the appropriate three-digit numerical code in the proper sequence on a digital panel.

a. Compute the total number of possible code combinations if no digit may be used twice.
b. Compute the total number of possible code combinations if digits may be used more than once.

3.67 *Replacing cutting tools.* In high-volume machining centers, cutting tools are replaced at regular, heuristically chosen intervals. These intervals are generally untimely, i.e., either the tool is replaced too early or too late. The *Journal of Engineering for Industry* (Aug. 1993) reported on an automated real-time diagnostic system designed to replace the cutting tool of a drilling machine at optimum times. To test the system, data were collected over a broad range of machining conditions. The experimental variables were as follows:

1. Two workpiece materials (steel and cast iron)
2. Two drill sizes (.125 and .25 inch)
3. Six drill speeds (1,250, 1,800, 2,500, 3,000, 3,750, and 4,000 revolutions per minute)
4. Seven feed rates (.003, .005, .0065, .008, .009, .010, .011 inches per revolution)

a. How many different machining conditions are possible?
b. The eight machining conditions actually employed in the study are described in the table on pg. 123. Suppose one (and only one) of the machining combinations in part **a** will detect a flaw in the system. What is the probability that the experiment conducted in the study will detect the system flaw?

*Hansen, J. L., and Best, D. C. "How to Pick a Winner." Paper presented at Joint Statistical Meetings, American Statistical Association and Biometric Society, Aug. 1986, Chicago, IL.

c. Refer to part **b**. Suppose the system flaw occurs when drilling steel material with a .25-inch drill size at a speed of 2,500 rpm. Find the probability that the experiment conducted in the actual study will detect the system flaw.

| Experiment | Workpiece Material | Drill Size (in.) | Drill Speed (rpm) | Feed Rate (ipr) |
|---|---|---|---|---|
| 1 | Cast iron | .25 | 1,250 | .011 |
| 2 | Cast iron | .25 | 1,800 | .005 |
| 3 | Steel | .25 | 3,750 | .003 |
| 4 | Steel | .25 | 2,500 | .003 |
| 5 | Steel | .25 | 2,500 | .008 |
| 6 | Steel | .125 | 4,000 | .0065 |
| 7 | Steel | .125 | 4,000 | .009 |
| 8 | Steel | .125 | 3,000 | .010 |

3.68 *Selecting gaskets.* Suppose you need to replace 5 gaskets in a nuclear-powered device. If you have a box of 20 gaskets from which to make the selection, how many different choices are possible; i.e., how many different samples of 5 gaskets can be selected from the 20?

3.69 *FAA task force.* To evaluate the traffic control systems of four facilities relying on computer-based equipment, the Federal Aviation Administration (FAA) formed a 16-member task force. If the FAA wants to assign 4 task force members to each facility, how many different assignments are possible?

## Optional Applied Exercises

3.70 *Poker hands.* What is the probability that you will be dealt a 5-card poker hand of four aces?

3.71 *Drawing blackjack.* Blackjack, a favorite game of gamblers, is played by a dealer and at least one opponent and uses a standard 52-card bridge deck. Each card is assigned a numerical value. Cards numbered from 2 to 10 are assigned the values shown on the card. For example, a 7 of spades has a value of 7; a 3 of hearts has a value of 3. Face cards (kings, queens, and jacks) are each valued at 10, and an ace can be assigned a value of either 1 or 11, at the discretion of the player holding the card. At the outset of the game, two cards are dealt to the player and two cards to the dealer. Drawing an ace and any card with a point value of 10 is called *blackjack*. In most casinos, if the dealer draws blackjack, he or she automatically wins.

a. What is the probability that the dealer will draw a blackjack?

b. What is the probability that a player will win with blackjack?

# 3.9 Probability and Statistics: An Example

We have introduced a number of new concepts in the preceding sections, and this makes the study of probability a particularly arduous task. It is, therefore, very important to establish clearly the connection between probability and statistics, which we will do in the remaining chapters. Although Bayes' rule demonstrates one way that probability can be used to make statistical inferences, traditional methods of statistical inference use probability in a slightly different way. In this section, we will present one brief example of this traditional approach to statistical inference so that you can begin to understand why some knowledge of probability is important in the study of statistics.

Suppose a firm that manufactures concrete studs is researching the hypothesis that its new chemically anchored studs achieve greater holding capacity and greater carrying load capacity than the more conventional, mechanically anchored studs. To test the hypothesis, three new chemical anchors are selected from a day's production and subjected to a durability test. Each of the three $\frac{1}{2}$-inch studs is drilled and set into a slab of 4,000 pounds-per-square-inch stone aggregate concrete, and their tensile load capacities (in pounds) are recorded. It is known from many previous durability tests of mechanically anchored studs that approximately 16% of mechanical anchors will have tensile strengths over 12,000 pounds. Suppose that all three of the chemically anchored studs tested have tensile strengths greater than 12,000 pounds. What can researchers for the firm conclude?

To answer these questions, define the events

$A_1$: {Chemically anchored stud 1 has tensile strength over 12,000 pounds}
$A_2$: {Chemically anchored stud 2 has tensile strength over 12,000 pounds}
$A_3$: {Chemically anchored stud 3 has tensile strength over 12,000 pounds}

We want to find $P(A_1 \cap A_2 \cap A_3)$, the probability that all three tested studs have tensile load capacities over 12,000 pounds.

Since the studs are selected by chance from a large production, it may be plausible to assume that the events $A_1$, $A_2$, and $A_3$ are independent. That is,

$$P(A_2 | A_1) = P(A_2)$$

In words, knowing that the first stud has a tensile strength over 12,000 pounds does not affect the probability that the second stud has a tensile strength over 12,000 pounds. With the assumption of independence, we can calculate the probability of the intersection by multiplying the individual probabilities:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

If the new chemically anchored studs are no stronger or no weaker than the mechanically anchored studs, that is, *if the relative frequency distribution of tensile strengths for chemically anchored studs is no different from that for mechanically anchored studs*, then we would expect about 16% of the new studs to have tensile strengths over 12,000 pounds. Consequently, our estimate of $P(A)$ is .16 for all three studs, and

$$P(A_1 \cap A_2 \cap A_3) \approx (.16)(.16)(.16) = .004096$$

Thus, the probability that the firm's researchers will observe all three studs with tensile load capacity over 12,000 pounds is only about .004. If this event were to occur, the researchers might conclude that it lends credence to the theory that chemically anchored studs achieve greater carrying load capacity than mechanically anchored studs, *since it is so unlikely to occur if the distributions of tensile strength are the same*. Such a conclusion would be an application of the rare event approach to statistical inference. You can see that the basic principles of probability play an important role.

## Applied Exercises

3.72 *Brightness of stars.* *Sky & Telescope* (May 1993) reported that Noah Brosch of Tel Aviv University, Israel, discovered a new asterism in Virgo. "Five stars, all appearing brighter than about the 13th magnitude, comprise a diamond-shaped area with sides only 42 seconds long. The probability is small that five stars with similar brightnesses could be so closely aligned by chance, and Brosch suggests that the stars of the diamond . . . are physically associated." Assuming the "probability" mentioned in the article is small (say, less than .01), do you agree with the inference made by the astronomer?

3.73 *Defecting CDs.* Experience has shown that a manufacturer of rewritable CDs produces, on the average, only 1 defective CD in 100. Suppose that of the next 4 CDs manufactured, at least 1 is defective. What would you infer about the claimed defective rate of .01? Explain.

3.74 *Oil leasing rights.* Since 1961, parcels of land that may contain oil have been placed in a lottery, with the winner receiving leasing rights (at \$1 per acre per year) for a period of 10 years. United States citizens 21 years or older are eligible and are entitled to one entry per lottery by paying a \$10 filing fee to the Bureau of Land Management (see *The Federal Oil & Gas Leasing System*, Federal Resource Registry, 1993). For several months in 1980, however, the lottery was suspended to investigate a player who won three parcels of land in 1 month. The numbers of entries for the three lotteries were 1,836, 1,365, and 495, respectively. An Interior Department audit stated that "federal workers did a poor job of shaking the drum before the drawing." Based on your knowledge of probability and rare events, would you make the same inference as that made by the auditor?

3.75 *Antiaircraft gun aiming errors.* At the beginning of World War II, a group of British engineers and statisticians was formed in London to investigate the problem of the lethality of antiaircraft weapons.* One of the main goals of the research team was to assess the probability that a single shell would destroy (or cripple) the aircraft at which it was fired. Although a great deal of data existed at the time on ground-to-ground firing with artillery shells, little information was available on the accuracy of antiaircraft guns. Consequently, a series of trials was run in 1940 in which gun crews shot at free-flying (unpiloted) aircraft. When German aircraft began to bomb England later in that same year, however, the researchers found that the aiming errors of antiaircraft guns under battle stress were considerably greater than those estimated from trials. Let $p$ be the probability that an antiaircraft shell strikes within a 30-foot radius of its target. Assume that under simulated conditions, $p = .45$.

a. In an actual attack by a single German aircraft, suppose that 3 antiaircraft shells are fired and all 3 miss their target by more than 30 feet. Is it reasonable to conclude that in battle conditions $p$ differs from .45?
b. Answer part **a** assuming that you observe 10 consecutive shots that all miss their target by more than 30 feet.

*Pearson, E. S. "Statistics and probability applied to problems of antiaircraft fire in World War II." In *Statistics: A Guide to the Unknown*, 2nd ed. San Francisco: Holden-Day, 1978, pp. 474–482.

## *STATISTICS IN ACTION REVISITED*

### Assessing Predictors of Software Defects in NASA Spacecraft Instrument Code

We now return to the problem posed in the SIA (p. 77) at the beginning of this chapter, namely, assessing different methods of detecting defects in software code written for NASA spacecraft instruments. Recall that the data for this application, saved in the **SWDEFECTS** file, contains 498 modules of software code written in "C" language. For each module, the software code was evaluated, line-by-line (a very time consuming process), for defects and classified as "true" (i.e., module has defective code) or "false" (i.e., module has correct code). In addition, several methods of predicting whether or not a module has defects were applied. The algorithms for four methods utilized in this study—lines of code, cyclomatic complexity, essential complexity, and design complexity — are described in Table SIA3.1 (p. 77). The **SWDEFECTS** file contains a variable that corresponds to each method. When the method predicts a defect, the corresponding variable's value is "yes". Otherwise, it is "no".

A standard approach to evaluating a software defect prediction algorithm is to form a two-way summary table similar to Table SIA3.2. In the table, $a$, $b$, $c$, and $d$ represent the number of modules in each cell. Software engineers use these table entries to compute several probability measures, called accuracy, detection rate, false alarm rate, and precision. These measures are defined as follows:

**TABLE SIA3.2 Summary Table for Evaluating Defect Prediction Algorithms**

| | | *Module Has Defects* | |
|---|---|---|---|
| | | False | True |
| *Algorithm Predicts Defects* | No | $a$ | $b$ |
| | Yes | $c$ | $d$ |

*Accuracy:* $P(\text{Algorithm is correct}) = \dfrac{(a+d)}{(a+b+c+d)}$

*Detection rate:* $P(\text{predict defect} \mid \text{module has defect}) = \dfrac{d}{(b+d)}$

*False alarm rate:* $P(\text{predict defect} \mid \text{module has no defect}) = \dfrac{c}{(a+c)}$

*Precision:* $P(\text{module has defect} \mid \text{predict defect}) = \dfrac{d}{(c+d)}$

You can see that each of these probabilities uses one of the probability rules defined in this chapter. For example, the detection rate is the probability that the algorithm predicts a defect, given that the module actually is a defect. This conditional probability is found by limiting the sample space to the given event, "module has defects". Thus, the denominator is simply the number of modules with defects, $b + d$.

We used SPSS to create summary tables like Table SIA3.1 for the data in the **SWDEFECTS** file. The SPSS printouts are shown in Figure SIA3.1. Consider, first, prediction model that uses the algorithm lines of code (LOC) $> 50$. The results in the top table of Figure SIA3.1 yield the following probability measures for the LOC predictor:

*Accuracy*: $P$(algorithm is correct) $= (400 + 20)/(400 + 29 + 40 + 20) = 420/498 = .843$

*Detection rate*: $P$(predict defect | module has defect) $= 20/(29 + 20) = 20/49 = .408$

*False alarm rate*: $P$(predict defect | module has no defect) $= 40/(400 + 40) = 40/440 = .091$

*Precision*: $P$(module has defect | predict defect) $= 20/(40 + 20) = 20/60 = .333$

**FIGURE SIA3.1**
SPSS Two-way summary tables for predicting software defects

**PRED_LOC * DEFECT Crosstabulation**

Count

| | | DEFECT | | |
|---|---|---|---|---|
| | | false | true | Total |
| PRED_LOC | no | 400 | 29 | 429 |
| | yes | 49 | 20 | 69 |
| Total | | 449 | 49 | 498 |

**PRED_VG * DEFECT Crosstabulation**

Count

| | | DEFECT | | |
|---|---|---|---|---|
| | | false | true | Total |
| PRED_VG | no | 397 | 35 | 432 |
| | yes | 52 | 14 | 66 |
| Total | | 449 | 49 | 498 |

**PRED_EVG * DEFECT Crosstabulation**

Count

| | | DEFECT | | |
|---|---|---|---|---|
| | | false | true | Total |
| PRED_EVG | no | 441 | 47 | 488 |
| | yes | 8 | 2 | 10 |
| Total | | 449 | 49 | 498 |

**PRED_IVG * DEFECT Crosstabulation**

Count

| | | DEFECT | | |
|---|---|---|---|---|
| | | false | true | Total |
| PRED_IVG | no | 422 | 38 | 460 |
| | yes | 27 | 11 | 38 |
| Total | | 449 | 49 | 498 |

**TABLE SIA3.3 Probability Measures for Evaluating Defect Prediction Algorithms**

| Method | Accuracy | Detection Rate | False Alarm Rate | Precision |
|---|---|---|---|---|
| Lines of code | .843 | **.408** | .091 | **.333** |
| Cyclomatic complexity | .825 | .286 | .116 | .212 |
| Essential complexity | **.990** | .041 | **.018** | .200 |
| Design complexity | .869 | .224 | .060 | .289 |

The probability that the algorithm correctly predicts a defect is .843 – a fairly high probability. Also, the false alarm probability is only .091; that is, there is only about a 9% chance that the algorithm will predict a defect when no defect exists. However, the other probability measures, the detection rate and precision, are only .408 and .333, respectively. The fairly low detection rate of this algorithm could be of concern for software engineers; given the module has a defect, there is only about a 40% chance that the algorithm will detect the defect.

Similar calculations were made for the other three prediction algorithms. These probability measures are shown in Table SIA3.3. For comparison purposes, we bolded the probability measure in each column that is "best". You can see that the essential complexity method (predict a defect if ev(g) $\geq$ 14.5) has the highest accuracy and lowest false alarm probability, and the LOC method (predict a defect if lines of code $>$ 50) has the highest detection probability and highest precision. The researchers showed that no single method will yield the optimal value for all four probability measures.

Analyses like this on these and other, more complex, detection algorithms led the researchers to make several recommendations on the choice of a defect prediction algorithm. They ultimately demonstrated that a good defect detector can be found for *any* software project.

# Quick Review

## Key Terms

[*Note: Items marked with an asterisk (*) are from the optional section in this chapter.*]

Additive rule of probability 100
*Bayes' rule 109
*Bayesian statistical methods 109
Combinations rule 117
Complementary events 90
Compound event 88
Conditional probability 94
Counting rule 85
Decision tree 112
Dependent events 104
Event 78
Experiment 78
Independent events 104
Intersection 88
Law of large numbers 81
Multiplicative counting rule 118
Multiplicative rule of probability 102
Mutually exclusive events 100
Partitions rule 116
Permutations rule 114
Probability (steps for calculating) 83
Probability (definition) 78
Probability of an event 82
Probability rules (simple events) 79
Sample space 79
Simple event 79
Unconditional probabilities 94
Union 88
Venn diagram 80

## Key Formulas

| | |
|---|---|
| $P(A) + P(A^c) = 1$ | Rule of Complements 92 |
| $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ | Additive rule of probability 100 |
| $P(A \cap B) = 0$ | Mutually exclusive events 101 |

| | |
|---|---|
| $P(A \cup B) = P(A) + P(B)$ | Additive rule of probability for mutually exclusive events 101 |
| $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ | Conditional probability 95 |
| $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$ | Multiplicative rule of probability 102 |
| $P(A \mid B) = P(A)$ | Independent events 104 |
| $P(A \cap B) = P(A)P(B)$ | Multiplicative rule of probability for independent events 104 |
| $P(A_i \mid E) = \dfrac{P(A_i)P(E \mid A_i)}{P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) + \cdots + P(A_k)P(E \mid A_k)}$ | *Bayes' rule 109 |

*Note: For a summary of counting rules, see p. 118*

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $S$ | | Sample space |
| $S$: $\{1, 2, 3, 4, 5\}$ | | Set of sample points, 1,2,3,4,5, in sample space |
| $A$: $\{1, 2\}$ | | Set of sample points, 1,2, in event $A$ |
| $P(A)$ | Probability of $A$ | Probability that event $A$ occurs |
| $A \cup B$ | $A$ union $B$ | Union of events $A$ and $B$ (either $A$ or $B$ or both occur) |
| $A \cap B$ | $A$ intersect $B$ | Intersection of events $A$ and $B$ (both $A$ and $B$ occur) |
| $A^c$ | $A$ complement | Complement of event $A$ (the event that $A$ does not occur) |
| $P(A\|B)$ | Probability of $A$ given $B$ | Conditional probability that event $A$ occurs given that event $B$ occurs |
| $\binom{N}{n}$ | $N$ choose $n$ | Number of combinations of $N$ elements taken $n$ at a time |
| $N!$ | $N$ factorial | Multiply $N(N-1)(N-2)\cdots(2)(1)$ |

## Chapter Summary Notes

- Probability rules for $k$ **sample points:**

  (1) $0 \le P(S_i) \le 1$ and (2) $\sum_{i=1}^{k} P(S_i) = 1$
- If $A = \{S_1, S_3, S_4\}$, then

  $P(A) = P(S_1) + P(S_3) + P(S_4)$
- The number of samples of size $n$ that can be selected from $N$ elements is $\binom{N}{n}$
- **Union:** $(A \cup B)$ implies that either $A$ or $B$ or both will occur.
- **Intersection:** $(A \cap B)$ implies that both $A$ and $B$ will occur.
- **Complement:** $A^c$ is all the sample points not in $A$.
- **Conditional:** $(A \mid B)$ is the event that $A$ occurs, given $B$ has occurred.
- **Independent:** $B$ occurring does not change the probability that $A$ occurs.

## Supplementary Exercises

3.76 *Awarding road contracts.* A state Department of Transportation (DOT) recently claimed that each of five bidders received equal consideration in the awarding of two road construction contracts and that, in fact, the two contract recipients were randomly selected from among the five bidders. Three of the bidders were large construction conglomerates and two were small specialty contractors. Suppose that both contracts were awarded to large construction conglomerates.

a. What is the probability of this event occurring if, in fact, the DOT's claim is true?

b. Is the probability computed in part **a** inconsistent with the DOT's claim that the selection was random?

3.77 *Environmentalism classifications.* Environmental engineers classify U.S. consumers into five groups based on consumers' feelings about environmentalism:

1. *Basic browns* claim they don't have the knowledge to understand environmental problems.
2. *True-blue greens* use biodegradable products.
3. *Greenback greens* support requiring new cars to run on alternative fuel.
4. *Sprouts* recycle newspapers regularly.
5. *Grousers* believe industries, not individuals, should solve environmental problems.

Assume the proportion of consumers in each group is shown in the table below. Suppose a U.S. consumer is selected at random and his (her) feelings about environmentalism determined.

| | |
|---|---|
| Basic browns | .28 |
| True-blue greens | .11 |
| Greenback greens | .11 |
| Sprouts | .26 |
| Grousers | .24 |

a. List the simple events for the experiment.

b. Assign reasonable probabilities to the simple events.

c. Find the probability that the consumer is either a basic brown or a grouser.

d. Find the probability that the consumer supports environmentalism in some fashion (i.e., the consumer is a true-blue green, a greenback green, or a sprout).

3.78 *Management system failures.* Refer to the *Process Safety Progress* (Dec. 2004) study of 83 industrial accidents caused by management system failures, Exercise 2.6 (p. 27). A summary of the root causes of these 83 incidents is reproduced in the table next table.

a. Find and interpret the probability that an industrial accident is caused by faulty engineering and design.

b. Find and interpret the probability that an industrial accident is caused by something other than faulty procedures and practices.

| Management System Cause Category | Number of Incidents |
|---|---|
| Engineering & Design | 27 |
| Procedures & Practices | 24 |
| Management & Oversight | 22 |
| Training & Communication | 10 |
| **Total** | 83 |

*Source:* Blair, A. S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board." *Process Safety Progress*, Vol. 23, No. 4, Dec. 2004 (Table 1).

3.79 *Unmanned watching system.* An article in *IEEE Computer Applications in Power* (April 1990) describes "an unmanned watching system to detect intruders in real time without spurious detections, both indoors and outdoors, using video cameras and microprocessors." The system was tested outdoors under various weather conditions in Tokyo, Japan. The numbers of intruders detected and missed under each condition are provided in the table.

| | Weather Condition | | | | |
|---|---|---|---|---|---|
| | Clear | Cloudy | Rainy | Snowy | Windy |
| Intruders detected | 21 | 228 | 226 | 7 | 185 |
| Intruders missed | 0 | 6 | 6 | 3 | 10 |
| **Totals** | 21 | 234 | 232 | 10 | 195 |

*Source*: Kaneda, K., et al. "An unmanned watching system using video cameras." *IEEE Computer Applications in Power*, Apr. 1990, p. 24.

a. Under cloudy conditions, what is the probability that the unmanned system detects an intruder?

b. Given that the unmanned system missed detecting an intruder, what is the probability that the weather condition was snowy?

3.80 *Acidic Adirondack lakes.* Based on a study of acid rain, the National Acid Precipitation Assessment Program (NAPAP) estimates the probability at .14 of an Adirondack lake being acidic. Given that the Adirondack lake is acidic, the probability that the lake comes naturally by its acidity is .25 (*Science News*, Sept. 15, 1990). Use this information to find the probability that an Adirondack lake is naturally acidic.

3.81 *Species hot spots.* Biologists define a "hot spot" as a species-rich geographical area (10-kilometer square). *Nature* (Sept. 1993) reported on a study of hot spots for several rare British species, including butterflies, dragonflies, and breeding birds. The table on pg. 130 gives the proportion of a particular species found in a hot spot for that or another species. For example, the value in the lower left corner, .70, implies that 70% of all British bird species inhabit a butterfly hot spot. (*Note*: It is possible for species hotspots to overlap.)

| | Proportion Found in | | |
|---|---|---|---|
| Species | Butterfly Hot Spots | Dragonfly Hot Spots | Bird Hot Spots |
| Butterflies | .91 | .91 | 1.00 |
| Dragonflies | .82 | .92 | .92 |
| Birds | .70 | .73 | .87 |

*Source*: Prendergast, J. R., et al. "Rare species, the coincidence of diversity hotspots and conservation strategies." *Nature*, Vol. 365, No. 6444, Sept. 23, 1993, p. 337 (Table 2c).

a. What is the probability that a dragonfly species will inhabit a dragonfly hot spot?
b. What is the probability that a butterfly species will inhabit a bird hot spot?
c. Explain why all butterfly hot spots are also bird hot spots.

3.82 *ATV injury rate.* The *Journal of Risk and Uncertainty* (May 1992) published an article investigating the relationship of injury rate of drivers of all-terrain vehicles (ATVs) to a variety of factors. One of the more interesting factors studied, age of the driver, was found to have a strong relationship to injury rate. The article reports that prior to a safety-awareness program, 14% of the ATV drivers were under age 12; another 13% were 12–15, and 48% were under age 25. Suppose an ATV driver is selected at random prior to the installation of the safety-awareness program.

a. Find the probability that the ATV driver is 15 years old or younger.
b. Find the probability that the ATV driver is 25 years old or older.
c. Given that the ATV driver is under age 25, what is the probability the driver is under age 12?
d. Are the events Under age 25 and Under age 12 mutually exclusive? Why or why not?
e. Are the events Under age 25 and Under age 12 independent? Why or why not?

3.83 *Testing a sustained-release tablet.* Researchers at the Upjohn Company have developed a sustained-release tablet for a prescription drug. To determine the effectiveness of the tablet, the following experiment was conducted. Six tablets were randomly selected from each of 30 production lots. Each tablet was submersed in water and the percent dissolved was measured at 2, 4, 6, 8, 10, 12, 16, and 20 hours.

a. Find the total number of measurements (percent dissolved) recorded in the experiment.
b. For each lot, the measurements at each time period are averaged. How many averages are obtained?

3.84 *Tensile strength of ingots.* A study was conducted to examine the relationship between the cost structure and the mechanical properties of equiaxed grains in unidirectionally solidified ingots (*Metallurgical Transactions*, May 1986). Ingots composed of copper alloys were poured into one of three mold types (columnar, mixed, or equiaxed) with either a transverse or a longitudinal orientation. From each ingot, five tensile specimens were obtained at varying distances (10, 35, 60, 85, and 100 millimeters) from the ingot chill face and yield strength was determined.

a. How many strength measurements will be obtained if the experiment includes one ingot for each mold type–orientation combination?
b. Suppose three of the ingots will be selected for further testing at the 100-mm distance. How many samples of three ingots can be selected from the total number of ingots in the experiment?
c. Use Table 6 of Appendix II to randomly select the three ingots for further testing.
d. Calculate the probability that the sample selected includes the three highest tensile strengths among all the ingots in the experiment.
e. Calculate the probability that the sample selected includes at least two of the three ingots with the highest tensile strengths.

3.85 *Critical items in shuttle flights.* According to NASA, each space shuttle in the U.S. fleet has 1,500 "critical items" that could lead to catastrophic failure if rendered inoperable during flight. NASA estimates that the chance of at least one critical-item failure within the shuttle's main engines is about 1 in 60 for each mission. To build space station *Freedom*, suppose NASA plans to fly 8 shuttle missions a year for the next decade.

a. Find the probability that at least 1 of the 8 shuttle flights scheduled next year results in a critical-item failure.
b. Find the probability that at least 1 of the 40 shuttle missions scheduled over the next 5 years results in a critical-item failure.

3.86 *Data-communication systems.* A company specializing in data-communications hardware markets a computing system with two types of hard disk drives, four types of display stations, and two types of interfacing. How many systems would the company have to distribute if it received one order for each possible combination of hard disk drive, display station, and interfacing?

3.87 *Reliability of a bottling process.* A brewery utilizes two bottling machines, but they do not operate simultaneously. The second machine acts as a backup system to the first machine and operates only when the first breaks down during operating hours. The probability that the first machine breaks down during operating hours is .20. If, in fact, the first breaks down, then the second machine is turned on and has a probability of .30 of breaking down.

a. What is the probability that the brewery's bottling system is not working during operating hours?
b. The *reliability* of the bottling process is the probability that the system is working during operating hours. Find the reliability of the bottling process at the brewery.

3.88 *Solar-powered batteries.* Recently, the National Aeronautics and Space Administration (NASA) purchased a new solar-powered battery guaranteed to have a failure rate of only 1 in 20. A new system to be used in a space vehicle operates on one of these batteries. To increase the reliability of the system, NASA installed three batteries, each designed to operate if the preceding batteries in the chain fail. If the system is operated in a practical situation, what is the probability that all three batteries would fail?

*3.89 *Repairing a computer system.* The local area network (LAN) for the College of Business computing system at a large university is temporarily shut-down for repairs. Previous shutdowns have been due to hardware failure, software failure, or power failure. Maintenance engineers have determined that the probabilities of hardware, software, and power problems are .01, .05, and .02, respectively. They have also determined that if the system experiences hardware problems, it shuts down 73% of the time. Similarly, if software problems occur, the system shuts down 12% of the time; and, if power failure occurs, the system shuts down 88% of the time. What is the probability that the current shutdown of the LAN is due to hardware failure? Software failure? Power failure?

*3.90 *Electric fuse production.* A manufacturing operation utilizes two production lines to assemble electronic fuses. Both lines produce fuses at the same rate and generally produce 2.5% defective fuses. However, production line 1 recently suffered mechanical difficulty and produced 6.0% defectives during a 3-week period. This situation was not known until several lots of electronic fuses produced in this period were shipped to customers. If one of two fuses tested by a customer was found to be defective, what is the probability that the lot from which it came was produced on malfunctioning line 1? (Assume all the fuses in the lot were produced on the same line.)

3.91 *Raw material supplier options.* To ensure delivery of its raw materials, a company has decided to establish a pattern of purchases with at least two potential suppliers. If five suppliers are available, how many choices (options) are available to the company?

3.92 *Illegal access to satellite TV.* A recent court case involved a claim of satellite television subscribers obtaining illegal access to local TV stations. The defendant (the satellite TV company) wanted to sample TV markets nationwide and determine the percentage of its subscribers in each sampled market who have illegal access to local TV stations. To do this, defendant's expert witness drew a rectangular grid over the continental United States, with horizontal and vertical grid lines every .02 degrees of latitude and longitude, respectively. This created a total of 500 rows and 1,000 columns, or $(500)(1{,}000) = 500{,}000$ intersections. The plan was to randomly sample 900 intersection points and include the TV market at each intersection in the sample. Explain how you could use a random number generator to obtain a random sample of 900 intersections.

3.93 *Random-digit dialing.* To ascertain the effectiveness of their advertising campaigns, firms frequently conduct telephone interviews with consumers using *random-digit dialing*. With this method, a random number generator mechanically creates the sample of phone numbers to be called.

a. Explain how the random number table (Table 1 of Appendix B) or a computer could be used to generate a sample of 7-digit telephone numbers.
b. Use the procedure you described in part **a** to generate a sample of ten 7-digit telephone numbers.
c. Use the procedure you described in part **a** to generate five 7-digit telephone numbers whose first three digits are 373.

## Optional Supplementary Exercises

3.94 *Modem supplier.* An assembler of computer routers and modems uses parts from two sources. Company A supplies 80% of the parts and company B supplies the remaining 20% of the parts. From past experience, the assembler knows that 5% of the parts supplied by company A are defective and 3% of the parts supplied by company B are defective. An assembled modem selected at random is found to have a defective part. Which of the two companies is more likely to have supplied the defective part?

3.95 *Bidding on DOT contracts.* Five construction companies each offer bids on three distinct Department of Transportation (DOT) contracts. A particular company will be awarded at most one DOT contract.

a. How many different ways can the bids be awarded?
b. Under the assumption that the simple events are equally likely, find the probability that company 2 is awarded a DOT contract.
c. Suppose that companies 4 and 5 have submitted noncompetitive bids. If the contracts are awarded at random by the DOT, find the probability that both these companies receive contracts.

3.96 *Writing a $C^{++}$ program.*

a. A professor asks his class to write a $C^{++}$ computer program that prints all three-letter sequences involving the five letters A, B, E, T, and O. How many different three-letter sequences will need to be printed?
b. Answer part **a** if the program is to be modified so that each three-letter sequence has at least one vowel, and no repeated letters.

3.97 *Rare poker hands.* Consider 5-card poker hands dealt from a standard 52-card bridge deck. Two important events are

$A$: {You draw a flush}
$B$: {You draw a straight}

a. Find $P(A)$.
b. Find $P(B)$.
c. The event that both $A$ and $B$ occur, i.e., $A \cap B$, is called a *straight flush*. Find $P(A \cap B)$.

(*Note*: A *flush* consists of any five cards of the same suit. A *straight* consists of any five cards with values in sequence. In a straight, the cards may be of any suit and an ace may be considered as having a value of 1 or a value higher than a king.)

3.98 *Intrusion detection system.* Refer to Example 3.19. (p. 109).

a. Find the probability that an intruder is detected, given a clear day.
b. Find the probability that an intruder is detected, given a cloudy day.

3.99 *Flawed Pentium computer chip.* In October 1994, a flaw was discovered in the Pentium microchip installed in personal computers. The chip produced an incorrect result when dividing two numbers. Intel, the manufacturer of the Pentium chip, initially announced that such an error would occur once in 9 billion divides, or "once in every 27,000 years" for a typical user; consequently, it did not immediately offer to replace the chip.

Depending on the procedure, statistical software packages (e.g., SAS) may perform an extremely large number of divisions to produce the required output. For heavy users of the software, 1 billion divisions over a short time frame is not unusual. Will the flawed chip be a problem for a heavy SAS user? (*Note*: Two months after the flaw was discovered, Intel agreed to replace all Pentium chips free of charge.)

CHAPTER 4

# Discrete Random Variables

**OBJECTIVE**

*To explain what is meant by a discrete random variable, its probability distribution, and corresponding numerical descriptive measures; to present some useful discrete probability distributions and show how they can be used to solve practical problems*

**CONTENTS**

## *STATISTICS IN ACTION*
### The Reliability of a "One-Shot" Device

The *reliability* of a product, system, weapon, or piece of equipment can be defined as the ability of the device to perform as designed, or, more simply, as the probability that the device does not fail when used. (See Chapter 17.) Engineers assess reliability by repeatedly testing the device and observing its failure rate. Certain products, called "one-shot" devices, make this approach challenging. One-shot devices can only be used once; after use, the device is either destroyed or must be rebuilt. Some examples of one-shot devices are nuclear weapons, space shuttles, automobile air bags, fuel injectors, disposable napkins, heat detectors, and fuses.

The destructive nature of a one-shot device makes repeated testing either impractical or too costly. Hence, the reliability of such a device must be determined with minimal testing. Design engineers need to determine the minimum number of tests to conduct on the device in order to demonstrate a desired reliability. For example, when Honda began an evaluation of its new automobile airbag system, the company set a goal of 99.999% reliability. Honda design engineers then visited McDonnell Douglas Aerospace Center (MDAC)—where NASA tests its space systems—to learn about the best techniques for determining the reliability of one-shot devices.

The current trend in determining the reliability of a one-shot device utilizes *acceptance sampling*—a statistical approach that employs the *binomial probability distribution*—(a probability distribution covered in this chapter)—to determine if the device has an acceptable defective rate at some acceptable level of risk. The methodology applies the "rare event" approach illustrated in Example 3.9 (p. 93). In the *Statistics in Action Revisited* at the end of this chapter, we demonstrate this approach.

## 4.1 Discrete Random Variables

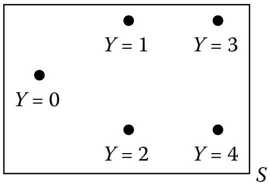
As we noted in Chapter 1, the experimental events of greatest interest are often numerical, i.e., we conduct an experiment and observe the numerical value of some variable. If we repeat the experiment $n$ times, we obtain a sample of quantitative data. To illustrate, suppose a manufactured product (e.g., a mechanical part) is sold in lots of 20 boxes of 12 items each. As a check on the quality of the product, a process control engineer randomly selects 4 from among the 240 items in a lot and checks to determine whether the items are defective. If more than 1 sampled item is found to be defective, the entire lot will be rejected.

The selection of 4 manufactured items from among 240 produces a sample space $S$ that contains $\binom{240}{4}$ simple events, one corresponding to each possible combination of 4 items that might be selected from the lot. Although a description of a specific simple event would identify the 4 items acquired in a particular sample, the event of interest to the process control engineer is an observation on the variable $Y$, the number of defective items among the 4 items that are tested. To each simple event in $S$, there corresponds one and only one value of the variable $Y$. Therefore, a functional relation exists between the simple events in $S$ and the values that $Y$ can assume. The event $Y = 0$ is the collection of all simple events that contain no defective items. Similarly, the event $Y = 1$ is the collection of all simple events in which 1 defective item is observed. Since the value that $Y$ can assume is a numerical event (i.e., an event defined by some number that varies in a random manner from one repetition of the experiment to another), it is called a **random variable**.

**Definition 4.1**

A **random variable** $Y$ is a numerical-valued function defined over a sample space. Each simple event in the sample space is assigned a value of $Y$.

**FIGURE 4.1**
Venn Diagram for Number ($Y$) of Defective Items in a Sample of 4 Items




The number $Y$ of defective items in a selection of 4 items from among 240 is an example of a **discrete random variable**, one that can assume a countable number of values. For our example, the random variable $Y$ may assume any of the five values, $Y = 0, 1, 2, 3$, or 4, as shown in Figure 4.1. As another example, the number $Y$ of lines of code in a $C^{++}$ software program is also a discrete random variable that could, theoretically, assume a value that is large beyond all bound. The possible values for this discrete random variable correspond to the nonnegative integers, $Y = 0, 1, 2, 3, \ldots, \infty$, and the number of such values is countable.

Random variables observed in nature often possess similar characteristics and consequently can be classified according to type. In this chapter, we will study seven different types of discrete random variables and will use the methods of Chapter 3 to derive the probabilities associated with their possible values. We will also begin to develop some intuitive ideas about how the probabilities of observed sample data can be used to make statistical inferences.

**Definition 4.2**
A **discrete random variable** $Y$ is one that can assume only a countable number of values.

## 4.2 The Probability Distribution for a Discrete Random Variable

Since the values that a random variable $Y$ can assume are numerical events, we will want to calculate their probabilities. A table, formula, or graph that gives these probabilities is called the **probability distribution** for the random variable $Y$. The usual convention in probability theory is to *use uppercase letters (e.g., Y) to denote random variables and lowercase letters (e.g., y) to denote particular numerical values a random variable may assume*. Therefore, we want to find a table, graph, or formula that gives the probability, $P(Y = y)$, for each possible value of $y$. To simplify notation, we will sometimes denote $P(Y = y)$ by $p(y)$. We will illustrate this concept using a simple coin-tossing example.

**Example 4.1**
Probability Distribution for Coin Tossing Experiment

A balanced coin is tossed twice, and the number $Y$ of heads is observed. Find the probability distribution for $Y$.

Solution

Let $H_i$ and $T_i$ denote the observation of a head and a tail, respectively, on the $i$th toss, for $i = 1, 2$. The four simple events and the associated values of $Y$ are shown in Table 4.1. You can see that $Y$ can take on the values 0, 1, or 2.

**TABLE 4.1 Outcomes of Coin-Tossing Experiment**

| Simple Event | Description | $P(E_i)$ | Number of Heads $Y = y$ |
|---|---|---|---|
| $E_1$ | $H_1H_2$ | $\frac{1}{4}$ | 2 |
| $E_2$ | $H_1T_2$ | $\frac{1}{4}$ | 1 |
| $E_3$ | $T_1H_2$ | $\frac{1}{4}$ | 1 |
| $E_4$ | $T_1T_2$ | $\frac{1}{4}$ | 0 |

The event $Y = 0$ is the collection of all simple events that yield a value of $Y = 0$, namely, the single simple event $E_4$. Therefore, the probability that $Y$ assumes the value 0 is

$$P(Y = 0) = p(0) = P(E_4) = \frac{1}{4}$$

The event $Y = 1$ contains two simple events, $E_2$ and $E_3$. Therefore,

$$P(Y = 1) = p(1) = P(E_2) + P(E_3) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Finally,

$$P(Y = 2) = p(2) = P(E_1) = \frac{1}{4}$$

The probability distribution $p(y)$ is displayed in tabular form in Table 4.2 and as a line graph in Figure 4.2. Note that in Figure 4.2, the probabilities associated with $y$ are illustrated with vertical lines; the height of the line is proportional to the value of $p(y)$. We show in Section 4.6 that this probability distribution can also be given by the formula

$$p(y) = \frac{\binom{2}{y}}{4}$$

where

$$p(0) = \frac{\binom{2}{0}}{4} = \frac{1}{4}$$

$$p(1) = \frac{\binom{2}{1}}{4} = \frac{2}{4} = \frac{1}{2}$$

$$p(2) = \frac{\binom{2}{2}}{4} = \frac{1}{4}$$

Any of these techniques—a table, graph, or formula—can be used to describe the probability distribution of a discrete random variable $y$.

**TABLE 4.2 Probability Distribution for *Y*, the Number of Heads in Two Tosses of a Coin**

| $Y = y$ | $p(y)$ |
|---|---|
| 0 | $\frac{1}{4}$ |
| 1 | $\frac{1}{2}$ |
| 2 | $\frac{1}{4}$ |
| | $\sum_y p(y) = 1$ |

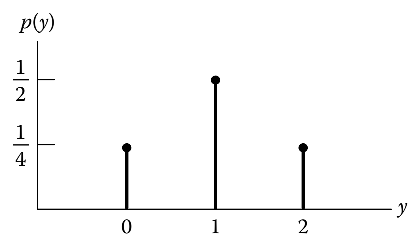



**FIGURE 4.2**
Probability distribution for $Y$, the number of heads in two tosses of a coin

**Definition 4.3**

The **probability distribution** for a discrete random variable $Y$ is a table, graph, or formula that gives the probability $p(y)$ associated with each possible value of $Y = y$.

The probability distribution $p(y)$ for a discrete random variable must satisfy two properties. First, because $p(y)$ is a probability, it must assume a value in the interval $0 \leq p(y) \leq 1$. Second, the sum of the values of $p(y)$ over all values of $Y$ must equal 1. This is true because we assigned one and only one value of $Y$ to each simple event in $S$. It follows that the values that $Y$ can assume represent different sets of simple events and are, therefore, mutually exclusive events. Summing $p(y)$ over all possible values of $Y$ is then equivalent to summing the probabilities of all simple events in $S$, and from Section 3.2, $P(S)$ is known to be equal to 1.

### Requirements for a Discrete Probability Distribution

1. $0 \leq p(y) \leq 1$
2. $\sum_{\text{all } y} p(y) = 1$

**Example 4.2**

Probability Distribution for Driver-Side Crash Ratings

CRASH

The National Highway Traffic Safety Administration (NHTSA) has developed a driver-side "star" scoring system for crash-testing new cars. Each crash-tested car is given a rating ranging from one star (*) to five stars (*****); the more stars in the rating, the better the level of crash protection in a head-on collision. Recent data for 98 new cars are saved in the **CRASH** file. A summary of the driver-side star ratings for these cars is reproduced in the MINITAB printout, Figure 4.3. Assume that one of the 98 cars is selected at random and let $Y$ equal the number of stars in the car's driver-side star rating. Use the information in the printout to find the probability distribution for $Y$. Then find $P(Y \leq 3)$.

**FIGURE 4.3**
MINITAB summary of driver-side star ratings

**Tally for Discrete Variables: DRIVSTAR**

| DRIVSTAR | Count | Percent |
|---|---|---|
| 2 | 4 | 4.08 |
| 3 | 17 | 17.35 |
| 4 | 59 | 60.20 |
| 5 | 18 | 18.37 |
| N= | 98 | |

Solution

Since driver-side star ratings range from 1 to 5, the discrete random variable $Y$ can take on the values 1, 2, 3, 4, or 5. The MINITAB printout gives the percentage of the 98 cars in the **CRASH** file that fall into each star category. These percentages represent the probabilities of a randomly selected car having one of the star ratings. Since none of the 98 cars has a star rating of 1, $P(Y = 1) = p(1) = 0$. The remaining probabilities for $Y$ are as follows: $p(2) = .0408$, $p(3) = .1735$, $p(4) = .6020$, and $p(5) = .1837$. Note that these probabilities sum to 1.

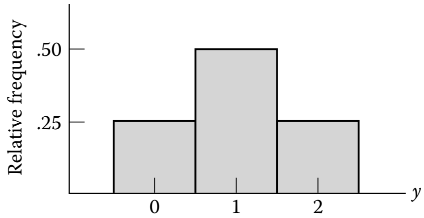
To find $P(Y \leq 3)$, we sum the values $p(1)$, $p(2)$, and $p(3)$.

$$P(Y \leq 3) = p(1) + p(2) + p(3) = 0 + .0408 + .1735 = .2143$$

Thus, about 21% of the driver-side star ratings have three or fewer stars.

To conclude this section, we will discuss the relationship between the probability distribution for a discrete random variable and the relative frequency distribution of data (discussed in Section 2.2). Suppose you were to toss two coins over and over again a very large number of times and record the number $Y$ of heads observed for each toss. A relative frequency histogram for the resulting collection of 0s, 1s, and 2s

**FIGURE 4.4**
Theoretical relative frequency histogram for $Y$, the number of heads in two tosses of a coin




would have bars with heights of approximately $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. In fact, if it were possible to repeat the experiment an infinitely large number of times, the distribution would appear as shown in Figure 4.4. Thus, the probability histogram of Figure 4.3 provides a **model** for a conceptual population of values of $Y$—the values of $Y$ that would be observed if the experiment were to be repeated an infinitely large number of times.

Beginning with Section 4.5, we will introduce a number of models for discrete random variables that occur in the physical, biological, social, and information sciences.

## Applied Exercises

4.1 *Solar energy cells.* According to *Wired* (June, 2008), 35% of the world's solar energy cells are manufactured in China. Consider a random sample of 5 solar energy cells, and let $Y$ represent the number in the sample that are manufactured in China. In Section 4.6, we show that the probability distribution for $Y = y$ is given by the formula,

$$p(y) = \frac{(5!)(.35)^y(.65)^{5-y}}{(y!)(5-y)!}, \text{ where n!}$$
$$= (n)(n-1)(n-2)...(2)(1)$$

a. Explain why $Y$ is a discrete random variable.
b. Find $p(y)$ for $y = 0, 1, 2, 3, 4$, and 5.
c. Show that the properties for a discrete probability distribution are satisfied.
d. Find the probability that at least 4 of the 5 solar energy cells in the sample are manufactured in China.

4.2 *Dust mite allergies.* A dust mite allergen level that exceeds 2 micrograms per gram ($\mu$g/g) of dust has been associated with the development of allergies. Consider a random sample of four homes and let $Y$ be the number of homes with a dust mite level that exceeds 2 $\mu$g/g. The probability distribution for $Y = y$, based on a study by the National Institute of Environmental Health Sciences, is shown in the following table.

| $y$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $p(y)$ | .09 | .30 | .37 | .20 | .04 |

a. Verify that the probabilities for $Y$ in the table sum to 1.
b. Find the probability that three or four of the homes in the sample have a dust mite level that exceeds 2 $\mu$g/g.
c. Find the probability that fewer than two homes in the sample have a dust mite level that exceeds 2 $\mu$g/g.

4.3 *Controlling water hyacinth.* Entomological engineers are continually searching for new biological agents to control one of the world's worst aquatic weeds, the water hyacinth. An insect that naturally feeds on water hyacinth is the delphacid. Female delphacids lay anywhere from one to four eggs onto a water hyacinth blade. The *Annals of the Entomological Society of America* (Jan. 2005) published a study of the life cycle of a South American delphacid species. The accompanying table gives the percentages of water hyacinth blades that have one, two, three, and four delphacid eggs.

| | One Egg | Two Eggs | Three Eggs | Four Eggs |
|---|---|---|---|---|
| Percentage of Blades | 40 | 54 | 2 | 4 |

*Source:* Sosa, A. J., et al. "Life history of *Megamelus scutellaris* with description of immature stages," *Annals of the Entomological Society of America*, Vol. 98, No. 1, Jan. 2005 (adapted from Table 1).

a. One of the water hyacinth blades in the study is randomly selected and $Y$, the number of delphacid eggs on the blade, is observed. Give the probability distribution of $Y$.
b. What is the probability that the blade has at least three delphacid eggs?

4.4 *Beach erosional hot spots.* Refer to the U.S. Army Corps of Engineers study of beach erosional hot spots, Exercise 2.5 (p. 27). The data on the nearshore bar condition for six beach hot spots are reproduced in the table on page 139. Suppose you randomly select two of these six beaches and count $Y$, the total number in the sample with a planar nearshore bar condition.

a. List all possible pairs of beach hot spots that can be selected from the six.
b. Assign probabilities to the outcomes in part **a**.

c. For each outcome in part **a**, determine the value of $Y$.
d. Form a probability distribution table for $Y$.
e. Find the probability that at least one hot spot in the sample has a planar nearshore bar condition.

| Beach Hot spot | Nearshore Bar Condition |
|---|---|
| Miami Beach, FL | Single, shore parallel |
| Coney Island, NY | Other |
| Surfside, CA | Single, shore parallel |
| Monmouth Beach, NJ | Planar |
| Ocean City, NJ | Other |
| Spring Lake, NJ | Planar |

*Source:* "Identification and characterization of erosional hotspots." William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March, 18, 2002.

4.5 *Contaminated gun cartridges.* A weapons manufacturer uses a liquid propellant to produce gun cartridges. During the manufacturing process, the propellant can get mixed with another liquid to produce a contaminated cartridge. A University of South Florida statistician, hired by the company to investigate the level of contamination in the stored cartridges, found that 23% of the cartridges in a particular lot were contaminated. Suppose you randomly sample (without replacement) gun cartridges from this lot until you find a contaminated one. Let $Y = y$ be the number of cartridges sampled until a contaminated one is found. It is known that the probability distribution for $Y = y$ is given by the formula:

$$p(y) = (.23)(.77)^{y-1}, \qquad y = 1, 2, 3, \ldots$$

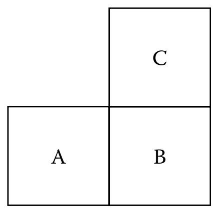
a. Find $p(1)$. Interpret this result.
b. Find $p(5)$. Interpret this result.
c. Find $P(Y \geq 2)$. Interpret this result.

4.6 *Reliability of a manufacturing network.* A team of industrial management university professors investigated the reliability of a manufacturing system that involves multiple production lines (*Journal of Systems Sciences & Systems Engineering*, March 2013). An example of such a network is a system for producing integrated circuit (IC) cards with two production lines set up in sequence. Items (IC cards) first pass through Line 1, then are processed by Line 2. The probability distribution of the maximum capacity level $Y$ of each line is shown in the next table. Assume the lines operate independently.

a. Verify that the properties of discrete probability distributions are satisfied for each line in the system.
b. Find the probability that the maximum capacity level for Line 1 will exceed 30 items.
c. Repeat part b for Line 2.
d. Now consider the network of two production lines. What is the probability that a maximum capacity level of 30 items is maintained throughout the network?

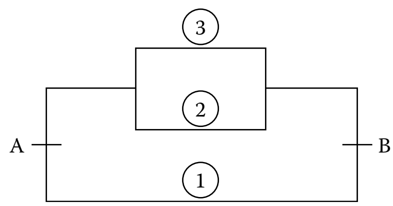
| Line | Maximum Capacity, $y$ | $p(y)$ |
|---|---|---|
| 1 | 0 | .01 |
| | 12 | .02 |
| | 24 | .02 |
| | 36 | .95 |
| 2 | 0 | .002 |
| | 35 | .002 |
| | 70 | .996 |

4.7 *Robot-sensor system configuration.* Engineers at Broadcom Corp. and Simon Fraser University collaborated on research involving a robot-sensor system in an unknown environment. (*The International Journal of Robotics Research*, Dec. 2004.) As an example, the engineers presented the three-point, single-link robotic system shown in the accompanying figure. Each point (A, B, or C) in the physical space of the system has either an "obstacle" status or a "free" status. There are two single links in the system: $A \leftrightarrow B$ and $B \leftrightarrow C$. A link has a "free" status if and only if both points in the link are "free." Otherwise, the link has an "obstacle" status. Of interest is the random variable $Y$, the total number of links in the system that are "free."




a. List the possible values of $Y$ for the system.
b. The researchers stated that the probability of any point in the system having a "free" status is .5. Assuming the three points in the system operate independently, find the probability distribution for $Y$.

4.8 *Electric circuit with relays.* Consider the segment of an electric circuit with three relays shown here. Current will flow from A to B if there is at least one closed path when the switch is thrown. Each of the three relays has an equally likely chance of remaining open or closed when the switch is thrown. Let $Y$ represent the number of relays that close when the switch is thrown.




a. Find the probability distribution for $Y$ and display it in tabular form.
b. What is the probability that current will flow from A to B?

4.9 *Variable speed limit control for freeways.* A common transportation problem in large cities is congestion on the freeways. In the *Canadian Journal of Civil Engineering* (Jan., 2013), civil engineers investigated the use of variable speed limits (VSL) to control the congestion problem. The study site was an urban freeway in Edmonton, Canada. A portion of the freeway was equally divided into three sections, and variable speed limits posted (independently) in each section. Simulation was used to find the optimal speed limits based on various traffic patterns and weather conditions. Probability distributions of the speed limits for the three sections were determined. For example, one possible set of distributions is as follows (probabilities in parentheses). *Section 1*: 30 mph (.05), 40 mph (.25), 50 mph (.25), 60 mph (.45); *Section 2*: 30 mph (.10), 40 mph (.25), 50 mph (.35), 60 mph (.30); *Section 3*: 30 mph (.15), 40 mph (.20), 50 mph (.30), 60 mph (.35).

a. Verify that the properties of discrete probability distributions are satisfied for each individual section of the freeway.
b. Consider a vehicle that will travel through the three sections of the freeway at a steady (fixed) speed. Let $Y$ represent this speed. Find the probability distribution for $Y$.
c. Refer to part b. What is the probability that the vehicle can travel at least 50 mph through the three sections of the freeway?

4.10 *Confidence of feedback information for improving quality.* Refer to the *Engineering Applications of Artificial Intelligence* (Vol. 26, 2013) study of the confidence level of feedback information generated by a semiconductor, Exercise 3.41 (p.104). Recall that at any point in time during the production process, a report can be generated indicating the system is either "OK" or "not OK". Assume that the probability of an "OK" report at one time period $(t + 1)$, given an "OK" report in the previous time period $(t)$, is .20. Also, the probability of an "OK" report at one time period $(t + 1)$, given a "not OK" report in the previous time period $(t)$, is .55. Now consider the results of reports generated over four consecutive time periods, where the first time period resulted in an "OK" report. Let $Y$ represent the number of "OK" reports in the next three time periods. Derive an expression for the probability distribution of $Y$.

4.11 *Acceptance sampling of firing pins.* A quality control engineer samples five from a large lot of manufactured firing pins and checks for defects. Unknown to the inspector, three of the five sampled firing pins are defective. The engineer will test the five pins in a randomly selected order until a defective is observed (in which case the entire lot will be rejected). Let $Y$ be the number of firing pins the quality control engineer must test. Find and graph the probability distribution of $Y$.

## 4.3 Expected Values for Random Variables

The data we analyze in engineering and the sciences often result from observing a process. For example, in quality control a production process is monitored and the number of defective parts produced per hour is recorded. As noted earlier, a probability distribution for a random variable $Y$ is a model for a population relative frequency distribution, i.e., a model for the data produced by a process. Consequently, we can describe process data with numerical descriptive measures, such as its mean and standard deviation, and we can use the Empirical Rule to identify improbable values of $Y$.

The **expected value** (or **mean**) of a random variable $Y$, denoted by the symbol $E(Y)$, is defined as follows:

**Definition 4.4**

Let $Y$ be a discrete random variable with probability distribution $p(y)$. Then the **mean** or **expected value** of $Y$ is

$$\mu = E(Y) = \sum_{\text{all } y} yp(y)$$

**Example 4.3**

Expected Value in Coin-Tossing Experiment

Refer to the coin-tossing experiment of Example 4.1 (p. 135) and the probability distribution for the random variable $Y$, shown in Table 4.1. Demonstrate that the formula for $E(Y)$ yields the mean of the probability distribution for the discrete random variable $Y$.

Solution

If we were to repeat the coin-tossing experiment a large number of times—say, 400,000 times—we would expect to observe $Y = 0$ heads approximately 100,000 times, $Y = 1$ head approximately 200,000 times, and $Y = 2$ heads approximately 100,000 times. If we calculate the mean value of these 400,000 values of $Y$, we obtain

$$\mu \approx \frac{\sum y}{n} = \frac{100{,}000(0) + 200{,}000(1) + 100{,}000(2)}{400{,}000}$$

$$= 0\left(\frac{100{,}000}{400{,}000}\right) + 1\left(\frac{200{,}000}{400{,}000}\right) + 2\left(\frac{100{,}000}{400{,}000}\right)$$

$$= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) = \sum_{\text{all } y} yp(y)$$

If $Y$ is a random variable, so also is any function $g(Y)$ of $Y$. The **expected value of $g(Y)$** is defined as follows:

**Definition 4.5**

Let $Y$ be a discrete random variable with probability distribution $p(y)$ and let $g(Y)$ be a function of $Y$. Then the **mean** or **expected value of $g(Y)$** is

$$E[g(Y)] = \sum_{\text{all } y} g(y)p(y)$$

One of the most important functions of a discrete random variable $Y$ is its **variance**, i.e., the expected value of the squared deviation of $Y$ from its mean $\mu$.

**Definition 4.6**

Let $Y$ be a discrete random variable with probability distribution $p(y)$. Then the **variance of $Y$** is

$$\sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2$$

The **standard deviation of $Y$** is the positive square root of the variance of $Y$:

$$\sigma = \sqrt{\sigma^2}$$

**Example 4.4**

Expected Values for Driver-Side Crash Ratings

Refer to the NHTSA driver-side crash ratings of Example 4.2 (p. 137). The probability distribution for $Y$, the number of stars in the rating of each car, is shown in Table 4.3. Find the mean and standard deviation of $Y$.

Solution

Using the formulas in Definitions 4.5 and 4.6, we obtain the following:

$$\mu = E(Y) = \sum_{y=1}^{5} y\,p(y) = (1)(0) + (2)(.0408)$$

$$+ (3)(.1735) + (4)(.6020) + (5)(.1837) = 3.93$$

$$\sigma^2 = E[(Y - \mu)^2] = \sum_{y=1}^{5}(y - \mu)^2 p(y)$$

$$= (1 - 3.93)^2(0) + (2 - 3.93)^2(.0408)$$

$$+ (3 - 3.93)^2(.1735) + (4 - 3.93)^2(.6020)$$

**TABLE 4.3 Probability Distribution of Driver-Side Crash Rating, *Y***

| Number of Stars in Rating, $y$ | $p(y)$ |
|---|---|
| 1 | 0 |
| 2 | .0408 |
| 3 | .1735 |
| 4 | .6020 |
| 5 | .1837 |

$$+ (5 - 3.93)^2(.1837) = .51$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{.51} = .71$$

**Example 4.5**

Empirical Rule Application—Driver-Side Crash Ratings

Refer to Example 4.4 and find the probability that a value of $Y$ will fall in the interval $\mu \pm 2\sigma$.

Solution

From Example 4.4 we know that $\mu = 3.93$ and $\sigma = .71$. Then the interval $\mu \pm 2\sigma$ is

$$\mu \pm 2\sigma = 3.93 \pm 2(.71) = 3.93 \pm 1.42 = (2.51, 5.35)$$

Now $Y$ can assume the values 1, 2, 3, 4, and 5. Note that only the values 3, 4, and 5 fall within the interval. Thus, the probability that a value of $Y$ will fall in the interval $\mu \pm 2\sigma$ is

$$p(3) + p(4) + p(5) = .1735 + .6020 + .1837 = .9592$$

That is, about 95.9% of the star ratings fall between 2.50 and 5.36 stars. Clearly, the Empirical Rule (used in Chapter 2 to describe the variation for a finite set of data) provides an adequate description of the spread or variation in the probability distribution for $Y$.

**Example 4.6**

Empirical Rule Application—Hurricane Evacuation

A panel of meteorological and civil engineers studying emergency evacuation plans for Florida's Gulf Coast in the event of a hurricane has estimated that it would take between 13 and 18 hours to evacuate people living in low-lying land with the probabilities shown in Table 4.4.

a. Calculate the mean and standard deviation of the probability distribution of the evacuation times.
b. Within what range would you expect the time to evacuate to fall?

Solution

a. Let $Y$ represent the time required to evacuate people in low-lying land. Using Definitions 4.4 and 4.6, we compute

**TABLE 4.4 Estimated Probability Distribution of Hurricane Evacuation Time**

| Time to Evacuate (nearest hour) | Probability |
|---|---|
| 13 | .04 |
| 14 | .25 |
| 15 | .40 |
| 16 | .18 |
| 17 | .10 |
| 18 | .03 |

$$\mu = E(Y) = \sum yp(y)$$
$$= 13(.04) + 14(.25) + 15(.40) + 16(.18) + 17(.10) + 18(.03)$$
$$= 15.14 \text{ hours}$$
$$\sigma^2 = E[(Y - \mu)^2] = \sum (y - \mu)^2 p(y)$$
$$= (13 - 15.14)^2(.04) + (14 - 15.14)^2(.25) + \cdots$$
$$+ (18 - 15.14)^2(.03)$$
$$= 1.2404$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{1.2404} = 1.11 \text{ hours}$$

b. Based on the Empirical Rule, we expect about 95% of the observed evacuation times ($y$'s) to fall within $\mu \pm 2\sigma$, where

$$\mu \pm 2\sigma = 15.14 \pm 2(1.11) = 15.14 \pm 2.22 = (12.92, 17.36)$$

Consequently, we expect the time to evacuate to be between 12.92 hours and 17.36 hours. Based on the estimated probability distribution in Table 4.4, the actual probability that $Y$ falls between 12.92 and 17.36 is

$$\begin{aligned} P(12.92 \leq Y \leq 17.36) &= p(13) + p(14) + p(15) + p(16) + p(17) \\ &= .04 + .25 + .40 + .18 + .10 \\ &= .97 \end{aligned}$$

Once again, the Empirical Rule provides a good approximation to the probability of a value of the random variable $Y$ falling in the interval $\mu \pm 2\sigma$, especially when the distribution is approximately round-shaped.

## Applied Exercises

4.12 *Downloading "apps" to your cell phone.* According to an August, 2011 survey by the Pew Internet & American Life Project, nearly 40% of adult cell phone owners have downloaded an application ("app") to their cell phone. The accompanying table gives the probability distribution for $Y$, the number of "apps" used at least once a week by cell phone owners who have downloaded an "app" to their phone. (The probabilities in the table are based on information from the Pew Internet & American Life Project survey.)

| Number of "apps" used, $y$ | $p(y)$ |
|---|---|
| 0 | .17 |
| 1 | .10 |
| 2 | .11 |
| 3 | .11 |
| 4 | .10 |
| 5 | .10 |
| 6 | .07 |
| 7 | .05 |
| 8 | .03 |
| 9 | .02 |
| 10 | .02 |
| 11 | .02 |
| 12 | .02 |
| 13 | .02 |
| 14 | .01 |
| 15 | .01 |
| 16 | .01 |
| 17 | .01 |
| 18 | .01 |
| 19 | .005 |
| 20 | .005 |

a. Show that the properties of a probability distribution for a discrete random variable are satisfied.
b. Find $P(Y \geq 10)$.
c. Find the mean and variance of $Y$.
d. Give an interval that will contain the value of $Y$ with a probability of at least .75.

4.13 *Dust mite allergies.* Exercise 4.2 (p. 138) gives the probability distribution for the number $Y$ of homes with high dust mite levels.

a. Find $E(Y)$. Give a meaningful interpretation of the result.
b. Find $\sigma$.
c. Find the exact probability that $Y = y$ is in the interval $\mu \pm 2\sigma$. Compare to Chebyshev's Rule and the Empirical Rule.

4.14 *Controlling water hyacinth.* Refer to Exercise 4.3 (p. 138) and the probability distribution for the number of delphacid eggs on a blade of water hyacinth. Find the mean of the probability distribution and interpret its value.

4.15 *Hurricane evacuation times.* Refer to Example 4.6 (p. 142). The probability distribution for the time to evacuate in the event of a hurricane, Table 4.4, is reproduced here. Weather forecasters say they cannot accurately predict a hurricane landfall more than 14 hours in advance. If the Gulf Coast Civil Engineering Department waits until the 14-hour warning before beginning evacuation, what is the probability that all residents of low-lying areas are evacuated safely (i.e., before the hurricane hits the Gulf Coast)?

| Time to Evacuate (nearest hour) | Probability |
|---|---|
| 13 | .04 |
| 14 | .25 |
| 15 | .40 |
| 16 | .18 |
| 17 | .10 |
| 18 | .03 |

4.16 *Reliability of a manufacturing network.* Refer to the *Journal of Systems Sciences & Systems Engineering* (March, 2013) study of the reliability of a manufacturing system that involves multiple production lines, Exercise 4.6 (p. 139). Consider, again, the network for producing integrated circuit (IC) cards with two production lines set up in sequence. The probability distribution of the maximum capacity level of each line is reproduced below.

a. Find the mean maximum capacity for each line. Interpret the results practically.
b. Find the standard deviation of the maximum capacity for each line. Interpret the results practically.

| Line | Maximum Capacity, $y$ | $p(y)$ |
|---|---|---|
| 1 | 0 | .01 |
| | 12 | .02 |
| | 24 | .02 |
| | 36 | .95 |
| 2 | 0 | .002 |
| | 35 | .002 |
| | 70 | .996 |

4.17 *Mastering a computer program.* The number of training units that must be passed before a complex computer software program is mastered varies from one to five, depending on the student. After much experience, the software manufacturer has determined the probability distribution that describes the fraction of users mastering the software after each number of training units:

| Number of Units | Probability of Mastery |
|---|---|
| 1 | .1 |
| 2 | .25 |
| 3 | .4 |
| 4 | .15 |
| 5 | .1 |

a. Calculate the mean number of training units necessary to master the program. Calculate the median. Interpret each.
b. If the firm wants to ensure that at least 75% of the students master the program, what is the minimum number of training units that must be administered? At least 90%?
c. Suppose the firm develops a new training program that increases the probability that only one unit of training is needed from .1 to .25, increases the probability that only two units are needed to .35, leaves the probability that three units are needed at .4, and completely eliminates the need for four or five units. How do your answers to parts **a** and **b** change for this new program?

4.18 *Acceptance sampling of firing pins.* Refer to Exercise 4.11 (p. 140). Suppose the cost of testing a single firing pin is \$200.

a. What is the expected cost of inspecting the lot?
b. What is the variance?
c. Within what range would you expect the inspection cost to fall?

## 4.4 Some Useful Expectation Theorems

We now present three theorems that are especially useful in finding the expected value of a function of a random variable. We will leave the proofs of these theorems as theoretical exercises.

**THEOREM 4.1**

Let $Y$ be a discrete random variable with probability distribution $p(y)$ and let $c$ be a constant. Then the expected value (or mean) of $c$ is

$$E(c) = c$$

**THEOREM 4.2**

Let $Y$ be a discrete random variable with probability distribution $p(y)$ and let $c$ be a constant. Then the expected value (or mean) of $cY$ is

$$E(cY) = cE(Y)$$

**THEOREM 4.3**

Let $Y$ be a discrete random variable with probability distribution $p(y)$, and let $g_1(Y), g_2(Y), \ldots, g_k(Y)$ be functions of $Y$. Then

$$E[g_1(Y) + g_2(Y) + \cdots + g_k(Y) = E[g_1(Y)] + E[g_2(Y)] + \cdots + E[g_k(Y)]$$

Theorems 4.1–4.3 can be used to derive a simple formula for computing the variance of a random variable, as given by Theorem 4.4.

**THEOREM 4.4**

Lets $Y$ be a discrete random variable with probability distribution $p(y)$ and mean $\mu$. Then the variance of $Y$ is:

$$\sigma^2 = E(Y^2) - \mu^2$$

**Proof of Theorem 4.4** From Definition 4.6, we have the following expression for $\sigma^2$:

$$\sigma^2 = E[(Y - \mu)^2] = E(Y^2 - 2\mu Y + \mu^2)$$

Applying Theorem 4.3 yields

$$\sigma^2 = E(Y^2) + E(-2\mu Y) + E(\mu^2)$$

We now apply Theorems 4.1 and 4.2 to obtain

$$\begin{aligned}\sigma^2 &= E(Y^2) - 2\mu E(Y) + \mu^2 = E(Y^2) - 2\mu(\mu) + \mu^2 \\ &= E(Y^2) - 2\mu^2 + \mu^2 \\ &= E(Y^2) - \mu^2\end{aligned}$$

We will use Theorem 4.4 to derive the variances for some of the discrete random variables presented in the following sections. The method is demonstrated in Example 4.7.

**Example 4.7**

**Finding a Variance Using Theorem 4.4**

Refer to Example 4.4 and Table 4.3 (p. 141). Use Theorem 4.4 to find the variance for the random variable $Y$ = number stars in the rating of each car.

**Solution**

In Example 4.4, we found the variance of $Y$, the number of stars, by finding $\sigma^2 = E[(Y - \mu)^2]$ directly. Since this can be a tedious procedure, it is usually easier to find $E(Y^2)$ and then use Theorem 4.4 to compute $\sigma^2$. For our example,

$$\begin{aligned}E(Y^2) = \sum_{\text{all } y} y^2 p(y) &= (1)^2(0) + (2^2)(.0408) + (3)^2(.1735) \\ &\quad + (4)^3(.6020) + (5)^2(.1837) = 15.95\end{aligned}$$

Substituting the value $\mu = 3.93$ (obtained in Example 4.4) into the statement of Theorem 4.4, we have

$$\begin{aligned}\sigma^2 &= E(Y^2) - \mu^2 \\ &= 15.95 - (3.93)^2 = .51\end{aligned}$$

Note that this is the value of $\sigma^2$ that we obtained in Example 4.4.

In Sections 4.6–4.10, we will present some useful models of discrete probability distributions and will state without proof the mean, variance, and standard deviation for each. Some of these quantities will be derived in optional examples; other derivations will be left as optional exercises.

## Applied Exercises

4.19 *Dust mite allergies.* Refer to Exercises 4.2 (p. 138) and 4.13 (p. 143). Each home with a dust mite level that exceeds $2\mu$g/g will spend $2,000 for an allergen air purification system. Find the mean and variance of the total amount spent by the four sampled homes. Give a range where this total is likely to fall.

4.20 *Beach erosionol hot spots.* Use Theorem 4.4 to calculate the variance of the probability distribution in Exercise 4.4 (p. 138). Interpret the result.

4.21 *Downloading "apps" to your cellphone.* Use Theorem 4.4 to calculate the variance of the probability distribution in Exercise 4.12 (p. 143). Verify that your result agrees with Exercise 4.12.

4.22 *Acceptance sampling of firing pins.* Refer to Exercise 4.11 (p. 140), where $Y$ is the number of firing pins tested in a sample of five selected from a large lot. Suppose the cost of inspecting a single pin is $300 if the pin is defective and $100 if not. Then the total cost $C$ (in dollars) of the inspection is given by the equation $C = 200 + 100Y$. Find the mean and variance of $C$.

## Theoretical Exercises

4.23 Prove Theorem 4.1. [*Hint*: Use the fact that $\Sigma_{\text{all } y}\, p(y) = 1$.]

4.24 Prove Theorem 4.2. [*Hint*: The proof follows directly from Definition 4.5.]

4.25 Prove Theorem 4.3.

## 4.5 Bernoulli Trials

Several of the discrete probability distributions discussed in this chapter are based on experiments or processes in which a sequence of trials, called **Bernoulli trials**, are conducted.

A Bernoulli trial results in one of two mutually exclusive outcomes, typically denoted $S$ (for Success) and $F$ (for Failure). For example, tossing a coin is a Bernoulli trial since only one of two different outcomes can occur, head ($H$) or tail ($T$). The characteristics of a Bernoulli trial are stated in the box.

### Characteristics of a Bernoulli Trial

1. The trial results in one of two mutually exclusive outcomes. (We denote one outcome by $S$ and the other by $F$.)
2. The outcomes are exhaustive, i.e., no other outcomes are possible.
3. The probabilities of $S$ and $F$ are denoted by $p$ and $q$, respectively. That is, $P(S) = p$ and $P(F) = q$. Note that $p + q = 1$.

A **Bernoulli random variable** $Y$ is defined as the numerical outcome of a Bernoulli trial, where $Y = 1$ if a success occurs and $Y = 0$ if a failure occurs. Consequently, the probability distribution for $Y = y$ is shown in Table 4.5 and the next box.

**TABLE 4.5 Bernoulli Probability Distribution**

| Outcome | $Y = y$ | $p(y)$ |
|---|---|---|
| $S$ | 1 | $p$ |
| $F$ | 0 | $q$ |

### The Bernoulli Probability Distribution

Consider a Bernoulli trial where

$$Y = \begin{cases} 1 & \text{if a success } (S) \text{ occurs} \\ 0 & \text{if a failure } (F) \text{ occurs} \end{cases}$$

The probability distribution for the Bernoulli random variable $Y$ is given by

$$p(y) = p^y q^{1-y} \quad (y = 0, 1)$$

where

$p$ = Probability of a success for a Bernoulli trial
$q = 1 - p$

The mean and variance of the Bernoulli random variable are, respectively,

$$\mu = p \quad \text{and} \quad \sigma^2 = pq$$

In the Bernoulli coin-tossing experiment, define $H$ as a success and $T$ as a failure. Then $Y = 1$ if $H$ occurs and $Y = 0$ if $T$ occurs. Since $P(H) = P(T) = .5$ if the coin is balanced, the probability distribution for $Y$ is

$$p(1) = p = .5$$
$$p(0) = q = .5$$

**Example 4.8**

$\mu$ and $\sigma$ for a Bernoulli Random Variable

Show that for a Bernoulli random variable $Y$, $\mu = p$ and $\sigma = \sqrt{pq}$

**Solution**

We know that $P(Y = 1) = p(1) = p$ and $P(Y = 0) = p(0) = q$. Then, from Definition 4.4,

$$\mu = E(Y) = \sum yp(y) = (1)p(1) + (0)p(0) = p(1) = p$$

Also, from Definition 4.5 and Theorem 4.4,

$$\sigma^2 = E(Y^2) - \mu^2 = \sum y^2 p(y) - \mu^2 = (1)^2 p(1) + (0)^2 p(0) - \mu^2$$
$$= p(1) - \mu^2 = p - p^2 = p(1 - p) = pq$$

Consequently, $\sigma = \sqrt{\sigma^2} = \sqrt{pq}$.

A Bernoulli random variable, by itself, is of little interest in engineering and science applications. Conducting a *series* of Bernoulli trials, however, leads to some well-known and useful discrete probability distributions. One of these is described in the next section.

## 4.6 The Binomial Probability Distribution

Many real-life experiments result from conducting a series of Bernoulli trials and are analogous to tossing an *unbalanced* coin a number $n$ of times. Suppose that 30% of the private wells that provide drinking water to a metropolitan area contain impurity A. Then selecting a random sample of 10 wells and testing for impurity A would be analogous to tossing an unbalanced coin 10 times, with the probability of tossing a head (detecting impurity A) on a single trial equal to .30. Public opinion or consumer preference polls that elicit one of two responses—yes or no, approve or disapprove, etc.—are also analogous to the unbalanced coin tossing experiment if the number $N$ in the population is large and if the sample size $n$ is relatively small, say, $.10N$ or less. All these experiments are particular examples of a **binomial experiment**. Such experiments and the resulting binomial random variables possess the characteristics stated in the box.

### Characteristics That Define a Binomial Random Variable

1. The experiment consists of $n$ identical Bernoulli trials.
2. There are only two possible outcomes on each trial: $S$ (for Success) and $F$ (for Failure).
3. $P(S) = p$ and $P(F) = q$ remain the same from trial to trial. (Note that $p + q = 1$.)
4. The trials are independent.
5. The binomial random variable $Y$ is the number of $S$'s in $n$ trials.

The binomial probability distribution, its mean, and its variance are shown in the next box. Figure 4.5 shows the relative frequency histograms of binomial distributions for a sample of $n = 10$ and different values of $p$. Note that the probability distribution is skewed to the right for small values of $p$, skewed to the left for large values of $p$, and symmetric for $p = .5$.

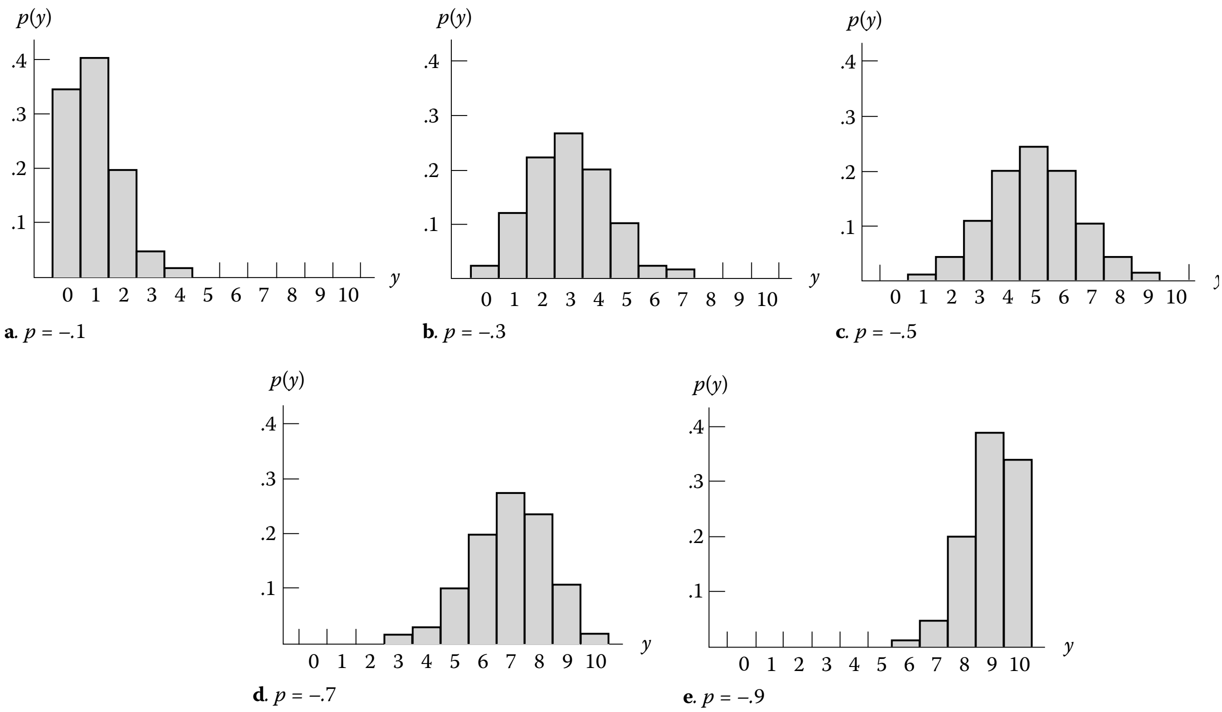



**FIGURE 4.5**
Binomial probability distributions for $n = 10$, $p = .1, .3, .5, .7, .9$

### The Binomial Probability Distribution

The probability distribution for a binomial random variable $Y$ is given by

$$p(y) = \binom{n}{y} p^y q^{n-y} \quad (y = 0, 1, 2, \ldots, n)$$

where

$$\begin{aligned} p &= \text{Probability of a success on a single trial} \\ q &= 1 - p \\ n &= \text{Number of trials} \\ y &= \text{Number of successes in } n \text{ trials} \\ \binom{n}{y} &= \frac{n!}{y!(n-y)!} \end{aligned}$$

The mean and variance of the binomial random variable are, respectively,

$$\mu = np \quad \text{and} \quad \sigma^2 = npq$$

The binomial probability distribution is derived as follows. A simple event for a binomial experiment consisting of $n$ Bernoulli trials can be represented by the symbol

$$SFSFFFSSSF \ldots SFS$$

where the letter in the $i$th position, proceeding from left to right, denotes the outcome of the $i$th trial. Since we want to find the probability $p(y)$ of observing $y$ successes in the $n$ trials, we will need to sum the probabilities of all simple events that contain $y$ successes ($S$'s) and $(n - y)$ failures ($F$'s). Such simple events would appear symbolically as

$$\overbrace{SSSS \ldots S}^{y} \; \overbrace{FF \ldots F}^{(n-y)}$$

or some different arrangement of these symbols.

Since the trials are independent, the probability of a *particular* simple event implying $y$ successes is

$$\overbrace{P(SSS \ldots S}^{y} \; \overbrace{FF \ldots F)}^{(n-y)} = p^y q^{n-y}$$

The *number* of these equiprobable simple events is equal to the number of ways we can arrange the $y$ $S$'s and the $(n - y)$ $F$'s in $n$ positions corresponding to the $n$ trials. This is equal to the number of ways of selecting $y$ positions (trials) for the $y$ $S$'s from a total of $n$ positions. This number, given by Theorem 3.4, is

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

We have determined the probability of each simple event that results in $y$ successes, as well as the number of such events. We now sum the probabilities of these simple events to obtain

$$p(y) = \left(\begin{array}{c}\text{Number of simple events} \\ \text{implying } y \text{ successes}\end{array}\right)\left(\begin{array}{c}\text{Probability of one of these} \\ \text{equiprobable simple events}\end{array}\right)$$

or

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

**Example 4.9**

**Binomial Application—Computer Power Loads**

Electrical engineers recognize that high neutral current in computer power systems is a potential problem. A survey of computer power system load currents at U.S. sites found that 10% of the sites had high neutral to full-load current ratios (*IEEE Transactions on Industry Applications*). If a random sample of five computer power systems is selected from the large number of sites in the country, what is the probability that

a. Exactly three will have a high neutral to full-current load ratio?
b. At least three?
c. Fewer than three?

**Solution**

The first step is to confirm that this experiment possesses the characteristics of a binomial experiment. The experiment consists of $n = 5$ Bernoulli trials, one corresponding to each randomly selected site. Each trial results in an $S$ (the site has a computer power system with a high neutral to full-load current ratio) or an $F$ (the system does not have a high ratio). Since the total number of sites with computer power systems in the country is large, the probability of drawing a single site and finding that it has a high neutral to full-load current ratio is .1, and this probability will remain approximately the same (for all practical purposes) for each of the five selected sites. Further, since the sampling was random, we assume that the outcome on any one site is unaffected by the outcome of any other and that the trials are independent. Finally, we are interested in the number $Y$ of sites in the sample of $n = 5$ that have high neutral to full-load current ratios. Therefore, the sampling procedure represents a binomial experiment with $n = 5$ and $p = .1$.

a. The probability of drawing exactly $Y = 3$ sites containing a high ratio is

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

where $n = 5$, $p = .1$, and $y = 3$. Thus,

$$p(3) = \frac{5!}{3!2!}(.1)^3(.9)^2 = .0081$$

b. The probability of observing at least three sites with high ratios is

$$P(Y \geq 3) = p(3) + p(4) + p(5)$$

where

$$p(4) = \frac{5!}{4!1!}(.1)^4(.9)^1 = .00045$$

$$p(5) = \frac{5!}{5!0!}(.1)^5(.9)^0 = .00001$$

Since we found $p(3)$ in part **a**, we have

$$\begin{aligned} P(Y \geq 3) &= p(3) + p(4) + p(5) \\ &= .0081 + .00045 + .00001 = .00856 \end{aligned}$$

c. Although $P(Y < 3) = p(0) + p(1) + p(2)$, we can avoid calculating these probabilities by using the complementary relationship and the fact that $\sum_{y=0}^{n} p(y) = 1$.

Therefore,

$$P(Y < 3) = 1 - P(Y \geq 3) = 1 - .00856 = .99144$$

Tables that give partial sums of the form

$$\sum_{y=0}^{k} p(y)$$

for binomial probabilities—called **cumulative binomial probabilities**—are given in Table 2 of Appendix B, for $n = 5, 10, 15, 20$, and 25. For example, you will find that the partial sum given in the table for $n = 5$, in the row corresponding to $k = 2$ and the column corresponding to $p = .1$, is

$$\sum_{y=0}^{2} p(y) = p(0) + p(1) + p(2) = .991$$

This answer, correct to three decimal places, agrees with our answer to part **c** of Example 4.9.

## Example 4.10

**Mean and Variance of a Binomial Random Variable**

Find the mean, variance, and standard deviation for a binomial random variable with $n = 20$ and $p = .6$. Construct the interval $\mu \pm 2\sigma$ and compute $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.

**Solution** Applying the formulas given previously, we have

$$\mu = np = 20(.6) = 12$$
$$\sigma^2 = npq = 20(.6)(.4) = 4.8$$
$$\sigma = \sqrt{4.8} = 2.19$$

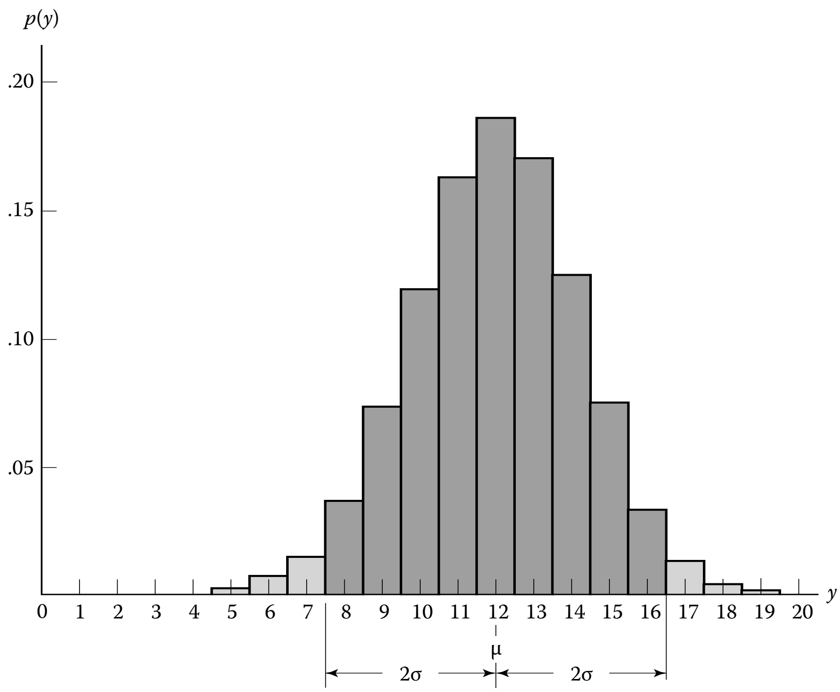



**FIGURE 4.6**
Binomial probability distribution for $Y$ in Example 4.10 $(n = 20, p = .6)$

The binomial probability distribution for $n = 20$ and $p = .6$ and the interval $\mu \pm 2\sigma$, or 7.62 to 16.38, are shown in Figure 4.6 (p. 151). The values of $Y$ that lie in the interval $\mu \pm 2\sigma$ (highlighted) are 8, 9, . . . , 16. Therefore,

$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = P(Y = 8, 9, 10, \ldots, \text{or } 16)$$
$$= \sum_{y=0}^{16} p(y) - \sum_{y=0}^{7} p(y)$$

We obtain the values of these partial sums from Table 2 of Appendix B:

$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = \sum_{y=0}^{16} p(y) - \sum_{y=0}^{7} p(y)$$
$$= .984 - .021 = .963$$

You can see that this result is close to the value of .95 specified by the Empirical Rule, discussed in Chapter 2.

**Example 4.11** *(optional)*

Derivation of Binomial Expected Value

Derive the formula for the expected value for the binomial random variable, $Y$.

Solution

By Definition 4.4,

$$\mu = E(Y) = \sum_{\text{all } y} yp(y) = \sum_{y=0}^{n} y \frac{n!}{y!(n-y)!} p^y q^{n-y}$$

The easiest way to sum these terms is to convert them into binomial probabilities and then use the fact that $\Sigma_{y=0}^{n} p(y) = 1$. Noting that the first term of the summation is equal to 0 (since $Y = 0$), we have

$$\mu = \sum_{y=1}^{n} y \frac{n!}{[y(y-1)\cdots 3 \cdot 2 \cdot 1](n-y)!} p^y q^{n-y}$$
$$= \sum_{y=1}^{n} \frac{n!}{(y-1)!(n-y)!} p^y q^{n-y}$$

Because $n$ and $p$ are constants, we can use Theorem 4.2 to factor $np$ out of the sum:

$$\mu = np \sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!(n-y)!} p^{y-1} q^{n-y}$$

Let $Z = (Y - 1)$. Then when $Y = 1$, $Z = 0$ and when $Y = n$, $Z = (n - 1)$; thus,

$$\mu = np \sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!(n-y)!} p^{y-1} q^{n-y}$$
$$= np \sum_{z=0}^{n-1} \frac{(n-1)!}{z![(n-1)-z]!} p^z q^{(n-1)-z}$$

The quantity inside the summation sign is $p(z)$, where $Z$ is a binomial random variable based on $(n - 1)$ Bernoulli trials. Therefore,

$$\sum_{z=0}^{n-1} p(z) = 1$$

and

$$\mu = np \sum_{z=0}^{n-1} p(z) = np(1) = np$$

## Applied Exercises

4.26 *STEM experiences for girls.* Refer to the March, 2013 National Science Foundation study of girls participation in science, technology, engineering or mathematics (STEM) programs, Exercise 1.1 (p.5). Recall that the study found that of girls surveyed who participated in a STEM program, 27% felt that the program increased their interest in science. Assume that this figure applies to all girls who have participated in a STEM program. Now, consider a sample of 20 girls randomly selected from all girls who have participated in a STEM program. Let $Y$ represent the number of girls in the sample who feel that the program increased their interest in science.

a. Demonstrate that $Y$ is a binomial random variable.
b. Give the mean and variance of $Y$, and interpret the results practically.
c. Find the probability that fewer than half of the 20 sampled girls feel that the STEM program increased their interest in science.

4.27 *Chemical signals of mice.* Refer to the *Cell* (May 14, 2010) study of the ability of a mouse to recognize the odor of a potential predator, Exercise 3.18 (p.91). Recall that the source of these odors is typically major urinary proteins (Mups). In an experiment, 40% of lab mice cells exposed to chemically produced cat Mups responded positively (i.e., recognized the danger of the lurking predator). Consider a sample of 100 lab mice cells, each exposed to chemically produced cat Mups. Let $Y$ represent the number of cells that respond positively.

a. Explain why the probability distribution of $Y$ can be approximated by the binomial distribution.
b. Find $E(Y)$ and interpret its value, practically.
c. Find the variance of $Y$.
d. Give an interval that is likely to contain the value of $Y$.

4.28 *Ecotoxicological survival study.* In the *Journal of Agricultural, Biological and Environmental Sciences* (Sept., 2000), researchers evaluated the risk posed by hazardous pollutants using an ecotoxicological survival model. For one experiment, 20 guppies (all the same age and size) were released into a tank of natural seawater polluted with the pesticide dieldrin. Of interest is $Y$, the number of guppies surviving after 5 days. The researchers estimated that the probability of any single guppy surviving was .60.

a. Demonstrate that $Y$ has a binomial probability distribution. What is the value of $p$?
b. Find the probability that $Y = 7$.
c. Find the probability that at least 10 guppies survive.

4.29 *Analysis of bottled water.* Is the bottled water you're drinking really purified water? A 4-year study of bottled water brands conducted by the Natural Resources Defense Council found that 25% of bottled water is just tap water packaged in a bottle. (*Scientific American*, July 2003.) Consider a sample of five bottled water brands, and let $Y$ equal the number of these brands that use tap water.

a. Give the probability distribution for $Y$ as a formula.
b. Find $P(Y = 2)$.
c. Find $P(Y \leq 1)$.

4.30 *Bridge inspection ratings.* According to the National Bridge Inspection Standard (NBIS), public bridges over 20 feet in length must be inspected and rated every 2 years. The NBIS rating scale ranges from 0 (poorest rating) to 9 (highest rating). University of Colorado engineers used a probabilistic model to forecast the inspection ratings of all major bridges in Denver. (*Journal of Performance of Constructed Facilities*, Feb. 2005.) For the year 2020, the engineers forecast that 9% of all major Denver bridges will have ratings of 4 or below.

a. Use the forecast to find the probability that in a random sample of 10 major Denver bridges, at least 3 will have an inspection rating of 4 or below in 2020.
b. Suppose that you actually observe 3 or more of the sample of 10 bridges with inspection ratings of 4 or below in 2020. What inference can you make? Why?

4.31 *Detecting a computer virus attack.* *Chance* (Winter 2004) presented basic methods for detecting virus attacks (e.g., Trojan programs or worms) on a network computer that are sent from a remote host. These viruses reach the network through requests for communication (e.g., e-mail, web chat, or remote login) that are identified as "packets." For example, the "SYN flood" virus ties up the network computer by "flooding" the network with multiple packets. Cybersecurity experts can detect this type of virus attack if at least one packet is observed by a network sensor. Assume the probability of observing a single packet sent from a new virus is only .001. If the virus actually sends 150 packets to a network computer, what is the probability that the virus is detected by the sensor?

4.32 *Fingerprint expertise.* Refer to the *Psychological Science* (August, 2011) study of fingerprint identification, Exercise 3.36 (p.103). The study found that when presented with prints from the same individual, a fingerprint expert will correctly identify the match 92% of the time. In contrast, a novice will correctly identify the match 75% of the time. Consider a sample of five different pairs of fingerprints, where each pair is a match.

a. What is the probability that an expert will correctly identify the match in all five pairs of fingerprints?
b. What is the probability that a novice will correctly identify the match in all five pairs of fingerprints?

4.33 *Chickens with fecal contamination.* The United States Department of Agriculture (USDA) reports that, under its standard inspection system, 1 in every 100 slaughtered chickens passes inspection with fecal contamination. In Exercise 3.8 (p. 85), you found the probability that a randomly selected slaughtered chicken passes inspection with fecal contamination. Now find the probability that, in a random sample of 5 slaughtered chickens, at least 1 passes inspection with fecal contamination.

4.34 *PhD's in engineering.* The National Science Foundation reports that 70% of the U.S. graduate students who earn PhD degrees in engineering are foreign nationals. Consider the number $Y$ of foreign students in a random sample of 25 engineering students who recently earned their PhD.

a. Find $P(Y = 10)$.
b. Find $P(Y \leq 5)$.
c. Find the mean $\mu$ and standard deviation $\sigma$ for $Y$.
d. Interpret the results, part **c**.

4.35 *Network forensic analysis.* A network forensic analyst is responsible for identifying worms, viruses, and infected nodes in the computer network. A new methodology for finding patterns in data that signify infections was investigated in *IEEE Transactions on Information Forensics and Security* (May, 2013). The method uses multiple filters to check strings of information. For this exercise, consider a data string of length 4 bytes (positions), where each byte is either a 0 or a 1 (e.g., 0010). Also, consider two possible strings, named $S_1$ and $S_2$. In a simple single filter system, the probability that $S_1$ and $S_2$ differ in any one of the bytes is .5. Derive a formula for the probability that the two strings differ on exactly $Y$ of the 4 bytes. Do you recognize this probability distribution?

4.36 *Reflection of neutron particles.* Refer to the neutral particle transport problem described in Exercise 3.42 (p. 84). Recall that particles released into an evacuated duct collide with the inner duct wall and are either scattered (reflected) with probability .16 or absorbed with probability .84.

a. If 4 particles are released into the duct, what is the probability that all 4 will be absorbed by the inner duct wall? Exactly 3 of the 4?
b. If 20 particles are released into the duct, what is the probability that at least 10 will be reflected by the inner duct wall? Exactly 10?

4.37 *Premature aging gene.* Ataxia-telangiectasia (A-T) is a neurological disorder that weakens immune systems and causes premature aging. *Science News* reports that when both members of a couple carry the A-T gene, their children have a 1 in 5 chance of developing the disease.

a. Consider 15 couples in which both members of each couple carry the A-T gene. What is the probability that more than 8 of 15 couples have children that develop the neurological disorder?
b. Consider 10,000 couples in which both members of each couple carry the A-T gene. Is it likely that fewer than 3,000 will have children that develop the disease?

## Theoretical Exercises

4.38 For the binomial probability distribution $p(y)$, show that $\Sigma_{y=0}^{n} p(y) = 1$ [*Hint*: The binomial theorem, which pertains to the expansion of $(a + b)^n$, states that

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n}b^n$$

Let $a = q$ and $b = p$.]

4.39 Show that, for a binomial random variable,

$$E[Y(Y - 1)] = npq + \mu^2 - \mu$$

[*Hint*: Write the expected value as a sum, factor out $y(y - 1)$, and then factor terms until each term in the sum is a binomial probability. Use the fact that $\Sigma_y p(y) = 1$ to sum the series.]

4.40 Use the results of Exercise 4.39 and the fact that

$$E[Y(Y - 1)] = E(Y^2 - Y) = E(Y^2) - E(Y) = E(Y^2) - \mu$$

to find $E(Y^2)$ for a binomial random variable.

4.41 Use the results of Exercises 4.39 and 4.40, in conjunction with Theorem 4.4, to show that $\sigma^2 = npq$ for a binomial random variable.

# 4.7 The Multinomial Probability Distribution

Many types of experiments result in observations on a qualitative variable with more than two possible outcomes. For example, suppose that sparkplugs for personal watercraft engines are manufactured on one of five different production lines, A, B, C, D, or E. To compare the proportions of defective sparkplugs that can be attributed to the five production lines, all defective plugs located by quality control engineers are classified each day according to the production line. Each sparkplug is an experimental unit and the observation is a letter that identifies the production line on which it was produced. Production line is clearly a qualitative variable.

**TABLE 4.6 Classification of the $n = 103$ Defective Sparkplugs According to Production Line**

| Production Line | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| 15 | 27 | 31 | 19 | 11 |

Suppose that $n = 103$ sparkplugs are found to be defective in a given week. The $n = 103$ qualitative observations, each resulting in an A, B, C, D, or E, produce counts giving the numbers of defectives emerging from the five production lines. For example, if there were $Y_1 = 15$ A's, $Y_2 = 27$ B's, $Y_3 = 31$ C's, $Y_4 = 19$ D's, and $Y_5 = 11$ E's, the classified data would appear as shown in Table 4.6, which shows the counts in each category of the classification. Note that the sum of the numbers of defective sparkplugs produced by the five lines must equal the total number of defectives:

$$n = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 = 15 + 27 + 31 + 19 + 11 = 103$$

The classification experiment that we have just described is called a **multinomial experiment** and represents an extension of the binomial experiment discussed in Section 4.6. Such an experiment consists of $n$ identical trials—that is, observations on $n$ experimental units. Each trial must result in one and only one of $k$ outcomes, the $k$ classification categories (for the binomial experiment, $k = 2$). The probability that the outcome of a single trial will fall in category $i$ is $p_i$ ($i = 1, 2, \ldots, k$). Finally, the trials are independent and we are interested in the numbers of observations, $Y_1, Y_2, \ldots, Y_k$, falling in the $k$ classification categories.

### Properties of the Multinomial Experiment

1. The experiment consists of $n$ identical trials.
2. There are $k$ possible outcomes to each trial.
3. The probabilities of the $k$ outcomes, denoted by $p_1, p_2, \ldots, p_k$, remain the same from trial to trial, where $p_1 + p_2 + \cdots + p_k = 1$.
4. The trials are independent.
5. The random variables of interest are the counts $Y_1, Y_2, \ldots, Y_k$ in each of the $k$ classification categories.

The multinomial distribution, its mean, and its variance are shown in the following box.

### The Multinomial Probability Distribution

$$p(y_1, y_2, \ldots, y_k) = \frac{n!}{y_1!y_2!\cdots y_k!}(p_1)^{y_1}(p_2)^{y_2}\cdots(p_k)^{y_k}$$

where

$p_i$ = Probability of outcome $i$ on a single trial
$p_1 + p_2 + \cdots + p_k = 1$
$n = y_1 + y_2 + \cdots + y_k$ = Number of trials

$y_i$ = Number of occurrences of outcome $i$ in $n$ trials

The mean and variance of the multinomial random variable $y_i$ are, respectively,

$$\mu_i = np_i \quad \text{and} \quad \sigma_i^2 = np_i(1 - p_i)$$

The procedure for deriving the **multinomial probability distribution** $p(y_1, y_2, \ldots, y_k)$ for the category counts, $y_1, y_2, \ldots, y_k$, is identical to the procedure employed for a binomial experiment. To simplify our notation, we will illustrate the procedure for $k = 3$ categories. The derivation of $p(y_1, y_2, \ldots, y_k)$ for $k$ categories is similar.

Let the three outcomes corresponding to the $k = 3$ categories be denoted as A, B, and C, with respective category probabilities $p_1$, $p_2$, and $p_3$. Then any observation of the outcome of $n$ trials will result in a simple event of the type shown in Table 4.7. The outcome of each trial is indicated by the letter that was observed. Thus, the simple event in Table 4.7 is one that results in C on the first trial, A on the second, A on the third, . . . , and B on the last.

**TABLE 4.7 A Typical Simple Event for a Multinomial Experiment (k = 3)**

| Trial | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | ... | $n$ |
| C | A | A | B | A | C | ... | B |

Now consider a simple event that will result in $y_1$ A outcomes, $y_2$ B outcomes, and $y_3$ C outcomes, where $y_1 + y_2 + y_3 = n$. One of these simple events is shown below:

$$\overbrace{\text{AAA}\ldots\text{A}}^{y_1} \quad \overbrace{\text{BBB}\ldots\text{B}}^{y_2} \quad \overbrace{\text{CCC}\ldots\text{C}}^{y_3}$$

The probability of this simple event, which results in $y_1$ A outcomes, $y_2$ B outcomes, and $y_3$ C outcomes, is

$$(p_1)^{y_1}(p_2)^{y_2}(p_3)^{y_3}$$

How many simple events will there be in the sample space $S$ that will imply $y_1$ A's, $y_2$ B's, and $y_3$ C's? This number is equal to the number of different ways that we can arrange the $y_1$ A's, $y_2$ B's, and $y_3$ C's in the $n$ distinct positions. The number of ways that we would assign $y_1$ positions to A, $y_2$ positions to B, and $y_3$ to C is given by Theorem 3.3 as

$$\frac{n!}{y_1!y_2!y_3!}$$

Therefore, there are $n!/(y_1!y_2!y_3!)$ simple events resulting in $y_1$ A's, $y_2$ B's, and $y_3$ C's, each with probability $(p_1)^{y_1}(p_2)^{y_2}(p_3)^{y_3}$. It then follows that the probability of observing $y_1$ A's, $y_2$ B's, and $y_3$ C's in $n$ trials is equal to the sum of the probabilities of these simple events:

$$p(y_1, y_2, y_3) = \frac{n!}{y_1!y_2!y_3!}(p_1)^{y_1}(p_2)^{y_2}(p_3)^{y_3}$$

You can verify that this is the expression obtained by substituting $k = 3$ into the formula for the multinomial probability distribution shown in the box.

The expected value, or mean, of the number of counts for a particular category, say, category $i$, follows directly from our knowledge of the properties of a binomial random variable. If we combine all categories other than category $i$ into a single category, then the multinomial classification becomes a binomial classification with $Y_i$ observations in category $i$ and $(n - Y_i)$ observations in the combined category. Then, from our knowledge of the expected value and variance of a binomial random variable, it follows that

$$E(Y_i) = np_i$$
$$V(Y_i) = np_i(1 - p_i)$$

**Example 4.12**

Multinomial Application-Computer Power Loads

Refer to the study of neutral to full-load current ratios in computer power systems, Example 4.9 (p. 150). Suppose that the electrical engineers found that 10% of the systems have high ratios, 30% have moderate ratios, and 60% have low ratios. Consider a random sample of $n = 40$ computer power system sites.

a. Find the probability that 10 sites have high neutral to full-load current ratios, 10 sites have moderate ratios, and 20 sites have low ratios.

b. Find the mean and variance of the number of sites that have high neutral to full-load current ratios. Use this information to estimate the number of sites in the sample of 40 that will have high ratios.

Solution

In the solution to Example 4.9, we verified that the properties of a binomial experiment were satisfied. This example is simply an extension of the binomial experiment to one involving $k = 3$ possible outcomes—high, neutral, or low ratio—for each site. Thus, the properties of a multinomial experiment are satisfied, and we may apply the formulas given in the box.

a. Define the following:

$Y_1$ = Number of sites with high ratios
$Y_2$ = Number of sites with moderate ratios
$Y_3$ = Number of sites with low ratios
$p_1$ = Probability of a site with a high ratio
$p_2$ = Probability of a site with a moderate ratio
$p_3$ = Probability of a site with a low ratio

Then we want to find the probability, $P(Y_1 = 10, Y_2 = 10, Y_3 = 20) = p(10, 10, 20)$, using the formula

$$p(y_1, y_2, y_3) = \frac{n!}{y_1!y_2!y_3!}(p_1)^{y_1}(p_2)^{y_2}(p_3)^{y_3}$$

where $n = 40$ and our estimates of $p_1$, $p_2$, and $p_3$ are .1, .3, and .6, respectively. Substituting these values, we obtain

$$p(10, 10, 20) = \frac{40!}{10!10!20!}(.1)^{10}(.3)^{10}(.6)^{20} = .0005498$$

b. We want to find the mean and variance of $Y_1$, the number of sites with high neutral to full-load current ratios. From the formula in the box, we have

$$\mu_1 = np_1 = 40(.1) = 4$$

and

$$\sigma_1^2 = np_1(1 - p_1) = 40(.1)(.9) = 3.6$$

From our knowledge of the Empirical Rule, we expect $Y_1$, the number of sites in the sample with high ratios, to fall within 2 standard deviations of its mean, i.e., between

$$\mu_1 - 2\sigma_1 = 4 - 2\sqrt{3.6} = .21$$

and

$$\mu_1 + 2\sigma_1 = 4 + 2\sqrt{3.6} = 7.80$$

Since $Y_1$ can take only whole-number values, 0, 1, 2, . . . , we expect the number of sites with high ratios to fall between 1 and 7.

## Applied Exercises

4.42 *Microsoft program security issues.* Refer to the *Computers & Security* (July, 2013) study of Microsoft program security issues, Exercise 2.4 (p. 27). Recall that Microsoft periodically issues a Security Bulletin that reports the software -- Windows, Explorer, or Office — affected by the vulnerability. The study discovered that 64% of the security bulletins reported an issue with Windows, 12% with Explorer, and 24% with Office. The researchers also categorized the security bulletins according to the expected repercussion of the vulnerability. Assume the categories (and associated percentages) are Denial of service (10%), Information disclosure (15%), Remote code execution (45%), Spoofing (5%), and Privilege elevation (25%). Now consider a random sample of 10 Microsoft security bulletins.

a. How many of these sampled bulletins would you expect to report an issue with Explorer?
b. How many of these sampled bulletins would you expect to report Remote code execution as a repercussion?

c. What is the likelihood that all 10 of the bulletins report an issue with Windows?

d. What is the likelihood that there are 2 sampled bulletins reporting repercussions for each of the five types, Denial of service, Information disclosure, Remote code execution, Spoofing, and Privilege elevation?

4.43 *Underwater acoustic communication.* A subcarrier is one telecommunication signal carrier that is carried on top of another carrier so that effectively two signals are carried at the same time. Subcarriers can be used for an entirely different purpose than main carriers. For example, data subcarriers are used for data transmissions; pilot subcarriers are used for channel estimation and synchronization; and, null subcarriers are used for direct current and guard banks transmitting no signal. In the *IEEE Journal of Oceanic Engineering* (April, 2013), researchers studied the characteristics of subcarriers for underwater acoustic communications. Based on an experiment conducted off the coast of Martha's Vineyard (MA), they estimated that 25% of subcarriers are pilot subcarriers, 10% are null subcarriers, and 65% are data subcarriers. Consider a sample of 50 subcarriers transmitted for underwater acoustic communications.

a. How many of the 50 subcarriers do you expect to be pilot subcarriers? Null subcarriers? Data subcarriers?

b. How likely is it to observe 10 pilot subcarriers, 10 null subcarriers, and 30 data subcarriers?

c. If you observe more than 25 pilot subcarriers, what would you conclude? Explain.

4.44 *Controlling water hyacinth.* Refer to the *Annals of the Entomological Society of America* (Jan. 2005) study of the life cycle of a South American delphacid species, Exercise 4.3 (p. 138). Recall that entomological engineers have found that the delphacid is a natural enemy of water hyacinth. The table giving the percentages of water hyacinth blades that have one, two, three, and four delphacid eggs is reproduced here. Consider a sample of 100 water hyacinth blades selected from an environment inhabited by delphacids. Let $Y_1$, $Y_2$, $Y_3$, and $Y_4$ represent the number of blades in the sample with one egg, two eggs, three eggs, and four eggs, respectively. Find the probability that half of the sampled blades have one egg, half have two eggs, and none of the blades has three or four eggs.

| | One Egg | Two Eggs | Three Eggs | Four Eggs |
|---|---|---|---|---|
| Percentage of Blades | 40 | 54 | 2 | 4 |

*Source:* Sosa, A. J., et al. "Life history of *Megamelus scutellaris* with description of immature stages." *Annals of the Entomological Society of America*, Vol. 98, No. 1, Jan. 2005 (adapted from Table 1).

4.45 *Dust explosions.* Dust explosions in the chemical process industry, although rare, pose a great potential for injury and equipment damage. *Process Safety Progress* (Sept., 2004) reported on the likelihood of dust explosion incidents. The table gives the proportion of incidents in each of several worldwide industries where dust explosions have occurred. Suppose 20 dust explosions occur worldwide next year.

| Industry | Proportion |
|---|---|
| Wood/Paper | .30 |
| Grain/Foodstuffs | .10 |
| Metal | .07 |
| Power Generation | .07 |
| Plastics/Mining/Textile | .08 |
| Miscellaneous | .38 |

*Source:* Frank, W. L. "Dust explosion prevention and the critical importance of housekeeping." *Process Safety Progress*, Vol. 23, No. 3, Sept. 2004 (adapted from Table 2).

a. Find the probability that 7 explosions occur in the wood/paper industry, 5 occur in the grain/foodstuffs industry, 2 occur in the metal industry, none occur in the power industry, 1 occurs in the plastics/mining/textile industry, and 5 occur in all other industries.

b. Find the probability that fewer than 3 occur in the wood/paper industry.

4.46 *Railway track allocation.* Refer to the *Journal of Transportation Engineering* (May, 2013) investigation of the assignment of tracks to trains at a busy railroad station, Exercise 2.8 (p. 28). Ideally, engineers will assign trains to tracks in order to minimize waiting time and bottlenecks. Assume there are 10 tracks at the railroad station and the trains will be randomly assigned to a track. Suppose that in a single day there are 50 trains that require track assignment.

a. What is the probability that exactly 5 trains are assigned to each of the 10 tracks?

b. A track is considered underutilized if fewer than 2 trains are assigned to the track during the day. Find the probability that Track #1 is underutilized.

4.47 *Color as body orientation clue.* To compensate for disorientation in zero gravity, astronauts rely heavily on visual information to establish a top-down orientation. The potential of using color brightness as a body orientation clue was studied in *Human Factors* (Dec. 1988). Ninety college students, reclining on their backs in the dark, were disoriented when positioned on a rotating platform under a slowly rotating disk that blocked their field of vision. The subjects were asked to say "stop" when they felt as if they were right-side up. The position of the brightness pattern on the disk in relation to each student's body orientation was then recorded. Subjects selected only three disk brightness patterns as subjective vertical clues: (1) brighter side up, (2) darker side up, and (3) brighter and darker side aligned on either side of the subjects' heads. Based on the study results, the probabilities of subjects selecting the three disk orientations are .65, .15, and .20, respectively. Suppose $n = 8$ subjects perform a similar experiment.

a. What is the probability that all eight subjects select the brighter-side-up disk orientation?
b. What is the probability that four subjects select the brighter-side-up orientation, three select the darker-side-up orientation, and one selects the aligned orientation?
c. On average, how many of the eight subjects will select the brighter-side-up orientation?

4.48 *Detecting overweight trucks.* Although illegal, overloading is common in the trucking industry. Minnesota Department of Transportation (MDOT) engineers monitored the movements of overweight trucks on an interstate highway using an unmanned, computerized scale that is built into the highway. Unknown to the truckers, the scale weighed their vehicles as they passed over it. One week, over 400 five-axle trucks were deemed to be overweight. The table at the bottom of the page shows the proportion of these overweight trucks that were detected each day of the week. Assume the daily distribution of overweight trucks remains the same from week to week.

a. If 200 overweight trucks are detected during a single week, what is the probability that 50 are detected on Monday, 50 on Tuesday, 30 on Wednesday, 30 on Thursday, 20 on Friday, 10 on Saturday, and 10 on Sunday?
b. If 200 overweight trucks are detected during a single week, what is the probability that at least 50 are detected on Monday?

4.49 *Repairing drill bits.* A sample of size $n$ is selected from a large lot of shear drill bits. Suppose that a proportion $p_1$ contains exactly one defect and a proportion $p_2$ contains more than one defect (with $p_1 + p_2 < 1$). The cost of replacing or repairing the defective drill bits is $C = 4Y_1 + Y_2$, where $Y_1$ denotes the number of bits with one defect and $Y_2$ denotes the number with two or more defects. Find the expected value of $C$.

4.50 *Electric current through a resistor.* An electrical current traveling through a resistor may take one of three different paths, with probabilities $p_1 = .25$, $p_2 = .30$, and $p_3 = .45$, respectively. Suppose we monitor the path taken in $n = 10$ consecutive trials.

a. Find the probability that the electrical current will travel the first path $Y_1 = 2$ times, the second path $Y_2 = 4$ times, and the third path $Y_3 = 4$ times.
b. Find $E(Y_2)$ and $V(Y_2)$. Interpret the results.

## Theoretical Exercise

4.51 For a multinomial distribution with $k = 3$ and $n = 2$, verify that

$$\sum_{y_1, y_2, y_3} p(y_1, y_2, y_3) = 1$$

[*Hint:* Use the binomial theorem (see Theoretical Exercise 4.38) to expand the sum $[a + (b + c)]^2$, then substitute the binomial expansion of $(b + c)^2$ in the resulting expression. Finally, substitute $a = p_1$, $b = p_2$, and $c = p_3$.]

| | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Proportion | .22 | .20 | .17 | .17 | .12 | .05 | .07 |

*Source:* Minnesota Department of Transportation.

# 4.8 The Negative Binomial and the Geometric Probability Distributions

Often we will be interested in measuring how long its takes before some event occurs—for example, the length of time a customer must wait in line until receiving service, or the number of assembly line items that are tested until one fails.

For the customer waiting time application, we view each unit of time as a Bernoulli trial that can result in a success ($S$) or a failure ($F$) and consider a series of trials identical to those described for the binomial experiment (Section 4.6). For the assembly line testing application, the tests are the Bernoulli trials. Unlike the binomial experiment where $Y$ is the total number of successes, the random variable of interest here is $Y$, the number of trials (or time units) until the $r$th success is observed.

The probability distribution for the random variable $Y$ is known as a **negative binomial distribution**. Its formula is given in the next box, together with the mean and variance for a negative binomial random variable.

### The Negative Binomial Probability Distribution

The probability distribution for a negative binomial random variable $Y$ is given by

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r} \qquad (y = r, r+1, r+2, \dots)$$

where

$p$ = Probability of success on a single Bernoulli trial
$q = 1 - p$
$y$ = Number of trials until the $r$th success is observed

The mean and variance of a negative binomial random variable are, respectively,

$$\mu = \frac{r}{p} \quad \text{and} \quad \sigma^2 = \frac{rq}{p^2}$$

From the box, you can see that the negative binomial probability distribution is a function of two parameters, $p$ and $r$. For the special case $r = 1$, the probability distribution of $Y$ is known as a **geometric probability distribution**.

### The Geometric Probability Distribution

$$p(y) = pq^{y-1} \quad (y = 1, 2, \ldots)$$

where

$Y$ = Number of trials until the first success is observed

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{q}{p^2}$$

To derive the negative binomial probability distribution, note that every simple event that results in $y$ trials until the $r$th success will contain $(y - r)$ $F$'s and $r$ $S$'s, as depicted here:

$$\overbrace{F\,F\,S\,F\,F\ldots S\,F}^{(y-r)F\text{'s and }(r-1)S\text{'s}} \quad \overbrace{\quad S \quad}^{r\text{th } S}$$

The number of different simple events that result in $(y - r)$ $F$'s before the $r$th $S$ is the number of ways that we can arrange the $(r - 1)$ $S$'s and $(y - r)$ $F$'s, namely,

$$\binom{(y-r)+(r-1)}{r-1} = \binom{y-1}{r-1}$$

Then, since the probability associated with each of these simple events is $p^r q^{y-r}$, we have

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r}$$

Examples 4.13 and 4.14 demonstrate the use of the negative binomial and the geometric probability distributions, respectively.

**Example 4.13**

Negative Binomial Application—Motor Assembly

To attach the housing on a motor, a production line assembler must use an electrical hand tool to set and tighten four bolts. Suppose that the probability of setting and tightening a bolt in any 1-second time interval is $p = .8$. If the assembler fails in the first second, the probability of success during the second 1-second interval is .8, and so on.

a. Find the probability distribution of $Y$, the number of 1-second time intervals until a complete housing is attached.
b. Find $p(6)$.
c. Find the mean and variance of $Y$.

Solution

a. Since the housing contains $r = 4$ bolts, we will use the formula for the negative binomial probability distribution. Substituting $p = .8$ and $r = 4$ into the formula for $p(y)$, we obtain

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r} = \binom{y-1}{3}(.8)^4(.2)^{y-4}$$

b. To find the probability that the complete assembly operation will require $Y = 6$ seconds, we substitute $y = 6$ into the formula obtained in part **a** and find

$$p(y) = \binom{5}{3}(.8)^4(.2)^2 = (10)(.4096)(.04) = .16384$$

c. For this negative binomial distribution,

$$\mu = \frac{r}{p} = \frac{4}{.8} = 5 \text{ seconds}$$

and

$$\sigma^2 = \frac{rq}{p^2} = \frac{4(.2)}{(.8)^2} = 1.25$$

**Example 4.14**

Geometric Application—Testing Fuses

A manufacturer uses electrical fuses in an electronic system. The fuses are purchased in large lots and tested sequentially until the first defective fuse is observed. Assume that the lot contains 10% defective fuses.

a. What is the probability that the first defective fuse will be one of the first five fuses tested?
b. Find the mean, variance, and standard deviation for $Y$, the number of fuses tested until the first defective fuse is observed.

Solution

a. The number $Y$ of fuses tested until the first defective fuse is observed is a geometric random variable with

$$\begin{aligned} p &= .1 \qquad \text{(probability that a single fuse is defective)} \\ q &= 1 - p = .9 \end{aligned}$$

and

$$\begin{aligned} p(y) &= pq^{y-1} \qquad (y = 1, 2, \ldots) \\ &= (.1)(.9)^{y-1} \end{aligned}$$

The probability that the first defective fuse is one of the first five fuses tested is

$$\begin{aligned} P(Y \le 5) &= p(1) + p(2) + \cdots + p(5) \\ &= (.1)(.9)^0 + (.1)(.9)^1 + \cdots + (.1)(.9)^4 = .41 \end{aligned}$$

b. The mean, variance, and standard deviation of this geometric random variable are

$$\mu = \frac{1}{p} = \frac{1}{.1} = 10$$

$$\sigma^2 = \frac{q}{p^2} = \frac{.9}{(.1)^2} = 90$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{90} = 9.49$$

## Applied Exercises

**4.52** *When to replace a maintenance system.* An article in the *Journal of Quality of Maintenance Engineering* (Vol. 19, 2013) studied the problem of finding the optimal replacement policy for a maintenance system. Consider a system that is tested every 12 hours. The test will determine whether there are any flaws in the system. Assume the probability of no flaw being detected is .85. If a flaw (failure) is detected, the system is repaired. Following the 5th failed test, the system is completely replaced. Now let $Y$ represent the number of tests until the system needs to be replaced.

a. Give the probability distribution for $Y$ as a formula. What is the name of this distribution?

b. Find the probability that the system needs to be replaced after 8 total tests.

**4.53** *Distribution of slugs.* The distributional pattern of pulmonate slugs inhabiting Libya was studied in the *AIUB Journal of Science and Engineering* (Aug. 2003). The number of slugs of a certain species found in the survey area was modeled using the negative binomial distribution. Assume that the probability of observing a slug of a certain species (say, *Milax rusticus*) in the survey area is .2. Let $Y$ represent the number of slugs that must be collected in order to obtain a sample of 10 *Milax rusticus* slugs.

a. Give the probability distribution for $Y$ as a formula.

b. What is the expected value of $Y$? Interpret this value.

c. Find $P(Y = 25)$.

**4.54** *Is a product "green"?* Refer to the *ImagePower Green Brands Survey* of international consumers, Exercise 3.4 (p. 84). Recall that a "green" product is one built from recycled materials that has minimal impact on the environment. The reasons why a consumer identifies a product as green are summarized in the next table. Consider interviewing consumers, at random. Let $Y$ represent the number of consumers who must be interviewed until one indicates something other than information given directly on the product's label or packaging as the reason a product is green.

a. Give a formula for the probability distribution of $Y$.

b. What is $E(Y)$? Interpret the result.

c. Find $P(Y = 1)$.

d. Find $P(Y > 2)$.

| Reason for saying a product is green | Percentage of consumers |
|---|---|
| Certification mark on label | 45 |
| Packaging | 15 |
| Reading information about the product | 12 |
| Advertisement | 6 |
| Brand website | 4 |
| Other | 18 |
| TOTAL | 100 |

*Source:* 2011 ImagePower Green Brands Survey

**4.55** *Chemical signals of mice.* Refer to the *Cell* (May 14, 2010) study of the ability of a mouse to recognize the odor of a potential predator, Exercise 4.27 (p. 153). Recall that 40% of lab mice cells exposed to major urinary proteins (Mups) chemically produced from a cat responded positively (i.e., recognized the danger of the lurking predator). Consider testing lab mice cells, each exposed to chemically produced cat Mups. Let $Y$ represent the number of cells that must be tested until one responds positively.

a. What is the name of the probability distribution for $Y$? Give the formula for $p(y)$.

b. Find $E(Y)$ and interpret its value, practically.

c. Find the variance of $Y$.

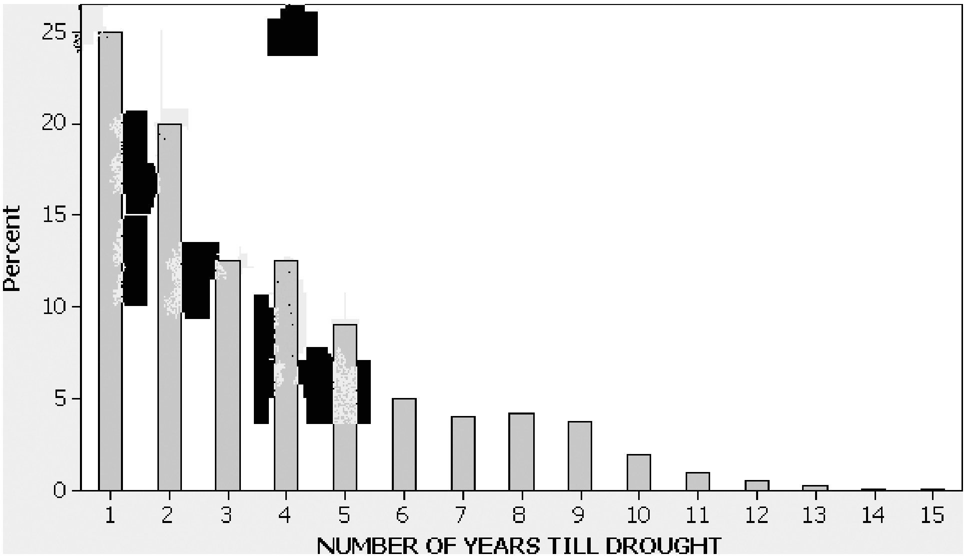
d. Give an interval that is likely to contain the value of $Y$.

**4.56** *Fingerprint expertise.* Refer to the *Psychological Science* (August, 2011) study of fingerprint identification, Exercise 4.32 (p. 153). Recall that the study found that when presented with prints from the same individual, a novice will correctly identify the match 75% of the time. Consider a novice presented with different pairs of fingerprints, one pair at a time, where each pair is a match. How many fingerprint pairs must be presented before the novice correctly identifies five pairs?

**4.57** *Drought recurrence in Texas.* A drought is a period of abnormal dry weather that causes a serious hydrologic imbalance in an area. The Palmer Drought Index (PDI) is designed to quantitatively measure the severity of a drought. A PDI value of −1 or less indicates a dry (drought) period. A PDI value greater than −1 indicates a wet (non-drought) period. Civil engineers at the University of Arizona used paleontology data and historical records to determine PDI values for each of the past 400 years in

**MINITAB output for Exercise 4.57**

*Source:* Gonzalez, J., and Valdes, J. B. "Bivariate drought recurrence analysis using tree ring reconstructions." *Journal of Hydrologic Engineering*, Vol. 8, No. 5, Sept./Oct. 2003 (Figure 5).




Texas (*Journal of Hydrologic Engineering*, Sept./Oct. 2003). The researchers discovered that the number $Y$ of years that need to be sampled until a dry (drought) year is observed follows an approximate geometric distribution. A graph of the distribution is shown at the top of the page.

a. From the graph, estimate $E(Y)$. (*Hint*: Use the formula in Section 4.3.)
b. Use the result, part **a**, to estimate the value of $p$ for the geometric distribution.
c. Estimate the probability that 7 years must be sampled in order to observe a drought year.

4.58 *Particles emitted from fusion reaction.* The carbon-nitrogen-oxygen (CNO) cycle is one type of fusion reaction by which stars convert hydrogen to helium. The distribution of high-energy charges resulting from proton–CNO interaction in space was investigated in the *Journal of Physics G: Nuclear and Particle Physics* (Nov. 1996). When a high-energy interaction occurs, charged particles are emitted. These particles are classified as shower or heavy particles. The number $Y$ of charged particles that must be observed in order to detect $r$ charged shower particles was shown to follow a negative binomial distribution with $p = .75$. Use this information to find the probability that five charged particles must be observed in order to detect three charged shower particles.

4.59 *Space shuttle failures.* Prior to the grounding of space shuttle program, the National Aeronautics and Space Administration (NASA) estimated that the chance of a "critical-item" failure within a space shuttle's main engine was approximately 1 in 63. The failure of a critical item during flight would lead directly to a shuttle catastrophe.

a. On average, how many shuttle missions would fly before a critical-item failure occurs?
b. What is the standard deviation of the number of missions before a critical-item failure occurs?
c. Give an interval that will capture the number of missions before a critical-item failure occurs with probability of approximately .95.

4.60 *Reflection of neutron particles.* Refer to the *Nuclear Science and Engineering* study, Exercise 4.36 (p. 154). If neutral particles are released one at a time into the evacuated duct, find the probability that more than five particles will need to be released until we observe two particles reflected by the inner duct wall.

4.61 *Probability of striking oil.* Assume that hitting oil at one drilling location is independent of another, and that, in a particular region, the probability of success at an individual location is .3.

a. What is the probability that a driller will hit oil on or before the third drilling?
b. If $Y$ is the number of drillings until the first success occurs, find the mean and standard deviation of $Y$.
c. Is it likely that $Y$ will exceed 10? Explain.
d. Suppose the drilling company believes that a venture will be profitable if the number of wells drilled until the second success occurs is less than or equal to 7. Find the probability that the venture will be successful.

## Theoretical Exercise

4.62 Let $Y$ be a negative binomial random variable with parameters $r$ and $p$. Then it can be shown that $W = Y - r$ is also a negative binomial random variable, where $W$ represents the number of failures before the $r$th success is observed. Use the facts that

$$E(Y) = \frac{r}{p} \quad \text{and} \quad \sigma_y^2 = \frac{rq}{p^2}$$

to show that

$$E(W) = \frac{rq}{p} \quad \text{and} \quad \sigma_w^2 = \frac{rq}{p^2}$$

(*Hint:* Use Theorems 4.1, 4.2, and 4.3.)

## 4.9 The Hypergeometric Probability Distribution

When we are sampling from a finite population of Successes and Failures (such as a finite population of consumer preference responses or a finite collection of observations in a shipment containing nondefective and defective manufactured products), the assumptions for a binomial experiment are satisfied exactly only if the result of each trial is observed and then replaced in the population before the next observation is made. This method of sampling is called **sampling with replacement**. However, in practice, we usually **sample without replacement**, i.e., we randomly select $n$ different elements from among the $N$ elements in the population. As noted in Section 4.6, when $N$ is large and $n/N$ is small (say, less than .05), the probability of drawing an $S$ remains approximately the same from one trial to another, the trials are (essentially) independent, and the probability distribution for the number of successes, $Y$, is *approximately* a binomial probability distribution. However, when $N$ is small or $n/N$ is large (say, greater than .05), we would want to use the exact probability distribution for $y$. This distribution, known as a **hypergeometric probability distribution**, is the topic of this section. The defining characteristics and probability distribution for a hypergeometric random variable are stated in the boxes.

### Characteristics That Define a Hypergeometric Random Variable

1. The experiment consists of randomly drawing $n$ elements without replacement from a set of $N$ elements, $r$ of which are $S$'s (for Success) and $(N - r)$ of which are $F$'s (for Failure).
2. The sample size $n$ is large relative to the number $N$ of elements in the population, i.e., $n/N > .05$.
3. The hypergeometric random variable $Y$ is the number of $S$'s in the draw of $n$ elements.

### The Hypergeometric Probability Distribution

The hypergeometric probability distribution is given by

$$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, \quad y = \begin{array}{l}\text{Maximum } [0, n - (N - r)], \ldots, \\ \text{Minimum } (r, n)\end{array}$$

where

$N =$ Total number of elements
$r =$ Number of S′s in the $N$ elements
$n =$ Number of elements drawn
$y =$ Number of S′s drawn in the $n$ elements

The mean and variance of a hypergeometric random variable are, respectively,

$$\mu = \frac{nr}{N} \qquad \sigma^2 = \frac{r(N-r)n(N-n)}{N^2(N-1)}$$

To derive the hypergeometric probability distribution, we first note that the total number of simple events in $S$ is equal to the number of ways of selecting $n$ elements from $N$, namely, $\binom{N}{n}$. A *simple event implying y successes* will be a selection of $n$ elements in which $y$ are $S$'s and $(n - y)$ are $F$'s. Since there are $r$ $S$'s from which to choose, the number of different ways of selecting $y$ of them is $a = \binom{r}{y}$. Similarly, the number of ways of selecting $(n - y)$ $F$'s from among the total of $(N - r)$ is $b = \binom{N-r}{n-y}$. We now apply Theorem 3.1 to determine the number of ways of selecting $y$ $S$'s and $(n - y)$ $F$'s—that is, the number of simple events implying $y$ successes:

$$a \cdot b = \binom{r}{y}\binom{N-r}{n-y}$$

Finally, since the selection of any one set of $n$ elements is as likely as any other, all the simple events are equiprobable and thus,

$$p(y) = \frac{\text{Number of simple events that imply } y \text{ successes}}{\text{Number of simple events}} = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

**Example 4.15**

**Hypergeometric Application—EDA Catalyst Selection**

An experiment is conducted to select a suitable catalyst for the commercial production of ethylenediamine (EDA), a product used in soaps. Suppose a chemical engineer randomly selects 3 catalysts for testing from among a group of 10 catalysts, 6 of which have low acidity and 4 of which have high acidity.

a. Find the probability that no highly acidic catalyst is selected.
b. Find the probability that exactly one highly acidic catalyst is selected.

**Solution**

Let $Y$ be the number of highly acidic catalysts selected. Then $Y$ is a hypergeometric random variable with $N = 10$, $n = 3$, $r = 4$, and

$$P(Y = y) = p(y) = \frac{\binom{4}{y}\binom{6}{3-y}}{\binom{10}{3}}$$

a. $$p(0) = \frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}} = \frac{(1)(20)}{120} = \frac{1}{6}$$

b. $$p(1) = \frac{\binom{4}{1}\binom{6}{2}}{\binom{10}{3}} = \frac{(4)(15)}{120} = \frac{1}{2}$$

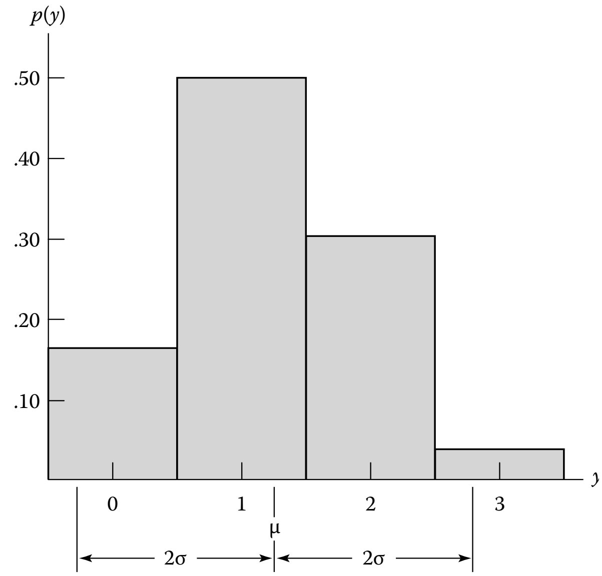
**Example 4.16**

**Hypergeometric Application—EDA Catalyst Selection Continued**

Refer to the EDA experiment, Example 4.15.

a. Find $\mu$, $\sigma^2$, and $\sigma$ for the random variable $Y$.
b. Find $P(\mu - 2\sigma < Y < \mu + 2\sigma)$. How does this result compare to the Empirical Rule?

**FIGURE 4.7**
Probability distribution for $y$ in Example 4.16




Solution

a. Since $Y$ is a hypergeometric random variable with $N = 10$, $n = 3$, and $r = 4$, the mean and variance are

$$\mu = \frac{nr}{N} = \frac{(3)(4)}{10} = 1.2$$

$$\sigma^2 = \frac{r(N - r)n(N - n)}{N^2(N - 1)} = \frac{4(10 - 4)3(10 - 3)}{(10)^2(10 - 1)}$$

$$= \frac{(4)(6)(3)(7)}{(100)(9)} = .56$$

The standard deviation is

$$\sigma = \sqrt{.56} = .75$$

b. The probability distribution and the interval $\mu \pm 2\sigma$, or $-.3$ to 2.7, are shown in Figure 4.7 (p. 166). The only possible value of $Y$ that falls outside the interval is $Y = 3$. Therefore,

$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = 1 - p(3) = 1 - \frac{\binom{4}{3}\binom{6}{0}}{\binom{10}{3}}$$

$$= 1 - \frac{4}{120} = .967$$

According to the Empirical Rule, we expect about 95% of the observed $Y$'s to fall in this interval. Thus, the Empirical Rule provides a good estimate of this probability.

**Example 4.17**

**Deriving the Mean of the Hypergeometric Distribution.**

Refer to Example 4.15. Find the mean, $\mu$, of the random variable $Y$ using Definition 4.4.

**Solution**

By Definition 4.4,

$$\mu = E(Y) = \sum_{\text{all } y} yp(y) = \sum_{y=0}^{3} y \frac{\binom{4}{y}\binom{6}{3-y}}{120}$$

Using the values of $p(y)$ calculated in Examples 4.15 and 4.16, and

$$p(2) = \frac{\binom{4}{2}\binom{6}{1}}{120} = \frac{(6)(6)}{120} = \frac{3}{10}$$

we obtain by substitution:

$$\begin{aligned}\mu &= 0p(0) + 1p(1) + 2p(2) + 3p(3)\\ &= 0 + 1\left(\frac{1}{2}\right) + 2\left(\frac{3}{10}\right) + 3\left(\frac{1}{30}\right) = 1.2\end{aligned}$$

Note that this is the value we obtained in Example 4.16 by applying the formula given in the previous box.

## Applied Exercises

4.63 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study of the trend in the design of social robots, Exercise 3.1 (p. 84). The study found that of 106 social robots, 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels. Suppose you randomly select 10 of the 106 social robots and count the number, $Y$, with neither legs nor wheels.

a. Demonstrate why the probability distribution for $Y$ should not be approximated by the binomial distribution.
b. Show that the properties of the hypergeometric probability distribution are satisfied for this experiment.
c. Find $\mu$ and $\sigma$ for the probability distribution for $Y$.
d. Calculate the probability that $Y = 2$.

4.64 *Mailrooms contaminated with anthrax.* During autumn 2001, there was a highly publicized outbreak of anthrax cases among U.S. Postal Service workers. In *Chance* (Spring 2002), research statisticians discussed the problem of sampling mailrooms for the presence of anthrax spores. Let $Y$ equal the number of mailrooms contaminated with anthrax spores in a random sample of $n$ mailrooms selected from a population of $N$ mailrooms. The researchers showed that $Y$ has a hypergeometric probability distribution. Let $r$ equal the number of contaminated mailrooms in the population. Suppose $N = 100$, $n = 3$, and $r = 20$.

a. Find $p(0)$.
b. Find $p(1)$.
c. Find $p(2)$.
d. Find $p(3)$.

4.65 *Establishing boundaries in academic engineering.* How academic engineers establish boundaries (e.g., differentiating between engineers and other scientists, defining the different disciplines within engineering, determining the quality of journal publications) was investigated in *Engineering Studies* (Aug., 2012). Participants were 10 tenured or tenure-earning engineering faculty members in a School of Engineering at a large research-oriented university. The table gives a breakdown of the department affiliation of the engineers. Each participant was interviewed at length and responses were used to establish boundaries. Suppose we randomly select 3 participants from the original 10 to form a university committee charged with developing boundary guidelines. Let $Y$ represent the number of committee members who are from the Department of Engineering Physics. Identify the probability distribution for $Y$ and give a formula for the distribution.

| Department | Number of Participants |
|---|---|
| Chemical Engineering | 1 |
| Civil Engineering | 2 |
| Engineering Physics | 4 |
| Mechanical Engineering | 2 |
| Industrial Engineering | 1 |

4.66 *On-site disposal of hazardous waste.* The Resource Conservation and Recovery Act mandates the tracking and disposal of hazardous waste produced at U.S. facilities. *Professional Geographer* (Feb. 2000) reported the hazardous waste generation and disposal characteristics of 209 facilities. Only eight of these facilities treated hazardous waste on-site.

a. In a random sample of 10 of the 209 facilities, what is the expected number in the sample that treats hazardous waste on-site? Interpret this result.

b. Find the probability that 4 of the 10 selected facilities treat hazardous waste on-site.

4.67 *Contaminated gun cartridges.* Refer to the investigation of contaminated gun cartridges at a weapons manufacturer, Exercise 4.5 (p. 139). In a sample of 158 cartridges from a certain lot, 36 were found to be contaminated, and 122 were "clean." If you randomly select 5 of these 158 cartridges, what is the probability that all 5 will be "clean"?

4.68 *Lot inspection sampling.* Imagine you are purchasing small lots of a manufactured product. If it is very costly to test a single item, it may be desirable to test a sample of items from the lot instead of testing every item in the lot. Suppose each lot contains 10 items. You decide to sample 4 items per lot and reject the lot if you observe 1 or more defective.

a. If the lot contains 1 defective item, what is the probability that you will accept the lot?

b. What is the probability that you will accept the lot if it contains 2 defective items?

**NZBIRDS**

4.69 *Extinct New Zealand birds.* Refer to the *Evolutionary Ecology Research* (July 2003) study of the patterns of extinction in the New Zealand bird population, Exercise 3.28 (p. 95). Of the 132 bird species saved in the **NZBIRDS** file, 38 are extinct. Suppose you randomly select 10 of the 132 bird species (without replacement) and record the extinct status of each.

a. What is the probability that exactly 5 of the 10 species you select are extinct?

b. What is the probability that at most 1 species is extinct?

4.70 *Cell phone handoff behavior.* Refer to the *Journal of Engineering, Computing and Architecture* (Vol. 3., 2009) study of cell phone handoff behavior, Exercise 3.15 (p. 91). Recall that a "handoff" describes the process of a cell phone moving from one base channel (identified by a color code) to another. During a particular driving trip a cell phone changed channels (color codes) 85 times. Color code "b" was accessed 40 times on the trip. You randomly select 7 of the 85 handoffs. How likely is it that the cell phone accesses color code "b" only twice for these 7 handoffs?

4.71 *Reverse cocaine sting.* An article in *The American Statistician* (May 1991) described the use of probability in a reverse cocaine sting. Police in a midsize Florida city seized 496 foil packets in a cocaine bust. To convict the drug traffickers, police had to prove that the packets contained genuine cocaine. Consequently, the police lab randomly selected and chemically tested 4 of the packets; all 4 tested positive for cocaine. This result led to a conviction of the traffickers.

a. Of the 496 foil packets confiscated, suppose 331 contain genuine cocaine and 165 contain an inert (legal) powder. Find the probability that 4 randomly selected packets will test positive for cocaine.

b. Police used the 492 remaining foil packets (i.e., those not tested) in a reverse sting operation. Two of the 492 packets were randomly selected and sold by undercover officers to a buyer. Between the sale and the arrest, however, the buyer disposed of the evidence. Given that 4 of the original 496 packets tested positive for cocaine, what is the probability that the 2 packets sold in the reverse sting did not contain cocaine? Assume the information provided in part **a** is correct.

c. *The American Statistician* article demonstrates that the conditional probability, part **b**, is maximized when the original 496 packets consist of 331 packets containing genuine cocaine and 165 containing inert powder. Recalculate the probability, part **b**, assuming that 400 of the original 496 packets contain cocaine.

## Theoretical Exercise

4.72 Show that the mean of a hypergeometric random variable $Y$ is $\mu = nr/N$. [*Hint:* Show that

$$\frac{y\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \frac{\frac{nr}{N}\binom{r-1}{y-1}\binom{N-1-(r-1)}{n-1-(y-1)}}{\binom{N-1}{n-1}}$$

and then use the fact that

$$\frac{\binom{r-1}{y-1}\binom{N-1-(r-1)}{n-1-(y-1)}}{\binom{N-1}{n-1}}$$

is the hypergeometric probability distribution for $Z = (Y-1)$, where $Z$ is the number of $S$'s in $(n-1)$ trials, with a total of $(r-1)$ $S$'s in $(N-1)$ elements.]

# 4.10 The Poisson Probability Distribution

The **Poisson probability distribution**, named for the French mathematician S. D. Poisson (1781–1840), provides a model for the relative frequency of the number of "rare events" that occur in a unit of time, area, volume, etc. The number of new jobs submitted to a computer in any one minute, the number of fatal accidents per month in a manufacturing plant, and the number of visible defects in a diamond are variables whose relative frequency distributions can be approximated well by Poisson probability distributions. The characteristics of a Poisson random variable are listed in the box.

### Characteristics of a Poisson Random Variable

1. The experiment consists of counting the number of times $Y$ a particular (rare) event occurs during a given unit of time or in a given area or volume (or weight, distance, or any other unit of measurement).
2. The probability that an event occurs in a given unit of time, area, or volume is the same for all the units. Also, units are mutually exclusive.
3. The number of events that occur in one unit of time, area, or volume is independent of the number that occur in other units.

The formulas for the probability distribution, the mean, and the variance of a Poisson random variable are shown in the next box. You will note that the formula involves the quantity $e = 2.71828\ldots$, the base of natural logarithms. Values of $e^{-y}$, needed to compute values of $p(y)$, are given in Table 3 of Appendix B.

### The Poisson Probability Distribution

The probability distribution* for a Poisson random variable $Y$ is given by

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad (y = 0, 1, 2, \ldots)$$

where

$\lambda =$ Mean number of events during a given unit of time, area, or volume

$e = 2.71828\ldots$

The mean and variance of a Poisson random variable are, respectively,

$$\mu = \lambda \quad \text{and} \quad \sigma^2 = \lambda$$

The shape of the Poisson distribution changes as its mean $\lambda$ changes. This fact is illustrated in Figure 4.8, which shows relative frequency histograms for a Poisson distribution with $\lambda = 1, 2, 3$, and 4.

**Example 4.18**

Poisson Application—Cracks in Concrete

Suppose the number $Y$ of cracks per concrete specimen for a particular type of cement mix has approximately a Poisson probability distribution. Furthermore, assume that the average number of cracks per specimen is 2.5.

a. Find the mean and standard deviation of $Y$, the number of cracks per concrete specimen.

b. Find the probability that a randomly selected concrete specimen has exactly five cracks.

c. Find the probability that a randomly selected concrete specimen has two or more cracks.

d. Find $P(\mu - 2\sigma < Y < \mu + 2\sigma)$. Does the result agree with the Empirical Rule?

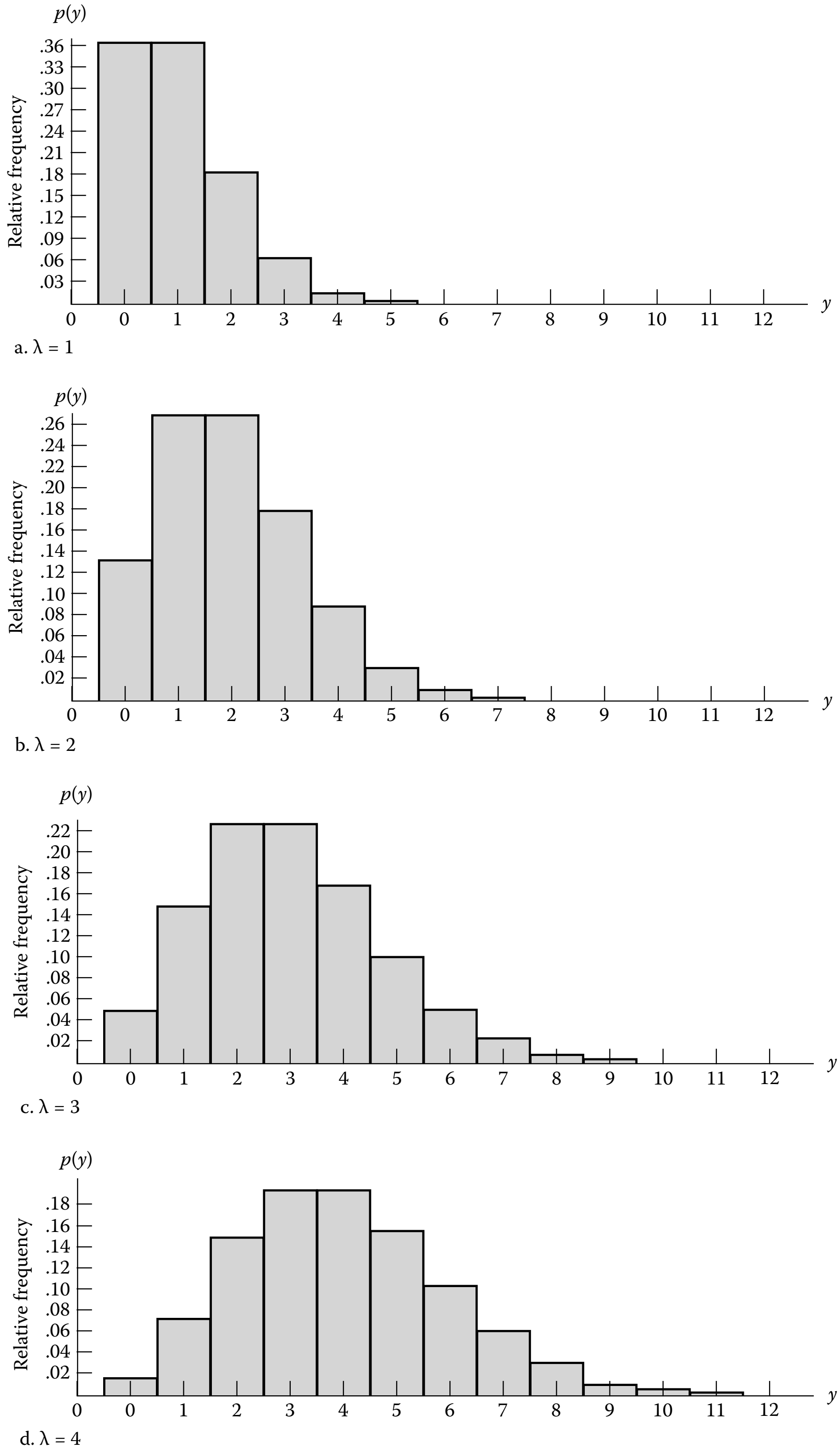
Solution

a. The mean and variance of a Poisson random variable are both equal to $\lambda$. Thus, for this example

$$\mu = \lambda = 2.5 \qquad \sigma^2 = \lambda = 2.5$$

*The derivation of this probability distribution is beyond the scope of this course. See *Mathematical Statistics with Application*, 7th ed., Wackerly, Mendenhall, and Scheaffer for a proof.

**FIGURE 4.8**
Histograms for the Poisson distribution for $\lambda = 1, 2, 3$, and 4



Then the standard deviation is

$$\sigma = \sqrt{2.5} = 1.58$$

b. We want the probability that a concrete specimen has exactly five cracks. The probability distribution for $Y$ is

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Then, since $\lambda = 2.5$, $Y = 5$, and $e^{-2.5} = .082085$ (from Table 3 of Appendix B),

$$p(5) = \frac{(2.5)^5 e^{-2.5}}{5!} = \frac{(2.5)^5(.082085)}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = .067$$

c. To find the probability that a concrete specimen has two or more cracks, we need to find

$$P(Y \geq 2) = p(2) + p(3) + p(4) + \cdots = \sum_{y=2}^{\infty} p(y)$$

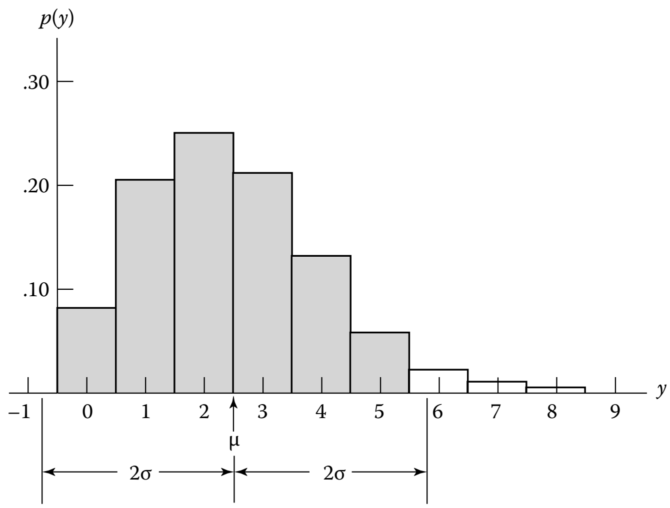
To find the probability of this event, we must consider the complementary event. Thus,

$$\begin{aligned} P(Y \geq 2) &= 1 - P(Y \leq 1) = 1 - [p(0) + p(1)] \\ &= 1 - \frac{(2.5)^0 e^{-2.5}}{0!} - \frac{(2.5)^1 e^{-2.5}}{1!} \\ &= 1 - \frac{1(.082085)}{1} - \frac{2.5(.082085)}{1} \\ &= 1 - .287 = .713 \end{aligned}$$

According to our Poisson model, the probability that a concrete specimen has two or more cracks is .713.

d. The probability distribution for $Y$ is shown in Figure 4.9 for $Y$ values between 0 and 9. The mean $\mu = 2.5$ and the interval $\mu \pm 2\sigma$, or $-.7$ to 5.7, are indicated.

**FIGURE 4.9**
Poisson probability distribution for $Y$ in Example 4.18



Consequently, $P(\mu - 2\sigma < Y < \mu + 2\sigma) = P(Y \leq 5)$. This probability is shaded in Figure 4.8.

The probabilities $p(0), p(1), \ldots, p(5)$ can be calculated and summed as in part **c**. However, we will use a table of cumulative Poisson probabilities to obtain the sum. Table 4 of Appendix B gives the partial sum, $\Sigma_{y=0}^{k} p(y)$, for different values of the Poisson mean $\lambda$. For $\lambda = 2.5$, the $\Sigma_{y=0}^{k} p(y) = p(0) + p(1) + \cdots + p(5)$ is given as .9581. Thus, $P(Y \leq 5) = .9581$; note that this probability agrees with the Empirical Rule's approximation of .95.

The Poisson probability distribution is related to and can be used to approximate a binomial probability distribution when $n$ is large and $\mu = np$ is small, say, $np \leq 7$. The proof of this fact is beyond the scope of this text, but it can be found in Feller (1968).

**Example 4.19**

Poisson Approximation to Binomial

Let $Y$ be a binomial random variable with $n = 25$ and $p = .1$.

a. Use Table 2 of Appendix B to determine the exact value of $P(Y \leq 1)$.
b. Find the Poisson approximation to $P(Y \leq 1)$. *(Note:* Although we would prefer to compare the Poisson approximation to binomial probabilities for larger values of $n$, we are restricted in this example by the limitations of Table 2.)

Solution

a. From Table 2 of Appendix B, with $n = 25$ and $p = .1$, we have

$$P(Y \leq 1) = \sum_{y=0}^{1} p(y) = .271$$

b. Since $n = 25$ and $p = .1$, we will approximate $p(y)$ using a Poisson probability distribution with mean

$$\lambda = np = (25)(.1) = 2.5$$

Locating $\lambda = 2.5$ in Table 4 of Appendix B, we obtain the partial sum

$$P(Y \leq 1) = \sum_{y=0}^{1} p(y) = .2873$$

This approximation, .2873, to the exact value of $P(Y \leq 1) = .271$ is reasonably good considering that the approximation procedure is usually applied to binomial probability distributions for which $n$ is much larger than 25.

**Example 4.20**

Derivation of Mean of Poisson Distribution

Show that the expected value of a Poisson random variable $Y$ is $\lambda$.

Solution

By Definition 4.4, we have

$$E(Y) = \sum_{\text{all } y} yp(y) = \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!}$$

The first term of this series will equal 0, because $y = 0$. Therefore,

$$E(Y) = \sum_{y=0}^{\infty} \frac{y\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y-1)!} = \sum_{y=1}^{\infty} \frac{\lambda \cdot \lambda^{y-1} e^{-\lambda}}{(y-1)!}$$

Factoring the constant $\lambda$ outside the summation and letting $Z = (Y - 1)$, we obtain

$$E(Y) = \lambda \sum_{z=0}^{\infty} \frac{\lambda^z e^{-\lambda}}{z!} = \lambda \sum_{z=0}^{\infty} p(z)$$

where $Z$ is a Poisson random variable with mean $\lambda$. Hence,

$$E(Y) = \lambda \sum_{z=0}^{\infty} p(z) = \lambda(1) = \lambda$$

## Applied Exercises

4.73 *Traffic fatalities and sporting events.* The relationship between close sporting events and game-day traffic fatalities was investigated in the *Journal of Consumer Research* (December, 2011). Transportation engineers found that closer football and basketball games are associated with more traffic fatalities. The methodology used by the researchers involved modeling the traffic fatality count for a particular game as a Poisson random variable. For games played at the winner's location (home court or home field), the mean number of traffic fatalities was .5. Use this information to find the probability that at least 3 game-day traffic fatalities will occur at the winning team's location.

4.74 *Rare planet transits.* A "planet transit" is a rare celestial event in which a planet appears to cross in front of its star as seen from Earth. The planet transit causes a noticeable dip in the star's brightness, allowing scientists to detect a new planet even though it is not directly visible. The National Aeronautics and Space Administration (NASA) recently launched its Kepler mission, designed to discover new planets in the Milky Way by detecting extrasolar planet transits. After one year of the mission in which 3,000 stars were monitored, NASA announced that 5 planet transits were detected. (NASA, American Astronomical Society, Jan. 4, 2010.) Assume that the number of planet transits discovered for every 3,000 stars follows a Poisson distribution with $\lambda=5$. What is the probability that, in the next 3,000 stars monitored by the Kepler mission, more than 10 planet transits will be seen?

4.75 *Flaws in plastic coated wire.* The British Columbia Institute of Technology provides on its website (*www.math.bcit.ca*) practical applications of statistics to mechanical engineering. The following is a Poisson application. A roll of plastic-coated wire has an average of .8 flaws per 4-meter length of wire. Suppose a quality control engineer will sample a 4-meter length of wire from a roll of wire 220 meters in length. If no flaws are found in the sample, the engineer will accept the entire roll of wire. What is the probability that the roll will be rejected? What assumption did you make to find this probability?

4.76 *Non-home-based trips.* In the *Journal of Transportation Engineering* (June 2005), the number of non-home-based trips per day taken by drivers in Korea was modeled using the Poisson distribution with $\lambda = 1.15$.

a. What is the probability that a randomly selected Korean driver will take no more than two non-home-based trips per day?
b. Find the variance of the number of non-home-based trips per day taken by a driver in Korea.
c. Use the information in part **b** to find a value for the number of non-home-based trips that the driver is not likely to exceed.

4.77 *Airline fatalities.* U.S. airlines average about 1.6 fatalities per month. (*Statistical Abstract of the United States: 2010.*) Assume the probability distribution for $Y$, the number of fatalities per month, can be approximated by a Poisson probability distribution.

a. What is the probability that no fatalities will occur during any given month?
b. What is the probability that one fatality will occur during any given month?
c. Find $E(Y)$ and the standard deviation of $Y$.

4.78 *Deep-draft vessel casualties.* Engineers at the University of New Mexico modeled the number of casualties (deaths or missing persons) experienced by a deep-draft U.S. flag vessel over a 3-year period as a Poisson random variable, $Y$. The researchers estimated $E(Y)$ to be .03. (*Management Science*, Jan. 1999.)

a. Find the variance of $Y$.
b. Discuss the conditions that would make the researchers' Poisson assumption plausible.
c. What is the probability that a deep-draft U.S. flag vessel will have no casualties in a 3-year time period?

4.79 *Vinyl chloride emissions.* The Environmental Protection Agency (EPA) limits the amount of vinyl chloride in plant air emissions to no more than 10 parts per million. Suppose the mean emission of vinyl chloride for a particular plant is 4 parts per million. Assume that the number of parts per million of vinyl chloride in air samples, $Y$, follows a Poisson probability distribution.

a. What is the standard deviation of $Y$ for the plant?
b. Is it likely that a sample of air from the plant would yield a value of $Y$ that would exceed the EPA limit? Explain.
c. Discuss conditions that would make the Poisson assumption plausible.

4.80 *Noise in laser imaging.* Penumbral imaging is a technique used by nuclear engineers for imaging objects (e.g., X-rays and lasers) that emit high-energy photons. In *IEICE Transactions on Information & Systems* (Apr. 2005), researchers demonstrated that penumbral images are always degraded by noise, where the number $Y$ of noise events occurring in a unit of time follows a Poisson process with mean $\lambda$. The signal-to-noise ratio (SNR) for a penumbral image is defined as $SNR = \mu/\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation, respectively, of the noise process. Show that the SNR for $Y$ is $\sqrt{\lambda}$.

4.81 *Ambient air quality.* The Environmental Protection Agency (EPA) has established national ambient air quality standards in an effort to control air pollution. Currently, the EPA limit on ozone levels in air is 12 parts per hundred million (pphm). A study examined the long-term trend in daily ozone levels in Houston, Texas.* One of the variables of interest is $Y$, the number of days in a year on which the ozone level exceeds the EPA 12 pphm threshold. The mean number of exceedances in a year is estimated to be 18. Assume that the probability distribution for $Y$ can be modeled with the Poisson distribution.

a. Compute $P(Y \leq 20)$.
b. Compute $P(5 \leq Y \leq 10)$.
c. Estimate the standard deviation of $Y$. Within what range would you expect $Y$ to fall in a given year?
d. The study revealed a decreasing trend in the number of exceedances of the EPA threshold level over the past several years. The observed values of $Y$ for the past 6 years were 24, 22, 20, 15, 14, and 16. Explain why this trend casts doubt on the validity of the Poisson distribution as a model for $Y$. (*Hint:* Consider characteristic #3 of the Poisson random variable.)

4.82 *Unplanned nuclear scrams.* The nuclear industry has made a concerted effort to significantly reduce the number of unplanned rapid emergency shutdowns of a nuclear reactor—called *scrams*. A decade ago, the mean annual number of unplanned scrams at U.S. nuclear reactor units was four (see Exercise 2.81, p. 72). Assume that the annual number of unplanned scrams that occur at a nuclear reactor unit follows, approximately, a Poisson distribution.

a. If the mean has not changed, compute the probability that a nuclear reactor unit will experience 10 or more unplanned scrams this year.
b. Suppose a randomly selected nuclear reactor actually experiences 10 or more unplanned scrams this year. What can you infer about the true mean annual number of unplanned scrams? Explain.

4.83 *Elevator passenger arrivals.* A study of the arrival process of people using elevators at a multi-level office building was conducted and the results reported in *Building Services Engineering Research and Technology* (Oct., 2012). Suppose that at one particular time of day, elevator passengers arrive in batches of size 1 or 2 (i.e., either 1 or 2 people arriving at the same time to use the elevator). The researchers assumed that the number of batches, $N$, arriving over a specific time period follows a Poisson process with mean $\lambda = 1.1$. Now let $X_N$ represent the number of passengers (either 1 or 2) in batch $N$ and assume the batch size has a Bernoulli distribution with $p = P(X_N = 1) = .4$ and $q = P(X_N = 2) = .6$. Then, the total number of passengers arriving over a specific time period is $Y = \sum_{i=1}^{N} = X_i$. The researchers showed that if $X_1, X_2, \ldots, X_N$ are independent and identically distributed random variables and also independent of $N$, then $Y$ follows a compound Poisson distribution.

a. Find $P(Y = 0)$, i.e., the probability of no arrivals during the time period. [*Hint:* $Y = 0$ only when $N = 0$.]
b. Find $P(Y = 1)$, i.e., the probability of only 1 arrival during the time period. [*Hint:* $Y = 1$ only when $N = 1$ and $X_1 = 1$.]
c. Find $P(Y = 2)$, i.e., the probability of 2 arrivals during the time period. [*Hint:* $Y = 2$ when $N = 1$ and $X_1 = 2$, or, when $N = 2$ and $X_1 + X_2 = 2$. Also, use the fact that the sum of Bernoulli random variables is a binomial random variable.]
d. Find $P(Y = 3)$, i.e., the probability of 3 arrivals during the time period. [*Hint:* $Y = 3$ when $N = 2$ and $X_1 + X_2 = 3$, or, when $N = 3$ and $X_1 + X_2 + X_3 = 3$.]

## Theoretical Exercises

4.84 Show that for a Poisson random variable $Y$,

a. $0 \leq p(y) \leq 1$
b. $\sum_{y=0}^{\infty} p(y) = 1$
c. $E(Y^2) = \lambda^2 + \lambda$

[*Hint:* First derive the result $E[Y(Y-1)] = \lambda^2$ from the fact that

$$E[Y(Y-1)] = \sum_{y=0}^{\infty} y(y-1)\frac{\lambda^y e^{-\lambda}}{y!}$$
$$= \lambda^2 \sum_{y=2}^{\infty} \frac{\lambda^{y-2}e^{-\lambda}}{(y-2)!} = \lambda^2 \sum_{z=0}^{\infty} \frac{\lambda^z e^{-\lambda}}{z!}$$

Then apply the result $E[Y(Y-1)] = E(Y)^2 - E(Y)$.]

4.85 Show that for a Poisson random variable $Y$, $\sigma^2 = \lambda$. (*Hint:* Use the result of Exercise 4.83 and Theorem 4.4.)

*Shively, Thomas S. "An analysis of the trend in ozone using nonhomogeneous Poisson processes." Paper presented at annual meeting of the American Statistical Association, Anaheim, Calif., Aug. 1990.

## 4.11 Moments and Moment Generating Functions (*Optional*)

The **moments** of a random variable can be used to completely describe its probability distribution.

**Definition 4.7**

The ***k*th moment** of a random variable *Y*, **taken about the origin**, is denoted by the symbol $\mu'_k$ and defined to be

$$\mu'_k = E(Y^k) \qquad (k = 1, 2, \dots)$$

**Definition 4.8**

The ***k*th moment** of a random variable *Y*, **taken about its mean**, is denoted by the symbol $\mu_k$ and defined to be

$$\mu_k = E[(Y - \mu)^k]$$

You have already encountered two important moments of random variables. The mean of a random variable is $\mu'_1 = \mu$ and the variance is $\mu_2 = \sigma^2$. Other moments about the origin or about the mean can be used to measure the lack of symmetry or the tendency of a distribution to possess a large peak near the center. In fact, if all of the moments of a discrete random variable exist, they completely define its probability distribution. This fact is often used to prove that two random variables possess the same probability distributions. For example, if two discrete random variables, $X$ and $Y$, possess moments about the origin, $\mu'_{1x}, \mu'_{2x}, \mu'_{3x}, \dots$ and $\mu'_{1y}, \mu'_{2y}, \mu'_{3y}, \dots$, respectively, and if all corresponding moments are equal, i.e., if $\mu'_{1x} = \mu'_{1y}$, $\mu'_{2x} = \mu'_{2y}$, etc., then the two discrete probability distributions, $p(x)$ and $p(y)$, are identical.

The moments of a discrete random variable can be found directly using Definition 4.7, but as Examples 4.11 and 4.20 indicate, summing the series needed to find $E(Y)$, $E(Y^2)$, etc., can be tedious. Sometimes the difficulty in finding the moments of a random variable can be alleviated by using the **moment generating function** of the random variable.

**Definition 4.9**

The **moment generating function**, $m(t)$, of a discrete random variable $Y$ is defined to be

$$m(t) = E(e^{tY})$$

The moment generating function of a discrete random variable is simply a mathematical expression that condenses all the moments into a single formula. To extract specific moments from it, we first note that, by Definition 4.9.

$$E(e^{tY}) = \sum_{\text{all } y} e^{ty} p(y)$$

where

$$e^{ty} = 1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \frac{(ty)^4}{4!} + \cdots$$

Then, if $\mu'_i$ is finite for $i = 1, 2, 3, 4, \dots,$

$$m(t) = E(e^{tY}) = \sum_{\text{all } y} e^{ty} p(y) = \sum_{\text{all } y} \left[ 1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \cdots \right] p(y)$$

$$= \sum_{\text{all } y} \left[ p(y) + typ(y) + \frac{t^2}{2!} y^2 p(y) + \frac{t^3}{3!} y^3 p(y) + \cdots \right]$$

Now apply Theorems 4.2 and 4.3 to obtain

$$m(t) = \sum_{\text{all } y} p(y) + t \sum_{\text{all } y} yp(y) + \frac{t^2}{2!} \sum_{\text{all } y} y^2 p(y) + \cdots$$

But, by Definition 4.7, $\sum_{\text{all } y} y^k p(y) = \mu'_k$. Therefore,

$$m(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots$$

This indicates that if we have the moment generating function of a random variable and can expand it into a power series in $t$, i.e.,

$$m(t) = 1 + a_1 t + a_2 t^2 + a_3 t^3 + \cdots$$

then it follows that the coefficient of $t$ will be $\mu'_1 = \mu$, the coefficient of $t^2$ will be $\mu'_2/2!$, and, in general, the coefficient of $t^k$ will be $\mu'_k/k!$.

If we cannot easily expand $m(t)$ into a power series in $t$, we can find the moments of $y$ by differentiating $m(t)$ with respect to $t$ and then setting $t$ equal to 0. Thus,

$$\begin{aligned}\frac{dm(t)}{dt} &= \frac{d}{dt}\left(1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots\right) \\ &= \left(0 + \mu'_1 + \frac{2t}{2!}\mu'_2 + \frac{3t^2}{3!}\mu'_3 + \cdots\right)\end{aligned}$$

Letting $t = 0$, we obtain

$$\left.\frac{dm(t)}{dt}\right]_{t=0} = (\mu'_1 + 0 + 0 + \cdots) = \mu'_1 = \mu$$

Taking the second derivative of $m(t)$ with respect to $t$ yields

$$\frac{d^2m(t)}{dt^2} = \left(0 + \mu'_2 + \frac{3!}{3!}t\mu'_3 + \cdots\right)$$

Then, letting $t = 0$, we obtain

$$\left.\frac{d^2m(t)}{dt^2}\right]_{t=0} = (\mu'_2 + 0 + 0 + \cdots) = \mu'_2$$

Theorem 4.5 describes how to extract $\mu'_k$ from the moment generating function $m(t)$.

**THEOREM 4.5**

If $m(t)$ exists, then the $k$th moment about the origin is equal to

$$\mu'_k = \left.\frac{d^k m(t)}{dt^k}\right]_{t=0}$$

To illustrate the use of the moment generating function (MGF), consider the following examples.

**Example 4.21**

MGF for a Binomial Random Variable

Derive the moment generating function for a binomial random variable $Y$.

Solution

The moment generating function is given by

$$m(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty}p(y) = \sum_{y=0}^{n} e^{ty}\binom{n}{y}p^y q^{n-y} = \sum_{y=0}^{n}\binom{n}{y}(pe^t)^y q^{n-y}$$

We now recall the binomial theorem (see Exercise 4.36, p. 154).

$$(a + b)^n = \sum_{y=0}^{n}\binom{n}{y}a^y b^{n-y}$$

Letting $a = pe^t$ and $b = q$ yields the desired result:

$$m(t) = (pe^t + q)^n$$

**Example 4.22**

First Two Moments for a Binomial Random Variable

Use Theorem 4.5 to derive $\mu_1' = \mu$ and $\mu_2'$ for the binomial random variable.

Solution

From Theorem 4.5,

$$\mu_1' = \mu = \left.\frac{dm(t)}{dt}\right]_{t=0} = n(pe^t + q)^{n-1}(pe^t)\Big]_{t=0}$$

$$= n(pe^0 + q)^{n-1}(pe^0)$$

But $e^0 = 1$. Therefore,

$$\mu_1' = \mu = n(p + q)^{n-1}p = n(1)^{n-1}p = np$$

Similarly,

$$\mu_2' = \left.\frac{d^2m(t)}{dt^2}\right]_{t=0} = np\,\frac{d}{dt}\,[e^t(pe^t + q)^{n-1}]\Big]_{t=0}$$

$$= np[e^t(n - 1)(pe^t + q)^{n-2}pe^t + (pe^t + q)^{n-1}e^t]\Big]_{t=0}$$

$$= np[(1)(n - 1)(1)p + (1)(1)] = np[(n - 1)p + 1]$$

$$= np(np - p + 1) = np(np + q) = n^2p^2 + npq$$

**Example 4.23**

Using Moments to Derive the Variance of a Binomial Random Variable

Use the results of Example 4.22, in conjunction with Theorem 4.4, to derive the variance of a binomial random variable.

Solution

By Theorem 4.4,

$$\sigma^2 = E(Y^2) - \mu^2 = \mu_2' - (\mu_1')^2$$

Substituting the values of $\mu_2'$ and $\mu_1' = \mu$ from Example 4.22 yields

$$\sigma^2 = n^2p^2 + npq - (np)^2 = npq$$

As demonstrated in Examples 4.22 and 4.23, it is easier to use the moment generating function to find $\mu_1'$ and $\mu_2'$ for a binomial random variable than to find

$\mu_1' = E(Y)$ and $\mu_2' = E(Y^2)$ separately via the binomial formula. You have to sum only a single series to find $m(t)$. This is also the best method for finding $\mu_1'$ and $\mu_2'$ for many other random variables, but not for all.

The probability distributions, means, variances, and moment generating functions for some useful discrete random variables are summarized in the **Key Formulas** of the **Quick Review** given at the end of this chapter.

## Theoretical Exercises

4.86 Derive the moment generating function of the Poisson random variable $Y$. [*Hint:* Write

$$m(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty}\frac{\lambda^y e^{-\lambda}}{y!}$$

$$= e^{-\lambda}\sum_{y=0}^{\infty}\frac{(\lambda e^t)^y}{y!} = e^{-\lambda}e^{\lambda e^t}\sum_{y=0}^{\infty}\frac{(\lambda e^t)^y e^{-\lambda e^t}}{y!}$$

Then note that the quantity being summed is a Poisson probability with parameter $\lambda e^t$.]

4.87 Use the result of Exercise 4.86 to derive the mean and variance of the Poisson distribution.

4.88 Use the moment generating function given in the **Key Formulas** table at the end of this chapter to derive the mean and variance of a geometric random variable.

## *STATISTICS IN ACTION REVISITED*

### The Reliability of a "One-Shot" Device

We now return to the problem of assessing the reliability of a "one-shot" device outlined in the *Statistics in Action* (p. 134) that introduced this chapter. Recall that a "one-shot" device can only be used once; consequently, design engineers need to determine the minimum number of tests to conduct in order to demonstrate a desired reliability level. As stated previously, the current trend in determining the reliability of a one-shot device utilizes *acceptance sampling*, the binomial probability distribution, and the "rare event" approach of Example 3.9 (p. 93) to determine if the device has an acceptable defective rate at some acceptable level of risk.

The basic methodology can be outlined as follows. Consider a one-shot device that has some probability, $p$, of failure. Of course, the true value of $p$ is unknown, so design engineers will specify a value of $p$ that is the largest defective rate that they are willing to accept. (This value of $p$ is often called the *Lot Tolerance Percent Defective*—LTPD.) Engineers will conduct $n$ tests of the device and determine the success or failure of each test. If the number of observed failures is less than or equal to some specified value, $k$, then the engineers will conclude that the device will perform as designed. Consequently, the engineers want to know the minimum sample size $n$ needed so that observing $k$ or fewer defectives in the sample will demonstrate that the true probability of failure for the one-shot device is no greater than $p$.

If we let $Y$ represent the observed number of failures in the sample, then, from our discussion in this chapter, $Y$ has a binomial distribution with parameters $n$ and $p$. The probability of observing $k$ or fewer defectives, i.e., $P(Y \le k)$, is found using either the binomial formula of Section 4.6 or the binomial table in Appendix B. If this probability is small (say, less than .05), then either a rare event has been observed, or, more likely, the true value of $p$ for the device is smaller than the specified LTPD value.

To illustrate, suppose the desired failure rate for a one-shot device is $p = .10$. Also, suppose engineers will conduct $n = 20$ tests of the device and conclude that the device is performing to specifications if $k = 1$, i.e., if 1 or no failure is observed in the sample. The probability of interest is

$$P(Y \le 1) = P(Y = 0) + P(Y = 1) = p(0) + p(1)$$

Using $p = .10$ and $n = 20$ in Table 2, Appendix B, we obtain the probability .392. Since this probability is not small (i.e., not a rare event), engineers will be unlikely to conclude that the device has a failure rate no greater than .10 in the population.

In reliability analysis, $1 - P(Y \le k)$ is often called the "level of confidence" for concluding that the true failure rate is less than or equal to $p$. In the above example, $1 - P(Y \le 1) = .608$; thus, an engineer would only be 60.8% "confident" that the one-shot device has a failure rate of .10 or less. There are several ways to increase the confidence level. One way is to increase the sample size. Another way is to decrease the number $k$ of failures allowed in the sample.

For example, suppose we decrease the number of failures allowed in the sample from $k = 1$ to $k = 0$. Then, the probability of $k$ or fewer failures in the sample is now $P(Y = 0)$. For $p = .10$ and $n = 20$, we find $P(Y = 0) = .122$. This yields a "level of confidence" of 87.8%. You can see that the level of confidence has increased. However, most engineers who conduct acceptance sampling want a confidence level of .90, .95, or .99. These values correspond to rare event probabilities of $P(Y \leq k) = .10$, $P(Y \leq k) = .05$, or $P(Y \leq k) = .01$.

Now, suppose we increase the sample size from $n = 20$ to $n = 30$. Applying the binomial formula of Section 4.6 with $n = 30$ and $p = .10$, we obtain

$$P(Y \leq k) = P(Y = 0) = \binom{30}{0}(.10)^0(.90)^{30} = .042 \qquad \text{and} \qquad 1 - P(Y \leq k) = 1 - .042 = .958$$

Now $P(Y \leq k)$ is less than .05 and the level of confidence is greater than 95%. Consequently, if no failures are observed in the sample, the engineers will conclude (with 95.8% confidence) that the failure rate for the one-shot device is no greater than $p = .10$.

You can see that by trial and error, manipulating the values of $p$, $n$, and $k$ in the binomial formula, you can determine a desirable confidence level. The U.S. Department of Defense Reliability Analysis Center (DoD RAC) provides engineers with free access to tables and toolboxes that give the minimum sample size $n$ required to obtain a desired confidence level for a specified number of observed failures in the sample. This information is invaluable for assessing the reliability of one-shot devices.

A table of required sample sizes for acceptance sampling with a LTPD set at $p = .10$ is shown in Table SIA4.1. The table shows that if design engineers want a confidence level of 99% with a sample that allows $k = 0$ failures, they need a sample size of $n = 45$ tests. Similarly, if engineers want a confidence level of 95% with a sample that allows up to $k = 10$ failures, they need a sample size of $n = 168$ tests.

**TABLE SIA4.1 Sample Size Required for $p = .1$ to Achieve a Desired Confidence Level**

| | Confidence Levels | | | | |
|---|---|---|---|---|---|
| No. of Failures | 60% | 80% | 90% | 95% | 99% |
| | Sample Size | | | | |
| 0 | 9 | 16 | 22 | 29 | 45 |
| 1 | 20 | 29 | 38 | 47 | 65 |
| 2 | 31 | 42 | 52 | 63 | 83 |
| 3 | 41 | 55 | 65 | 77 | 98 |
| 4 | 52 | 67 | 78 | 92 | 113 |
| 5 | 63 | 78 | 91 | 104 | 128 |
| 6 | 73 | 90 | 104 | 116 | 142 |
| 7 | 84 | 101 | 116 | 129 | 158 |
| 8 | 95 | 112 | 128 | 143 | 170 |
| 9 | 105 | 124 | 140 | 156 | 184 |
| 10 | 115 | 135 | 152 | 168 | 197 |
| 11 | 125 | 146 | 164 | 179 | 210 |
| 12 | 135 | 157 | 176 | 191 | 223 |
| 13 | 146 | 169 | 187 | 203 | 236 |
| 14 | 156 | 178 | 198 | 217 | 250 |
| 15 | 167 | 189 | 210 | 228 | 264 |
| 16 | 177 | 200 | 223 | 239 | 278 |

**TABLE SIA4.1 Sample Size Required for P = .1 to Achieve a Desired Confidence Level** (*continued*)

| No. of Failures | Confidence Levels 60% | 80% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| | Sample Size | | | | |
| 17 | 188 | 211 | 234 | 252 | 289 |
| 18 | 198 | 223 | 245 | 264 | 301 |
| 19 | 208 | 233 | 256 | 276 | 315 |
| 20 | 218 | 244 | 267 | 288 | 327 |
| 22 | 241 | 266 | 290 | 313 | 342 |
| 24 | 262 | 286 | 312 | 340 | 378 |
| 26 | 282 | 308 | 330 | 364 | 395 |
| 28 | 303 | 331 | 354 | 385 | 430 |
| 30 | 319 | 354 | 377 | 408 | 448 |
| 35 | 374 | 403 | 430 | 462 | 505 |
| 40 | 414 | 432 | 490 | 512 | 565 |
| 45 | 478 | 510 | 550 | 580 | 620 |
| 50 | 513 | 534 | 595 | 628 | 675 |

*Source*: Department of Defense Reliability Analysis Center, *START: Analysis of "One-Shot" Devices*, Vol. 7, No. 4, 2000 (Table 2).

# Quick Review

## Key Terms

*Note: Starred (∗) terms are from the optional section in this chapter.*

Bernoulli distribution 174
Bernoulli random variable 146
Bernoulli trials 146
Binomial distribution 159
Binomial experiment 148
Binomial random variable 174
Discrete random variable 135
Expected value 140
Geometric distribution 163
Hypergeometric distribution 167
Hypergeometric random variable 164
Mean 140
*Moments 175
*Moment generating function 175
Multinomial distribution 155
Multinomial experiment 155
Negative binomial distribution 159
Poisson distribution 159
Poisson random variable 169
Probability distribution 135
Random variable 134
Sampling with replacement 164
Standard deviation 141
Variance 141

## Key Formulas

*Note: Starred (∗) formulas are from the optional section in this chapter.*

| Random Variable | $p(y)$ | $\mu$ | $\sigma^2$ | $*m(t)$ |
|---|---|---|---|---|
| Discrete (general) | $p(y)$ | $E(Y) = \Sigma yp(y)$ | $E(Y^2) - \mu^2$ | |
| Bernoulli | $p(y) = p^y q^{1-y}$ where $q = 1 - p$, $y = 0, 1$ | $p$ | $pq$ | $pe^t + q$ |
| Binomial | $p(y) = \binom{n}{y} p^y q^{n-y}$ where $q = 1 - p$, $y = 0, 1, \ldots, n$ | $np$ | $npq$ | $(pe^t + q)^n$ |

| Random Variable | $p(y)$ | $\mu$ | $\sigma^2$ | $*m(t)$ |
|---|---|---|---|---|
| Hypergeometric | $p(y) = \dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$ | $\dfrac{nr}{N}$ | $\dfrac{r(N-r)n(N-n)}{N^2(N-1)}$ | Not given |
| Poisson | $p(y) = \dfrac{\lambda^y e^{-\lambda}}{y!} \quad y = 1, 2, \ldots$ | $\lambda$ | $\lambda$ | $e^{\lambda(e^t-1)}$ |
| Geometric | $p(y) = p(1-p)^{y-1} \quad y = 1, 2, \ldots$ | $\dfrac{1}{p}$ | $\dfrac{1-p}{p^2}$ | $\dfrac{pe^t}{1-(1-p)e^t}$ |
| Negative binomial | $p(y) = \binom{y-1}{r-1} p^r(1-p)^{y-r} \quad y = r, r+1, \ldots$ | $\dfrac{r}{p}$ | $\dfrac{r(1-p)}{p^2}$ | $\left(\dfrac{pe^t}{1-(1-p)e^t}\right)^r$ |
| Multinomial | $p(y_1, y_2, \ldots, y_k) = \dfrac{n!}{y_1!y_2!\ldots y_k!}(p_1)^{y_1}(p_2)^{y_2}\cdots(p_k)^{y_k}$ | $np_i$ | $np_i(1-p_i)$ | Not given |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $p\,(y)$ | | Probability distribution of the random variable $Y$ |
| $E\,(Y)$ | Expected value of $Y$ | Mean of the probability distribution for $Y$ |
| $S$ | | The outcome of a Bernoulli trial denoted a “success” |
| $F$ | | The outcome of a Bernoulli trial denoted a “failure” |
| $p$ | | The probability of success ($S$) in a Bernoulli trial |
| $q$ | | The probability of failure ($F$) in a Bernoulli trial, where $q = 1 - p$ |
| $\lambda$ | lambda | The mean (or expected) number of events for a Poisson random variable |
| $e$ | | A constant used in the Poisson probability distribution, where $e = 2.71828\ldots$ |
| $m(t)$ | “m” of “t” | Moment generating function |

## Chapter Summary Notes

- A **discrete random variable** can assume only a countable number of values.
- Requirements for a discrete probability distribution: $p(y) \geq 0$ and $\Sigma p(y) = 1$
- Probability models for discrete random variables: **Bernoulli, binomial, multinomial, negative binomial, geometric, hypergeometric**, and **Poisson**
- Characteristics of a **Bernoulli random variable:** (1) two mutually exclusive outcomes, $S$ and $F$, in a trial, (2) outcomes are exhaustive, (3) $P(S) = p$ and $P(F) = q$, where $p + q = 1$
- Characteristics of a **binomial random variable:** (1) $n$ identical trials, (2) two possible outcomes, $S$ and $F$, per trial, (3) $P(S) = p$ and $P(F) = q$ remain the same from trial to trial, (4) trials are independent, (5) $Y =$ number of $S$’s in $n$ trials
- Characteristics of a **multinomial random variable:** (1) $n$ identical trials, (2) $k$ possible outcomes per trial, (3) probabilities of $k$ outcomes remain the same from trial to trial, (4) trials are independent, (5) $Y_1, Y_2, \ldots, Y_k$ are counts of outcomes in $k$ categories

- Characteristics of a **negative binomial random variable:** (1) identical trials, (2) two possible outcomes, $S$ and $F$, per trial, (3) $P(S) = p$ and $P(F) = q$ remain the same from trial to trial, (4) trials are independent, (5) $Y =$ number of trials until $r$th $S$ is observed
- Characteristics of a **geometric random variable:** (1) identical trials, (2) two possible outcomes, $S$ and $F$, per trial, (3) $P(S) = p$ and $P(F) = q$ remain the same from trial to trial, (4) trials are independent, (5) $Y =$ number of trials until 1st $S$ is observed
- Characteristics of a **hypergeometric random variable:** (1) draw $n$ elements without replacement from a set of $N$ elements, $r$ of which have outcome $S$ and $(N - r)$ of which have outcome $F$, (2) $Y =$ number of $S$'s in $n$ trials
- Characteristics of a **Poisson random variable:** (1) $Y =$ number of times a rare event, $S$, occurs in a unit of time, area, or volume, (2) $P(S)$ remains the same for all units, (3) value of $Y$ in one unit is independent of value in another unit

## Supplementary Exercises

4.89 *Management system failures.* Refer to the *Process Safety Progress* (Dec., 2004) study of industrial accidents caused by management system failures, Exercise 3.78 (p. 127). The table listing the four root causes of system failures (and associated proportions) is reproduced below. Suppose three industrial accidents are randomly selected (without replacement) from among all industrial accidents caused by management system failures. Find and graph the probability distribution of $Y$, the number of accidents caused by Engineering and Design failure.

| Cause Category | Proportion |
|---|---|
| Engineering & Design | .32 |
| Procedures & Practices | .29 |
| Management & Oversight | .27 |
| Training & Communication | .12 |
| Total | 1.00 |

4.90 *Unmanned watching system.* Refer to the *IEEE Computer Applications in Power* study of an outdoor unmanned watching system designed to detect trespassers, Exercise 3.79 (p. 127). In snowy weather conditions, the system detected 7 out of 10 intruders; thus, the researchers estimated the system's probability of intruder detection in snowy conditions at .70.

a. Assuming the probability of intruder detection in snowy conditions is only .50, find the probability that the unmanned system detects at least 7 of the 10 intruders.
b. Based on the result, part **a**, comment on the reliability of the researcher's estimate of the system's detection probability in snowy conditions.
c. Suppose two of the 10 intruders had criminal intentions. What is the probability that both of these intruders were detected by the system.

4.91 *Classifying environmentalists.* Environmental engineers classify consumers into one of five categories (see Exercise 3.77, (p. 127), for a description of each group). The probabilities associated with the groups follow:

| | |
|---|---|
| Basic browns | .28 |
| True-blue greens | .11 |
| Greenback greens | .11 |
| Sprouts | .26 |
| Grousers | .24 |

*Source*: The *Orange Country Register,* Aug. 7, 1990.

Let $Y$ equal the number of consumers that must be sampled until the first environmentalist is found. (*Note:* From Exercise 3.3, an environmentalist is a true-blue green, greenback green, or sprout.)

a. Specify the probability distribution for $Y$ in table form.
b. Give a formula for $p(y)$.
c. Find $\mu$ and $\sigma$, the mean and standard deviation of $Y$.
d. Use the information, part **a**, to form an interval that will include $Y$ with a high probability.

4.92 *Mercury in seafood.* An issue of *Consumer Reports* found widespread contamination and mislabeling of seafood in supermarkets in New York City and Chicago. The study revealed one alarming statistic: 40% of the swordfish pieces available for sale had a level of mercury above the Food and Drug Administration (FDA) maximum amount. For a random sample of three swordfish pieces, find the probability that

a. All three swordfish pieces have mercury levels above the FDA maximum.
b. Exactly one swordfish piece has a mercury level above the FDA maximum.
c. At most one swordfish piece has a mercury level above the FDA maximum.

4.93 *Gastroenteritis outbreak.* A waterborne nonbacterial gastroenteritis outbreak occurred in Colorado as a result of a long-standing filter deficiency and malfunction of a sewage treatment plant. A study was conducted to determine whether the incidence of gastrointestinal disease during the epidemic was related to water consumption (*American*

*Water Works Journal*, Jan. 1986). A telephone survey of households yielded the accompanying information on daily consumption of 8-ounce glasses of water for a sample of 40 residents who exhibited gastroenteritis symptoms during the epidemic.

| | Daily Consumption of 8-Ounce Glasses of Water | | | | |
|---|---|---|---|---|---|
| | 0 | 1–2 | 3–4 | 5 or more | Total |
| Number of respondents with symptoms | 6 | 11 | 13 | 10 | 40 |

*Source*: Hopkins, R. S., et. al. "Gastroenteritis: Case study of a Colorado outbreak." *Journal American Water Works Association*, Vol. 78, No. 1, Jan. 1986, p. 42, Table 1. Copyright © 1986, American Water Works Association. Reprinted with permission.

a. If the number of respondents with symptoms *does not depend* on the daily amount of water consumed, assign probabilities to the four categories shown in the table.
b. Use the information, part **a**, to find the probability of observing the sample result shown in the table.

4.94 *Lifelength of solar heating panel.* An engineering development laboratory conducted an experiment to investigate the life characteristics of a new solar heating panel, designed to have a useful life of at least 5 years with probability $p = .95$. A random sample of 20 such solar panels was selected, and the useful life of each was recorded.

a. What is the probability that exactly 18 will have a useful life of at least 5 years?
b. What is the probability that at most 10 will have a useful life of at least 5 years?
c. If only 10 of the 20 solar panels have a useful life of at least 5 years, what would you infer about the true value of $p$?

4.95 *Steam turbine power plant.* Two of the five mechanical engineers employed by the county sanitation department have experience in the design of steam turbine power plants. You have been instructed to choose randomly two of the five engineers to work on a project for a new power plant.

a. What is the probability that you will choose the two engineers with experience in the design of steam turbine power plants?
b. What is the probability that you will choose at least one of the engineers with such experience?

4.96 *Rail system shutdowns.* Lesser-developed countries experiencing rapid population growth often face severe traffic control problems in their large cities. Traffic engineers have determined that elevated rail systems may provide a feasible solution to these traffic woes. Studies indicate that the number of maintenance-related shutdowns of the elevated rail system in a particular country has a mean equal to 6.5 per month.

a. Find the probability that at least five shutdowns of the elevated rail system will occur next month in the country.
b. Find the probability that exactly four shutdowns will occur next month.

4.97 *Species hot spots.* "Hot spots" are species-rich geographical areas (see Exercise 3.81, p. 127). A *Nature* (Sept. 1993) study estimated the probability of a bird species in Great Britain inhabiting a butterfly hot spot at .70. Consider a random sample of 4 British bird species selected from a total of 10 tagged species. Assume that 7 of the 10 tagged species inhabit a butterfly hot spot.

a. What is the probability that exactly half of the 4 bird species sampled inhabit a butterfly hot spot?
b. What is the probability that at least 1 of the 4 bird species sampled inhabits a butterfly hot spot?

4.98 *Pollution control regulations.* A task force established by the Environmental Protection Agency was scheduled to investigate 20 industrial firms to check for violations of pollution control regulations. However, budget cutbacks have drastically reduced the size of the task force, and they will be able to investigate only 3 of the 20 firms. If it is known that 5 of the firms are actually operating in violation of regulations, find the probability that

a. None of the three sampled firms will be found in violation of regulations.
b. All three firms investigated will be found in violation of regulations.
c. At least 1 of the 3 firms will be operating in violation of pollution control regulations.

4.99 *Use of road intersection.* The random variable $Y$, the number of cars that arrive at an intersection during a specified period of time, often possesses (approximately) a Poisson probability distribution. When the mean arrival rate $\lambda$ is known, the Poisson probability distribution can be used to aid a traffic engineer in the design of a traffic control system. Suppose you estimate that the mean number of arrivals per minute at the intersection is one car per minute.

a. What is the probability that in a given minute, the number of arrivals will equal three or more?
b. Can you assure the engineer that the number of arrivals will rarely exceed three per minute?

4.100 *Tapeworms in fish.* The negative binomial distribution was used to model the distribution of parasites (tapeworms) found in several species of Mediterranean fish (*Journal of Fish Biology*, Aug. 1990). Assume the event of interest is whether a parasite is found in the digestive tract of brill fish, and let $Y$ be the number of brill that must be sampled until a parasitic infection is found. The researchers estimate the probability of an infected fish at .544. Use this information to estimate the following probabilities:

a. $P(Y = 3)$
b. $P(Y \leq 2)$
c. $P(Y > 2)$

4.101 *Major rockslides in Canada.* A study of natural rock slope movements in the Canadian Rockies over the past 5,000 years revealed that the number of major rockslides per 100 square kilometers had an expected value of 1.57 (*Canadian Geotechnical Journal*, Nov. 1985).

a. Find the mean and standard deviation of $Y$, the number of major rockslides per 100 square kilometers in the Canadian Rockies over a 5,000-year period.
b. What is the probability of observing 3 or more major rockslides per 100 square kilometers over a 5,000-year period?

4.102 *Optical scanner errors.* The manufacturer of a price-reading optical scanner claims that the probability it will misread the price of any product by misreading the "bar code" on a product's label is .001. At the time one of the scanners was installed in a supermarket, the store manager tested its performance. Let $Y$ be the number of trials (i.e., the number of prices read by the scanner) until the first misread price is observed.

a. If the manufacturer's claim is correct, find the probability distribution for $Y$. (Assume the trials represent independent events.)
b. If the manufacturer's claim is correct, what is the probability that the scanner will not misread a price until after the fifth price is read?
c. If in fact the third price is misread, what inference would you make about the manufacturer's claim? Explain.

4.103 *Fungi in beech forest trees.* Refer to the *Applied Ecology and Environmental Research* (Vol. 1, 2003) study of beech trees damaged by fungi, Exercise 3.9 (p. 85). The researchers found that 25% of the beech trees in East Central Europe have been damaged by fungi. Consider a sample of 20 beech trees from this area.

a. What is the probability that fewer than half are damaged by fungi?
b. What is the probability that more than 15 are damaged by fungi?
c. How many of the sampled trees would you expect to be damaged by fungi?

4.104 *Use of acceleration lane.* A study of vehicle flow characteristics on acceleration lanes (i.e., merging ramps) at a major freeway in Israel found that one out of every six vehicles uses less than one-third of the acceleration lane before merging into traffic (*Journal of Transportation Engineering*, Nov. 1985). Suppose we monitor the location of the merge for the next five vehicles that enter the acceleration lane.

a. What is the probability that none of the vehicles will use less than one-third of the acceleration lane?
b. What is the probability that exactly two of the vehicles will use less than one-third of the acceleration lane?

4.105 *Use of acceleration lane (continued).* Refer to Exercise 4.103. Suppose that the number of vehicles using the acceleration lane per minute has a mean equal to 1.1.

a. What is the probability that more than two vehicles will use the acceleration lane in the next minute?
b. What is the probability that exactly three vehicles will use the acceleration lane in the next minute?

4.106 *Breakdowns of industrial robots.* Industrial robots are programmed to operate through microprocessors. The probability that one such computerized robot breaks down during any one 8-hour shift is .2. Find the probability that the robot will operate for at most five shifts before breaking down twice.

4.107 *Level of benzene at petrochemical plants.* Benzene, a solvent commonly used to synthesize plastics and found in consumer products such as paint strippers and high-octane unleaded gasoline, has been classified by scientists as a leukemia-causing agent. Let $Y$ be the level (in parts per million) of benzene in the air at a petrochemical plant. Then $Y$ can take on the values $0, 1, 2, 3, \dots, 1{,}000{,}000$ and can be approximated by a Poisson probability distribution. In 1978, the federal government lowered the maximum allowable level of benzene in the air at a workplace from 10 parts per million (ppm) to 1 ppm. Any industry in violation of these government standards is subject to severe penalties, including implementation of expensive measures to lower the benzene level.

a. Suppose the mean level of benzene in the air at petrochemical plants is $\mu = 5$ ppm. Find the probability that a petrochemical plant exceeds the government standard of 1 ppm.
b. Repeat part **a**, assuming that $\mu = 2.5$.
c. A study by Gulf Oil revealed that 88% of benzene-using industries expose their workers to 1 ppm or less of the solvent. Suppose you randomly sampled 55 of the benzene-using industries in the country and determined $Y$, the number in violation of government standards. Use the Poisson approximation to the binomial to find the probability that none of the sampled industries violates government standards. Compare this probability to the exact probability computed using the binomial probability distribution.
d. Refer to part **c**. Use the fact that 88% of benzene-using industries expose their workers to 1 ppm or less of benzene to approximate $\mu$, the mean level of benzene in the air at these industries. [*Hint*: Search Table 4 of Appendix B for the value of $\mu$ that yields $P(Y \le 1)$ closest to .88.]

4.108 *Auditory nerve fibers.* A discharge (or response) rate of auditory nerve fibers [recorded as the number of spikes per 200 milliseconds (ms) of noise burst] is used to measure the effect of acoustic stimuli in the auditory nerve. An empirical study of auditory nerve fiber response rates in cats resulted in a mean of 15 spikes/ms (*Journal of the Acousti-*

*cal Society of America*, Feb. 1986). Let $Y$ represent the auditory nerve fiber response rate for a randomly selected cat in the study.

a. If $Y$ is approximately a Poisson random variable, find the mean and standard deviation of $Y$.
b. Assuming $Y$ is Poisson, what is the approximate probability that $Y$ exceeds 27 spikes/ms?
c. In the study, the variance of $Y$ was found to be "substantially smaller" than 15 spikes/ms. Is it reasonable to expect $Y$ to follow a Poisson process? How will this affect the probability computed in part **b**?

## Theoretical Exercises

4.109 Suppose the random variable $Y$ has a moment generating function given by

$$m(t) = \frac{1}{5}e^{t} + \frac{2}{5}e^{2t} + \frac{2}{5}e^{3t}$$

a. Find the mean of $Y$.
b. Find the variance of $Y$.

4.110 Let $Y$ be a geometric random variable. Show that $E(Y) = 1/p$. [*Hint:* Write

$$E(Y) = p\sum_{y=1}^{\infty} yq^{y-1} \quad \text{where } q = 1 - p$$

and note that

$$\frac{dq^{y}}{dq} = yq^{y-1}$$

Thus,

$$E(Y) = p\sum_{y=1}^{\infty} yq^{y-1} = p\frac{d}{dq}\left(\sum_{y=1}^{\infty} q^{y}\right)$$

Then use the fact that

$$\sum_{y=1}^{\infty} q^{y} = \frac{q}{1 - q}$$

(The sum of this infinite series is given in most mathematical handbooks.)]

4.111 The probability generating function $P(t)$ for a discrete random variable $Y$ is defined to be

$$P(t) = E(t^{Y}) = p_0 + p_1 t + p_2 t^2 + \cdots$$

where $p_i = P(Y = i)$.

a. Find $P(t)$ for the Poisson distribution. [*Hint*: Write

$$E(t^{Y}) = \sum_{y=0}^{\infty}\frac{(\lambda t)^{y}e^{-\lambda}}{y!} = e^{\lambda(t-1)}\sum_{y=0}^{\infty}\frac{(\lambda t)^{y}e^{-\lambda t}}{y!}$$

and note that the quantity being summed is a Poisson probability with mean $\lambda t$.]

b. Use the facts that

$$E(Y) = \left.\frac{dP(t)}{dt}\right]_{t=1} \quad \text{and} \quad E[Y(Y-1)] = \left.\frac{d^2P(t)}{dt^2}\right]_{t=1}$$

to derive the mean and variance of a Poisson random variable.

CHAPTER 5

# Continuous Random Variables

## OBJECTIVE

*To distinguish between continuous and discrete random variables and their respective probability distributions; to present some useful continuous probability distributions and show how they can be used to solve some practical problems*

## CONTENTS



*STATISTICS IN ACTION*

Super Weapons Development—Optimizing the Hit Ratio

## *STATISTICS IN ACTION*

### Super Weapons Development—Optimizing the Hit Ratio

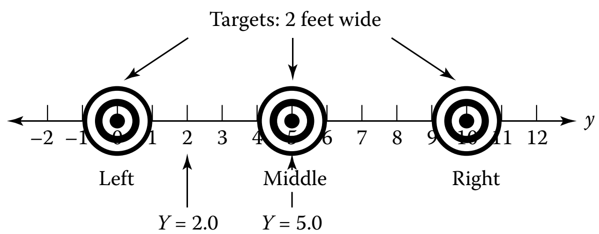
The U.S. Army is working with a major defense contractor to develop a "super" weapon. The weapon is designed to fire a large number of sharp tungsten bullets—called flechettes—with a single shot that will destroy a large number of enemy soldiers. Flechettes are about the size of an average nail, with small fins at one end to stabilize them in flight. Since World War I, when France dropped them in large quantities from aircraft on masses of ground troops, munitions experts have experimented with using flechettes in a variety of guns. The problem with using flechettes as ammunition is accuracy—current weapons that fire large quantities of flechettes have unsatisfactory hit ratios when fired at long distances.

The defense contractor (not named here for both confidentiality and security reasons) has developed a prototype gun that fires 1,100 flechettes with a single round. In range tests, three 2-feet-wide targets were set up a distance of 500 meters (approximately 1,500 feet) from the weapon. Using a number line as a reference, the centers of the three targets were at 0, 5, and 10 feet, respectively, as shown in Figure SIA5.1. The prototype gun was aimed at the middle target (center at 5 feet) and fired once. The point $Y$ where each of the 1,100 flechettes landed at the 500-meter distance was measured using a horizontal grid. The 1,100 measurements on the random variable $Y$ are saved in the **MOAGUN** file. (The data are simulated for confidentiality reasons.) For example, a flechette with a horizontal value of $Y = 5.5$ (shown in Figure SIA5.1) hit the middle target, but a flechette with a horizontal value of $Y = 2.0$ (also shown in the figure) did not hit any of the three targets.

The defense contractor is interested in the likelihood of any one of the targets being hit by a flechette, and in particular wants to set the gun specifications to maximize the number of target hits. The weapon is designed to have a mean horizontal value, $E(Y)$, equal to the aim point (e.g., $\mu = 5$ feet when aimed at the center target). By changing specifications, the contractor can vary the standard deviation, $\sigma$. The **MOAGUN** file contains flechette measurements for three different range tests—one with a standard deviation of $\sigma = 1$ foot, one with $\sigma = 2$ feet, and one with $\sigma = 4$ feet.

In the *Statistics in Action Revisited* at the end of this chapter, we demonstrate how to utilize one of the probability distributions covered in this chapter to aid the defense contractor in developing its "super" weapon.

**FIGURE SIA5.1**




## 5.1 Continuous Random Variables

Many random variables observed in real life are not discrete random variables because the number of values that they can assume is not countable. For example, the waiting time $Y$ (in minutes) at a traffic light could, in theory, assume any of the uncountably infinite number of values in the interval $0 < Y < \infty$. The daily rainfall at some location, the strength (in pounds per square inch) of a steel bar, and the intensity of sunlight at a particular time of the day are other examples of random variables that can assume any one of the uncountably infinite number of points in one or more intervals on the real line. In contrast to discrete random variables, such variables are called **continuous random variables**.

The preceding discussion identifies the difference between discrete and continuous random variables, but it fails to point to a practical problem. It is impossible to assign a finite amount of probability to each of the uncountable number of points in a line interval in such a way that the sum of the probabilities is 1. Therefore, the distinction between discrete and continuous random variables is usually based on the difference in their **cumulative distribution functions**.

**Definition 5.1**

The **cumulative distribution function $F(y_0)$** for a random variable $Y$ is equal to the probability

$$F(y_0) = P(Y \leq y_0), \; -\infty < y_0 < \infty$$

For a discrete random variable, the cumulative distribution function is the cumulative sum of $p(y)$, from the smallest value that $Y$ can assume, to a value of $y_0$. For example, from the cumulative sums in Table 2 of Appendix B, we obtain the following values of $F(y)$ for a binomial random variable with $n = 5$ and $p = .5$:

$$F(0) = P(Y \leq 0) = \sum_{y=0}^{0} p(y) = p(0) = .031$$

$$F(1) = P(Y \leq 1) = \sum_{y=0}^{1} p(y) = .188$$

$$F(2) = P(Y \leq 2) = \sum_{y=0}^{2} p(y) = .500$$

$$F(3) = P(Y \leq 3) = .812$$

$$F(4) = P(Y \leq 4) = .969$$

$$F(5) = P(Y \leq 5) = 1$$

A graph of $p(y)$ is shown in Figure 5.1. The value of $F(y_0)$ is equal to the sum of the areas of the probability rectangles from $Y = 0$ to $Y = y_0$. The probability $F(3)$ is shaded in the figure.

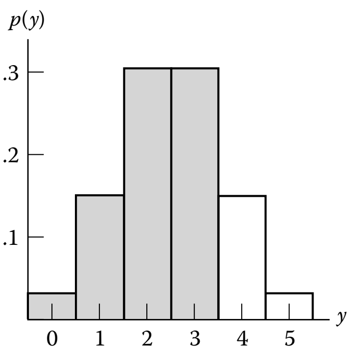



**FIGURE 5.1**
Probability distribution for a binomial random variable ($n = 5$, $p = 5$); shaded area corresponds to $F(3)$

A graph of the cumulative distribution function for the binomial random variable with $n = 5$ and $p = .5$, shown in Figure 5.2, illustrates an important property of the cumulative distribution functions for all discrete random variables: *They are step functions*. For example, $F(y)$ is equal to .031 until, as $Y$ increases, it reaches $Y = 1$. Then $F(y)$ jumps abruptly to $F(1) = .188$. The value of $F(Y)$ then remains constant as $Y$ increases until $Y$ reaches $Y = 2$. Then $F(y)$ rises abruptly to $F(2) = .500$. Thus, $F(y)$ is a discontinuous function that jumps upward at a countable number of points ($Y = 0, 1, 2, 3$, and 4).

In contrast to the cumulative distribution function for a discrete random variable, the cumulative distribution function $F(y)$ for a continuous random variable is a **monotonically increasing continuous** function of $Y$. This means that $F(y)$ is a continuous function such that if $y_a < y_b$, then $F(y_a) \leq F(y_b)$, i.e., as $Y$ increases, $F(y)$ never decreases. A graph of the cumulative distribution function for a continuous random variable might appear as shown in Figure 5.3.

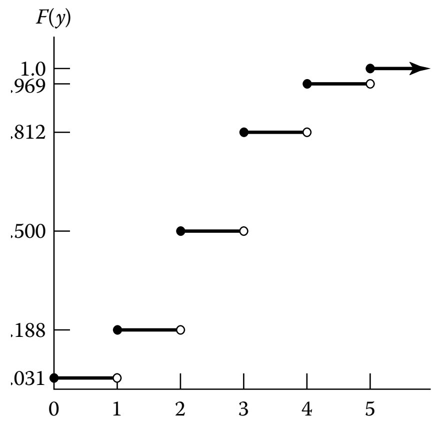



**FIGURE 5.2**
Cumulative distribution function $F(y)$ for a binomial random variable ($n = 5$, $p = 5$)

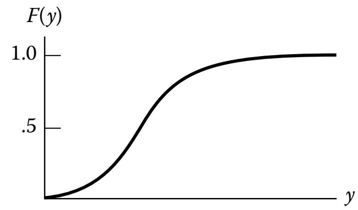



**FIGURE 5.3**
Cumulative distribution function for a continuous random variable

**Definition 5.2**

A **continuous random variable** $Y$ is one that has the following three properties:

1. $Y$ takes on an uncountably infinite number of values in the interval $(-\infty, \infty)$.
2. The cumulative distribution function, $F(y)$, is continuous.
3. The probability that $Y$ equals any one particular value is 0.

## 5.2 The Density Function for a Continuous Random Variable

In Chapter 1, we described a large set of data by means of a relative frequency distribution. If the data represent measurements on a continuous random variable and if the amount of data is very large, we can reduce the width of the class intervals until the distribution appears to be a smooth curve. A **probability density function** is a theoretical model for this distribution.

**Definition 5.3**

If $F(y)$ is the cumulative distribution function for a continuous random variable $Y$, then the **density function $f(y)$** for $Y$ is

$$f(y) = \frac{dF(y)}{dy}$$

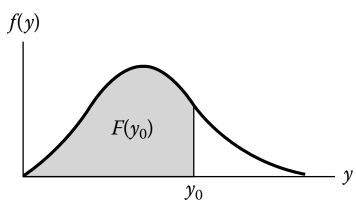



**FIGURE 5.4**
Density function $f(y)$ for a continuous random variable

The density function for a continuous random variable $y$, the model for some real-life population of data, will usually be a smooth curve, as shown in Figure 5.4. It follows from Definition 5.3 that

$$F(y) = \int_{-\infty}^{y} f(t)\, dt$$

Thus, the cumulative area under the curve between $-\infty$ and a point $y_0$ is equal to $F(y_0)$.

The density function for a continuous random variable must always satisfy the three properties given in the following box.

**Properties of a Density Function for a Continuous Random Variable $Y$**

1. $f(y) \geq 0$
2. $\int_{-\infty}^{\infty} f(y)\,dy = F(\infty) = 1$
3. $P(a < Y < b) = \int_a^b f(y)\,dy = F(b) - F(a)$, where $a$ and $b$ are constants

**Example 5.1**

**Density Function Application — Microwave Magnetrons**

A cavity magnetron is a high-powered vacuum tube commonly used in microwave ovens. One brand of microwave uses a new type of magnetron that can be unstable when not installed properly. Let $Y$ be a continuous random variable that represents the proportion of these new magnetrons in a large shipment of microwave ovens that are improperly installed. Let $c$ be a constant and consider the following probability density function for $Y$:

$$f(y) = \begin{cases} cy & \text{if } 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$.

b. Find $P(.2 < Y < .5)$. Interpret the result.

**Solution**

a. Since $\int_{-\infty}^{\infty} f(y)\,dy$ must equal 1, we have

$$\int_{-\infty}^{\infty} f(y)\,dy = \int_0^1 cy\,dy = c\frac{y^2}{2}\Bigg]_0^1 = c\left(\frac{1}{2}\right) = 1$$

Solving for $c$ yields $c = 2$, and thus, $f(y) = 2y$. A graph of $f(y)$ is shown in Figure 5.5.

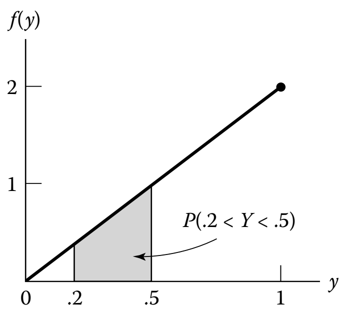



**FIGURE 5.5**
Graph of the density function $f(y)$ for Example 5.1

b.
$$\begin{aligned} P(.2 < Y < .5) &= \int_{.2}^{.5} f(y)\,dy \\ &= \int_{.2}^{.5} 2y\,dy \\ &= y^2\Bigg]_{.2}^{.5} = (.5)^2 - (.2)^2 \\ &= .25 - .04 = .21 \end{aligned}$$

This probability, shaded in Figure 5.5, is the area under the density function between $Y = .2$ and $Y = .5$. Since Y represents the proportion of improperly installed magnetrons, we can say that the probability that between 20% and 50% of the magnetrons are improperly installed is .21.

**Example 5.2**

**Finding a Cumulative Distribution Function**

Refer to Example 5.1. Find the cumulative distribution function for the random variable $Y$. Then find $F(.2)$ and $F(.7)$. Interpret the results.

**Solution**

By Definition 5.3, it follows that

$$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(t)\,dt = \int_0^y 2t\,dt \\ &= 2\left(\frac{t^2}{2}\right)\Bigg]_0^y = y^2 \end{aligned}$$

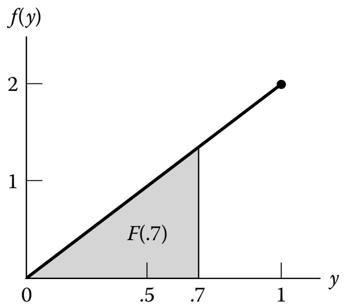



**FIGURE 5.6**
Graph of the density function $f(y)$ for Example 5.2; shaded area corresponds to $F(.7)$

Then

$$F(.2) = P(Y \leq .2) = (.2)^2 = .04$$
$$F(.7) = P(Y \leq .7) = (.7)^2 = .49$$

The value of $F(y)$ when $Y = .7$—i.e., $F(.7)$—is the shaded area in Figure 5.6. This implies that the probability of finding 70% or fewer improperly installed magnetrons is .49.

Many of the continuous random variables with applications in statistics have density functions whose integrals cannot be expressed in closed form. They can only be approximated by numerical methods. Tables of areas under several such density functions are presented in Appendix B and will be introduced as required.

## Theoretical Exercises

5.1 Let $c$ be a constant and consider the density function for the random variable $Y$:

$$f(y) = \begin{cases} cy^2 & \text{if } 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(1)$.
d. Compute $F(.5)$.
e. Compute $P(1 \leq Y \leq 1.5)$.

5.2 Let $c$ be a constant and consider the density function for the random variable $Y$:

$$f(y) = \begin{cases} c(2 - y) & \text{if } 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(.4)$.
d. Compute $P(.1 \leq Y \leq .6)$.

5.3 Let $c$ be a constant and consider the density function for the random variable $Y$:

$$f(y) = \begin{cases} c + y & \text{if } -1 < y < 0 \\ c - y & \text{if } 0 \leq y < 1 \end{cases}$$

a. Find the value of $c$.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(-.5)$.
d. Compute $P(0 \leq Y \leq .5)$.

5.4 Let $c$ be a constant and consider the density function for the random variable $Y$:

$$f(y) = \begin{cases} (1/c)e^{-y/2} & \text{if } y \geq 0 \\ (1/c)e^{y/2} & \text{if } y < 0 \end{cases}$$

a. Find the value of $c$.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(1)$.
d. Compute $P(Y > .5)$.

## Applied Exercises

5.5 *Time a train is late.* The amount of time $Y$ (in minutes) that a commuter train is late is a continuous random variable with probability density

$$f(y) = \begin{cases} \dfrac{c}{500}(25 - y^2) & \text{if } -5 < y < 5 \\ 0 & \text{elsewhere} \end{cases}$$

[*Note*: A negative value of $Y$ means that the train is early.]

a. Find the value of $c$ for this probability distribution.
b. Find the cumulative distribution function, $F(y)$.
c. What is the probability that the train is no more than 3 minutes late?

5.6 *Coastal sea level rise.* Rising sea levels are a threat to coastal cities in the United States. Consequently, for planning purposes it is important to have accurate forecasts of the future rise in sea level. The *Journal of Waterway, Port, Coastal, and Ocean Engineering* (March/April, 2013) published a study which used a statistical probability distribution to model the projected rise in sea level. The acceleration $Y$ in sea level rise (standardized between 0 and 1) was modeled using the following density function:

$$f(y) = cy(1 - y), 0 < y < 1$$

(This distribution is called the *Beta* distribution.)

a. Find the value of $c$ for this probability distribution.
b. Find the cumulative distribution function, $F(y)$.
c. Compute $F(.5)$ and interpret this value.

5.7 *Earthquake recurrence in Iran.* The *Journal of Earthquake Engineering* (Vol. 17, 2013) modeled the time $Y$ (in years) between major earthquakes occurring in the Iranian Plateau. One of the models considered had the following density function:

$$f(y) = ce^{-cy}, y > 0$$

a. Show that the properties of a density function for a continuous random variable are satisfied for any constant $c > 0$.
b. For one area of Iran, $c$ was estimated to be $c = .04$. Using this value, give the equation of the cumulative distribution function, $F(y)$.
c. The earthquake system reliability at time $t$, $R(t)$, is defined as $R(t) = 1 - F(t)$. Find $R(5)$ and interpret this probability.

5.8 *Extreme value distributions.* Extreme value distributions are used to model values of a continuous random variable that represent extremely rare events. For example, an oceanic engineer may want to model the size of a freak wave from a tsunami, or an environmental engineer might want to model the probability of the hottest temperature exceeding a certain threshhold. The journal *Extremes* (March, 2013) investigated several probability distributions for extreme values.

a. The cumulative distribution function for a Type I extreme value distribution with mean 0 and variance 1 takes the form: $F(y) = \exp\{-\exp(-y)\}, y > 0$ (This is known as the *Gumbel* distribution.) Show that the property, $F(\infty) = 1$, is satisfied.
b. Refer to part a. Find $F(2)$ and interpret the result.
c. The cumulative distribution function for a Type II extreme value distribution with mean 0 and variance 1 takes the form: $F(y) = \exp\{-y^{-1}\}, y > 0$ (This is known as the *Frechet* distribution.) Show that the property, $F(\infty) = 1$, is satisfied.)
d. Refer to part c. Find $F(2)$ and interpret the result.
e. For which extreme value distribution, Type I or Type II, is it more likely that the extreme value exceeds 2?

## Optional Theoretical Exercise

5.9 *New better than used.* Continuous probability distributions provide theoretical models for the lifelength of a component (e.g., computer chip, lightbulb, automobile, air-conditioning unit, and so on). Often, it is important to know whether or not it is better to periodically replace an old component with a new component. For example, for certain types of lightbulbs, an old bulb that has been in use for a while tends to have a longer lifelength than a new bulb. Let $Y$ represent the lifelength of some component with cumulative distribution function $F(y)$. Then the "life" distribution $F(y)$ is considered **new better than used** (NBU) if

$$\overline{F}(x + y) \le \overline{F}(x)\overline{F}(y) \quad \text{for all } x, y \ge 0$$

where $\overline{F}(y) = 1 - F(y)$ (*Microelectronics and Reliability*, Jan. 1986). Alternatively, a "life" distribution $F(y)$ is **new worse than used** (NWU) if

$$\overline{F}(x + y) \ge \overline{F}(x)\overline{F}(y) \quad \text{for all } x, y \ge 0$$

a. Consider the density function

$$f(y) = \begin{cases} y/2 & \text{if } 0 < y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the "life" distribution, $F(y)$.
b. Determine whether the "life" distribution $F(y)$ is NBU or NWU.

## 5.3 Expected Values for Continuous Random Variables

You will recall from your study of calculus that integration is a summation process. Thus, finding the integral

$$F(y_0) = \int_{-\infty}^{y_0} f(t)\, dt$$

for a continuous random variable is analogous to finding the sum

$$F(y_0) = \sum_{y \le y_0} p(y)$$

for a discrete random variable. Then it is natural to employ the same definitions for the expected value of a continuous random variable $Y$, for the expected value of a function $g(Y)$, and for the variance of $Y$ that were given for a discrete random variable in Section 4.3. The only difference is that we will substitute the integration symbol for the summation symbol. It also can be shown (proof omitted) that the expectation theorems of Section 4.4 hold for continuous random variables. We now summarize these definitions and theorems, and present some examples of their use.

**Definition 5.4**

Let $Y$ be a continuous random variable with density function $f(y)$, and let $g(Y)$ be any function of $Y$. Then the **expected values** of $Y$ and $g(Y)$ are

$$E(Y) = \int_{-\infty}^{\infty} yf(y)\,dy$$

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)\,dy$$

**THEOREM 5.1**

Let $c$ be a constant, let $Y$ be a continuous random variable, and let $g_1(Y), g_2(Y), \ldots, g_k(Y)$ be $k$ functions of $Y$. Then,

$$E(c) = c$$

$$E(cY) = cE(Y)$$

$$E[g_1(Y) + g_2(Y) + \cdots + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + \cdots + E[g_k(Y)]$$

**THEOREM 5.2**

Let $Y$ be a continuous random variable with $E(Y) = \mu$. Then

$$\sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2$$

**Example 5.3**

Finding $\mu$ and $\sigma$—Microwave Magnetrons

Refer to Example 5.1 (p. 190). Find the mean and standard deviation for the proportion $Y$ of magnetrons in a large shipment of microwave ovens that are improperly installed. Give a practical interpretation of the mean.

Solution

Recall that $f(y) = 2y$. Therefore,

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)\,dy = \int_0^1 y(2y)\,dy = \int_0^1 2y^2\,dy = \left.\frac{2y^3}{3}\right]_0^1 = \frac{2}{3}$$

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y)\,dy = \int_0^1 y^2(2y)\,dy = \int_0^1 2y^3\,dy = \left.\frac{2y^4}{4}\right]_0^1 = \frac{1}{2}$$

Then, by Theorem 5.2,

$$\sigma^2 = E(y^2) - \mu^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = .0556$$

and thus

$$\sigma = \sqrt{.0556} = .24$$

Our interpretation of $\mu = E(Y)$ is that on average, 2/3 of the magnetrons in a large shipment will be improperly installed. We interpret $\sigma$ in the next example.

## Example 5.4

**Finding a Probability—Microwave Magnetrons**

Refer to Examples 5.1 and 5.3. The interval $\mu \pm 2\sigma$ is shown on the graph of $f(y)$ in Figure 5.7. Find $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.

**Solution**

From Example 5.3, we have $\mu = \frac{2}{3} \approx .67$ and $\sigma = .24$. Therefore, $\mu - 2\sigma = .19$ and $\mu + 2\sigma = 1.15$. Since $P(Y > 1) = 0$, we want to find the probability $P(.19 < Y < 1)$, corresponding to the shaded area in Figure 5.7:

$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = P(.19 < Y < 1) = \int_{.19}^{1} f(y)\,dy$$

$$= \int_{.19}^{1} 2y\,dy = y^2\Big]_{.19}^{1} = 1 - (.19)^2 = .96$$

Therefore, the probability that a large shipment contains a proportion of improperly installed magnetrons between .19 and 1.0 is .96.

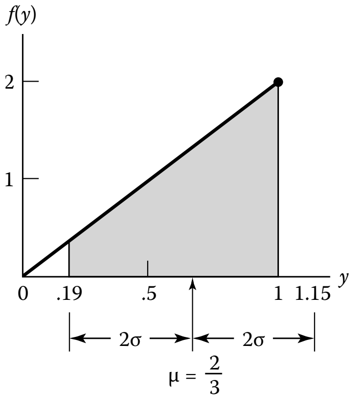



**FIGURE 5.7**
Graph showing the interval $\mu \pm 2\sigma$ for $f(y) = 2y$

In Chapter 1, we applied the Empirical Rule to mound-shaped relative frequency distributions of data. The Empirical Rule may also be applied to mound-shaped theoretical—i.e., probability—distributions. As examples in the preceding chapters demonstrate, the percentage (or proporti on) of a data set in the interval $\mu \pm 2\sigma$ is usually very close to .95, the value specified by the Empirical Rule. This is certainly true for the probability distribution considered in Example 5.4.

## Example 5.5

**Finding $\mu$ and $\sigma$—Extracting Lead from a Shredder**

Suppose the amount $Y$ of extractable lead (measured in miligrams per liter) in a metal shredder residue is a continuous random variable with probability density function

$$f(y) = \begin{cases} \dfrac{e^{-y/2}}{2} & \text{if } 0 \le y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Find the mean, variance, and standard deviation of $Y$. (This density function is known as the **exponential probability distribution**.)

**Solution**

The mean of the random variable $Y$ is given by

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)\,dy = \int_{0}^{\infty} \frac{ye^{-y/2}}{2}\,dy$$

To compute this definite integral, we use the following general formula, found in most mathematical handbooks:*

$$\int ye^{ay}\,dy = \frac{e^{ay}}{a^2}(ay - 1)$$

By substituting $a = -\frac{1}{2}$, we obtain

$$\mu = \frac{1}{2}(4) = 2$$

Thus, the average amount of extractable lead is 2 milligrams per liter of metal shredder residue.

*See, for example, *Standard Mathematical Tables* (1969). Otherwise, the result can be derived using integration by parts:

$$\int ye^{ay}\,dy = \frac{ye^{ay}}{a} - \int \frac{e^{ay}}{a}\,dy$$

To find $\sigma^2$, we will first find $E(Y^2)$ by making use of the general formula†

$$\int y^m e^{ay}\,dy = \frac{y^m e^{ay}}{a} - \frac{m}{a}\int y^{m-1}e^{ay}\,dy$$

Then with $a = -\frac{1}{2}$ and $m = 2$, we can write

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y)\,dy = \int_0^{\infty} \frac{y^2 e^{-y/2}}{2}\,dy = \frac{1}{2}(16) = 8$$

Thus, by Theorem 5.2,

$$\sigma^2 = E(Y^2) - \mu^2 = 8 - (2)^2 = 4$$

and

$$\sigma = \sqrt{4} = 2$$

## Example 5.6

**Finding a Probability—Extracting Lead from a Shredder**

A graph of the density function of Example 5.5 is shown in Figure 5.8. Find $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.

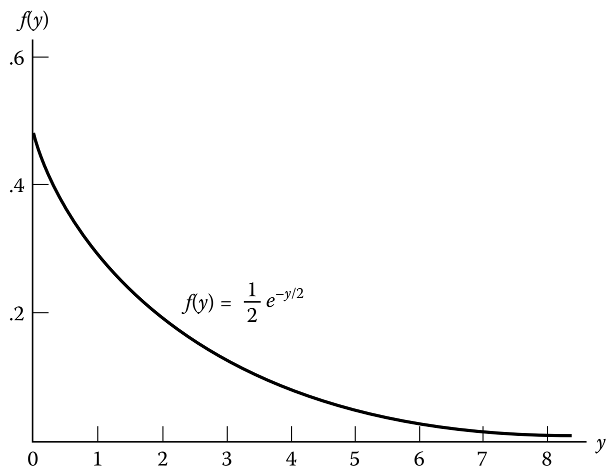
**Solution**

We showed in Example 5.5 that $\mu = 2$ and $\sigma = 2$. Therefore, $\mu - 2\sigma = 2 - 4 = -2$ and $\mu + 2\sigma = 6$. Since $f(y) = 0$ for $y < 0$,

$$\begin{aligned} P(\mu - 2\sigma < Y < \mu + 2\sigma) &= \int_0^6 f(y)\,dy = \int_0^6 \frac{e^{-y/2}}{2}\,dy \\ &= -e^{-y/2}\Big]_0^6 = 1 - e^{-3} \\ &= 1 - .049787 = .950213 \end{aligned}$$

The Empirical Rule of Chapter 2 would suggest that a good approximation to this probability is .95. You can see that for the exponential density function, the approximation is very close to the exact probability, .950213.

**FIGURE 5.8**
Graph of the density function of Example 5.5




†This result is also derived using integration by parts.

In many practical situations, we will know the variance (or standard deviation) of a random variable $Y$ and will want to find the standard deviation of $(c + Y)$ or $cY$, where $c$ is a constant. For example, we might know the standard deviation of the weight $Y$ in ounces of a particular type of computer chip and want to find the standard deviation of the weight in grams. Since 1 ounce = 28.35 grams, we would want to find the standard deviation of $cy$, where $c = 28.35$. The variances of $(c + Y)$ and $cY$ are given by Theorem 5.3.

**THEOREM 5.3**

Let $Y$ be a random variable* with mean $\mu$ and variance $\sigma^2$. Then the variances of $(c + Y)$ and $cY$ are

$$V(c + Y) = \sigma^2_{(c+Y)} = \sigma^2 \quad \text{and} \quad V(cY) = \sigma^2_{cY} = c^2\sigma^2$$

**Proof of Theorem 5.3** From Theorem 5.1, we know that $E(cY) = cE(Y) = c\mu$. Using the definition of the variance of a random variable, we can write

$$V(cY) = \sigma^2_{cY} = E[(cY - c\mu)^2] = E\{[c(Y - \mu)]^2\} = E[c^2(Y - \mu)^2]$$

Then, by Theorem 5.1,

$$\sigma^2_{cY} = c^2E[(Y - \mu)^2]$$

But, $E[(Y - \mu)^2] = \sigma^2$. Therefore,

$$\sigma^2_{cY} = c^2\sigma^2$$

Now, let's apply Theorem 5.3 to the computer chip example. Suppose the variance of the weight $Y$ of the chip is 1.1 (ounces)$^2$. Then the variance of the weight in grams is equal to $(28.35)^2(1.1) = 884.1$ (grams)$^2$. Also, the standard deviation of the weight in grams is $\sqrt{884.1} = 29.7$ grams.

## Applied Exercises

5.10 *Time a train is late.* Refer to Exercise 5.5 (p. 191) The amount of time $Y$ (in minutes) that a commuter train is late is a continuous random variable with probability density

$$f(y) = \begin{cases} \frac{3}{500}(25 - y^2) & \text{if } -5 < y < 5 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the mean and variance of the amount of time in minutes the train is late.
b. Find the mean and variance of the amount of time in hours the train is late.
c. Find the mean and variance of the amount of time in seconds the train is late.

5.11 *Coastal sea level rise.* Refer to the *Journal of Waterway, Port, Coastal, and Ocean Engineering* (March/April, 2013) study of coastal sea level rise, Exercise 5.6 (p. 191). Recall that the acceleration $Y$ in sea level rise (standardized between 0 and 1) was modeled using the following density function:

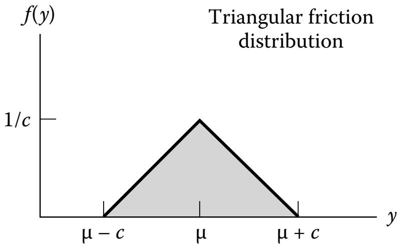
$$f(y) = 6y(1 - y), 0 < y < 1$$

a. Find $E(Y)$. Interpret the result.
b. Find the variance of $Y$.
c. Use the Empirical Rule to estimate $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.
d. Find the actual probability, $P(\mu - 2\sigma < Y < \mu + 2\sigma)$. How does the answer compare to the result in part c?

5.12 *Photocopier friction.* Researchers at the University of Rochester studied the friction that occurs in the paper-feeding process of a photocopier (*Journal of Engineering*

*This theorem applies to discrete or continuous random variables.

*for Industry*, May 1993). The coefficient of friction is a proportion that measures the degree of friction between two adjacent sheets of paper in the feeder stack. In one experiment, a triangular distribution was used to model the friction coefficient, $Y$. (See the accompanying figure.)




The density function for the triangular friction distribution is given by

$$f(y) = \begin{cases} \dfrac{(c-\mu)+y}{c^2} & \text{if } \mu - c < y < \mu \\ \dfrac{(c+\mu)-y}{c^2} & \text{if } \mu < y < \mu + \text{c} \\ 0 & \text{elsewhere} \end{cases}$$

where $c > 0$.

a. Show that $\int_{-\infty}^{\infty} f(y)\,dy = 1$
b. Find the mean of the triangular friction distribution.
c. Find the variance of the triangular friction distribution.

5.13 *Earthquake recurrence in Iran.* Refer to the *Journal of Earthquake Engineering* (Vol. 17, 2013) study of the time $Y$ (in years) between major earthquakes occurring in the Iranian Plateau, Exercise 5.7 (p. 191). Recall that $Y$ has the following density function: $f(y) = .04e^{-.04y}$, $y > 0$.
a. Find $E(Y)$. Interpret the result.
b. Find the variance of $Y$.
c. Use the Empirical Rule to estimate $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.
d. Find the actual probability, $P(\mu - 2\sigma < Y < \mu + 2\sigma)$. How does the answer compare to the result in part c?

## Theoretical Exercises

5.14 For each of the following exercises, find $\mu$ and $\sigma^2$. Then compute $P(\mu - 2\sigma < Y < \mu + 2\sigma)$ and compare to the Empirical Rule.
a. Exercise 5.1
b. Exercise 5.2
c. Exercise 5.3
d. Exercise 5.4

5.15 Prove Theorem 5.1.

5.16 Prove Theorem 5.2.

# 5.4 The Uniform Probability Distribution

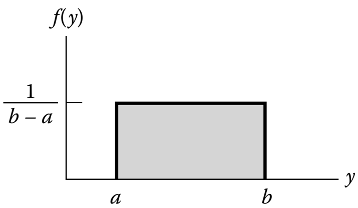



**FIGURE 5.9**
Uniform density function

Suppose you were to randomly select a number $Y$ represented by a point in the interval $a \le Y \le b$. The density function of $Y$ is represented graphically by a rectangle, as shown in Figure 5.9. Notice that the height of the rectangle is $1/(b - a)$ to ensure that the area under the rectangle equals 1.

A random variable of the type shown in Figure 5.9 is called a **uniform random variable**; its density function, mean, and variance are shown in the next box.

### The Uniform Probability Distribution

The probability density function for a **uniform random variable**, $Y$, is given by

$$f(y) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le y \le b \\ 0 & \text{elsewhere} \end{cases}$$

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

**Example 5.7**

**Uniform Distribution—Steel Sheet Thickness**

Suppose the research department of a steel manufacturer believes that one of the company's rolling machines is producing sheets of steel of varying thickness. The thickness $Y$ is a uniform random variable with values between 150 and 200 millimeters. Any sheets less than 160 millimeters thick must be scrapped, since they are unacceptable to buyers.

a. Calculate the mean and standard deviation of $Y$, the thickness of the sheets produced by this machine. Then graph the probability distribution, and show the mean on the horizontal axis. Also show 1 and 2 standard deviation intervals around the mean.

b. Calculate the fraction of steel sheets produced by this machine that have to be scrapped.

Solution

a. To calculate the mean and standard deviation for $Y$, we substitute 150 and 200 millimeters for $a$ and $b$, respectively, in the formulas. Thus,

$$\mu = \frac{a+b}{2} = \frac{150+200}{2} = 175 \text{ millimeters}$$

and

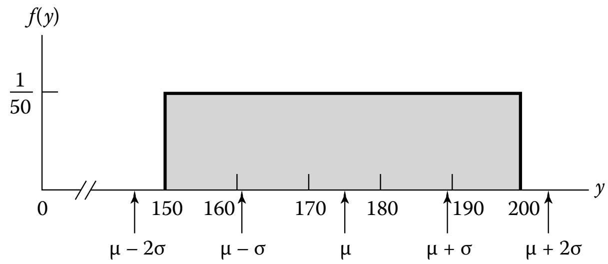
$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{200-150}{\sqrt{12}} = \frac{50}{3.464} = 14.43 \text{ millimeters}$$

The uniform probability distribution is

$$f(y) = \frac{1}{b-a} = \frac{1}{200-150} = \frac{1}{50}$$

The graph of this function is shown in Figure 5.10. The mean and 1 and 2 standard deviation intervals around the mean are shown on the horizontal axis.

**FIGURE 5.10**
Frequency function for $Y$ in Example 5.7

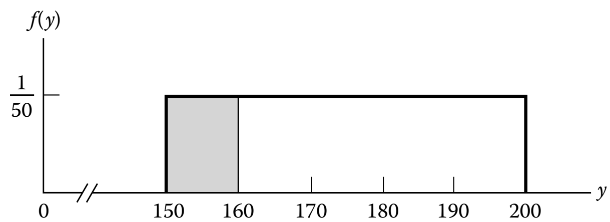



b. To find the fraction of steel sheets produced by the machine that have to be scrapped, we must find the probability that $y$, the thickness, is less than 160 millimeters. As indicated in Figure 5.11, we need to calculate the area under the frequency function $f(y)$ between the points $a = 150$ and $c = 160$. This is the area of a rectangle with base $160 - 150 = 10$ and height $\frac{1}{50}$. The fraction that has to be scrapped is then

$$P(Y < 160) = \text{(Base)(Height)} = (10)\left(\frac{1}{50}\right) = \frac{1}{5}$$

That is, 20% of all the sheets made by this machine must be scrapped.

**FIGURE 5.11**
The probability that the sheet thickness, $Y$, is between 150 and 160 millimeters



The random numbers in Table 1 of Appendix B were generated by a computer program that randomly selects values of $y$ from a uniform distribution. (However, the random numbers are terminated at some specified decimal place.) One of the most important applications of the uniform distribution is described in Chapter 7, where, along with a computer program that generates random numbers, we will use it to simulate the sampling of many other types of random variables.

## Applied Exercises

5.17 *Uranium in the Earth's crust.* The *American Mineralogist* (October 2009) published a study of the evolution of uranium minerals in the Earth's crust. Researchers estimate that the trace amount of uranium $Y$ in reservoirs follows a uniform distribution ranging between 1 and 3 parts per million.

a. Find $E(Y)$ and interpret its value.
b. Compute $P(2 < Y < 2.5)$.
c. Compute $P(Y \leq 1.75)$.

5.18 *Requests to a Web server.* According to Brighton Webs LTD, a British company that specializes in data analysis, the arrival time of requests to a Web server within each hour can be modeled by a uniform distribution. (www.brighton-webs.co.uk.) Specifically, the number of seconds $Y$ from the start of the hour that the request is made is uniformly distributed between 0 and 3,600 seconds. Find the probability that a request is made to a Web server sometime during the last 15 minutes of the hour.

5.19 *Load on timber beams.* Timber beams are widely used in home construction. When the load (measured in pounds) per unit length has a constant value over part of a beam the load is said to be uniformly distributed over that part of the beam. Uniformly distributed beam loads were used to derive the stiffness distribution of the beam in the *American Institute of Aeronautics and Astronautics Journal* (May, 2013). Consider a cantilever beam with a uniformly distributed load between 100 and 115 pounds per linear foot. Find a value $L$ such that the probability that the beam load exceeds $L$ is only .1.

5.20 *Maintaining pipe wall temperature.* Maintaining a constant pipe wall temperature in some hot process applications is critical. A new technique that utilizes bolt-on trace elements to maintain temperature was presented in the *Journal of Heat Transfer* (November 2000). Without bolt-on trace elements, the pipe wall temperature of a switch condenser used to produce plastic has a uniform distribution ranging from 260° to 290°F. When several bolt-on trace elements are attached to the piping, the wall temperature is uniform from 278° to 285°F.

a. Ideally, the pipe wall temperature should range between 280° and 284°F. What is the probability that the temperature will fall in this ideal range when no bolt-on trace elements are used? When bolt-on trace elements are attached to the pipe?
b. When the temperature is 268°F or lower, the hot liquid plastic hardens (or plates), causing a buildup in the piping. What is the probability of plastic plating when no bolt-on trace elements are used? When bolt-on trace elements are attached to the pipe?

5.21 *Cycle availability of a system.* In the jargon of system maintenance, "cycle availability" is defined as the probability that the system is functioning at any point in time. The United States Department of Defense developed a series of performance measures for assessing system cycle availability (*START*, Vol. 11, 2004). Under certain assumptions about the failure time and maintenance time of a system, cycle availability is shown to be uniformly distributed between 0 and 1. Find the following parameters for cycle availability: mean, standard deviation, 10th percentile, lower quartile, and upper quartile. Interpret the results.

5.22 *Trajectory of an electric circuit.* Researchers at the University of California–Berkeley have designed, built, and tested a switched-capacitor circuit for generating random signals (*International Journal of Circuit Theory and Applications*, May–June 1990). The circuit's trajectory was shown to be uniformly distributed on the interval (0, 1).

a. Give the mean and variance of the circuit's trajectory.
b. Compute the probability that the trajectory falls between .2 and .4.
c. Would you expect to observe a trajectory that exceeds .995? Explain.

5.23 *Gouges on a spindle.* A tool-and-die machine shop produces extremely high-tolerance spindles. The spindles are 18-inch slender rods used in a variety of military equipment. A piece of equipment used in the manufacture of the spindles malfunctions on occasion and places a single gouge somewhere on the spindle. However, if the spindle can be cut so that it has 14 consecutive inches without a gouge, then the spindle can be salvaged for other purposes. Assuming that the location of the gouge along the spindle is best described by a uniform distribution, what is the probability that a defective spindle can be salvaged?

5.24 *Reliability of a robotic device.* The *reliability* of a piece of equipment is frequently defined to be the probability, $P$, that the equipment performs its intended function successfully for a given period of time under specific conditions. (Render and Heizer, *Principles of Operations Management*, 2013.) Because $P$ varies from one point in time to another, some reliability analysts treat $P$ as if it were a random variable.

Suppose an analyst characterizes the uncertainty about the reliability of a particular robotic device used in an automobile assembly line using the following distribution:

$$f(p) = \begin{cases} 1 & 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a. Graph the analyst's probability distribution for $P$.
b. Find the mean and variance of $P$.
c. According to the analyst's probability distribution for $P$, what is the probability that $P$ is greater than .95? Less than .95?
d. Suppose the analyst receives the additional information that $P$ is definitely between .90 and .95, but that there is complete uncertainty about where it lies between these values. Describe the probability distribution the analyst should now use to describe $P$.

## Theoretical Exercises

5.25 Statistical software packages, such as SAS and MINITAB, are capable of generating random numbers from a uniform distribution. For example, the SAS function RANUNI uses a prime modulus multiplicative generator with modulus $2^{31} - 1$ and multiplier 397,204,094 to generate a random variable $Y$ from a uniform distribution on the interval (0, 1). This function can be used to generate a uniform random variable on any interval $(a, b)$, where $a$ and $b$ are constants, using an appropriate transformation.

a. Show that the random variable $W = bY$ is uniformly distributed on the interval $(0, b)$.
b. Find a function of $Y$ that will be uniformly distributed on the interval $(a, b)$.

5.26 Assume that the random variable $Y$ is uniformly distributed over the interval $a \le Y \le b$. Verify the following:

a. $\mu = \dfrac{a + b}{2}$ and $\sigma^2 = \dfrac{(b - a)^2}{12}$

b. $F(y) = \begin{cases} \dfrac{y - a}{b - a} & \text{if } a \le y \le b \\ 0 & \text{if } y < a \\ 1 & \text{if } y > b \end{cases}$

5.27 Show that the uniform distribution is *new better than used* (NBU) over the interval (0, 1). (See Optional Exercise 5.9, p. 191, for the definition of NBU.)

5.28 Assume that $Y$ is uniformly distributed over the interval $0 \le Y \le 1$. Show that, for $a \ge 0$, $b \ge 0$, and $(a + b) \le 1$,

$$P(a < Y < a + b) = b$$

# 5.5 The Normal Probability Distribution

The **normal** (or **Gaussian**) **density function** was proposed by C. F. Gauss (1777–1855) as a model for the relative frequency distribution of *errors*, such as errors of measurement. Amazingly, this bell-shaped curve provides an adequate model for the relative frequency distributions of data collected from many different scientific areas and, as we will show in Chapter 7, it models the probability distributions of many statistics that we will use for making inferences. For example, driver reaction time to a brake signal (transportation engineering), concrete cover depth of a bridge column (civil engineering), offset voltage of an amplifier (electrical engineering), transmission delay of a wireless device (computer engineering), and the friction produced from a feed paper copier (industrial engineering) are all random variables that have been shown by researchers to have an approximately normal distribution.

The normal random variable possesses a density function characterized by two parameters. This density function, its mean, and its variance are shown in the box.

### The Normal Probability Distribution

The density function for a **normal random variable**, $Y$, is given by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)} \qquad -\infty < y < \infty$$

The parameters $\mu$ and $\sigma^2$ are the mean and variance, respectively, of the normal random variable $Y$.

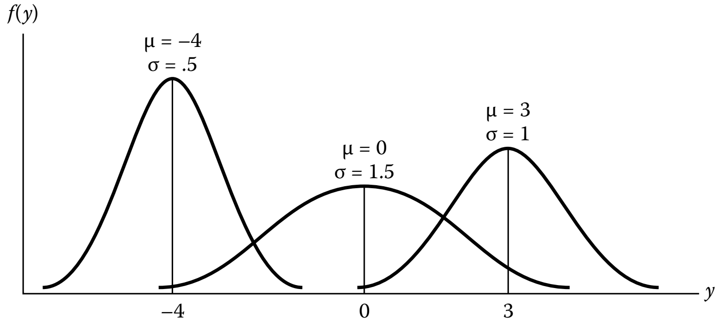



**FIGURE 5.12**
Several normal distributions, with different means and standard deviations

There is an infinite number of normal density functions—one for each combination of $\mu$ and $\sigma$. The mean $\mu$ measures the location of the distribution, and the standard deviation $\sigma$ measures its spread. Several different normal density functions are shown in Figure 5.12.

A closed-form expression cannot be obtained for the integral of the normal density function. However, areas under the normal curve can be obtained by using approximation procedures and Theorem 5.4.

**THEOREM 5.4**

If $Y$ is a normal random variable with mean $\mu$ and variance $\sigma^2$, then $Z = (Y - \mu)/\sigma$ is a normal random variable with mean 0 and variance 1.* The random variable $Z$ is called a **standard normal variable**.

The areas for the standard normal variable,

$$Z = \frac{Y - \mu}{\sigma}$$

are given in Table 5 of Appendix B. Recall from Section 2.6 that $Z$ is the distance between the value of the normal random variable $Y$ and its mean $\mu$, measured in units of its standard deviation $\sigma$.

The entries in Table 5 of Appendix B are the areas under the normal curve between the mean, $Z = 0$, and a value of $Z$ to the right of the mean (see Figure 5.13). To find the area under the normal curve between $Z = 0$ and, say, $Z = 1.33$, move down the left column of Table 5 to the row corresponding to $Z = 1.3$. Then move

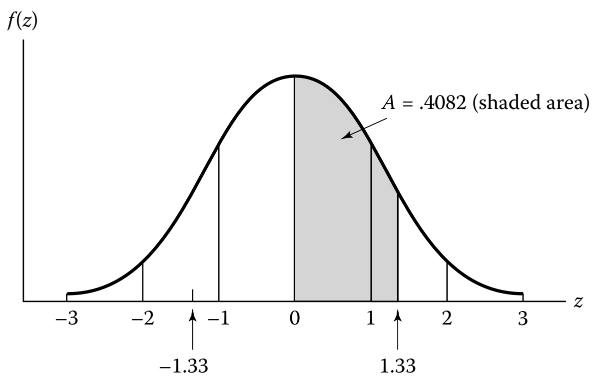



**FIGURE 5.13**
Standard normal density function showing the tabulated areas given in Table 5 of Appendix B

*The proof that the mean and variance of $Z$ are 0 and 1, respectively, is left as a theoretical exercise.

across the top of the table to the column marked .03. The entry at the intersection of this row and column gives the area $A = .4082$. Because the normal curve is symmetric about the mean, areas to the left of the mean are equal to the corresponding areas to the right of the mean. For example, the area $A$ between the mean $Z = 0$ and $Z = -.68$ is equal to the area between $Z = 0$ and $Z = .68$. This area will be found in Table 5 at the intersection of the row corresponding to 0.6 and the column corresponding to .08 as $A = .2517$.

**Example 5.8**

Finding Normal Probabilities

Suppose $Y$ is a normally distributed random variable with mean 10 and standard deviation 2.1.

a. Find $P(Y \geq 11)$.

b. Find $P(7.6 \leq Y \leq 12.2)$.

Solution

a. The value $Y = 11$ corresponds to a $Z$ value of

$$Z = \frac{Y - \mu}{\sigma} = \frac{11 - 10}{2.1} = .48$$

and thus, $P(Y \geq 11) = P(Z \geq .48)$. The area under the standard normal curve corresponding to this probability is shaded in Figure 5.14. Since the normal curve is symmetric about $Z = 0$ and the total area beneath the curve is 1, the area to the right of $Z = 0$ is equal to .5. Thus, the shaded area is equal to $(.5 - A)$, where $A$ is the tabulated area corresponding to $z = .48$. The area $A$, given in Table 5 of Appendix B, is .1844. Therefore,

$$P(Y \geq 11) = .5 - A = .5 - .1844 = .3156$$

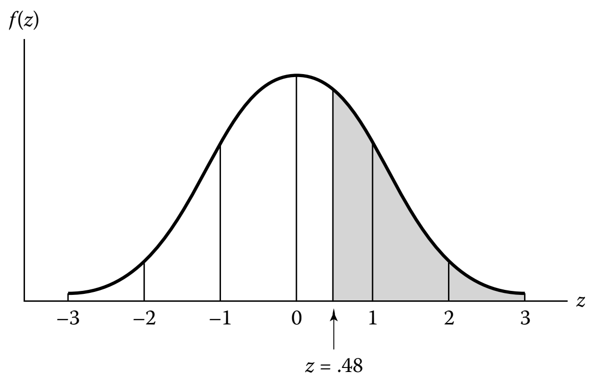



**FIGURE 5.14**
Standard normal distribution for Example 5.8; shaded area is $P(Y \geq 11)$

b. The values $Y_1 = 7.6$ and $Y_2 = 12.2$ correspond to the $Z$ values

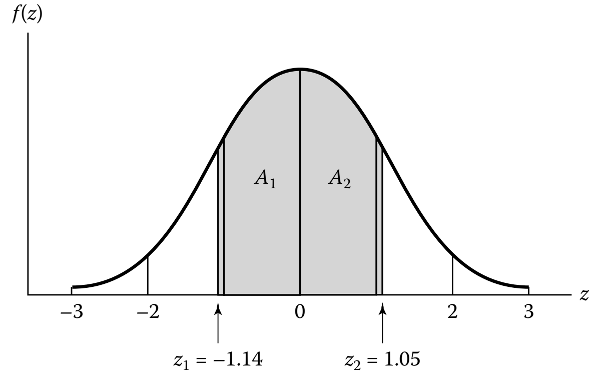
$$Z_1 = \frac{Y_1 - \mu}{\sigma} = \frac{7.6 - 10}{2.1} = -1.14$$

$$Z_2 = \frac{Y_2 - \mu}{\sigma} = \frac{12.2 - 10}{2.1} = 1.05$$

The probability $P(7.6 \leq Y \leq 12.2) = P(-1.14 \leq Z \leq 1.05)$ is the shaded area shown in Figure 5.15. It is equal to the sum of $A_1$ and $A_2$, the areas corresponding to $Z_1$ and $Z_2$, respectively, where $A_1 = .3729$ and $A_2 = .3531$. Therefore,

$$P(7.6 \leq Y \leq 12.2) = A_1 + A_2 = .3729 + .3531 = .7260$$

**FIGURE 5.15**
Standard normal distribution for Example 5.8




**Example 5.9**

Normal Probability—Bitterness Removed from Citrus

The U.S. Department of Agriculture (USDA) patented a process that uses a bacterium for removing bitterness from citrus juices (*Chemical Engineering*, Feb. 3, 1986). In theory, almost all the bitterness could be removed by the process, but for practical purposes the USDA aims at 50% overall removal. Suppose a USDA spokesman claims that the percentage of bitterness removed from an 8-ounce glass of freshly squeezed citrus juice is normally distributed with mean 50.1 and standard deviation 10.4. To test this claim, the bitterness removal process is applied to a randomly selected 8-ounce glass of citrus juice. Assuming the claim is true, find the probability that the process removes less than 33.7% of the bitterness.

Solution

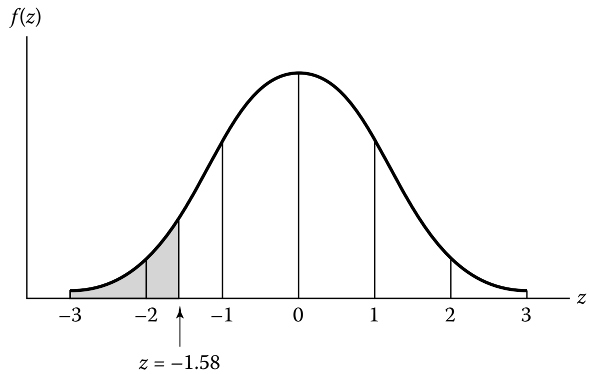
The value $Y = 33.7$ corresponds to the value of the standard normal random variable:

$$Z = \frac{Y - \mu}{\sigma} = \frac{33.7 - 50.1}{10.4} = -1.58$$

Therefore, $P(Y \leq 33.7) = P(Z \leq -1.58)$, the shaded area in Figure 5.16, is equal to .5 minus the area $A$ that corresponds to $Z = 1.58$. Then, the probability that the process removes less than 33.7% of the bitterness is

$$P(Y \leq 33.7) = .5 - .4429 = .0571$$

**FIGURE 5.16**
The probability that percentage of bitterness removed is less than 33.7% in Example 5.9




**Example 5.10**

Normal Probability Inference—Bitterness Removed from Citrus

Refer to Example 5.9. If the test on the single glass of citrus juice yielded a bitterness removal percentage of 33.7, would you tend to doubt the USDA spokesman's claim?

Solution

Given the sample information, we have several choices. We could conclude that the spokesman's claim is true, i.e., that the mean percentage of bitterness removed for the new process is 50.1% and that we have just observed a *rare event*, one that would occur with a probability of only .0571. Or, we could conclude that the spokesman's claim for the mean percentage is too high, i.e., that the true mean is less than 50.1%. Or, perhaps the assumed value of $\sigma$ or the assumption of normality may be in error. Given a choice, we think you will agree that there is reason to doubt the USDA spokesman's claim.

In the last example of this section, we demonstrate how to find a specific value of a normal random variable based on a given probability.

**Example 5.11**

Finding a Value of the Normal Random Variable—Six Sigma Application

*Six Sigma* is a comprehensive approach to quality goal setting that involves statistics. The use of the normal distribution in Six Sigma goal setting at Motorola Corp. was demonstrated in *Aircraft Engineering and Aerospace Technology* (Vol. 76, 2004). Motorola discovered that the defect rate, $Y$, for parts produced on an assembly line varies according to a normal distribution with $\mu = 3$ defects per million and $\sigma = .5$ defect per million. Assume that Motorola's quality engineers want to find a target defect rate, $t$, such that the actual defect rate will be no greater than $t$ on 90% of the runs. Find the value of $t$.

Solution

Here, we want to find $t$ such that $P(Y < t) = .90$. Rewriting the probability as a function of the standard normal random variable Z and substituting the values of $\mu$ and $\sigma$, we have

$$P(Y < t) = P\{(Y - \mu)/\sigma < (t - 3)/.5\} = P\{Z < (t - 3)/.5\} = .90$$

This probability is illustrated in Figure 5.17. Note that the value of $Z$ we need to find, $z_t$, cuts off an area of .10 in the upper tail of the standard normal distribution. This corresponds to an area of .40 in Table 5 of Appendix B. Searching the areas in Table 5 for a probability of approximately .40, we find the corresponding standard normal value $z_t = 1.28$. Consequently,

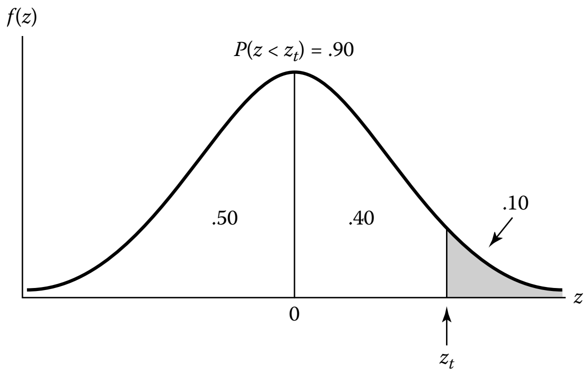
$$z_t = (t - 3)/.5 = 1.28$$

Solving for $t$, we have

$$t = 3 + .5(1.28) = 3.64$$

Therefore, on 90% of the runs the actual defect rate will be no greater than the target defect rate of $t = 3.64$ defects per million.

**FIGURE 5.17**
Probability of Defect Rate Less than Target, Example 5.11




## Applied Exercises

5.29 *Tomato as a taste modifier.* Miraculin—a protein naturally produced in a rare tropical fruit—can convert a sour taste into a sweet taste. Consequently, miraculin has the potential to be an alternative low-calorie sweetener. In *Plant Science* (May, 2010), a group of Japanese environmental engineers investigated the ability of a hybrid tomato plant to produce miraculin. For a particular generation of the tomato plant, the amount $Y$ of miraculin produced (measured in micro-grams per gram of fresh weight) had a mean of 105.3 and a standard deviation of 8.0. Assume that $Y$ is normally distributed.

a. Find $P(Y > 120)$.
b. Find $P(100 < Y < 110)$.
c. Find the value $a$ for which $P(Y < a) = .25$.

5.30 *Voltage sags and swells.* Refer to the *Electrical Engineering* (Vol. 95, 2013) study of the power quality of a transformer, Exercise 2.53 (p. 54). Recall that two causes of poor power quality are "sags" and "swells". (A sag is an unusual dip and a swell is an unusual increase in the voltage level of a transformer.) For Turkish transformers built for heavy industry, the mean number of sags per week was 353 and the mean number of swells per week was 184. As in Exercise 2.53, assume the standard deviation of the sag distribution is 30 sags per week and the standard deviation of the swell distribution is 25 swells per week. Also, assume that the number of sags and number of swells are both normally distributed. Suppose one of the transformers is randomly selected and found to have 400 sags and 100 swells in a week.

a. What is the probability that the number of sags per week is less than 400?
b. What is the probability that the number of swells per week is greater than 100?

5.31 *Transmission delays in wireless technology.* Resource reservation protocol (RSVP) was originally designed to establish signaling links for stationary networks. In *Mobile Networks and Applications* (Dec. 2003), RSVP was applied to mobile wireless technology (e.g., a PC notebook with wireless LAN card for Internet access). A simulation study revealed that the transmission delay (measured in milliseconds) of an RSVP-linked wireless device has an approximate normal distribution with mean $\mu = 48.5$ milliseconds and $\sigma = 8.5$ milliseconds.

a. What is the probability that the transmission delay is less than 57 milliseconds?
b. What is the probability that the transmission delay is between 40 and 60 milliseconds?

5.32 *Natural gas consumption and temperature.* The *Transactions of the ASME* (June 2004) presented a model for predicting daily natural gas consumption in urban areas. A key component of the model is the distribution of daily temperatures in the area. Based on daily July temperatures collected in Buenos Aires, Argentina, from 1944 to 2000, researchers demonstrated that the daily July temperature is normally distributed with $\mu = 11°\text{C}$ and $\sigma = 3.1°\text{C}$. Suppose you want to use temperature to predict natural gas consumption on a future July day in Buenos Aires.

a. An accurate prediction can be obtained if you know the chance of the July temperature falling below 9°C. Find the probability of interest.
b. Give a temperature value that is exceeded on only 5% of the July days in Buenos Aires.

5.33 *Seismic ground noise.* Seismic ground noise describes the persistent vibration of the ground due to surface waves generated from traffic, heavy machinery, winds, ocean waves, and earthquakes. A group of civil engineers investigated the structural damage to a three-story building caused by seismic ground noise in *Earthquake Engineering and Engineering Vibration* (March, 2013). The methodology involved modeling the acceleration $Y$ (in meters per second-squared) of the seismic ground noise using a normal probability distribution. Consider a normal distribution for $Y$ with $\mu = .5\text{ m/s}^2$ and $\sigma = 0.1\text{ m/s}^2$. Find a value of acceleration, $Y = a$, such that $P(Y > a) = .70$.

5.34 *Maintenance of a wind turbine system.* As part of a risk assessment, quality engineers monitored the corrosion rate (millimeters per year) of a wind turbine system susceptible to corrosion. (*Journal of Quality in Maintenance Engineering,* Vol. 18, 2013). For demonstration purposes, the corrosion rate $Y$ was modeled as a normal distribution with $\mu = .4$ mm/year and $\sigma = .1$ mm/year. Would you expect the corrosion rate for a similar wind turbine system to exceed .75 mm/year? Explain.

5.35 *Deep mixing of soil.* Deep mixing is a ground improvement method developed for soft soils like clay, silt, and peat. Swedish civil engineers investigated the properties of soil improved by deep mixing with lime-cement columns in the journal *Giorisk* (Vol. 7, 2013). The mixed soil was tested by advancing a cylindrical rod with a cone tip down into the soil. During penetration, the cone penetrometer measures the cone tip resistance (megapascals, MPa). The researchers established that tip resistance for the deep mixed soil followed a normal distribution with $\mu = 2.2$ MPa and $\sigma = .9$ MPa.

a. Find the probability that the tip resistance will fall between 1.3 and 4.0 MPa.
b. Find the probability that the tip resistance will exceed 1.0 MPa.

5.36 *Alkalinity of river water.* The alkalinity level of water specimens collected from the Han River in Seoul, Korea, has a mean of 50 milligrams per liter and a standard deviation of 3.2 milligrams per liter. (*Environmental Science & Engineering*, Sept. 1, 2000.) Assume the distribution of alkalinity levels is approximately normal and find the probability that a water specimen collected from the river has an alkalinity level

a. exceeding 45 milligrams per liter.
b. below 55 milligrams per liter.
c. between 51 and 52 milligrams per liter.

5.37 *Flicker in an electrical power system.* An assessment of the quality of the electrical power system in Turkey was the topic of an article published in *Electrical Engineering* (March, 2013). One measure of quality is the degree to which voltage fluctuations cause light flicker in the system. The perception of light flicker $Y$ when the system is set at 380 kV was measured periodically (over 10-minute intervals). For transformers supplying heavy industry plants, the light flicker distribution was found to follow (approximately) a normal distribution with $\mu = 2.2\%$ and $\sigma = .5\%$. If the perception of light flicker exceeds 3%, the transformer is shut down and the system is reset. How likely is it for a transformer supplying a heavy industry plant to be shut down due to light flicker?

**CRASH**

5.38 *NHTSA crash safety tests.* Refer to the National Highway Traffic Safety Administration (NHTSA) crash test data for new cars, introduced in Exercise 2.74 (p. 70) and saved in the **CRASH** file. One of the variables measured is the severity of a driver's head injury when the car is in a head-on collision with a fixed barrier while traveling at 35 miles per hour. The more points assigned to the head injury rating, the more severe the injury. The head injury ratings can be shown to be approximately normally distributed with a mean of 605 points and a standard deviation of 185 points. One of the crash-tested cars is randomly selected from the data and the driver's head injury rating is observed.

a. Find the probability that the rating will fall between 500 and 700 points.
b. Find the probability that the rating will fall between 400 and 500 points.
c. Find the probability that the rating will be less than 850 points.
d. Find the probability that the rating will exceed 1,000 points.
e. What rating will only 10% of the crash-tested cars exceed?

5.39 *Industrial filling process.* The characteristics of an industrial filling process in which an expensive liquid is injected into a container were investigated in *Journal of Quality Technology* (July 1999). The quantity injected per container is approximately normally distributed with mean 10 units and standard deviation .2 units. Each unit of fill costs \$20. If a container contains less than 10 units (i.e., is underfilled), it must be reprocessed at a cost of \$10. A properly filled container sells for \$230.

a. Find the probability that a container is underfilled.
b. A container is initially underfilled and must be reprocessed. Upon refilling it contains 10.6 units. How much profit will the company make on this container?
c. The operations manager adjusts the mean of the filling process upward to 10.5 units in order to make the probability of underfilling approximately zero. Under these conditions, what is the expected profit per container?

5.40 *Rock displacement.* Paleomagnetic studies of Canadian volcanic rock known as the Carmacks Group have recently been completed. The studies revealed that the northward displacement of the rock units has an approximately normal distribution with standard deviation of 500 kilometers (*Canadian Journal of Earth Sciences*, Vol. 27, 1990). One group of researchers estimated the mean displacement at 1,500 kilometers, whereas a second group estimated the mean at 1,200 kilometers.

a. Assuming the mean is 1,500 kilometers, what is the probability of a northward displacement of less than 500 kilometers?
b. Assuming the mean is 1,200 kilometers, what is the probability of a northward displacement of less than 500 kilometers?
c. If, in fact, the northward displacement is less than 500 kilometers, which is the more plausible mean, 1,200 or 1,500 kilometers?

## Theoretical Exercise

5.41 Let $Y$ be a normal random variable with mean $\mu$ and variance $\sigma^2$. Show that

$$Z = \frac{Y - \mu}{\sigma}$$

has mean 0 and variance 1. (*Hint*: Apply Theorems 5.1–5.2.)

# 5.6 Descriptive Methods for Assessing Normality

In the chapters that follow, we learn how to make inferences about the population based on information in the sample. Several of these techniques are based on the assumption that the population is approximately normally distributed. Consequently, it will be important to determine whether the sample data come from a normal population before we can properly apply these techniques.

Several descriptive methods can be used to check for normality. In this section, we consider the three methods summarized in the box.

### Determining Whether the Data Are from an Approximately Normal Distribution

1. Construct either a **histogram** or a **stem-and-leaf** display for the data. If the data are approximately normal, the shape of the graph will be similar to the normal curve, Figure 5.12 (i.e., mound-shaped and symmetric around the mean with thin tails).
2. Find the **interquartile range, IQR**, and **standard deviation**, $s$, for the sample, then calculate the ratio IQR/s. If the data are approximately normal, then $\text{IQR}/s \approx 1.3$.
3. Construct a **normal probability plot** for the data. (See the following example.) If the data are approximately normal, the points will fall (approximately) on a straight line.

**Example 5.12**

Assessing Normality—EPA Gas Mileages

The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings (miles per gallon). Table 5.1 lists mileage ratings obtained from 100 tests on a certain new car model. (The data are saved in the **EPAGAS** file.) Numerical and graphical descriptive measures for the 100 mileage ratings are shown in the SAS printouts, Figures 5.18a–c. Determine whether the data have an approximately normal distribution.

**EPAGAS**

**TABLE 5.1 EPA Gas Mileage Ratings for 100 Cars (miles per gallon)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 36.3 | 41.0 | 36.9 | 37.1 | 44.9 | 36.8 | 30.0 | 37.2 | 42.1 | 36.7 |
| 32.7 | 37.3 | 41.2 | 36.6 | 32.9 | 36.5 | 33.2 | 37.4 | 37.5 | 33.6 |
| 40.5 | 36.5 | 37.6 | 33.9 | 40.2 | 36.4 | 37.7 | 37.7 | 40.0 | 34.2 |
| 36.2 | 37.9 | 36.0 | 37.9 | 35.9 | 38.2 | 38.3 | 35.7 | 35.6 | 35.1 |
| 38.5 | 39.0 | 35.5 | 34.8 | 38.6 | 39.4 | 35.3 | 34.4 | 38.8 | 39.7 |
| 36.3 | 36.8 | 32.5 | 36.4 | 40.5 | 36.6 | 36.1 | 38.2 | 38.4 | 39.3 |
| 41.0 | 31.8 | 37.3 | 33.1 | 37.0 | 37.6 | 37.0 | 38.7 | 39.0 | 35.8 |
| 37.0 | 37.2 | 40.7 | 37.4 | 37.1 | 37.8 | 35.9 | 35.6 | 36.7 | 34.5 |
| 37.1 | 40.3 | 36.7 | 37.0 | 33.9 | 40.1 | 38.0 | 35.2 | 34.8 | 39.5 |
| 39.9 | 36.9 | 32.9 | 33.8 | 39.8 | 34.0 | 36.8 | 35.0 | 38.1 | 36.9 |

Solution

As a first check, we examine the relative frequency histogram of the data shown in Figure 5.18a. A normal curve is superimposed on the graph. Clearly, the mileage ratings fall in an approximately mound-shaped, symmetric distribution centered around the mean of about 37 mpg.

Check #2 in the box requires that we find the interquartile range (i.e., the difference between the 75th and 25th percentiles) and the standard deviation of the data set and compute the ratio of these two numbers. The ratio IQR/$s$ for a sample from a normal distribution will approximately equal 1.3.* The values of IQR and $s$, shaded in Figure 5.18b, are IQR $= 2.7$ and $s = 2.42$. Then the ratio is

$$\frac{\text{IQR}}{s} = \frac{2.7}{2.42} = 1.12$$

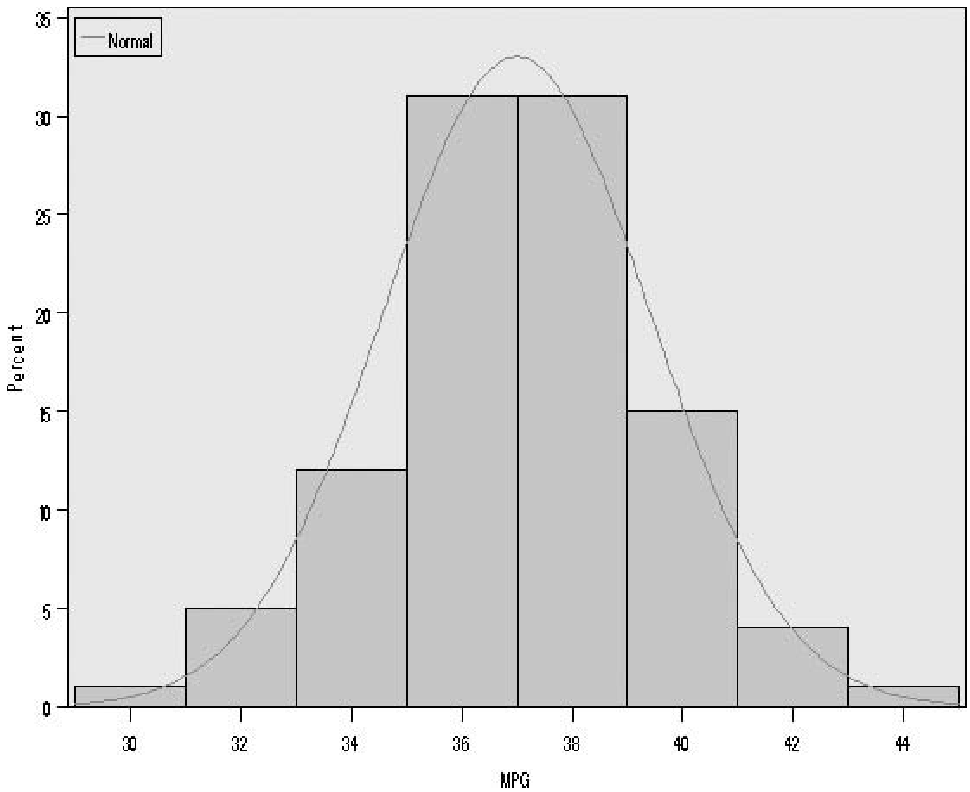
Since this value is approximately equal to 1.3, we have further confirmation that the data are approximately normal.

A third descriptive technique for checking normality is a **normal probability plot**. In a normal probability plot, the observations in the data set are ordered and then plotted against the standardized expected values ($Z$-scores) of the observations under the assumption that the data are normally distributed. When the data are, in fact, normally distributed, an observation will approximately equal its expected value. Thus, a linear (straight-line) trend on the normal probability plot suggests that the data are from an approximate normal distribution, while a nonlinear trend indicates that the data are nonnormal.

Although normal probability plots can be constructed by hand, the process is tedious. It is easier to generate these plots using statistical software. A SAS normal probability plot for the 100 mileage ratings is shown in Figure 5.18c. Notice that the ordered measurements fall reasonably close to a straight line. Thus, check #3 also suggests that the data are likely to be approximately normally distributed.

*You can see that this property holds for normal distributions by noting that the $Z$ values (obtained from Table 5 in Appendix B) corresponding to the 75th and 25th percentiles are .67 and $-.67$, respectively. Since $\sigma = 1$ for a standard normal ($Z$) distribution, IQR/$\sigma = [.67 - (-.67)]/1 = 1.34$.

**FIGURE 5.18**
SAS descriptive statistics for Example 5.12

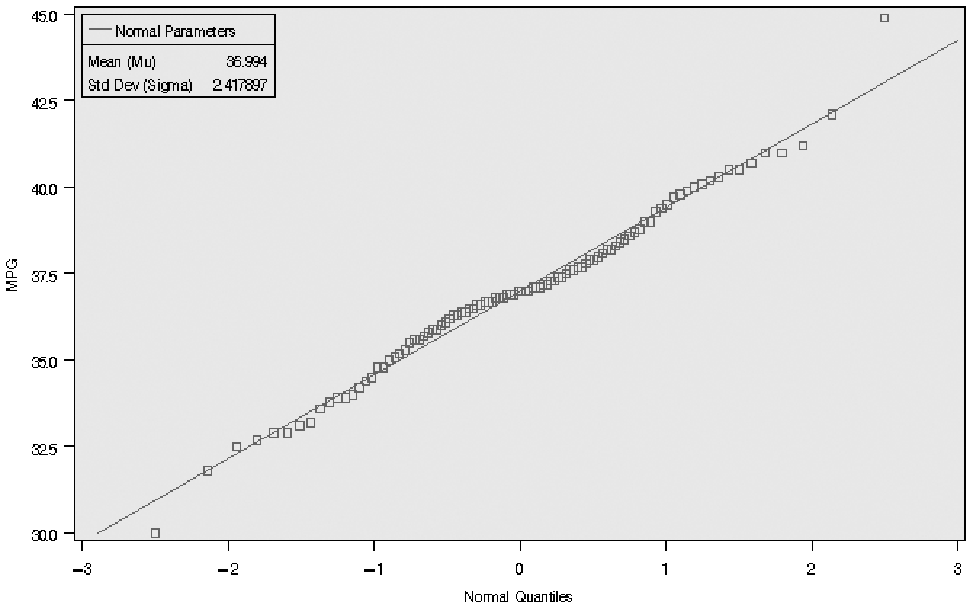



a. SAS histogram for mileage ratings

```
                        The UNIVARIATE Procedure
                           Variable:  MPG

                              Moments

N                         100    Sum Weights              100
Mean                   36.994    Sum Observations      3699.4
Std Deviation      2.41789707    Variance          5.84622626
Skewness           0.05090878    Kurtosis          0.76992269
Uncorrected SS      137434.38    Corrected SS        578.7764
Coeff Variation    6.53591684    Std Error Mean    0.24178971

                      Basic Statistical Measures

         Location                    Variability

     Mean     36.99400     Std Deviation              2.41790
     Median   37.00000     Variance                   5.84623
     Mode     37.00000     Range                     14.90000
                           Interquartile Range        2.70000

                    Quantiles (Definition 5)

                     Quantile        Estimate

                     100% Max         44.90
                     99%              43.50
                     95%              40.85
                     90%              40.15
                     75% Q3           38.35
                     50% Median       37.00
                     25% Q1           35.65
                     10%              33.85
                     5%               32.90
                     1%               30.90
                     0% Min           30.00
```

b. SAS summary statistics for mileage ratings

**FIGURE 5.18 (*Continued*)**
c. SAS normal probability plot for mileage ratings




**Definition 5.5**

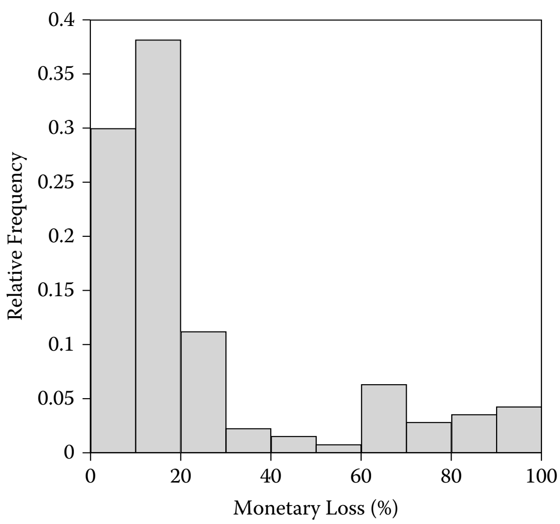
A **normal probability plot** for a data set is a scatterplot with the ranked data values on one axis and their corresponding expected *Z*-scores from a standard normal distribution on the other axis. (*Note*: Computation of the expected standard normal *Z*-scores is beyond the scope of this text. Therefore, we will rely on available statistical software packages to generate a normal probability plot.)

The checks for normality given in the example are simple, yet powerful, techniques to apply, but they are only descriptive in nature. It is possible (although unlikely) that the data are nonnormal even when the checks are reasonably satisfied. Thus, we should be careful not to claim that the 100 mileage ratings in Table 5.1 are, in fact, normally distributed. We can only state that it is reasonable to believe that the data are from a normal distribution.*

## Applied Exercises

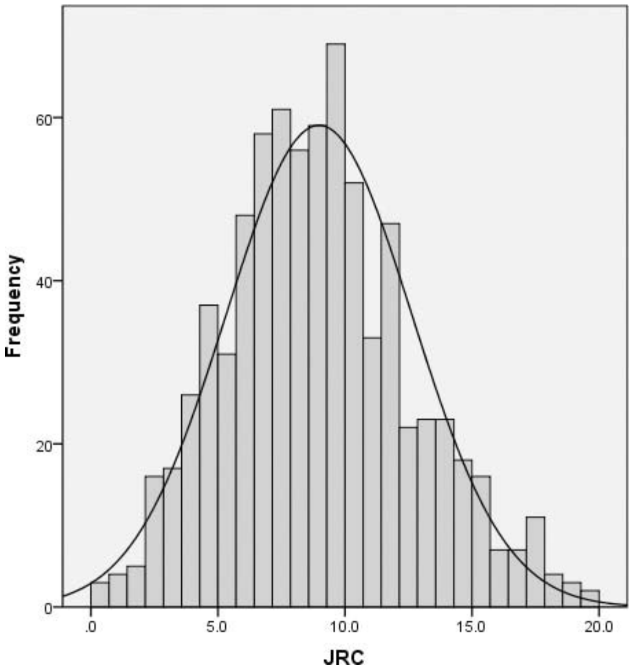
5.42 *Software file updates.* Software configuration management was used to monitor a software engineering team's performance at Motorola, Inc. (*Software Quality Professional*, Nov. 2004). One of the variables of interest was the number of updates to a file changed because of a problem report. Summary statistics for $n = 421$ files yielded the following results: $\bar{y} = 4.71$, $s = 6.09$, $Q_1 = 1$, and $Q_3 = 6$. Are these data approximately normally distributed? Explain.

5.43 *Annual survey of computer crimes.* Refer to the 2010 *CSI Computer Crime and Security Survey,* Exercise 2.13 (p. 35). Recall that the percentage of monetary losses attributable to malicious actions by individuals within the organization (i.e., malicious insider actions) was recorded for 144 firms. The histogram for the data is reproduced here. A researcher wants to analyze the data using a statistical method that is valid only if the data is normally distributed. Should the researcher apply this method to the data?




*Statistical tests of normality that provide a measure of reliability for the inference are available. However, these tests tend to be very sensitive to slight departures from normality, i.e., they tend to reject the hypothesis of normality for any distribution that is not perfectly symmetrical and mound-shaped. Consult the references if you want to learn more about these tests.

5.44 *Shear strength of rock fractures.* Understanding the characteristics of rock masses, especially the nature of the fractures, is essential when building dams and power plants. The shear strength of rock fractures was investigated in *Engineering Geology* (May 12, 2010). The Joint Roughness Coefficient (JRC) was used to measure shear strength. Civil engineers collected JRC data for over 750 rock fractures. The results (simulated from information provided in the article) are summarized in the SPSS histogram shown below. Should the engineers use the normal probability distribution to model the behavior of shear strength for rock fractures? Explain.




5.45 *Drug content assessment.* Scientists at GlaxoSmithKline Medicines Research Center used high-performance liquid chromatography (HPLC) to determine the amount of drug in a tablet produced by the company. (*Analytical Chemistry,* Dec. 15, 2009.) Drug concentrations (measured as a percentage) for 50 randomly selected tablets are listed in the accompanying table and saved in the **DRUGCON** file.

**DRUGCON**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 91.28 | 92.83 | 89.35 | 91.90 | 82.85 | 94.83 | 89.83 | 89.00 | 84.62 |
| 86.96 | 88.32 | 91.17 | 83.86 | 89.74 | 92.24 | 92.59 | 84.21 | 89.36 |
| 90.96 | 92.85 | 89.39 | 89.82 | 89.91 | 92.16 | 88.67 | 89.35 | 86.51 |
| 89.04 | 91.82 | 93.02 | 88.32 | 88.76 | 89.26 | 90.36 | 87.16 | 91.74 |
| 86.12 | 92.10 | 83.33 | 87.61 | 88.20 | 92.78 | 86.35 | 93.84 | 91.20 |
| 93.44 | 86.77 | 83.77 | 93.19 | 81.79 | | | | |

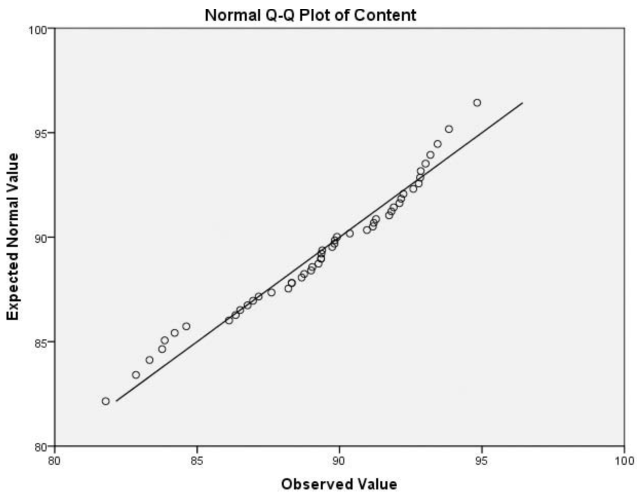
*Source:* Borman, P.J., Marion, J.C., Damjanov,I., & Jackson, P. "Design and analysis of method equivalence studies", *Analytical Chemistry,* Vol. 81, No. 24, December 15, 2009 (Table 3).

a. Descriptive statistics for the drug concentrations are shown at the top of the accompanying SPSS printout. Use this information to assess whether the data are approximately normal.

**Descriptives**

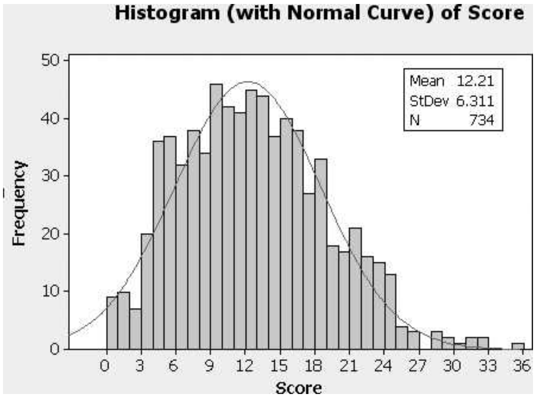
| | | | Statistic | Std. Error |
|---|---|---|---|---|
| Content | Mean | | 89.2906 | .45021 |
| | 95% Confidence Interval for Mean | Lower Bound | 88.3859 | |
| | | Upper Bound | 90.1953 | |
| | 5% Trimmed Mean | | 89.3963 | |
| | Median | | 89.3750 | |
| | Variance | | 10.134 | |
| | Std. Deviation | | 3.18344 | |
| | Minimum | | 81.79 | |
| | Maximum | | 94.83 | |
| | Range | | 13.04 | |
| | Interquartile Range | | 4.84 | |
| | Skewness | | -.544 | .337 |
| | Kurtosis | | -.389 | .662 |

b. An SPSS normal probability plot follows. Use this information to assess whether the data are approximately normal.




**HABITAT**

5.46 *Habitats of endangered species.* An evaluation of the habitats of endangered salmon species was performed in *Conservation Ecology* (December 2003). The researchers identified 734 sites (habitats) for Chinook, coho, or steelhead salmon species in Oregon, and assigned a habitat quality score to each. (Scores range from 0 to 36 points, with lower scores indicating poorly maintained or degraded habitats.) A MINITAB histogram for the data (saved in the **HABITAT** file) is displayed on the next page. Give your opinion on whether the data is normally distributed.




**MINITAB output for Exercise 5.46**

*Source*: Good, T. P., Harms, T. K., and Ruckelshaus, M. H. "Misuse of checklist assessments in endangered species recovery efforts." *Conservation Ecology*, Vol. 7, No. 2, Dec. 2003 (Figure 3).

**SILICA**

Without calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −47.1 | −53.0 | −50.8 | −54.4 | −57.4 | −49.2 | −51.5 | −50.2 | −46.4 | −49.7 |
| −53.8 | −53.8 | −53.5 | −52.2 | −49.9 | −51.8 | −53.7 | −54.8 | −54.5 | −53.3 |
| −50.6 | −52.9 | −51.2 | −54.5 | −49.7 | −50.2 | −53.2 | −52.9 | −52.8 | −52.1 |
| −50.2 | −50.8 | −56.1 | −51.0 | −55.6 | −50.3 | −57.6 | −50.1 | −54.2 | −50.7 |
| −55.7 | −55.0 | −47.4 | −47.5 | −52.8 | −50.6 | −55.6 | −53.2 | −52.3 | −45.7 |

With calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −9.2 | −11.6 | −10.6 | −8.0 | −10.9 | −10.0 | −11.0 | −10.7 | −13.1 | −11.5 |
| −11.3 | −9.9 | −11.8 | −12.6 | −8.9 | −13.1 | −10.7 | −12.1 | −11.2 | −10.9 |
| −9.1 | −12.1 | −6.8 | −11.5 | −10.4 | −11.5 | −12.1 | −11.3 | −10.7 | −12.4 |
| −11.5 | −11.0 | −7.1 | −12.4 | −11.4 | −9.9 | −8.6 | −13.6 | −10.1 | −11.3 |
| −13.0 | −11.9 | −8.6 | −11.3 | −13.0 | −12.2 | −11.3 | −10.5 | −8.8 | −13.4 |

5.47 *Breast height diameters of trees.* Foresters periodically "cruise" a forest to determine the size (usually measured as the diameter at breast height) of a certain species of trees. The breast height diameters (in meters) for a sample of 28 trembling aspen trees in British Columbia's boreal forest are listed here. Determine whether the sample data are from an approximately normal distribution.

**ASPENHTS**

| | | | | | | |
|---|---|---|---|---|---|---|
| 12.4 | 17.3 | 27.3 | 19.1 | 16.9 | 16.2 | 20.0 |
| 16.6 | 16.3 | 16.3 | 21.4 | 25.7 | 15.0 | 19.3 |
| 12.9 | 18.6 | 12.4 | 15.9 | 18.8 | 14.9 | 12.8 |
| 24.8 | 26.9 | 13.5 | 17.9 | 13.2 | 23.2 | 12.7 |

*Source*: Scholz, H. "Fish Creek Community Forest: Exploratory statistical analysis of selected data," working paper, Northern Lights College, British Columbia, Canada.

**CRASH**

5.48 *NHTSA crash tests.* Refer to the National Highway Traffic Safety Administration (NHTSA) crash test data for new cars. In Exercise 5.38 (p. 205), you assumed that the driver's head injury rating is approximately normally distributed. Apply the methods of this chapter to the data saved in the **CRASH** file to support this assumption.

**SHIPSANIT**

5.49 *Cruise ship sanitation scores.* Refer to the data on sanitation scores for 186 cruise ships, first presented in Exercise 2.19 (p. 37). The data are saved in the **SHIPSANIT** file. Assess whether the sanitation scores are approximately normally distributed.

5.50 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercise 2.23 (p. 38). Recall that 50 solutions of deionized water were prepared both with and without calcium/gypsum, and the level of flotation of silica in the solution was measured using a variable called *zeta potential* (measured in millivolts, mV). The data (simulated, based on information provided in the journal article) are reproduced in the table above and saved in the **SILICA** data file. Which of the two zeta potential distributions, without calcium/gypsum or with calcium/gypsum, is better approximated by a normal distribution?

**SANDSTONE**

5.51 *Permeability of sandstone during weathering.* Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercise 2.33 (p. 44). Recall that blocks of sandstone were cut into 300 equal-sized slices and the slices randomly divided into three groups of 100 slices each. Slices in group A were not exposed to any type of weathering; slices in group B were repeatedly sprayed with a 10% salt solution (to simulate wetting by driven rain) under temperate conditions; and, slices in group C were soaked in a 10% salt solution and then dried (to simulate blocks of sandstone exposed during a wet winter and dried during a hot summer). All sandstone slices were then tested for permeability, measured in milliDarcies (mD). The data for the study (simulated) are saved in the **SANDSTONE** file. Is it plausible to assume that the permeability measurements in any of the three experimental groups are approximately normally distributed?

## 5.7 Gamma-Type Probability Distributions

Many random variables, such as the length of the useful life of a laptop computer, can assume only nonnegative values. The relative frequency distributions for data of this type can often be modeled by **gamma-type density functions**. The formulas for a gamma density function, its mean, and its variance are shown in the box.

### The Gamma Probability Distribution

The probability density function for a gamma-type random variable $Y$ is given by

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & \text{if } 0 \leq y < \infty;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\,dy$$

The mean and variance of a gamma-type random variable are, respectively,

$$\mu = \alpha\beta \qquad \sigma^2 = \alpha\beta^2$$

It can be shown (proof omitted) that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ and that $\Gamma(\alpha) = (\alpha - 1)!$ when $\alpha$ is a positive integer. Values of $\Gamma(\alpha)$ for $1.0 \leq \alpha \leq 2.0$ are presented in Table 6 of Appendix B.

The formula for the gamma density function contains two parameters, $\alpha$ and $\beta$. The parameter $\beta$, known as a **scale parameter**, reflects the size of the units in which $Y$ is measured. (It performs the same function as the parameter $\sigma$ that appears in the formula for the normal density function.) The parameter $\alpha$ is known as a **shape parameter**. Changing its value changes the shape of the gamma distribution. This enables us to obtain density functions of many different shapes to model relative frequency distributions of experimental data. Graphs of the gamma density function for $\alpha = 1$, 3, and 5, with $\beta = 1$, are shown in Figure 5.19.

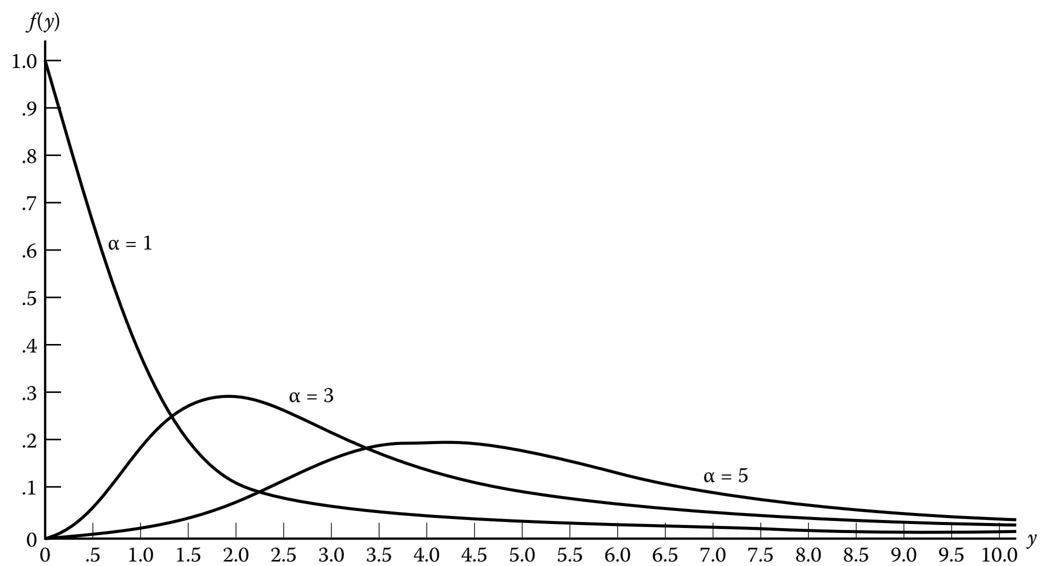



**FIGURE 5.19**
Graphs of gamma density functions for $\alpha = 1$, 3, and 5; $\beta = 1$

Except for the special case where $\alpha$ is an integer, we cannot obtain a closed-form expression for the integral of the gamma density function. Consequently, the cumulative distribution function for a gamma random variable, called an **incomplete gamma function**, must be obtained using approximation procedures with the aid of a computer. Values of this function are given in *Tables of the Incomplete Gamma Function* (1956).

A gamma-type random variable that plays an important role in statistics is the **chi-square random variable**. Chi-square values and corresponding areas under the chi-square density function are given in Table 9 of Appendix B. We will discuss the use of this table in Chapters 7 and 8, when we introduce inferential statistics.

### The Chi-Square Probability Distribution

A **chi-square random variable** is a gamma-type random variable $Y$ with $\alpha = \nu/2$ and $\beta = 2$:

$$f(y) = c(y)^{(\nu/2)-1}e^{-y/2} \quad (0 \leq y^2 < \infty)$$

where

$$c = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$$

The mean and variance of a chi-square random variable are, respectively,

$$\mu = \nu \qquad \sigma^2 = 2\nu$$

The parameter $\nu$ is called the **number of degrees of freedom** for the chi-square distribution.

When $\alpha = 1$, the gamma density function is known as an *exponential distribution.** This important density function is employed as a model for the relative frequency distribution of the length of time between random arrivals at a service counter (computer center, supermarket checkout counter, hospital clinic, etc.) when the probability of a customer arrival in any one unit of time is equal to the probability of arrival during any other. It is also used as a model for the length of life of industrial equipment or products when the probability that an "old" component will operate at least $t$ additional time units, given it is now functioning, is the same as the probability that a "new" component will operate at least $t$ time units. Equipment subject to periodic maintenance and parts replacement often exhibits this property of "never growing old."

The exponential distribution is related to the Poisson probability distribution. In fact, it can be shown (proof omitted) that if the number of arrivals at a service counter follows a Poisson probability distribution with the mean number of arrivals per unit time equal to $1/\beta$, then the density function for the length of time $y$ between any pair of successive arrivals will be an exponential distribution with mean equal to $\beta$, i.e.,

$$f(y) = \frac{e^{-y/\beta}}{\beta} \quad (0 \leq y < \infty)$$

*The exponential distribution was encountered in Examples 5.5 and 5.6 of Section 5.3.

### The Exponential Probability Distribution

An **exponential distribution** is a gamma density function with $\alpha = 1$:

$$f(y) = \frac{e^{-y/\beta}}{\beta} \qquad (0 \le y < \infty)$$

with mean and variance

$$\mu = \beta \qquad \sigma^2 = \beta^2$$

**Example 5.13**

Gamma Distribution Application—Customer Complaints

From past experience, a manufacturer knows that the relative frequency distribution of the length of time $Y$ (in months) between major customer product complaints can be modeled by a gamma density function with $\alpha = 2$ and $\beta = 4$. Fifteen months after the manufacturer tightened its quality control requirements, the first complaint arrived. Does this suggest that the mean time between major customer complaints may have increased?

Solution

We want to determine whether the observed value of $Y = 15$ months, or some larger value of $Y$ would be improbable if, in fact, $\alpha = 2$ and $\beta = 4$. We do not give a table of areas under the gamma density function in this text, but we can obtain some idea of the magnitude of $P(Y \geq 15)$ by calculating the mean and standard deviation for the gamma density function when $\alpha = 2$ and $\beta = 4$. Thus,

$$\mu = \alpha\beta = (2)(4) = 8$$
$$\sigma^2 = \alpha\beta^2 = (2)(4)^2 = 32$$
$$\sigma = 5.7$$

Since $Y = 15$ months lies barely more than 1 standard deviation beyond the mean ($\mu + \sigma = 8 + 5.7 = 13.7$ months), we would not regard 15 months as an unusually large value of $Y$. Consequently, we would conclude that there is insufficient evidence to indicate that the company's new quality control program has been effective in increasing the mean time between complaints. We will present formal statistical procedures for answering this question in later chapters.

**Example 5.14** *(optional)*

Deriving the Mean of a Gamma Random Variable

Show that the mean for a gamma-type random variable $Y$ is equal to $\mu = \alpha\beta$.

Solution

We first write

$$E(Y) = \int_{-\infty}^{\infty} yf(y)\,dy = \int_0^{\infty} y\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dy = \int_0^{\infty} \frac{y^{(\alpha+1)-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dy$$

Multiplying and dividing the integrand by $\alpha\beta$ and using the fact that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, we obtain

$$E(Y) = \alpha\beta \int_0^{\infty} \frac{y^{(\alpha+1)-1}e^{-y/\beta}}{(\alpha\beta)\beta^\alpha\Gamma(\alpha)}\,dy = \alpha\beta \int_0^{\infty} \frac{y^{(\alpha+1)-1}e^{-y/\beta}}{\beta^{\alpha+1}\Gamma(\alpha + 1)}\,dy$$

The integrand is a gamma density function with parameters $(\alpha + 1)$ and $\beta$. Therefore, since the integral of any density function over $-\infty < y < \infty$, is equal to 1, we conclude

$$E(Y) = \alpha\beta(1) = \alpha\beta$$

## Applied Exercises

5.52 *Preventative maintenance tests.* The optimal scheduling of preventative maintenance tests of some (but not all) of $n$ independently operating components was developed in *Reliability Engineering and System Safety* (Jan., 2006). The time (in hours) between failures of a component was approximated by an exponential distribution with mean $\beta$.

a. Suppose $\beta = 1{,}000$ hours. Find the probability that the time between component failures ranges between 1,200 and 1,500 hours.
b. Again, assume $\beta = 1{,}000$ hours. Find the probability that the time between component failures is at least 1,200 hours.
c. Given that the time between failures is at least 1,200 hours, what is the probability that the time between failures is less than 1,500 hours?

5.53 *Lead in metal shredder residue.* Based on data collected from metal shredders across the nation, the amount $Y$ of extractable lead in metal shredder residue has an approximate exponential distribution with mean $\beta = 2.5$ milligrams per liter (Florida Shredder's Association).

a. Find the probability that $Y$ is greater than 2 milligrams per liter.
b. Find the probability that $Y$ is less than 5 milligrams per liter.

**PHISHING**

5.54 *Phishing attacks to email accounts.* Refer to the *Chance* (Summer, 2007) article on phishing attacks at a company, Exercise 2.24 (p. 38). Recall that *phishing* describes an attempt to extract personal/financial information through fraudulent email. The company set up a publicized email account—called a "fraud box"—which enabled employees to notify them if they suspected an email phishing attack. If there is minimal or no collaboration or collusion from within the company, the interarrival times (i.e., the time between successive email notifications, in seconds) have an approximate exponential distribution with a mean of 95 seconds.

a. What is the probability of observing an interarrival time of at least 2 minutes?
b. Data for a sample of 267 interarrival times are saved in the **PHISHING** file. Do the data appear to follow an exponential distribution with $\beta = 95$?

5.55 *Product failure behavior.* An article in *Hotwire* (Dec., 2002) discussed the length of time till failure of a product produced at Hewlett-Packard. At the end of the product's lifetime, the time till failure (in thousands of hours) is modeled using a gamma distribution with parameters $\alpha = 1$ and $\beta = 500$. In reliability jargon this is known as the "wear-out" distribution for the product. During its normal (useful) life, assume the product's time till failure is uniformly distributed over the range 100 thousand to 1 million hours.

a. At the end of the product's lifetime, find the probability that the product fails before 700 thousand hours.
b. During its normal (useful) life, find the probability that the product fails before 700 thousand hours.
c. Show that the probability of the product failing before 830 thousand hours is approximately the same for both the normal (useful) life distribution and the wear-out distribution.

5.56 *Flood level analysis.* Researchers have discovered that the maximum flood level (in millions of cubic feet per second) over a 4-year period for the Susquehanna River at Harrisburg, Pennsylvania, follows approximately a gamma distribution with $\alpha = 3$ and $\beta = .07$ (*Journal of Quality Technology*, Jan. 1986).

a. Find the mean and variance of the maximum flood level over a 4-year period for the Susquehanna River.
b. The researchers arrived at their conclusions about the maximum flood level distribution by observing maximum flood levels over 4-year periods, beginning in 1890. Suppose that over the next 4-year period the maximum flood level was observed to be .60 million cubic feet per second. Would you expect to observe a value this high from a gamma distribution with $\alpha = 3$ and $\beta = .07$? What can you infer about the maximum flood level distribution for the 4-year period observed?

5.57 *Acceptance sampling of a product.* An essential tool in the monitoring of the quality of a manufactured product is *acceptance sampling*. An acceptance sampling plan involves knowing the distribution of the life length of the item produced and determining how many items to inspect from the manufacturing process. The *Journal of Applied Statistics* (Apr., 2010) demonstrated the use of the exponential distribution as a model for the life length $Y$ of an item (e.g., a bullet). The article also discussed the importance of using the median of the lifetime distribution as a measure of product quality, since half of the items in a manufactured lot will have life lengths exceeding the median. For an exponential distribution with mean $\beta$, give an expression for the median of the distribution. (Hint: Your answer will be a function of $\beta$.)

5.58 *Spare parts demand model.* Effective maintenance of equipment depends on the ability to accurately forecast the demand for spare parts. The *Journal of Quality in Maintenance Engineering* (Vol. 18, 2012) developed a statistical approach to forecasting spare parts demand. The methodology used the gamma distribution with parameters $\alpha$ and $\beta$ to model the failure rate, $Y$, of system components. The model was developed under the assumption that the actual failure rate does not exceed twice the theoretical mean failure rate, $\mu$. Assume $\alpha = 2$ and $\beta = 5$, then find $P(Y < 2\mu)$.

5.59 *Flexible manufacturing system.* A part processed in a flexible manufacturing system (FMS) is routed through a set of operations, some of which are sequential and some of which are parallel. In addition, an FMS operation can be processed by alternative machines. An article in *IEEE Transactions* (Mar. 1990) gave an example of an FMS with four machines operating independently. The repair rates for the machines (i.e., the time, in hours, it takes to repair a failed machine) are exponentially distributed with means $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = .5$, and $\mu_4 = .5$, respectively.

a. Find the probability that the repair time for machine 1 exceeds 1 hour.
b. Repeat part **a** for machine 2.
c. Repeat part a for machines 3 and 4.
d. If all four machines fail simultaneously, find the probability that the repair time for the entire system exceeds 1 hour.

5.60 *Reaction to tear gas.* The length of time $Y$ (in minutes) required to generate a human reaction to tear gas formula A has a gamma distribution with $\alpha = 2$ and $\beta = 2$. The distribution for formula B is also gamma, but with $\alpha = 1$ and $\beta = 4$.

a. Find the mean length of time required to generate a human reaction to tear gas formula A. Find the mean for formula B.
b. Find the variances for both distributions.
c. Which tear gas has a higher probability of generating a human reaction in less than 1 minute? (*Hint*: You may use the fact that

$$\int ye^{-y/2}\,dy = -2ye^{-y/2} + \int 2e^{-y/2}dy$$

This result is derived by integration by parts.)

5.61 *Reliability of CD-ROMs.* In *Reliability Ques* (March 2004), the exponential distribution was used to model the life lengths of CD-ROM drives in a two-drive system. The two CD-ROM drives operate independently, and at least one drive must be operating for the system to operate successfully. Both drives have a mean lifelength of 25,000 hours.

a. The reliability, R($t$), of a single CD-ROM drive is the probability that the drive has a lifelength exceeding $t$ hours. Give a formula for R($t$).
b. Use the result, part **a**, to find the probability that the single CD-ROM drive has a lifelength exceeding 8,760 hours (the number of hours of operation in a year).
c. The reliability of the two CD-ROM drive system, S($t$), is the probability that at least one drive has a lifelength exceeding $t$ hours. Give a formula for S($t$). (*Hint*: Use the rule of complements and the fact that the two drives operate independently.)
d. Use the result, part **c**, to find the probability that the two CD-ROM drive system has a lifelength exceeding 8,760 hours.
e. Compare the probabilities, parts **b** and **d**.

## Theoretical Exercises

5.62 Show that the variance of a gamma distribution with parameters $\alpha$ and $\beta$ is $\alpha\beta^2$.

5.63 Let $Y$ have an exponential distribution with mean $\beta$. Show that $P(Y > a) = e^{-a/\beta}$. [*Hint*: Find $F(a) = P(Y \leq a)$.]

5.64 Refer to the concepts of *new better than used* (NBU) and *new worse than used* (NWU) in Theoretical Exercise 5.6 (p. 191). Show that the exponential distribution satisfies both the NBU and NWU properties. (Such a "life" distribution is said to be *new same as used* or *memoryless*.)

5.65 Show that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

5.66 We have stated that a chi-square random variable has a gamma-type density with $\alpha = \nu/2$ and $\beta = 2$. Find the mean and variance of a chi-square random variable.

5.67 Suppose a random variable $Y$ has a probability distribution given by

$$f(Y) = \begin{cases} cy^2e^{-y/2} & \text{if } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the value of $c$ that makes $f(y)$ a density function.

# 5.8 The Weibull Probability Distribution

In Section 5.7, we noted that the gamma density function can be used to model the distribution of the length of life (failure time) of manufactured components, equipment, etc. Another distribution used by engineers for the same purpose is known as the **Weibull distribution**.*

*See Weibull (1951).

### The Weibull Probability Distribution

The probability density function for a Weibull random variable, $Y$ is given by

$$f(y) = \begin{cases} \dfrac{\alpha}{\beta} y^{\alpha-1} e^{-y^{\alpha}/\beta} & \text{if } 0 \leq y < \infty;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\mu = \beta^{1/\alpha}\Gamma\left(\frac{\alpha + 1}{\alpha}\right)$$

$$\sigma^2 = \beta^{2/\alpha}\left[\Gamma\left(\frac{\alpha + 2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha + 1}{\alpha}\right)\right]$$

The Weibull density function contains two parameters, $\alpha$ and $\beta$. The **scale parameter**, $\beta$, reflects the size of the units in which the random variable $y$ is measured. The parameter $\alpha$ is the **shape parameter**. By changing the value of the shape parameter $\alpha$, we can generate a widely varying set of curves to model real-life failure time distributions. For the case $\alpha = 1$, we obtain the exponential distribution of Section 5.7. The graphs of Weibull density functions for different values of $\alpha$ and $\beta$ are shown in Figure 5.20.

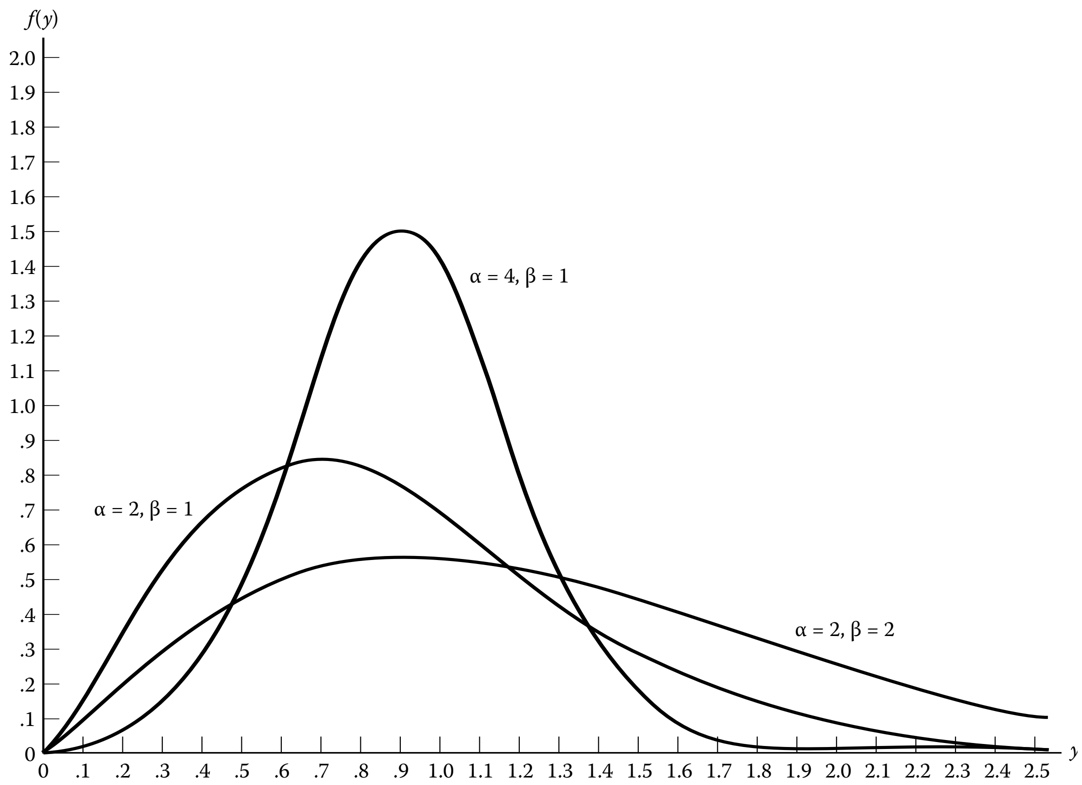



**FIGURE 5.20**
Graphs of Weibull density functions

In addition to providing a good model for the failure-time distributions of many manufactured items, the Weibull distribution is easy to use. A closed-form expression for its cumulative distribution function exists and can be used to obtain areas under the Weibull curve. Example 5.15 will illustrate the procedure.

**Example 5.15**

Weibull Distribution Application—Drill Bit Failure

The length of life $Y$ (in hours) of a drill bit used in a manufacturing operation has a Weibull distribution with $\alpha = 2$ and $\beta = 100$. Find the probability that a drill bit will fail before 8 hours of usage.

Solution

The cumulative distribution function for a Weibull distribution is

$$F(y_0) = \int_0^{y_0} f(y)\,dy = \int_0^{y_0} \frac{\alpha}{\beta} y^{\alpha-1} e^{-y^\alpha/\beta} dy$$

By making the transformation $Z = Y^\alpha$, we have $dz = \alpha y^{\alpha-1} dy$ and the integral reduces to

$$F(y_0) = 1 - e^{-z/\beta} = 1 - e^{-y_0^\alpha/\beta}$$

To find the probability that $y$ is less than 8 hours, we calculate

$$P(Y < 8) = F(8) = 1 - e^{-(8)^\alpha/\beta}$$
$$= 1 - e^{-(8)^2/100} = 1 - e^{-.64}$$

Interpolating between $e^{-.60}$ and $e^{-.65}$ in Table 3 of Appendix B or using a calculator with the $e$ function, we find $e^{-.64} \approx .527$. Therefore, the probability that a drill bit will fail before 8 hours is

$$P(Y < 8) = 1 - e^{-.64} = 1 - .527 = .473$$

**Example 5.16**

Finding the Mean of a Weibull Random Variable

Refer to Example 5.15. Find the mean life of the drill bits.

Solution

Substituting $\alpha = 2$ and $\beta = 100$ into the formula for the mean of a Weibull random variable yields

$$\mu = \beta^{1/\alpha}\Gamma\left(\frac{\alpha + 1}{\alpha}\right) = (100)^{1/2}\Gamma\left(\frac{2 + 1}{2}\right) = 10\Gamma(1.5)$$

From Table 6 of Appendix B, we find $\Gamma(1.5) = .88623$. Therefore, the mean life of the drill bits is

$$\mu = (10)\Gamma(1.5) = (10)(.88623) = 8.8623 \approx 8.86 \text{ hours}$$

## Applied Exercises

5.68 *Wind climate modeling.* Wind energy is fast becoming a major source of electricity in the U.S. Understanding the distribution of wind speeds is critical in the design of wind farms. Research published in the *International Journal of Engineering, Science and Technology* (Vol. 3, 2011) established the Weibull distribution as a good model for low to moderate wind speeds. For one area of India, wind speed $Y$ (meters per second) had a Weibull distribution with parameters $\alpha = 2$ and $\beta = 10$.

a. What is the value of $E(Y)$? Give a practical interpretation of this result.
b. What is the probability that the wind speed for this area of India will exceed 5 meters per second?

5.69 *Fracture toughness of materials.* Titanium diboride is an extremely hard ceramic material known for its resistance to mechanical erosion or fracture. The fracture toughness of the material, measured in megaPascals per meters-squared (MPa/m$^2$), was modeled using the Weibull distribution in *Quality Engineering* (Vol. 25, 2013). One possible set of Weibull parameters for this data is $\alpha = 6$ and $\beta = 1800$.

a. Find the mean and variance of the fracture toughness for this material.
b. Provide an estimate of the proportion of fracture toughness values that lie within 2 standard deviations of the mean.
c. Use the Weibull cumulative distribution function to calculate the exact probability that fracture toughness falls within 2 standard deviations of the mean. Compare the result with your answer in part b.

5.70 *Life cycle of a power plant.* Engineers studied the life cycle cost of a coal-fired power plant in the *Journal of Quality in Maintenance Engineering* (Vol. 19, 2013). The analysis used the Weibull distribution to model the probability distribution of time to failure, $Y$, measured in thousands of hours. The engineers used $\beta = 65$ as the value of the scale parameter in the Weibull distribution. Assume $\alpha = 2$ is the value of the shape parameter. Use this information to find a time to failure value, $Y = t$, such that the probability of failure before time $t$ is only .2. [Hint: Use the closed form of the distribution function, $F(t)$ given in Example 5.15.]

5.71 *Lifelength of avionic circuits.* University of Maryland engineers investigated the lifelengths of several commercial avionics, including flight control systems, autopilots, flight director systems, and symbol generators. (*The Journal of the Reliability Analysis Center*, 1st Quarter, 2005.) The lifelength of integrated circuits sold by a certain avionic vendor was modeled using the Weibull distribution with parameters $\alpha = 1$ and $\beta = 100{,}000$ hours. What is the probability that an integrated circuit sold by the vendor will fail before 50,000 hours of use?

5.72 *Lifelengths of drill bits.* Refer to Example 5.15 (p. 197) and the lifelength $Y$ of a drill bit. Recall that $Y$ has a Weibull distribution with $\alpha = 2$ and $\beta = 100$.

a. Calculate the values of $f(y)$ for $y = 2, 5, 8, 11, 14$, and 17. Plot the points $(y, f(y))$ and construct a graph of the failure time distribution of the drill bits.
b. Calculate the variance of the failure time distribution.
c. Find the probability that the length of life of a drill bit will fall within 2 hours of its mean.

5.73 *Doppler frequency magnitudes.* Japanese electrical engineers have developed a sophisticated radar system called the moving target detector (MTD), designed to reject ground clutter, rain clutter, birds, and other interference. The system was used to successfully detect aircraft embedded in ground clutter. (*Scientific and Engineering Reports of the National Defense Academy,* Vol. 39, 2001). The researchers show that the magnitude $Y$ of the Doppler frequency of a radar-received signal obeys a Weibull distribution with parameters $\alpha = 2$ and $\beta$.

a. Find $E(Y)$.
b. Find $\sigma^2$.
c. Give an expression for the probability that the magnitude $Y$ of the Doppler frequency exceeds some constant $C$.

5.74 *Washing machine repair time.* Based on extensive testing, a manufacturer of washing machines believes that the distribution of the time $Y$ (in years) until a major repair is required has a Weibull distribution with $\alpha = 2$ and $\beta = 4$.

a. If the manufacturer guarantees all machines against a major repair for 2 years, what proportion of all new washers will have to be repaired under the guarantee?
b. Find the mean and standard deviation of the length of time until a major repair is required.
c. Find $P(\mu - 2\sigma \le Y \le \mu + 2\sigma)$.
d. Is it likely that $Y$ will exceed 6 years?

5.75 *Bank surveillance failure.* The length of time (in months after maintenance) until failure of a bank's surveillance television equipment has a Weibull distribution with $\alpha = 2$ and $\beta = 60$. If the bank wants the probability of a breakdown before the next scheduled maintenance to be .05, how frequently should the equipment receive periodic maintenance?

## Theoretical Exercises

5.76 Show that the Weibull distribution with $\alpha = 2$ and $\beta > 0$ is *new better than used* (NBU). [See Optional Exercise 5.9, (p. 192) for the definition of NBU.]

5.77 Show that for the Weibull distribution,

$$\mu = \beta^{1/\alpha}\Gamma\left(\frac{\alpha + 1}{\alpha}\right)$$

5.78 Show that for the Weibull distribution,

$$E(Y^2) = \beta^{2/\alpha}\Gamma\left(\frac{\alpha + 2}{\alpha}\right)$$

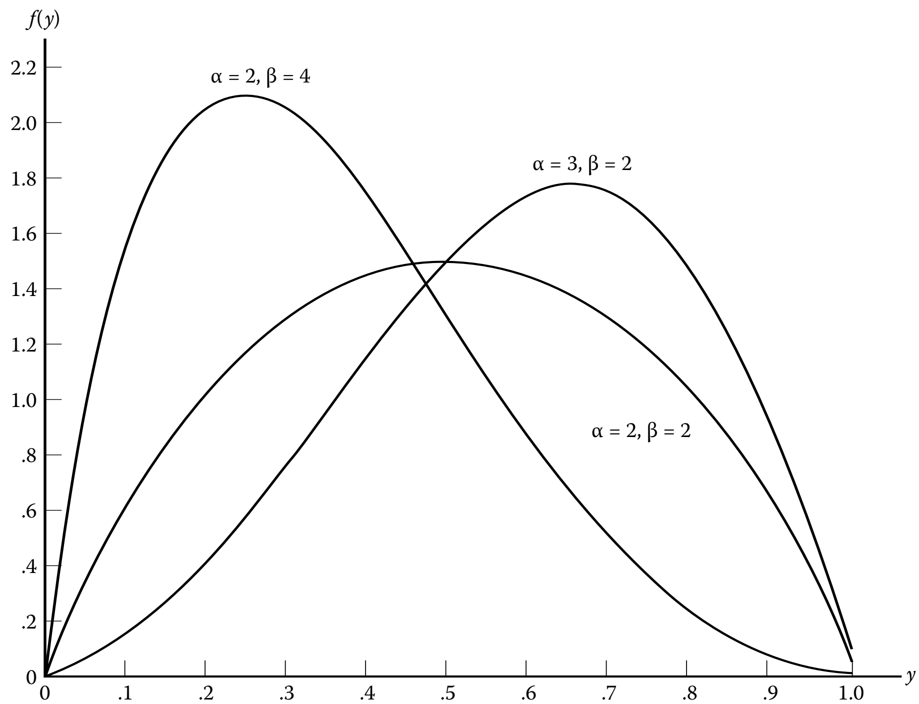
Then use the relationship $\sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2$ to show that

$$\sigma^2 = \beta^{2/\alpha}\left[\Gamma\left(\frac{\alpha + 2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha + 1}{\alpha}\right)\right]$$

## 5.9 Beta-Type Probability Distributions

Recall from Section 5.7 that the gamma density function provides a model for the relative frequency distribution of a random variable that possesses a fixed lower limit but that can become infinitely large. In contrast, the **beta density function**, also characterized by two parameters, possesses finite lower and upper limits. We will give these limits as 0 to 1, but the density function, with modification, can be defined over any specified finite interval. Graphs of beta density functions for ($\alpha = 2, \beta = 4$), ($\alpha = 2, \beta = 2$), and ($\alpha = 3, \beta = 2$), are shown in Figure 5.21. The probability density function, the mean, and the variance for a beta-type random variable are shown in the next box.

**FIGURE 5.21**
Graphs of beta density functions




### The Beta Probability Distribution

The probability density function for a beta-type random variable $Y$ is given by

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} & \text{if } 0 \le y \le 1;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The mean and variance of a beta random variable are, respectively,

$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

[Recall that

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$$

and $\Gamma(\alpha) = (\alpha - 1)!$ when $\alpha$ is a positive integer.]

**Example 5.17**

Beta Distribution Application—Robotic Sensors

Infrared sensors in a computerized robotic system send information to other sensors in different formats. The percentage $Y$ of signals sent that are directly compatible for all sensors in the system follows a beta distribution with $\alpha = \beta = 2$.

a. Find the probability that more than 30% of the infrared signals sent in the system are directly compatible for all sensors.

b. Find the mean and variance of $Y$.

Solution

a. From the box, the probability density function for $Y$ is given as

$$f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 \le y \le 1$$

Substituting $\alpha = \beta = 2$ into the expression for $f(y)$, we obtain

$$\begin{aligned} f(y) &= \frac{\Gamma(2+2)y^{2-1}(1-y)^{2-1}}{\Gamma(2)\Gamma(2)} = \frac{(3!)y(1-y)}{(1!)(1!)} \\ &= 6y(1-y) \end{aligned}$$

The probability we seek is $P(Y > .30)$. Integrating $f(y)$, we obtain

$$\begin{aligned} P(Y > .30) &= \int_{.30}^{1} 6y(1-y)\,dy = 6\int_{.30}^{1}(y - y^2)\,dy \\ &= 6\left\{\int_{.30}^{1} y\,dy - \int_{.30}^{1} y^2\,dy\right\} = 6\left\{y^2/2\Big]_{.30}^{1} - y^3/3\Big]_{.30}^{1}\right\} \\ &= 6\left[\frac{1}{2} - (.3)^2/2 - \left(\frac{1}{3} - (.3)^3/3\right)\right] \\ &= 6(.130667) = .784 \end{aligned}$$

b. From the box, the mean and variance of a beta random variable are

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Substituting $\alpha = \beta = 2$, we obtain:

$$\mu = \frac{2}{2+2} = \frac{2}{4} = .5$$

$$\sigma^2 = \frac{(2)(2)}{(2+2)^2(2+2+1)} = \frac{4}{(16)(5)} = .05$$

The cumulative distribution function $F(y)$ of a beta density function is called an **incomplete beta function**. Values of this function for various values of $y$, $\alpha$, and $\beta$ are given in *Tables of the Incomplete Beta Function* (1956). For the special case where $\alpha$ and $\beta$ are integers, it can be shown that

$$F(p) = \int_0^p \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}\,dy = \sum_{y=\alpha}^{n} p(y)$$

where $p(y)$ is a binomial probability distribution with parameters $p$ and $n = (\alpha + \beta - 1)$. Recall that tables giving the cumulative sums of binomial probabilities are given in Table 2 of Appendix B, for $n = 5, 10, 15, 20$, and 25. More extensive tables of these probabilities are listed in the references at the end of the chapter.

**Example 5.18**

Using the Binomial Distribution to Find a Beta Probability

Data collected over time on the utilization of a computer core (as a proportion of the total capacity) were found to possess a relative frequency distribution that could be approximated by a beta density function with $\alpha = 2$ and $\beta = 4$. Find the probability that the proportion of the core being used at any particular time will be less than .20.

Solution

The probability that the proportion of the core being utilized will be less than $p = .2$ is

$$F(p) = \int_0^p \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}\,dy = \sum_{y=\alpha}^{n} p(y)$$

where $p(y)$ is a binomial probability distribution with $n = (\alpha + \beta - 1) = (2 + 4 - 1) = 5$ and $p = .2$. Therefore,

$$F(.2) = \sum_{y=2}^{5} p(y) = 1 - \sum_{y=0}^{1} p(y)$$

From Table 2 of Appendix B for $n = 5$ and $p = .2$, we find that

$$\sum_{y=0}^{1} p(y) = .737$$

Therefore, the probability that the computer core will be less than 20% occupied at any particular time is

$$F(.2) = 1 - \sum_{y=0}^{1} p(y) = 1 - .737 = .263$$

## Applied Exercises

5.79 *Mineral composition of rocks.* Sedimentary rocks are a mixture of minerals and pores. Engineers employ well (or borehole) logging to make a detailed record of the mineral composition of the rock penetrated by a borehole. One important measure is shale density of a rock (measured as a fraction between 0 and 1). In the journal *SPE Reservoir Evaluation & Testing* (Dec., 2009), petroleum engineers modeled the shale density with the beta distribution. Assume shale density, $Y$, follows a beta distribution with $\alpha = 3$ and $\beta = 2$. Find the probability that the shale density of a rock is less than .5.

5.80 *Segmenting skin images.* At the 2006 *International Conference on Biomedical Engineering,* researchers presented an unsupervised learning technique for segmenting skin images. The method employed various forms of a beta distribution, one of which was symmetric. The parameters of a symmetric beta distribution are equal, i.e., $\alpha = \beta$. Consider the random variable $Y$ = proportion of pixel shift changes in a skin image following treatment and assume $Y$ follows a symmetric beta distribution.

a. Find $E(Y)$ and interpret the result.

b. Show that the variance of this beta distribution is $1/[4(2\alpha + 1)]$.

5.81 *Parameters of the beta distribution.* In the *Journal of Statistical Computation and Simulation* (Vol. 67, 2000), researchers presented a numerical procedure for estimating the parameters of a beta distribution. The method involves a re-parameterization of the probability density function.

Consider a beta distribution with parameters $\alpha^*$ and $\beta^*$, where $\alpha^* = \alpha\beta$ and $\beta^* = \beta(1 - \alpha)$.

a. Show that the mean of this beta distribution is $\alpha$.
b. Show that the variance of this beta distribution is $\alpha(1 - \alpha)/(\beta + 1)$.

5.82 *Breaking block ciphers.* A block cipher is an encryption algorithm that transforms a fixed-length block of unencrypted text data (called *plaintext*) into a block of encrypted text data (called *ciphertext*) of the same length for security purposes. A group of Korean communications engineers have designed a new linear approximation method for breaking a block cipher (*IEICE Transactions on Fundamentals*, Jan. 2005). The researchers showed that the success rate $Y$ of the new algorithm has a beta distribution with parameters $\alpha = n/2$ and $\beta = N/2n$, where $n$ is the number of linear approximations used and $N$ is the number of plaintexts in the encrypted data.

a. Give the probability density function for $Y$.
b. Find the mean and variance of $Y$.
c. Give an interval which is likely to contain the success rate, $Y$.

5.83 *Environmental shutdowns of plant.* An investigation into pollution control expenditures of industrial firms found that the annual percentage of plant capacity shutdown attributable to environmental and safety regulation has an approximate beta distribution with $\alpha = 1$ and $\beta = 25$.

a. Find the mean and variance of the annual percentage of plant capacity shutdown attributable to environmental and safety regulation.
b. Find the probability that more than 1% of plant capacity shutdown is attributable to environmental and safety regulation.

5.84 *Laser color printer repairs.* The proportion $Y$ of a data-processing company's yearly hardware repair budget allocated to repair its laser color printer has an approximate beta distribution with parameters $\alpha = 2$ and $\beta = 9$.

a. Find the mean and variance of $Y$.
b. Compute the probability that for any randomly selected year, at least 40% of the hardware repair budget is used to repair the laser color printer.
c. What is the probability that at most 10% of the yearly repair budget is used for the laser color printer?

5.85 *Internet start-up firms.* Suppose the proportion of Internet start-up firms that make a profit during their first year of operation possesses a relative frequency distribution that can be approximated by the beta density with $\alpha = 5$ and $\beta = 6$.

a. Find the probability that at most 60% of all Internet start-up firms make a profit during their first year of operation.
b. Find the probability that at least 80% of all Internet start-up firms make a profit during their first year of operation.

5.86 *Coarse cement granules.* An important property of certain products that are in powder or granular form is their particle size distribution. For example, refractory cements are adversely affected by too high a proportion of coarse granules, which can lead to weaknesses from poor packing. G. H. Brown (*Journal of Quality Technology*, July 1985) showed that the beta distribution provides an adequate model for the percentage $Y$ of refractory cement granules in bulk form that are coarse. Suppose you are interested in controlling the proportion $Y$ of coarse refractory cement in a lot, where $Y$ has a beta distribution with parameters $\alpha = \beta = 2$.

a. Find the mean and variance of $Y$.
b. If you will accept the lot only if less than 10% of refractory cement granules are coarse, find the probability of lot acceptance.

## Theoretical Exercises

5.87 A continuous random variable $Y$ has a beta distribution with probability density

$$f(y) = \begin{cases} cy^5(1 - y)^2 & \text{if } 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the value of $c$ that will make $f(y)$ a density function.

5.88 Verify that the mean of a beta density with parameters $\alpha$ and $\beta$ is given by $\mu = \alpha/(\alpha + \beta)$.

5.89 Show that if $Y$ has a beta density with $\alpha = 1$ and $\beta = 1$, then $Y$ is uniformly distributed over the interval $0 \le Y \le 1$.

5.90 Show that the beta distribution with $\alpha = 2$ and $\beta = 1$ is *new better than used* (NBU). (See Optional Exercise 5.9, p. 192 for the definition of NBU.)

# 5.10 Moments and Moment Generating Functions (*Optional*)

The **moments** and **moment generating functions** for continuous random variables are defined in exactly the same way as for discrete random variables, except that the expectations involve integration.* The relevance and applicability of a moment generating function $m(t)$ are the same in the continuous case, as we now illustrate with two examples.

*Moments and moment generating functions for discrete random variables are discussed in Optional Section 4.9.

**Example 5.19**

MGF for a Gamma Random Variable

Find the moment generating function for a gamma-type random variable $Y$.

Solution

The moment generating function is given by

$$m(t) = E(e^{tY}) = \int_0^\infty e^{ty}\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dy$$

$$= \int_0^\infty \frac{y^{\alpha-1}e^{-y(1/\beta-t)}}{\beta^\alpha\Gamma(\alpha)}\,dy = \int_0^\infty \frac{y^{\alpha-1}e^{-y/[\beta/(1-\beta t)]}}{\beta^\alpha\Gamma(\alpha)}\,dy$$

An examination of this integrand indicates that we can convert it into a gamma density function with parameters $\alpha$ and $\beta/(1 - \beta t)$, by factoring $1/\beta^\alpha$ out of the integral and multiplying and dividing by $[\beta/(1 - \beta t)]^\alpha$. Therefore,

$$m(t) = \frac{1}{\beta^\alpha}\left(\frac{\beta}{1-\beta t}\right)^\alpha \int_0^\infty \frac{y^{\alpha-1}e^{-y/[\beta/(1-\beta t)]}}{\left(\dfrac{\beta}{1-\beta t}\right)^\alpha \Gamma(\alpha)}\,dy$$

The integral of this gamma density function is equal to 1. Therefore,

$$m(t) = \frac{1}{(1-\beta t)^\alpha}(1) = \frac{1}{(1-\beta t)^\alpha}$$

**Example 5.20**

Using Moments to find Gamma Mean and Variance

Refer to Example 5.18. Use $m(t)$ to find $\mu_1'$ and $\mu_2'$. Use the results to derive the mean and variance of a gamma-type random variable.

Solution

The first two moments about the origin, evaluated at $t = 0$, are

$$\mu_1' = \mu = \left.\frac{d\,m(t)}{dt}\right]_{t=0} = \left.\frac{-\alpha(-\beta)}{(1-\beta t)^{\alpha+1}}\right]_{t=0} = \alpha\beta$$

and

$$\mu_2' = \left.\frac{d^2\,m(t)}{dt^2}\right]_{t=0}$$

$$= -\left.\frac{\alpha\beta(\alpha+1)(-\beta)}{(1-\beta t)^{\alpha+2}}\right]_{t=0} = \alpha(\alpha+1)\beta^2$$

Then, applying Theorem 5.4, we obtain

$$\sigma^2 = E(y^2) - \mu^2 = \mu_2' - \mu^2$$
$$= \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$$

Some useful probability density functions, with their means, variances, and moment generating functions, are summarized in the *Key Formulas* section of the *Quick Review* at the end of this chapter.

## Theoretical Exercises

5.91 Use the moment generating function $m(t)$ of the normal density to find $\mu_1'$ and $\mu_2'$. Then use these results to show that a normal random variable has mean $\mu$ and variance $\sigma^2$.

5.92 Verify that the moment generating function of a chi-square random variable with $\nu$ degrees of freedom is

$$m(t) = (1 - 2t)^{-\nu/2}$$

(*Hint*: Use the fact that a chi-square random variable has a gamma-type density function with $\alpha = \nu/2$ and $\beta = 2$.)

5.93 Verify that the moment generating function of a uniform random variable on the interval $a \le Y \le b$ is

$$m(t) = \frac{e^{tb} - e^{ta}}{t(b - a)}$$

5.94 Consider a continuous random variable $Y$ with density

$$f(y) = \begin{cases} e^y & \text{if } y < 0 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the moment generating function $m(t)$ of $Y$.
b. Use the result of part a to find the mean and variance of $Y$.

## *STATISTICS IN ACTION REVISITED*

### Super Weapons Development—Optimizing the Hit Ratio

Recall that a U.S. Army defense contractor has developed a prototype gun that fires 1,100 flechettes with a single round. In range tests, three 2-feet-wide targets were set up a distance of 500 meters (approximately 1,500 feet) from the weapon. The centers of the three targets were at 0, 5, and 10 feet, respectively, as shown in Figure SIA5.1 (p. 187). The prototype gun was aimed at the middle target (center at 5 feet) and fired once. The point $Y$ where each of the 1,100 flechettes landed at the 500-meter distance was measured using a horizontal grid. (The 1,100 measurements on the random variable $Y$ are saved in the **MOAGUN** file.)

The defense contractor wants to set the gun specifications to maximize the number of target hits. The weapon is designed to have a mean horizontal value, $E(Y)$, equal to the aim point (e.g., $\mu = 5$ feet when aimed at the center target). By changing specifications, the contractor can vary the standard deviation, $\sigma$. Recall that the **MOAGUN** file contains flechette measurements for three different range tests—one with a standard deviation of $\sigma = 1$ foot, one with $\sigma = 2$ feet, and one with $\sigma = 4$ feet.

From past experience, the defense contractor has found that the distribution of the horizontal flechette measurements is closely approximated by a normal distribution. MINITAB histograms of the horizontal hit measurements in the **MOAGUN** file are shown in Figures SIA5.2a–c. You can see that the normal curves superimposed on the histograms fit the data very well. Consequently, we'll use the normal distribution to find the probability that a single flechette shot from the weapon will hit any one of the three targets. Recall from Figure SIA5.1 that the three targets range from $-1$ to 1, 4 to 6, and 9 to 11 feet on the horizontal grid.

**FIGURE SIA5.2a**
MINITAB histogram for the horizontal hit measurements when $\sigma = 1$

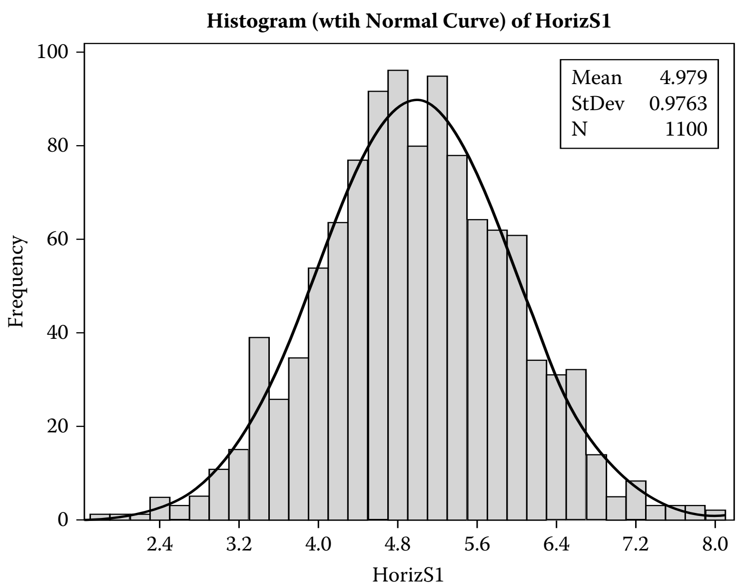


**FIGURE S1A5.2b**
MINITAB histogram for the horizontal hit measurements $\sigma = 2$

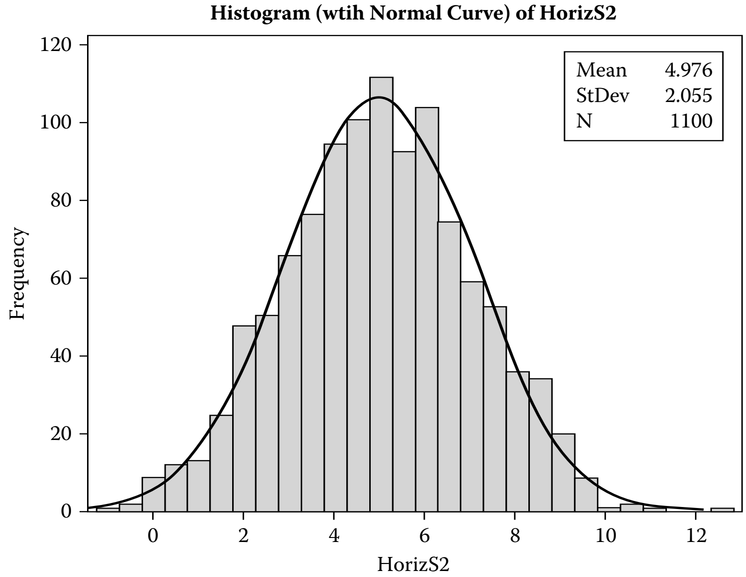
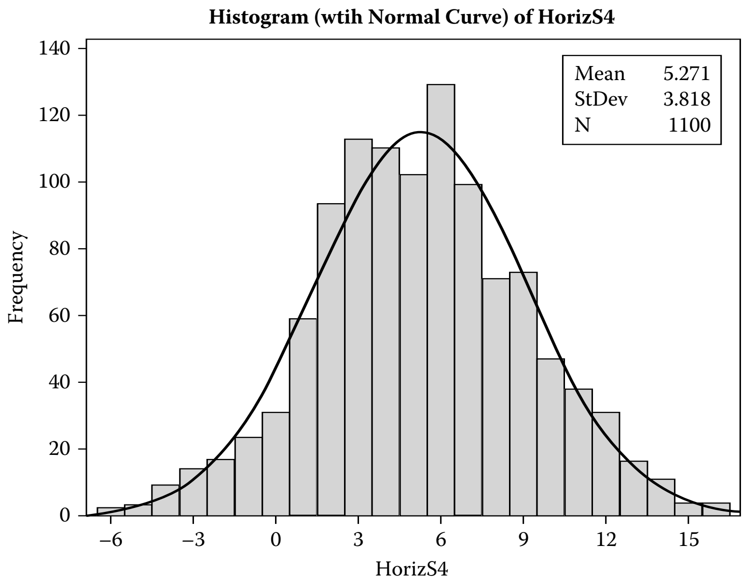



**FIGURE SIA5.2c**
MINITAB histogram for the horizontal hit measurements $\sigma = 4$




Consider, first, the middle target. Again, let $Y$ represent the horizontal measurement for a flechette shot from the gun. Then, the flechette will hit the target if $4 \leq Y \leq 6$. The probability that this flechette will hit the target when $\mu = 5$ and $\sigma = 1$ is, using the normal probability table (Table 5, Appendix B),

$$\textit{Middle}\ (\sigma = 1): P(4 \leq Y \leq 6) = P\left(\frac{4-5}{1} < Z < \frac{6-5}{1}\right) = P(-1 < Z < 1) = 2(.3413) = .6826$$

Similarly, we find the probability that the flechette hits the left and right targets shown in Figure SIA5.1.

$$\textit{Left}\ (\sigma = 1): P(-1 \leq Y \leq 1) = P\left(\frac{-1-5}{1} < Z < \frac{1-5}{1}\right) = P(-6 < Z < -4) \approx 0$$

$$\textit{Right}\ (\sigma = 1): P(9 \leq Y \leq 11) = P\left(\frac{9-5}{1} < Z < \frac{11-5}{1}\right) = P(4 < Z < 6) \approx 0$$

**FIGURE SIA5.3**
MINITAB worksheet with cumulative normal probabilities

| ↓ | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| | Y | Sigma1 | Sigma2 | Sigma4 | |
| 1 | -1 | 0.00000 | 0.001350 | 0.066807 | |
| 2 | 1 | 0.00003 | 0.022750 | 0.158655 | |
| 3 | 4 | 0.15866 | 0.308538 | 0.401294 | |
| 4 | 6 | 0.84134 | 0.691462 | 0.598706 | |
| 5 | 9 | 0.99997 | 0.977250 | 0.841345 | |
| 6 | 11 | 1.00000 | 0.998650 | 0.933193 | |
| 7 | | | | | |

You can see that there is about a 68% chance that a flechette will hit the middle target, but virtually no chance that one will hit the left and right targets when the standard deviation is set at 1 foot.

To find these three probabilities for $\sigma = 2$ and $\sigma = 4$, we use the normal probability function in MINITAB. Figure SIA5.3 is a MINITAB worksheet giving the cumulative probabilities of a normal random variable falling below the values of $Y$ in the first column. The cumulative probabilities for $\sigma = 2$ and $\sigma = 4$ are given in the columns named "Sigma2" and "Sigma4", respectively.

Using the cumulative probabilities in the figure to find the three probabilities when $\sigma = 2$, we have:

*Middle* $(\sigma = 2)$: $P(4 \leq Y \leq 6) = P(Y \leq 6) - P(Y \leq 4) = .6915 - .3085 = .3830$

*Left* $(\sigma = 2)$: $P(-1 \leq Y \leq 1) = P(Y \leq 1) - P(Y \leq -1) = .0227 - .0013 = .0214$

*Right* $(\sigma = 2)$: $P(9 \leq Y \leq 11) = P(Y \leq 11) - P(Y \leq 9) = .9987 - .9773 = .0214$

Thus, when $\sigma = 2$, there is about a 38% chance that a flechette will hit the middle target, a 2% chance that one will hit the left target, and a 2% chance that one will hit the right target. The probability that a flechette will hit either the middle or left or right target is simply the sum of these three probabilities (an application of the Additive Rule of probability). This sum is $.3830 + .0214 + .0214 = .4258$; consequently, there is about a 42% chance of hitting any one of the three targets when specifications are set so that $\sigma = 2$.

Now, we use the cumulative probabilities in Figure SIA5.3 to find the three hit probabilities when $\sigma = 4$:

*Middle* $(\sigma = 4)$: $P(4 \leq Y \leq 6) = P(Y \leq 6) - P(Y \leq 4) = .5987 - .4013 = .1974$

*Left* $(\sigma = 4)$: $P(-1 \leq Y \leq 1) = P(Y \leq 1) - P(Y \leq -1) = .1587 - .0668 = .0919$

*Right* $(\sigma = 4)$: $P(9 \leq Y \leq 11) = P(Y \leq 11) - P(Y \leq 9) = .9332 - .8413 = .0919$

Thus, when $\sigma = 4$, there is about a 20% chance that a flechette will hit the middle target, a 9% chance that one will hit the left target, and a 9% chance that one will hit the right target. The probability that a flechette will hit any one of the three targets is $.1974 + .0919 + .0919 = .3812$.

Table SIA5.1 shows the calculated normal probabilities of hitting the three targets for the different values of $\sigma$, as well as the actual results of the three range tests. (Recall that the actual data is saved in the MOAGUN file.) You can see that the proportion of the 1,100 flechettes that actually hit each target—called the hit ratio—agrees very well with the estimated probability of a hit using the normal distribution.

These probability calculations reveal a few patterns. First, the probability of hitting the middle target (the target where the gun is aimed) is reduced as the standard deviation is increased. Obviously, if the U.S. Army wants to maximize the chance of hitting the target that the prototype gun is aimed at, it will want specifications set with a small value of $\sigma$. But if the Army wants to hit multiple targets with a single shot of the weapon, $\sigma$ should be increased. With a larger $\sigma$, not as many of the flechettes will hit the target aimed at, but more will hit peripheral targets. Whether $\sigma$ should be set at 4 or 6 (or some other value) depends on how high of a hit rate is required for the peripheral targets.

MOAGUN **TABLE SIA5.1 Summary of Normal Probability Calculations and Actual Range Test Results**

| Target | Specification | Normal Probability | Actual Number of Hits | Hit Ratio (Hits/1,100) |
|---|---|---|---|---|
| Left ($-1$ to 1) | $\sigma = 1$ | .0000 | 0 | .000 |
| | $\sigma = 2$ | .0214 | 30 | .027 |
| | $\sigma = 4$ | .0919 | 73 | .066 |
| Middle (4 to 6) | $\sigma = 1$ | .6826 | 764 | .695 |
| | $\sigma = 2$ | .3820 | 409 | .372 |
| | $\sigma = 4$ | .1974 | 242 | .220 |
| Right (9 to 11) | $\sigma = 1$ | .0000 | 0 | .000 |
| | $\sigma = 2$ | .0214 | 23 | .021 |
| | $\sigma = 4$ | .0919 | 93 | .085 |

# Quick Review

## Key Terms

*Note: Starred (*) terms are from the optional section in this chapter.*

Beta density function 220
Beta distribution 221
Chi-square distribution 213
Chi-square random variable 213
Continuous random variable 188
Cumulative distribution function 188
Density function 188
Expected values 193
Exponential distribution 213
Exponential random variable 229
Gamma distribution 212
Gamma-type density function 212
Incomplete beta function 222
Incomplete gamma function 213
*Moment generating function 223
Monotonically increasing function 188
Normal density function 201
Normal distribution 204
*Normal probability plot 206
*Normal random variable 200
Scale parameter 212, 217
Shape parameter 212, 217
Standard normal random variable 203
Uniform distribution 199
Uniform random variable 197
Weibull distribution 216
Weibull random variable 217

## Key Formulas

| Random Variable | Probability Density Function | Mean | Variance | Moment Generating Function* | |
|---|---|---|---|---|---|
| Uniform | $f(y) = \dfrac{1}{b-a} \quad a \le y \le b$ | $\dfrac{a+b}{2}$ | $\dfrac{(b-a)^2}{12}$ | $\dfrac{e^{tb} - e^{ta}}{t(b-a)}$ | 197 |
| Normal | $f(y) = \dfrac{e^{-(y-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$ $-\infty < y < \infty$ | $\mu$ | $\sigma^2$ | $e^{\mu t + (t^2\sigma^2/2)}$ | 200 |
| Gamma | $f(y) = \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$ $0 \le y < \infty$ | $\alpha\beta$ | $\alpha\beta^2$ | $(1-\beta t)^{-\alpha}$ | 212 |
| Exponential | $f(y) = \dfrac{1}{\beta}e^{-y/\beta}$ $0 \le y < \infty$ | $\beta$ | $\beta^2$ | $\dfrac{1}{(1-\beta t)}$ | 194 |

| Random Variable | Probability Density Function | Mean | Variance | Moment Generating Function* | |
|---|---|---|---|---|---|
| Chi-square | $f(y) = \dfrac{(y)^{(\nu/2)-1}e^{-y/2}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$ $0 \le y < \infty$ | $\nu$ | $2\nu$ | $(1 - 2t)^{-\nu/2}$ | 213 |
| Weibull | $f(y) = \dfrac{\alpha}{\beta}y^{\alpha-1}e^{-y^{\alpha}/\beta}$ $0 \le y < \infty$ | $\beta^{1/\alpha}\Gamma\left(\dfrac{\alpha+1}{\alpha}\right)$ | $\beta^{2/\alpha}\left[\Gamma\left(\dfrac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\dfrac{\alpha+1}{\alpha}\right)\right]$ | $\beta^{t/\alpha}\Gamma(1 + t/\alpha)$ | 217 |
| Beta | $f(y) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$ $0 \le y \le 1$ | $\dfrac{\alpha}{\alpha+\beta}$ | $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ | Closed-form expression does not exist. | 220 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $f(y)$ | *f* of *y* | Probability density function for a continuous random variable $Y$ |
| $F(y)$ | cap *F* of *y* | Cumulative distribution function for a continuous random variable $Y$ |
| $\Gamma(\alpha)$ | gamma of alpha | Gamma function for a positive integer $\alpha$ |

## Chapter Summary

- Properties of a **density function** for a continuous random variable $Y$: (1) $f(y) \ge 0$, (2) $F(\infty) = 1$, $P(a < Y < b) = F(b) - F(a)$
- Types of continuous random variables: **uniform, normal, gamma-type, Weibull**, and **beta-type**.
- **Uniform probability distribution** is a model for continuous random variables that are evenly distributed over a certain interval.
- **Normal** (or **Gaussian**) **probability distribution** is a model for continuous random variables that have a bell-shaped curve with thin tails.
- Descriptive methods for assessing normality: **histogram, stem-and-leaf display**, **IQR/s ≈ 1.3**, and **normal probability plot**.
- **Gamma-type probability distribution** is a model for continuous random variables that are lifelengths or waiting times.
- Two special types of gamma random variables are: **chi-square** random variables and **exponential** random variables.
- **Weibull probability distribution** is a model for continuous random variables that represent failure times.
- **Beta-type probability distribution** is a model for continuous random variables that fall in the interval 0 to 1.

## Supplementary Exercises

5.95 *Shopping vehicle and judgment.* Refer to the *Journal of Marketing Research* (Dec., 2011) study of whether you are more likely to choose a vice product (e.g., a candy bar) when your arm is flexed (as when carrying a shopping basket) than when your arm is extended (as when pushing a shopping cart), Exercise 2.43 (p. 50). The study measured choice scores (on a scale of 0 to 100, where higher scores indicate a greater preference for vice options) for consumers shopping under each of the two conditions. Recall that the average choice score for consumers with a flexed arm was 59, while the average for consumers with an extended arm was 43. For both conditions, assume that the standard deviation of the choice scores is 5. Also assume that both distributions are approximately normally distributed.

a. In the flexed arm condition, what is the probability that a consumer has a choice score of 60 or greater?

b. In the extended arm condition, what is the probability that a consumer has a choice score of 60 or greater?

5.96 *Forest development following wildfires.* *Ecological Applications* (May 1995) published a study on the development of forests following wildfires in the Pacific Northwest. One variable of interest to the researcher was tree diameter at breast height 110 years after the fire. The population of Douglas fir trees was shown to have an approximately normal diameter distribution, with $\mu = 50$ centimeters (cm) and $\sigma = 12$ cm.

a. Find the diameter, $d$, such that 30% of the Douglas fir trees in the population have diameters that exceed $d$.
b. Another species of tree, western hemlock, was found to have a breast height diameter distribution that resembled an exponential distribution with $\beta = 30$ centimeters. Find the probability that a western hemlock tree growing in the forest damaged by wildfire 110 years ago has a diameter that exceeds 25 centimeters.

5.97 *Cleaning rate of pressure washer.* A manufacturing company has developed a fuel-efficient machine that combines pressure washing with steam cleaning. It is designed to deliver 7 gallons of cleaner per minute at 1,000 pounds per square inch for pressure washing. In fact, it delivers an amount at random anywhere between 6.5 and 7.5 gallons per minute. Assume that $Y$, the amount of cleaner delivered, is a uniform random variable with probability density

$$f(y) = \begin{cases} 1 & \text{if } 6.5 \le y \le 7.5 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the mean and standard deviation of $Y$. Then graph $f(y)$, showing the locations of the mean and 1 and 2 standard deviation intervals around the mean.
b. Find the probability that more than 7.2 gallons of cleaner are dispensed per minute.

5.98 *Waiting for a monorail.* The problem of passenger congestion prompted a large international airport to install a monorail connecting its main terminal to the three concourses, A, B, and C. The engineers designed the monorail so that the amount of time a passenger at concourse B must wait for a monorail car has a uniform distribution ranging from 0 to 10 minutes.

a. Find the mean and variance of $Y$, the time a passenger at concourse B must wait for the monorail. (Assume that the monorail travels sequentially from concourse A, to concourse B, to concourse C, back to concourse B, and then returns to concourse A. The route is then repeated.)
b. If it takes the monorail 1 minute to go from concourse to concourse, find the probability that a hurried passenger can reach concourse A less than 4 minutes after arriving at the monorail station at concourse B.

5.99 *Pacemaker specifications.* A pacemaker is made up of several biomedical components that must be of a high quality for the pacemaker to work. It is vitally important for manufacturers of pacemakers to use parts that meet specifications. One particular plastic part, called a connector module, mounts on the top of the pacemaker. Connector modules are required to have a length between .304 inch and .322 inch to work properly. Any modules with length outside these limits are "out-of-spec." *Quality* (Aug. 1989) reported on one supplier of connector modules that had been shipping out-of-spec parts to the manufacturer for 12 months.

a. The lengths of the connector modules produced by the supplier were found to follow an approximate normal distribution with mean $\mu = .3015$ inch and standard deviation $\sigma = .0016$ inch. Use this information to find the probability that the supplier produces an out-of-spec part.
b. Once the problem was detected, the supplier's inspection crew began to employ an automated data-collection system designed to improve product quality. After 2 months, the process was producing connector modules with mean $\mu = .3146$ inch and standard deviation $\sigma = .0030$ inch. Find the probability that an out-of-spec part will be produced. Compare your answer to part **a**.

5.100 *Spruce budworm infestation.* An infestation of a certain species of caterpillar, the spruce budworm, can cause extensive damage to the timberlands of the northern United States. It is known that an outbreak of this type of infestation occurs, on the average, every 30 years. Assuming that this phenomenon obeys an exponential probability law, what is the probability that catastrophic outbreaks of spruce budworm infestation will occur within 6 years of each other?

5.101 *Modeling an airport taxi service.* In an article published in the *European Journal of Operational Research* (Vol. 21, 1985), the vehicle-dispatching decisions of an airport-based taxi service were investigated. In modeling the system, the authors assumed travel times of successive taxi trips to and from the terminal to be independent exponential random variables. Assume $\beta = 20$ minutes.

a. What is the mean trip time for the taxi service?
b. What is the probability that a particular trip will take more than 30 minutes?
c. Two taxis have just been dispatched. What is the probability that both will be gone for more than 30 minutes? That at least one of the taxis will return within 30 minutes?

5.102 *Ambulance response time.* Ambulance response time is measured as the time (in minutes) between the initial call to emergency medical services (EMS) and when the patient is reached by ambulance. *Geographical Analysis* (Vol. 41, 2009) investigated the characteristics of ambulance response time for EMS calls in Edmonton, Alberta. For a particular EMS station (call it Station A), ambulance response time is known to be normally distributed with $\mu = 7.5$ minutes and $\sigma = 2.5$ minutes.

a. Regulations require that 90% of all emergency calls should be reached in 9 minutes or less. Are the regulations met at EMS station A? Explain.
b. A randomly selected EMS call in Edmonton has an ambulance response time of 2 minutes. Is it likely that this call was serviced by Station A? Explain.

5.103 *Modeling machine downtime.* The importance of modeling machine downtime correctly in simulation studies was discussed in *Industrial Engineering* (Aug. 1990). The paper presented simulation results for a single-machine-tool system with the following properties:

- The interarrival times of jobs are exponentially distributed with a mean of 1.25 minutes.
- The amount of time the machine operates before breaking down is exponentially distributed with a mean of 540 minutes.
- The repair time (in minutes) for the machine has a gamma distribution with parameters $\alpha = 2$ and $\beta = 30$.

a. Find the probability that two jobs arrive for processing at most 1 minute apart.
b. Find the probability that the machine operates for at least 720 minutes (12 hours) before breaking down.
c. Find the mean and variance of the repair time for the machine. Interpret the results.
d. Find the probability that the repair time for the machine exceeds 120 minutes.

5.104 *Defective modems.* Suppose that the fraction of defective modems shipped by a data-communications vendor has an approximate beta distribution with $\alpha = 5$ and $\beta = 21$.

a. Find the mean and variance of the fraction of defective modems per shipment.
b. What is the probability that a randomly selected shipment will contain at least 30% defectives?
c. What is the probability that a randomly selected shipment will contain no more than 5% defectives?

5.105 *Life of roller bearings.* W. Nelson (*Journal of Quality Technology*, July 1985) suggests that the Weibull distribution usually provides a better representation for the lifelength of a product than the exponential distribution. Nelson used a Weibull distribution with $\alpha = 1.5$ and $\beta = 110$ to model the lifelength $Y$ of a roller bearing (in thousands of hours).

a. Find the probability that a roller bearing of this type will have a service life of less than 12.2 thousand hours.
b. Recall that a Weibull distribution with $\alpha = 1$ is an exponential distribution. Nelson claims that very few products have an exponential life distribution, although such a distribution is commonly applied. Calculate the probability from part **a** using the exponential distribution. Compare your answer to that obtained in part **a**.

5.106 *Voltage readings.* The Harris Corporation data on voltage readings at two locations, Exercise 2.72 (p. 35), are reproduced at the bottom of the page. Determine whether the voltage readings at each location are approximately normal.

5.107 *Sedimentary deposits in reservoirs.* Geologists have successfully used statistical models to evaluate the nature of sedimentary deposits (called *facies*) in reservoirs. One of the model's key parameters is the proportion $P$ of facies bodies in a reservoir. An article in *Mathematical Geology* (Apr. 1995) demonstrated that the number of facies bodies that must be sampled to satisfactorily estimate $P$ is approximately normally distributed with $\mu = 99$ and $\sigma = 4.3$. How many facies bodies are required to satisfactorily estimate $P$ for 99% of the reservoirs evaluated?

5.108 *Water retention of soil cores.* A team of soil scientists investigated the water retention properties of soil cores sampled from an uncropped field consisting of silt loam (*Soil Science*, Jan. 1995). At a pressure of .1 megapascal (MPa), the water content of the soil (measured in cubic meters of water per cubic meter of soil) was determined to be approximately normally distributed with $\mu = .27$ and $\sigma = .04$. In addition to water content readings at a pressure of .1 MPa, measurements were obtained at pressures 0, .005, .01, .03, and 1.5 MPa. Consider a soil core with a water content reading of .14. Is it likely that this reading was obtained at a pressure of .1 MPa? Explain.

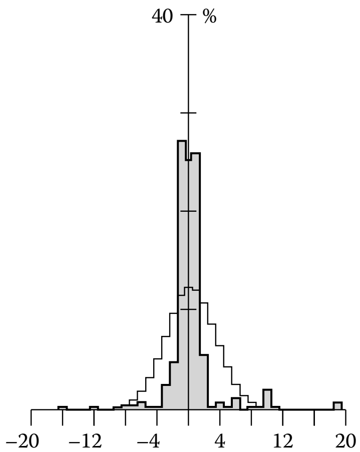
**VOLTAGE**

| Old Location | | | New Location | | |
|---|---|---|---|---|---|
| 9.98 | 10.12 | 9.84 | 9.19 | 10.01 | 8.82 |
| 10.26 | 10.05 | 10.15 | 9.63 | 8.82 | 8.65 |
| 10.05 | 9.80 | 10.02 | 10.10 | 9.43 | 8.51 |
| 10.29 | 10.15 | 9.80 | 9.70 | 10.03 | 9.14 |
| 10.03 | 10.00 | 9.73 | 10.09 | 9.85 | 9.75 |
| 8.05 | 9.87 | 10.01 | 9.60 | 9.27 | 8.78 |
| 10.55 | 9.55 | 9.98 | 10.05 | 8.83 | 9.35 |
| 10.26 | 9.95 | 8.72 | 10.12 | 9.39 | 9.54 |
| 9.97 | 9.70 | 8.80 | 9.49 | 9.48 | 9.36 |
| 9.87 | 8.72 | 9.84 | 9.37 | 9.64 | 8.68 |

*Source*: Harris Corporation, Melbourne, FL.

5.109 *Estimating glacier elevations.* Digital elevation models (DEM) are now used to estimate elevations and slopes of remote regions. In *Arctic, Antarctic, and Alpine Research* (May 2004), geographers analyzed reading errors from maps produced by DEM. Two readers of a DEM map of White Glacier (Canada) estimated elevations at 400 points in the area. The difference between the elevation estimates, $Y$, of the two readers had a mean of $\mu = .28$ meter and a standard deviation of $\sigma = 1.6$ meters. A histogram for $Y$ (with a normal histogram superimposed on the graph) is shown below.

a. Based on the histogram, the researchers concluded that $Y$ is not normally distributed. Why?
b. Will the interval $\mu \pm 2\sigma$ contain more than 95%, exactly 95%, or less than 95% of the 400 elevation differences? Explain.




*Source*: Cogley, J. G., and Jung-Rothenhausler, F. "Uncertainty in digital elevation models of Axel Heiberg Island, Arctic Canada." *Arctic, Antarctic, and Alpine Research*, Vol. 36, No. 2, May 2004 (Figure 3).

**LUMPYORE**

5.110 Refer to the data on percentage iron in 66 bulk specimens of Chilean lumpy iron ore, Exercise 2.79 (p. 71). The data are saved in the **LUMPYORE** file. Assess whether the data are approximately normal.

5.111 *Left-turn lane accidents.* Suppose we are counting events that occur according to a Poisson distribution, such as the number of automobile accidents at a left-turn lane. If it is known that exactly one such event has occurred in a given interval of time, say $(0, t)$, then the actual time of occurrence is uniformly distributed over this interval. Suppose that during a given 30-minute period, one accident occurred. Find the probability that the accident occurred during the last 5 minutes of the 30-minute period.

5.112 *Ship-to-shore transfer times.* Lack of port facilities or shallow water may require cargo on a large ship to be transferred to a pier using smaller craft. This process may require the smaller craft to cycle back and forth from ship to shore many times. Researchers G. Horne (Center for Naval Analysis) and T. Irony (George Washington University) developed models of this transfer process that provide estimates of ship-to-shore transfer times. (*Naval Research Logistics*, Vol. 41, 1994.) They modeled the time between arrivals of the smaller craft at the pier using an exponential distribution.

a. Assume the mean time between arrivals at the pier is 17 minutes. Give the value of $\alpha$ and $\beta$ for this exponential distribution. Graph the distribution.
b. Suppose there is only one unloading zone at the pier available for the small craft to use. If the first craft docks at 10:00 AM and doesn't finish unloading until 10:15 AM, what is the probability that the second craft will arrive at the unloading zone and have to wait before docking?

5.113 *Chemical impurity.* The percentage $Y$ of impurities per batch in a certain chemical product is a beta random variable with probability density

$$f(y) = \begin{cases} 90y^8(1 - y) & \text{if } 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. What are the values of $\alpha$ and $\beta$?
b. Compute the mean and variance of $Y$.
c. A batch with more than 80% impurities cannot be sold. What is the probability that a randomly selected batch cannot be sold because of excessive impurities?

5.114 *Lifelengths of memory chips.* The lifelength $Y$ (in years) of a memory chip in a laptop computer is a Weibull random variable with probability density

$$f(y) = \begin{cases} \frac{1}{8}ye^{-y^2/16} & \text{if } 0 \le y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

a. What are the values of $\alpha$ and $\beta$?
b. Compute the mean and variance of $Y$.
c. Find the probability that a new memory chip will not fail before 6 years.

5.115 *Auditory nerve fibers.* Refer to the *Journal of the Acoustical Society of America* (Feb. 1986) study of auditory nerve response rates in cats, discussed in Exercise 4.107 (p. 184). A key question addressed by the research is whether rate changes (i.e., changes in number of spikes per burst of noise) produced by tones in the presence of background noise are large enough to detect reliably. That is, can the tone be detected reliably when background noise is present? In the theory of signal detection, the problem involves a comparison of two probability distributions. Let $Y$ represent the auditory nerve response rate (i.e., the number of spikes observed) under two conditions: when the stimulus is background noise only (N) and when the stimulus is a tone plus background noise (T). The probability distributions for $Y$ under the two conditions are represented by the density functions, $f_N(y)$ and $f_T(y)$, respectively, where we assume that the mean response rate under the background-noise-only condition is less than the mean response rate under the tone-plus-noise condition, i.e., $\mu_N < \mu_T$. In this situation, an observer sets a threshold C and decides that a tone is present if $Y \geq C$ and decides that no tone is present if $Y < C$. Assume that $f_N(y)$ and $f_T(y)$ are both normal density functions with means $\mu_N = 10.1$ spikes per burst and $\mu_T = 13.6$ spikes per burst, respectively, and equal variances $\sigma_N^2 = \sigma_T^2 = 2$.

a. For a threshold of $C = 11$ spikes per burst, find the probability of detecting the tone given that the tone is present. (This is known as the *detection probability*.)
b. For a threshold of $C = 11$ spikes per burst, find the probability of detecting the tone given that only background noise is present. (This is known as the *probability of false alarm*.)
c. Usually, it is desirable to maximize detection probability while minimizing false alarm probability. Can you find a value of $C$ that will both increase the detection probability (part **a**) and decrease the probability of false alarm (part **b**)?

5.116 *Finding software coding errors.* In finding and correcting errors in a software code (*debugging*) and determining the code's reliability, computer software experts have noted the importance of the distribution of the time until the next coding error is found. Suppose that this random variable has a gamma distribution with parameter $\alpha = 1$. One computer programmer believes that the mean time between finding coding errors is $\beta = 24$ days. Suppose that a coding error is found today.

a. Assuming that $\beta = 24$, find the probability that it will take at least 60 days to discover the next coding error.
b. If the next coding error takes at least 60 days to find, what would you infer about the programmer's claim that the mean time between the detection of coding errors is $\beta = 24$ days? Why?

## Theoretical Exercises

5.117 Let $c$ be a constant and consider the density function for the random variable $Y$:

$$f(y) = \begin{cases} ce^{-y} & \text{if } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(2.6)$.
d. Show that $F(0) = 0$ and $F(\infty) = 1$.
e. Compute $P(1 \leq Y \leq 5)$.

5.118 The continuous random variable $Y$ has a probability distribution given by

$$f(y) = \begin{cases} cye^{-y^2} & \text{if } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$ that makes $f(y)$ a probability density.
b. Find $F(y)$.
c. Compute $P(Y > 2.5)$.

5.119 To aid engineers seeking to predict the efficiency of a solar-powered device, Olseth and Skartveit (*Solar Energy*, Vol. 33, No. 6, 1984) developed a model for daily insolation $Y$ at sea-level locations within the temperate storm belt. To account for both "clear sky" and "overcast" days, the researchers constructed a probability density function for $Y$ (measured as a percentage) using a linear combination of two modified gamma distributions:

$$f(y) = wg(y, \lambda_1) + (1 - w)g(1 - y, \lambda_2), \quad (0 < y < 1)$$

where

$$g(y) = \frac{(1 - y)e^{\lambda y}}{\int_0^1 (1 - y)e^{\lambda y}\,dy}$$

$\lambda_1$ = mean insolation of "clear sky" days

$\lambda_2$ = insolation of "cloudy" days

and $w$ is a weighting constant, $0 \leq w \leq 1$. Show that

$$\int_0^1 f(y)\,dy = 1$$

*5.120 Let $m_y(t)$ be the moment generating function of a continuous random variable $Y$. If $a$ and $b$ are constants, show that

a. $m_{y+a}(t) = E[e^{(y+a)t}] = e^{at}m_y(t)$

b. $m_{by}(t) = E[e^{(by)t}] = m_y(bt)$

c. $m_{[(y+a)/b]}(t) = E[e^{(y+a)t/b}] = e^{at/b}m_y\left(\frac{t}{b}\right)$

CHAPTER 6

# Bivariate Probability Distributions and Sampling Distributions

**OBJECTIVE**

*To introduce the concepts of a bivariate probability distribution, covariance, and independence; to show you how to find the expected value and variance of a linear function of random variables; to find and identify the probability distribution of a statistic (a sampling distribution)*

**CONTENTS**



*STATISTICS IN ACTION*

Availability of an Up/Down Maintained System

## *STATISTICS IN ACTION*

### Availability of an Up/Down Maintained System

In the Statistics in Action of Chapter 4 (p. 134), the reliability of a "one-shot" device or system was of interest. One-shot systems are non-maintained systems that either fulfill their objective by surviving beyond "mission" time or fail by perishing before the mission is accomplished. In contrast, maintained systems are systems that can be repaired and put back into operation when the system fails. The reliability of maintained systems was the topic of a United States Department of Defense publication (*START*, Vol. 11, 2004). The publication is intended to "help engineers better understand the meaning and implications of the statistical methods used to develop performance measures" for maintained systems.

During regular system maintenance, the system is typically "down" – that is, the system is unavailable for use while being repaired. Consequently, a key concept in system maintenance is "availability". By definition, "cycle availability" is the probability that the system is functioning at any point in time during the maintenance cycle. Cycle availability, $A$, can be expressed as a function of two continuous random variables. Let the random variable $X$ represent the time between failures of the system and let the random variable $Y$ represent the time to repair the system during a maintenance cycle. Thus, $X$ represents the "up" time of the system, $Y$ represents the "down" time, and $(X + Y)$ represents the total cycle time. Then,

$$A = X/(X + Y), \qquad X > 0 \quad \text{and} \quad Y > 0$$

The system maintenance engineer's goal is to understand the properties of availability, $A$. This includes expected (mean) availability and the 10th percentile of the availability values.

In the *Statistics in Action Revisited* section at the end of this chapter, we use the methods outlined in this chapter to find the probability density function for availability, $f(a)$, and its properties.

## 6.1 Bivariate Probability Distributions for Discrete Random Variables

Engineers responsible for estimating the cost of road construction utilize many variables to derive the estimate. For example, two important discrete random variables are $X$, the number of bridges that must be constructed, and $Y$, the number of structures that need to be leveled. Assessing the probability of $X$ and $Y$ taking specific values is key to developing an accurate estimate.

In Chapter 3, we learned that the probability of the intersection of two events (i.e., the event that both $A$ and $B$ occur) is equal to

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

If we assign two numbers to each point in the sample space—one corresponding to the value of a discrete random variable $X$ (e.g., the number of bridges built) and the second to a discrete random variable $Y$ (e.g., the number of structures leveled)—then specific values of $X$ and $Y$ represent two numerical events. The probability of the intersection of these two events is obtained by replacing the symbol $A$ by $X$ and the symbol $B$ by $Y$:

$$\begin{aligned} P(A \cap B) &= P(X = x, Y = y) = p(x, y) \\ &= p_1(x)p_2(y \mid x) \\ &= p_2(y)p_1(x \mid y) \end{aligned}$$

(*Note*: To distinguish between the probability distributions, we will always use the subscript 1 (as in $p_1$) when we refer to the probability distribution of $X$ and the subscript 2 (as in $p_2$) when we refer to the probability distribution of $Y$.)

A table, graph, or formula that gives the probability of the intersection $(x, y)$ for all values of $X$ and $Y$ is called the **joint probability distribution** of $X$ and $Y$. The probability distribution of $p_1(x)$ gives the probabilities of observing specific values of $X$; similarly, $p_2(y)$ gives the probabilities of the discrete random variable $Y$. Thus, $p_1(x)$ and $p_2(y)$, called **marginal probability distributions** for $X$ and $Y$, respectively, are the familiar unconditional probability distributions for discrete random variables encountered in Chapter 4.

**Definition 6.1**

The **joint probability distribution *p(x,y)*** for two discrete random variables, $X$ and $Y$—called a **bivariate distribution**—is a table, graph, or formula that gives the values of $p(x, y)$ for every combination of values of $X$ and $Y$.

### Requirements for a Discrete Bivariate Probability Distribution for *X* and *Y*

1. $0 \le p(x, y) \le 1$ for all values of $X$ and $Y$
2. $\sum_y \sum_x p(x, y) = 1$

(*Note*: The symbol $\sum_y \sum_x$ denotes summation over all values of both $X$ and $Y$.)

**Example 6.1**

Properties of a Discrete Bivariate Probability Distribution

Consider two discrete random variables, $X$ and $Y$, where $X = 1$ or $X = 2$, and $Y = 0$ or $Y = 1$. The bivariate probability distribution for $X$ and $Y$ is defined as follows:

$$p(x, y) = \frac{.25 + x - y}{5}$$

Verify that the properties (requirements) of a discrete bivariate probability distribution are satisfied.

Solution

Since $X$ takes on two values (1 or 2) and $Y$ takes on two values (0 or 1), there are $2 \times 2 = 4$ possible combinations of $X$ and $Y$. These four $(x, y)$ pairs are (1, 0), (1, 1), (2, 0), and (2, 1). Substituting these possible values of $X$ and $Y$ into the formula for $p(x, y)$, we obtain the following joint probabilities:

$$p(1, 0) = \frac{.25 + 1 - 0}{5} = .25$$

$$p(1, 1) = \frac{.25 + 1 - 1}{5} = .05$$

$$p(2, 0) = \frac{.25 + 2 - 0}{5} = .45$$

$$p(2, 1) = \frac{.25 + 2 - 1}{5} = .25$$

Note that each of these joint probabilities is between 0 and 1 (satisfying requirement 1 given in the box) and the sum of these four probabilities equals 1 (satisfying requirement 2).

## Example 6.2

Finding a Marginal Discrete Probability Distribution

Consider the bivariate joint probability distribution shown in Table 6.1. The values of $p(x, y)$ corresponding to pairs of values of the discrete random variables $X$ and $Y$, for $X = 1, 2, 3, 4$ and $Y = 0, 1, 2, 3$ are shown in the body of the table. For example, in a flexible manufacturing system, $X$ might represent the number of machines available and $Y$ might represent the number of sequential operations required to process a part. The table shows that the probability that 2 machines are available for a process that requires 1 operation is $P(2, 1) = .07$. Find the marginal probability distribution $p_1(x)$ for the discrete random variable $X$.

**TABLE 6.1 Bivariate Probability Distribution for *X* and *Y***

| | | $X = x$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | 0 | 0 | .10 | .20 | .10 |
| $Y = y$ | 1 | .03 | .07 | .10 | .05 |
| | 2 | .05 | .10 | .05 | 0 |
| | 3 | 0 | .10 | .05 | 0 |

Solution

To find the marginal probability distribution for $X$, we need to find $P(X = 1)$, $P(X = 2)$, $P(X = 3)$, and $P(X = 4)$. Since $X = 1$ can occur when $Y = 0, 1, 2$, or $3$ occurs, then $P(X = 1) = p_1(1)$ is calculated by summing the probabilities of four mutually exclusive events:

$$P(X = 1) = p_1(1) = p(1, 0) + p(1, 1) + p(1, 2) + p(1, 3)$$

Substituting the values for $p(x, y)$ given in Table 6.1, we obtain

$$P(X = 1) = p_1(1) = 0 + .03 + .05 + 0 = .08$$

Note that this marginal probability is obtained by summing the probabilities in the column $X = 1$ in Table 6.1.

Similarly,

$$P(X = 2) = p_1(2) = p(2, 0) + p(2, 1) + p(2, 2) + p(2, 3)$$
$$= .10 + .07 + .10 + .10 = .37$$
$$P(X = 3) = p_1(3) = p(3, 0) + p(3, 1) + p(3, 2) + p(3, 3)$$
$$= .20 + .10 + .05 + .05 = .40$$
$$P(X = 4) = p_1(4) = p(4, 0) + p(4, 1) + p(4, 2) + p(4, 3)$$
$$= .10 + .05 + 0 + 0 = .15$$

The marginal probability distribution $p_1(x)$ is given in the following table:

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $p_1(x)$ | .08 | .37 | .40 | .15 |

Note from the table that $\Sigma_{x=1}^{4} p_1(x) = 1$

Example 6.2 shows that the marginal probability distribution for a discrete random variable $X$ may be obtained by summing $p(x, y)$ over all values of $Y$. The result is summarized in the next box.

**Definition 6.2**

Let $X$ and $Y$ be discrete random variables and let $p(x, y)$ be their joint probability distribution. Then the **marginal (unconditional) probability distributions** of $X$ and $Y$ are, respectively,

$$p_1(x) = \sum_y p(x, y) \quad \text{and} \quad p_2(y) = \sum_x p(x, y)$$

(*Note*: We will use the symbol $\sum_y$ to denote summation over all values of $Y$.)

The probability of the numerical event $X$, given that the event $Y$ occurred, is the conditional probability of $X$ given $Y = y$. A table, graph, or formula that gives these probabilities for all values of $Y$ is called the **conditional probability distribution** for $X$ given $Y$ and is denoted by the symbol $p_1(x \mid y)$.

**Example 6.3**

**Finding a Conditional Discrete Probability Distribution**

Refer to Example 6.2. Find the conditional probability distribution of $X$ given $Y = 2$.

**Solution**

There are four conditional probability distributions of $X$—one for each value of $Y$. From Chapter 3, we know that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

If we let a value of $X$ correspond to the event $A$ and a value of $Y$ to the event $B$, then it follows that

$$p_1(x \mid y) = \frac{p(x, y)}{p_2(y)}$$

or, when $Y = 2$,

$$p_1(x \mid 2) = \frac{p(x, 2)}{p_2(2)}$$

Therefore,

$$p_1(1 \mid 2) = \frac{p(1, 2)}{p_2(2)}$$

From Table 6.1, we obtain $p(1, 2) = .05$ and $P(Y = 2) = p_2(2) = .2$. Therefore,

$$p_1(1 \mid 2) = \frac{p(1, 2)}{p_2(2)} = \frac{.05}{.20} = .25$$

Similarly,

$$p_1(2 \mid 2) = \frac{p(2, 2)}{p_2(2)} = \frac{.10}{.20} = .50$$

$$p_1(3 \mid 2) = \frac{p(3, 2)}{p_2(2)} = \frac{.05}{.20} = .25$$

$$p_1(4 \mid 2) = \frac{p(4, 2)}{p_2(2)} = \frac{.0}{.20} = 0$$

Therefore, the conditional probability distribution of $X$, given that $Y = 2$, is as shown in the following table:

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $p_1(x \mid 2)$ | .25 | .50 | .25 | 0 |

Note from Example 6.3 that the sum of the conditional probabilities $p_1(x \mid 2)$ over all values of $X$ is equal to 1. Thus, a conditional probability distribution satisfies the requirements that all probability distributions must satisfy:

$$p_1(x \mid y) \geq 0 \quad \text{and} \quad \sum_x p_1(x \mid y) = 1$$

Similarly,

$$p_2(y \mid x) \geq 0 \quad \text{and} \quad \sum_y p_2(y \mid x) = 1$$

**Definition 6.3**

Let $X$ and $Y$ be discrete random variables and let $p(x, y)$ be their joint probability distribution. Then the **conditional probability distributions** for $X$ and $Y$ are defined as follows:

$$p_1(x \mid y) = \frac{p(x, y)}{p_2(y)} \quad \text{and} \quad p_2(y \mid x) = \frac{p(x, y)}{p_1(x)}$$

In the preceding discussion, we defined the bivariate joint marginal and conditional probability distributions for two discrete random variables, $X$ and $Y$. The concepts can be extended to any number of discrete random variables. Thus, we could define a third random variable $W$ to each point in the sample space. The joint probability distribution $p(x, y, w)$ would be a table, graph, or formula that gives the values of $p(x, y, w)$, the event that the intersection $(x, y, w)$ occurs, for all combinations of values of $X$, $Y$, and $W$. In general, the joint probability distribution for two or more discrete random variables is called a **multivariate probability distribution**. Although the remainder of this chapter is devoted to bivariate probability distributions, the concepts apply to the general multivariate case also.

## Applied Exercises

6.1 *IF-THEN software code.* A software program is designed to perform two tasks, $A$ and $B$. Let $X$ represent the number of IF-THEN statements in the code for task $A$ and let $Y$ represent the number of IF-THEN statements in the code for task $B$. The joint probability distribution $p(x, y)$ for the two discrete random variables is given in the accompanying table.

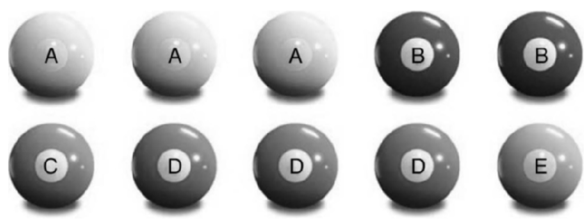
| | | $X$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 |
| | 0 | 0 | .050 | .025 | 0 | .025 | 0 |
| $Y$ | 1 | .200 | .050 | 0 | .300 | 0 | 0 |
| | 2 | .100 | 0 | 0 | 0 | .100 | .150 |

a. Verify that the properties of a joint probability distribution hold.
b. Find the marginal probability distribution $p_1(x)$ for $X$.
c. Find the marginal probability distribution $p_2(y)$ for $Y$.
d. Find the conditional probability distribution $p_1(x \mid y)$.
e. Find the conditional probability distribution $p_2(y \mid x)$.

6.2 *Tossing dice.* Consider the experiment of tossing a pair of dice. Let $X$ be the outcome (i.e., the number of dots appearing face up) on the first die and let $Y$ be the outcome on the second die.

a. Find the joint probability distribution $p(x, y)$.
b. Find the marginal probability distributions $p_1(x)$ and $p_2(y)$.
c. Find the conditional probability distributions $p_1(x \mid y)$ and $p_2(y \mid x)$.
d. Compare the probability distributions of parts **b** and **c**. What phenomenon have you observed?

6.3 *Cloning credit or debit cards.* Refer to the *IEEE Transactions on Information Forensics and Security* (March 2013) study of wireless identify theft using cloned credit or debit cards, Exercise 3.44 (p. 105). A cloning detection method was illustrated using a simple ball drawing game. Consider the group of 10 balls shown below. Of these, 5 represent genuine credit/debit cards and 5 represent clones of one or more of these cards. The 5 letters—A, B, C, D, and E—were used to distinguish among the different genuine cards. (Balls with the same letter represent either the genuine card or a clone of the card.) For this illustration, let $X$ represent the number of genuine balls selected and $Y$ represent the number of B-lettered balls selected when 2 balls are randomly drawn (without replacement) from the 10 balls.




a. Find the bivariate probability distribution, $p(x, y)$.
b. Find the marginal distribution, $p_1(x)$.
c. Find the marginal distribution, $p_2(y)$.
d. Recall that if two balls with the same letter are drawn from the 10 balls, then a cloning attack is detected. Consider a credit card identified by a B-lettered ball. Use your answer to part c to find the probability of a cloning attack for this card.

6.4 *Modeling the behavior of granular media.* Refer to the *Engineering Computations: International Journal for Computer-Aided Engineering and Software* (Vol. 30, No. 2, 2013) study of the properties of granular media (e.g., sand, rice, ball bearings, and flour), Exercise 3.62 (p. 120). The study assumes there is a system of $N$ non-interacting granular particles, where the particles are grouped according to energy level, $r$. For this problem (as in Exercise 3.62), assume that $N = 7$ and $r = 3$, then consider the scenario where there is one particle (of the total of 7 particles) at energy level 1, two particles at energy level 2, and four particles at energy level 3. Another feature of the particles studied was the position in time where the particle reached a certain entropy level during compression. All particles reached the desired entropy level at one of three time periods, 1, 2, or 3. Assume the 7 particles had the characteristics shown in the table. Consider a randomly selected particle and let $X$ represent the energy level and $Y$ the time period associated with particle.

| Particle ID | Energy Level | Time Period |
|---|---|---|
| 1 | 3 | 1 |
| 2 | 1 | 1 |
| 3 | 3 | 3 |
| 4 | 2 | 1 |
| 5 | 3 | 2 |
| 6 | 3 | 2 |
| 7 | 2 | 1 |

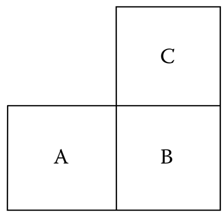
a. Find the bivariate probability distribution, $p(x, y)$.
b. Find the marginal distribution, $p_1(x)$.
c. Find the marginal distribution, $p_2(y)$.
d. Find the conditional distribution, $p_2(y \mid x)$.

6.5 *Variable speed limit control for freeways.* Refer to the *Canadian Journal of Civil Engineering* (Jan. 2013) investigation of the use of variable speed limits to control freeway traffic congestion, Exercise 4.9 (p. 140). Recall that the study site was an urban freeway divided into three sections with variable speed limits posted in each section. The probability distribution of the optimal speed limit for each of the three sections was determined. One possible set of distributions is as follows (probabilities in parentheses). *Section 1:* 30 mph (.06), 40 mph (.24), 50 mph (.24), 60 mph (.46); *Section 2:* 30 mph (.10), 40 mph (.24), 50 mph (.36), 60 mph (.30); *Section 3:* 30 mph (.15), 40 mph (.18), 50 mph (.30), 60 mph (.37). Consider a randomly selected vehicle traveling through the study site at a randomly selected time. For this vehicle, let $X$ represent the section and $Y$ represent the speed limit at the time of selection.

a. Which of the following probability distributions is represented by the given probabilities, $p(x,y)$, $p_1(x)$, $p_2(y)$, $p_1(x \mid y)$, or $p_2(y \mid x)$? Explain.
b. Assume that the sections are of equal length. Find $p_1(x)$. Justify your answer.
c. Find the bivariate distribution, $p(x, y)$.

6.6 *Robot-sensor system configuration.* Refer to *The International Journal of Robotics Research* (Dec. 2004) study of a robot-sensor system in an unknown environment, Exercise 4.7 (p. 139). In the three-point, single-link robotic system shown in the accompanying figure, each point (A, B, or C) in the system has either an "obstacle" status or a "free" status. Let $X = 1$ if the $A \leftrightarrow B$ link has an "obstacle" and $X = 0$ if the link is "free" (i.e., has no obstacles). Similarly, let $Y = 1$ if the $B \leftrightarrow C$ link has an "obstacle" and $Y = 0$ if the link is "free." Recall that the researchers assumed that the probability of any point in the system having a "free" status is .5 and that the three points in the system operate independently.

a. Give $p(x, y)$, the joint probability distribution of $X$ and $Y$, in table form.
b. Find the conditional probability distribution, $p_1(x \mid y)$.
c. Find the marginal probability distribution, $p_1(x)$.




6.7 *Red lights on truck route.* A special delivery truck travels from point A to point B and back over the same route each day. There are three traffic lights on this route. Let $X$ be the number of red lights the truck encounters on the way to delivery point B and let $Y$ be the number of red lights the truck encounters on the way back to delivery point A. A traffic engineer has determined the joint probability distribution of $X$ and $Y$ shown in the table.

| | | $X = x$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| | 0 | .01 | .02 | .07 | .01 |
| $Y = y$ | 1 | .03 | .06 | .10 | .06 |
| | 2 | .05 | .12 | .15 | .08 |
| | 3 | .02 | .09 | .08 | .05 |

a. Find the marginal probability distribution of $Y$.
b. Given that the truck encounters $X = 2$ red lights on the way to delivery point $B$, find the probability distribution of $Y$.

6.8 *Feasibility of a CPU cooler.* From a group of three data-processing managers, two senior systems analysts, and two quality control engineers, three people are to be randomly selected to form a committee that will study the feasibility of adding a dual-core CPU cooler at a consulting firm. Let $X$ denote the number of data-processing managers and $Y$ the number of senior systems analysts selected for the committee.

a. Find the joint probability distribution of $X$ and $Y$.
b. Find the marginal distribution of $X$.

6.9 *Face recognition technology.* The Face Recognition Technology (FERET) program, sponsored by the U.S. Department of Defense, was designed to develop automatic face recognition capabilities to assist homeland security. A biometric face "signature" of an unknown person (called the *probe*) is compared to a signature of a known individual from the "gallery" and a similarity score is measured. FERET includes algorithms for finding the gallery signature that best matches the probe signature by ranking similarity scores. In *Chance* (Winter 2004), the discrete *Copula probability distribution* was employed to compare algorithms.

Let $X$ represent the similarity score for a probe using algorithm A and $Y$ represent the similarity score for the same probe using algorithm B. Suppose that the gallery contains signatures for $n = 3$ known individuals, numbered 1, 2, and 3. Then $X_1$, $X_2$, and $X_3$ represent the similarity scores using algorithm A and $Y_1$, $Y_2$, and $Y_3$ represent the similarity scores using algorithm B. Now rank the $X$ values and define $X_{(i)}$ so that $X_{(1)} > X_{(2)} > X_{(3)}$. Similarly, rank the $Y$ values and define $Y_{(i)}$ so that $Y_{(1)} > Y_{(2)} > Y_{(3)}$. Then, the Copula distribution is the joint probability distribution of the ranked $X$'s and $Y$'s, given as follows:

$p(x, y) = 1/n$ if the pair $(X_{(x)}, Y_{(y)})$ is in the sample, 0 if not

where $x = 1, 2, 3, \ldots, n$ and $y = 1, 2, 3, \ldots, n$.

Suppose the similarity scores (measured on a 100-point scale) for a particular probe with $n = 3$ are $(X_1 = 75, Y_1 = 60)$, $(X_2 = 30, Y_2 = 80)$, and $(X_3 = 15, Y_3 = 5)$.

a. Give the Copula distribution, $p(x, y)$, for this probe in table form.
b. Demonstrate that if both algorithms agree completely on the signature match, then $p(1, 1) = p(2, 2) = p(3, 3) = 1/3$.

## Theoretical Exercises

6.10 The joint probability distribution for two discrete random variables, $X$ and $Y$, is given by the formula

$$p(x, y) = p^{x+y}q^{2-(x+y)}, \quad x = 0, 1, \quad 0 \le p \le 1, \quad q = 1 - p, \quad y = 0, 1$$

Verify that the properties of a bivariate probability distribution are satisfied.

6.11 Let $X$ and $Y$ be two discrete random variables with joint probability distribution $p(x, y)$. Define

$$F_1(a) = P(X \le a) \quad \text{and} \quad F_1(a \mid y) = P(X \le a \mid Y)$$

Verify each of the following:

a. $F_1(a) = \sum_{x \le a} \sum_{y} p(x, y)$

b. $F_1(a \mid y) = \dfrac{\sum_{x \le a} p(x, y)}{p_2(y)}$

# 6.2 Bivariate Probability Distributions for Continuous Random Variables

As we have noted in Chapters 4 and 5, definitions and theorems that apply to discrete random variables apply as well to continuous random variables. The only difference is that the probabilities for discrete random variables are summed, whereas those for continuous random variables are integrated. As we proceed through this chapter, we will define and develop concepts in the context of discrete random variables and will use them to justify equivalent definitions and theorems pertaining to continuous random variables.

**Definition 6.4**

The **bivariate joint probability density function** $f(x, y)$ for two continuous random variables $X$ and $Y$ is one that satisfies the following properties:

1. $f(x, y) \ge 0$ for all values of $X$ and $Y$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y) = 1$
3. $P(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f(x, y)\, dx\, dy$ for all constants $a$, $b$, $c$, and $d$

**Definition 6.5**

*Let* $f(x, y)$ be the joint density function for $X$ and $Y$. Then the **marginal density functions** for $X$ and $Y$ are

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{and} \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

**Definition 6.6**

Let $f(x, y)$ be the joint density function for $X$ and $Y$. Then the **conditional density functions** for $X$ and $Y$ are

$$f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} \quad \text{and} \quad f_2(y \mid x) = \frac{f(x, y)}{f_1(x)}$$

**Example 6.4**

Joint Density Function for Continuous Random Variables

Suppose the joint density function for two continuous random variables, $X$ and $Y$, is given by

$$f(x, y) = \begin{cases} cx & \text{if } 0 \le x \le 1;\, 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Determine the value of the constant $c$.

Solution

A graph of $f(x, y)$ traces a three-dimensional, wedge-shaped figure over the unit square $(0 \le x \le 1$ and $0 \le y \le 1)$ in the $(x, y)$-plane, as shown in Figure 6.1. The value of $c$ is chosen so that $f(x, y)$ satisfies the property

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

Performing this integration yields

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy &= \int_0^1\int_0^1 cx\, dx\, dy \\ &= c\int_0^1\int_0^1 x\, dx\, dy = c\int_0^1 \frac{x^2}{2}\bigg]_0^1 dy \\ &= c\int_0^1 \frac{1}{2}\, dy = \left(\frac{c}{2}\right)y\bigg]_0^1 = \frac{c}{2} \end{aligned}$$

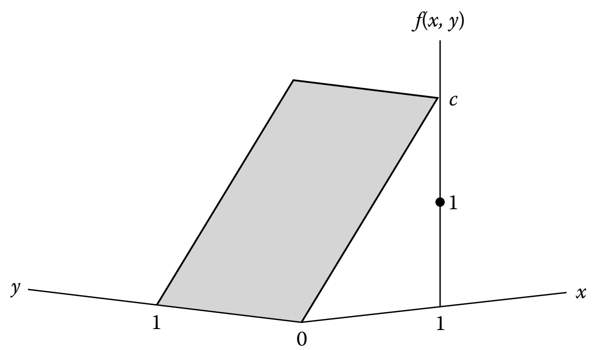



**FIGURE 6.1**
Graph of the joint density function for Example 6.4

Setting this quantity equal to 1 and solving for $c$, we obtain

$$1 = \frac{c}{2} \quad \text{or} \quad c = 2$$

Therefore,

$$f(x, y) = 2x \quad \text{for } 0 \le x \le 1 \quad \text{and} \quad 0 \le y \le 1$$

**Example 6.5**

Finding a Marginal Density Function

Refer to Example 6.4 and find the marginal density function for $X$. Show that

$$\int_{-\infty}^{\infty} f_1(x) = 1$$

Solution

By Definition 6.5,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = 2\int_0^1 x\, dy = 2xy\Big]_{y=0}^{y=1} = 2x, \quad 0 \le x \le 1$$

Thus,

$$\int_{-\infty}^{\infty} f_1(x)\, dx = 2\int_0^1 x\, dx = 2\left(\frac{x^2}{2}\right)\Big]_0^1 = 1$$

**Example 6.6**

Finding a Marginal Density Function

Refer to Example 6.4 and show that the marginal density function for $Y$ is a uniform distribution.

Solution

The marginal density function for $Y$ is given by

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = 2\int_0^1 x\, dx = 2\left(\frac{x^2}{2}\right)\Big]_0^1 = 1, \quad 0 \le y \le 1$$

Thus, $f_2(y)$ is a uniform distribution defined over the interval $0 \le y \le 1$.

**Example 6.7**

Finding a Continuous Density Function

Refer to Examples 6.4–6.6. Find the conditional density function for $X$ given $Y$, and show that it satisfies the property

$$\int_{-\infty}^{\infty} f_1(x \mid y)\, dx = 1$$

Solution

Using the marginal density function $f_2(y) = 1$ (obtained in Example 6.6) and Definition 6.6, we derive the conditional density function as follows:

$$f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{2x}{1} = 2x, \quad 0 \le x \le 1$$

We now show that the integral of $f_1(x \mid y)$ over all values of $X$ is equal to 1:

$$\int_0^1 f_1(x \mid y)\, dx = 2\int_0^1 x\, dx = 2\left(\frac{x^2}{2}\right)\Big]_0^1 = 1$$

**Example 6.8**

Joint Density Function—Range of $X$ Depends on $Y$

Suppose the joint density function for $X$ and $Y$ is

$$f(x, y) = \begin{cases} cx & \text{if } 0 \le x \le y; 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the value of $c$.

**Solution** Refer to Figure 6.1. If we pass a plane through the wedge, diagonally between the points (0, 0) and (1, 1), and perpendicular to the $(x, y)$-plane, then the slice lying along the $y$-axis will have a shape similar to that of the given density function (graphed in Figure 6.2). The value of $c$ will be larger than the value found in Example 6.4 because the volume of the solid shown in Figure 6.2 must equal 1.

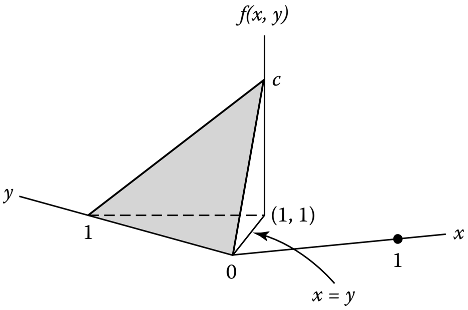



**FIGURE 6.2**
Graph of the joint density function for Example 6.8

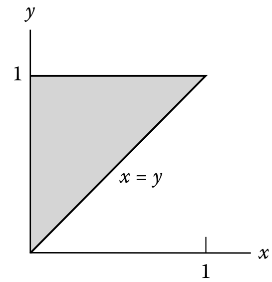



**FIGURE 6.3**
Region of integration for Example 6.8

We find $c$ by integrating $f(x, y)$ over the triangular region (shown in Figure 6.3) defined by $0 \le x \le y$ and $0 \le y \le 1$, setting this integral equal to 1, and solving for $c$:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = \int_0^1\int_0^y cx\,dx\,dy = c\int_0^1 \frac{x^2}{2}\Big]_0^y\,dy$$
$$= c\int_0^1 \frac{y^2}{2}\,dy = c\left(\frac{y^3}{6}\right)\Big]_0^1 = \frac{c}{6}$$

Setting this quantity equal to 1 and solving for $c$ yields $c = 6$; thus, $f(x, y) = 6x$ over the region of interest.

The joint density function for more than two random variables, say, $Y_1, Y_2, \ldots, Y_n$, is denoted by the symbol $f(y_1, y_2, \ldots, y_n)$. Marginal and conditional density functions are defined in a manner similar to that employed for the bivariate case.

## Applied Exercises

6.12 *Distribution of low bids.* The Department of Transportation (DOT) monitors sealed bids for new road construction. For new access roads in a certain state, let $X =$ low bid (thousands of dollars) and let $Y =$ DOT estimate of fair cost of building the road (thousands of dollars). The joint probability density of $X$ and $Y$ is

$$f(x, y) = \frac{e^{-y/10}}{10y}, \qquad 0 < y < x < 2y$$

a. Find $f(y)$, the marginal density function for $Y$. Do you recognize this distribution?
b. What is the mean DOT estimate, $E(Y)$?

6.13 *Characteristics of a truss subjected to loads.* In building construction, a truss is a structure comprised of triangular units whose ends are connected at joints referred to as nodes. The *Journal of Engineering Mechanics* (Dec. 2009) published a study of the characteristics of a 10-bar truss subjected to loads at four different nodes. Two random variables measured were stiffness index (pounds per square inch) and load (thousand pounds). One possible joint probability distribution of stiffness index $X$ and load $Y$ is given by the formula,

$$f(x, y) = (1/40)\, e^{-x}, \quad 0 < x < \infty, 80 < y < 120$$

a. Show that $\iint f(x, y)dydx = 1$.
b. Find the marginal density function, $f_1(x)$. Do you recognize this distribution?
c. Find the marginal density function, $f_2(y)$. Do you recognize this distribution?

6.14 *Servicing an automobile.* The joint density of $X$, the total time (in minutes) between an automobile's arrival in the service queue and its leaving the system after servicing, and $Y$, the time (in minutes) the car waits in the queue before being serviced, is

$$f(x, y) = \begin{cases} ce^{-x^2} & \text{if } 0 \le y \le x; 0 \le x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$ that makes $f(x, y)$ a probability density function.
b. Find the marginal density for $X$ and show that

$$\int_{-\infty}^{\infty} f_1(x)\, dx = 1$$

c. Show that the conditional density for $Y$ given $X$ is a uniform distribution over the interval $0 \le Y \le X$.

6.15 *Photocopier friction.* Refer to the *Journal of Engineering for Industry* (May 1993) study of friction feed paper separation, Exercise 5.12 (p. 196). Consider a system that utilizes two interrelated feed paper separators. The joint density of $X$ and $Y$, the friction coefficients of the two machines, is given by

$$f(x, y) = \begin{cases} xy & \text{if } 0 \le x \le 1; 0 \le y \le 1 \\ (2 - x)y & \text{if } 1 \le x \le 2; 0 \le y \le 1 \\ x(2 - y) & \text{if } 0 \le x \le 1; 1 \le y \le 2 \\ (2 - x)(2 - y) & \text{if } 1 \le x \le 2; 1 \le y \le 2 \end{cases}$$

a. Verify that $f(x, y)$ is a bivariate joint probability distribution function. [Show that Definition 6.4 holds for $f(x, y)$.]
b. Find the probability that both friction coefficients exceed .8.

## Theoretical Exercises

6.16 Let $X$ and $Y$ have the joint density

$$f(x, y) = \begin{cases} cxy & \text{if } 0 \le x \le 1; 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$ that makes $f(x, y)$ a probability density function.
b. Find the marginal densities $f_1(x)$ and $f_2(y)$.
c. Find the conditional densities $f_1(x \mid y)$ and $f_2(y \mid x)$.

6.17 Let $X$ and $Y$ have the joint density

$$f(x, y) = \begin{cases} x + cy & \text{if } 1 \le x \le 2; 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

where $c$ is a constant.

a. Find the value of $c$ that makes $f(x, y)$ a probability density function.
b. Find the marginal density for $y$ and show that

$$\int_{-\infty}^{\infty} f_2(y)\, dy = 1$$

c. Find $f_1(x \mid y)$, the conditional density for $X$ given $Y$.

6.18 Let $X$ and $Y$ be two continuous random variables with joint probability density

$$f(x, y) = \begin{cases} ce^{-(x+y)} & \text{if } 0 \le x < \infty; 0 \le y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$.
b. Find $f_1(x)$.
c. Find $f_2(y)$.
d. Find $f_1(x \mid y)$.
e. Find $f_2(y \mid x)$.
f. Find $P(X \le 1 \text{ and } Y \le 1)$.

6.19 Let $X$ and $Y$ be two continuous random variables with joint probability density $f(x, y)$. The joint distribution function $F(a, b)$ is defined as follows:

$$F(a, b) = P(X \le a, Y \le b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)\, dy\, dx$$

Verify each of the following:

a. $F(-\infty, -\infty) = F(-\infty, y) = F(x, -\infty) = 0$
b. $F(\infty, \infty) = 1$
c. If $a_2 \ge a_1$ and $b_2 \ge b_1$, then

$$F(a_2, b_2) - F(a_1, b_2) \ge F(a_2, b_1) - F(a_1, b_1)$$

# 6.3 The Expected Value of Functions of Two Random Variables

The statistics that we will subsequently use for making inferences are computed from the data contained in a sample. The sample measurements can be viewed as observations on $n$ random variables, $Y_1, Y_2, \ldots, Y_n$, where $Y_1$ represents the first measurement in the sample, $Y_2$ represents the second measurement, etc. Since the sample statistics are functions of the random variables $Y_1, Y_2, \ldots, Y_n$, they also will be random variables and will possess probability distributions. To describe these distributions, we will define the expected value (or mean) of functions of two or more random variables and present three expectation theorems that correspond to those given in Chapter 5. The definitions and theorems will be given in the bivariate context, but they can be written in general for any number of random variables by substituting corresponding multivariate functions and notation.

**Definition 6.7**

Let $g(X, Y)$ be a function of the random variables $X$ and $Y$. Then the **expected value (mean)** of $g(X, Y)$ is defined to be

$$E[g(X, Y)] = \begin{cases} \sum_y \sum_x g(x, y)p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,dx\,dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Suppose $g(X, Y)$ is a function of only one of the random variables, say, $X$. We will show that, in the discrete situation, the expected value of this function possesses the same meaning as in Chapter 5. Let $g(X, Y)$ be a function of $X$ only, i.e., $g(X, Y) = g(X)$. Then

$$E[g(X)] = \sum_x \sum_y g(x)p(x, y)$$

Summing first over $Y$ (in which case, $X$ is regarded as a constant that can be factored outside the summation sign), we obtain

$$E[g(X)] = \sum_x g(x) \sum_y p(x, y)$$

However, by Definition 6.2, $\sum_y p(x, y)$ is the marginal probability distribution for $X$. Therefore,

$$E[g(X)] = \sum_x g(x)p_1(x)$$

You can verify that this is the same expression given for $E[g(X)]$ in Definition 4.5. An analogous result holds (proof omitted) if $X$ and $Y$ are continuous random variables. Thus, if $(\mu_x, \sigma_x^2)$ and $(\mu_y, \sigma_y^2)$ denote the means and variances of $X$ and $Y$, respectively, then the bivariate expectations for functions of either $x$ or $y$ will equal the corresponding expectations given in Chapter 5, i.e., $E(X) = \mu_x$, $E[(x - \mu_x)^2] = \sigma_x^2$, etc.

It can be shown (proof omitted) that the three expectation theorems of Chapter 5 hold for bivariate and, in general, for multivariate probability distributions. We will use these theorems in Sections 6.5 and 6.6.

**THEOREM 6.1**

Let $c$ be a constant. Then the expected value of $c$ is

$$E(c) = c$$

**THEOREM 6.2**

Let $c$ be a constant and let $g(X, Y)$ be a function of the random variables $X$ and $Y$. Then the expected value of $cg(X, Y)$ is

$$E[cg(X, Y)] = cE[g(X, Y)]$$

**THEOREM 6.3**

Let $g_1(X, Y), g_2(X, Y), \ldots, g_k(X, Y)$ be $k$ functions of the random variables $X$ and $Y$. Then the expected value of the sum of these functions is

$$E[g_1(X, Y) + g_2(X, Y) + \cdots + g_k(X, Y)]$$
$$= E[g_1(X, Y)] + E[g_2(X, Y)] + \cdots + E[g_k(X, Y)]$$

## Applied Exercises

6.20 *Cloning credit or debit cards.* Refer to the *IEEE Transactions on Information Forensics and Security* (March 2013) study of wireless identify theft using cloned credit or debit cards, Exercise 6.3 (p. 239). On average, how many genuine balls will be drawn when 2 balls are randomly selected from the 10 balls?

6.21 *Variable speed limit control for freeways.* Refer to the *Canadian Journal of Civil Engineering* (Jan. 2013) study, Exercise 6.5 (p. 240).

a. Find $E(X)$. Interpret this result.
b. Find $E(Y)$. Interpret this result.

6.22 *Red lights on truck route.* Refer to Exercise 6.7 (p. 240).

a. On the average, how many red lights should the truck expect to encounter on the way to delivery point B, i.e., what is $E(X)$?
b. The total number of red lights encountered over the entire route—that is, going to point B and back to point A—is $(X + Y)$. Find $E(X + Y)$.

6.23 *Distribution of low bids.* Refer to Exercise 6.12 (p. 244).

a. Find $E(Y - 10)$.
b. Find $E(3Y)$.

## Theoretical Exercises

6.24 Refer to Exercise 6.16 (p. 245).

a. Find $E(X)$.
b. Find $E(Y)$.
c. Find $E(X + Y)$.
d. Find $E(XY)$.

6.25 Refer to Exercise 6.17 (p. 245).

a. Find $E(X)$.
b. Find $E(Y)$.
c. Find $E(X + Y)$.
d. Find $E(XY)$.

6.26 Let $X$ and $Y$ be two continuous random variables with joint probability distribution $f(x, y)$. Consider the function $g(X)$. Show that

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_1(x)\,dx$$

6.27 Prove Theorems 6.1–6.3 for discrete random variables $X$ and $Y$.

6.28 Prove Theorems 6.1–6.3 for continuous random variables $X$ and $Y$.

# 6.4 Independence

In Chapter 3 we learned that two events $A$ and $B$ are said to be independent if $P(A \cap B) = P(A)P(B)$. Then, since the values assumed by two discrete random variables, $X$ and $Y$, represent two numerical events, it follows that $X$ and $Y$ are **independent** if $p(x, y) = p_1(x)p_2(y)$. Two continuous random variables are said to be independent if they satisfy a similar criterion.

**Definition 6.8**

Let $X$ and $Y$ be discrete random variables with joint probability distribution $p(x, y)$ and marginal probability distributions $p_1(x)$ and $p_2(y)$. Then $X$ and $Y$ are said to be **independent** if and only if

$$p(x, y) = p_1(x)p_2(y) \quad \text{for all pairs of values of } x \text{ and } y$$

**Definition 6.9**

Let $X$ and $Y$ be continuous random variables with joint density function $f(x, y)$ and marginal density functions $f_1(x)$ and $f_2(y)$. Then $X$ and $Y$ are said to be **independent** if and only if

$$f(x, y) = f_1(x)f_2(y) \quad \text{for all pairs of values of } x \text{ and } y$$

**Example 6.9**

Demonstrating Independence

Refer to Example 6.4 and determine whether $X$ and $Y$ are independent.

Solution

From Examples 6.4–6.6, we have the following results:

$$f(x, y) = 2x \qquad f_1(x) = 2x \qquad f_2(y) = 1$$

Therefore,

$$f_1(x)f_2(y) = (2x)(1) = 2x = f(x, y)$$

and, by Definition 6.9, $X$ and $Y$ are independent random variables.

**Example 6.10**

Demonstrating Dependence

Refer to Example 6.8 and determine whether $X$ and $Y$ are independent.

Solution

From Example 6.8, we determined that $f(x, y) = 6x$ when $0 \le x \le y$ and $0 \le y \le 1$. Therefore,

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy = \int_x^1 6x\, dy = 6xy\Big]_x^1 \\ &= 6x(1 - x) \quad \text{where } 0 \le x \le 1 \end{aligned}$$

Similarly,

$$\begin{aligned} f_2(y) &= \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^y 6x\, dx = \frac{6x^2}{2}\Big]_0^y \\ &= 3y^2 \quad \text{where } 0 \le y \le 1 \end{aligned}$$

You can see that $f_1(x)f_2(y) = 18x(1 - x)y^2$ is *not* equal to $f(x, y)$. Therefore, $X$ and $Y$ are *not* independent random variables.

Theorem 6.4 points to a useful consequence of independence.

**THEOREM 6.4**

If $X$ and $Y$ are independent random variables, then

$$E(XY) = E(X)E(Y)$$

**Proof of Theorem 6.4** We will prove the theorem for the discrete case. The proof for the continuous case is identical, except that integration is substituted for summation. By the definition of expected value, we have

$$E(XY) = \sum_y \sum_x xyp(x, y)$$

But, since $X$ and $Y$ are independent, we can write $p(x, y) = p_1(x)p_2(y)$. Therefore,

$$E(XY) = \sum_y \sum_x xyp_1(x)p_2(y)$$

If we sum first with respect to $X$, then we can treat $Y$ and $p_2(y)$ as constants and apply Theorem 6.2 to factor them out of the sum as follows:

$$E(XY) = \sum_y yp_2(y) \sum_x xp_1(x)$$

But,

$$\sum_x xp_1(x) = E(x) \quad \text{and} \quad \sum_y yp_2(y) = E(y)$$

Therefore,

$$E(XY) = E(X)E(Y)$$

## Applied Exercises

6.29 *IF-THEN software code.* Refer to Exercise 6.1 (p. 215). Are $X$ and $Y$ independent?

6.30 *Tossing dice.* Refer to Exercise 6.2 (p. 239). Are $X$ and $Y$ independent?

6.31 *Cloning credit or debit cards.* Refer to Exercise 6.3 (p. 239). Are $X$ and $Y$ independent?

6.32 *Robot-sensor system configuration.* Refer to Exercise 6.6 (p. 240). Are $X$ and $Y$ independent?

6.33 *Reliability of a manufacturing network.* Refer to the *Journal of Systems Sciences & Systems Engineering* (March 2013) study of the reliability of a manufacturing system for producing integrated circuit (IC) cards that involves two production lines, Exercise 4.6 (p. 139). Recall that items (IC cards) first pass through Line 1, then are processed by Line 2. The probability distribution of the maximum capacity level of each line is reproduced below. Consider an IC card randomly selected at some point in the production process. Let $X$ represent the line number and $Y$ represent the maximum capacity of the line at the time of selection. Assuming the lines operate independently, find the bivariate probability distribution $p(x, y)$.

| Line, $X$ | Maximum Capacity, $Y$ | $p(y)$ |
|---|---|---|
| 1 | 0 | .01 |
| | 12 | .02 |
| | 24 | .02 |
| | 36 | .95 |
| 2 | 0 | .002 |
| | 35 | .002 |
| | 70 | .996 |

6.34 *Modeling annual rainfall and peaks.* In the *Journal of Hydrological Sciences* (April 2000), the bivariate normal distribution was used to model the joint distribution of annual storm peak (i.e., maximum rainfall intensity) and total yearly rainfall amount in Tokushima, Japan. Let $X$ represent storm peak and $Y$ represent total amount of rainfall. Then the bivariate normal distribution is given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

where $\mu_x$ and $\mu_y$ are the means for $X$ and $Y$, respectively, and $\sigma_x$ and $\sigma_y$ are the standard deviations for $X$ and $Y$, respectively. Use the properties of density functions to show that when $\rho = 0$, $X$ and $Y$ are independent.

6.35 *Lifelengths of fuses.* The lifelength $Y$ (in hundreds of hours) for fuses used in a televideo computer terminal has an exponential distribution with mean $\beta = 5$. Each terminal requires two such fuses—one acting as a backup that comes into use only when the first fuse fails.

a. If two such fuses have independent lifelengths $X$ and $Y$, find the joint density $f(x, y)$.
b. The total effective lifelength of the two fuses is $(X + Y)$. Find the expected total effective lifelength of a pair of fuses in a televideo computer terminal.

6.36 *Lifetimes of components.* Let $X$ and $Y$ denote the lifetimes of two different types of components in an electronic system. The joint density of $X$ and $Y$ is given by

$$f(x, y) = \begin{cases} \frac{1}{8}xe^{-(x+y)/2} & \text{if } x > 0; y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Show that $X$ and $Y$ are independent. [*Hint*: A theorem in multivariate probability theory states that $X$ and $Y$ are independent if we can write

$$f(x, y) = g(x)h(y)$$

where $g(X)$ is a nonnegative function of $X$ only and $h(Y)$ is a nonnegative function of $Y$ only.]

6.37 *Photocopier friction.* Refer to Exercise 6.15. Show that $X$ and $Y$ are independent. [Use the hint, Exercise 6.36.]

## Theoretical Exercises

6.38 Refer to Exercise 6.16 (p. 245). Are $X$ and $Y$ independent?

6.39 Refer to Exercise 6.17 (p. 245). Are $X$ and $Y$ independent?

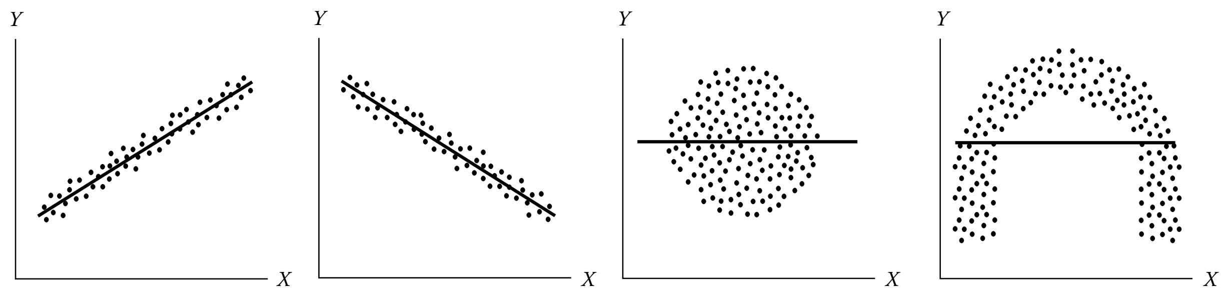
6.40 Prove Theorem 6.4 for the continuous case.

## 6.5 The Covariance and Correlation of Two Random Variables

When we think of two variables $X$ and $Y$ being related, we usually imagine a relationship in which $Y$ increases as $X$ increases or $Y$ decreases as $X$ increases. In other words, we tend to think in terms of **linear relationships**.

If $X$ and $Y$ are random variables and we collect a sample of $n$ pairs of values $(x, y)$, it is unlikely that the plotted data points would fall exactly on a straight line. If the points lie very close to a straight line, as in Figures 6.4a and 6.4b, we think of the linear relationship between $X$ and $Y$ as being very strong. If they are widely scattered about a line, as in Figures 6.4c and 6.4d, we think of the linear relationship as weak. (Note that the relationship between $X$ and $Y$ in Figure 6.4d is strong in a curvilinear manner.) How can we measure the strength of the linear relationship between two random variables, $X$ and $Y$?

One way to measure the strength of a linear relationship is to calculate the cross-product of the deviations $(x - \mu_x)(y - \mu_y)$ for each data point. These cross-products will be positive when the data points are in the upper right or lower left quadrant of Figure 6.5 and negative when the points are in the upper left or lower right quadrant. If all the points lie close to a line with positive slope, as in Figure 6.4a, almost all the cross-products $(x - \mu_x)(y - \mu_y)$ will be positive and their mean value will be relatively large and positive. Similarly, if all the points lie close to a line with a negative slope, as in Figure 6.4b, the mean value of $(x - \mu_x)(y - \mu_y)$ will be a relatively large negative number. However, if the linear relationship between $X$ and $Y$ is relatively weak, as in Figure 6.4c, the points will fall in all four quadrants, some cross-products $(x - \mu_x)(y - \mu_y)$ will be positive, some will be negative, and their mean value will be relatively small—perhaps very close to 0. This leads to the following definition of a measure of the strength of the linear relationship between two random variables.




a. Strong positive relationship

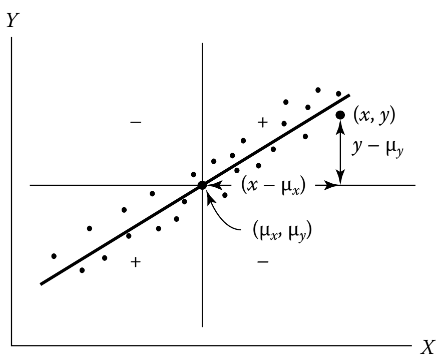
b. Strong negative relationship

c. Weak relationship

d. Weak linear relationship (strong nonlinear relationship)

**FIGURE 6.4**
Linear relationships between $X$ and $Y$

**FIGURE 6.5**
Signs of the cross-products $(x - \mu_x)(y - \mu_y)$



**Definition 6.10**

The **covariance** of two random variables, $X$ and $Y$, is defined to be

$$\mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

**THEOREM 6.5**

$$\mathrm{Cov}(X, Y) = E(XY) - \mu_x\mu_y$$

**Proof of Theorem 6.5** By Definition 6.10, we can write

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[(X - \mu_x)(Y - \mu_y)]\\ &= E(XY - \mu_x Y - \mu_y X + \mu_x\mu_y)\end{aligned}$$

Applying Theorems 6.1, 6.2, and 6.3 yields

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x\mu_y\\ &= E(XY) - \mu_x\mu_y - \mu_x\mu_y + \mu_x\mu_y\\ &= E(XY) - \mu_x\mu_y\end{aligned}$$

**Example 6.11**

Finding Covariance

Find the covariance of the random variables $X$ and $Y$ of Example 6.4.

Solution

The variables have joint density function $f(x, y) = 2x$ when $0 \le x \le 1$ and $0 \le y \le 1$. Then,

$$\begin{aligned}E(XY) &= \int_0^1 \int_0^1 (xy)2x\,dx\,dy\\ &= \int_0^1 2\left(\frac{x^3}{3}\right)\Bigg]_0^1 y\,dy = \frac{2}{3}\int_0^1 y\,dy = \frac{2}{3}\left(\frac{y^2}{2}\right)\Bigg]_0^1 = \frac{1}{3}\end{aligned}$$

In Examples 6.5 and 6.6, we obtained the marginal density functions $f_1(x) = 2x$ and $f_2(y) = 1$. Therefore,

$$\mu_x = E(X) = \int_0^1 x f_1(x)\,dx = \int_0^1 x(2x)\,dx = 2\left(\frac{x^3}{3}\right)\Bigg]_0^1 = \frac{2}{3}$$

Furthermore, since $Y$ is a uniform random variable defined over the interval $0 \le y \le 1$ (see Example 6.6), it follows from Section 5.4 that $\mu_y = \frac{1}{2}$. Then,

$$\mathrm{Cov}(X, Y) = E(XY) - \mu_x\mu_y = \frac{1}{3} - \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) = 0$$

Example 6.11 demonstrates an important result: If $X$ and $Y$ are independent, then their covariance will equal 0. However, *the converse is generally not true*.*

**THEOREM 6.6**

If two random variables $X$ and $Y$ are independent, then

$$\mathrm{Cov}(X, Y) = 0$$

*It can be shown (proof omitted) that if $X$ and $Y$ are jointly normally distributed, the converse is true.

The proof of Theorem 6.6, which follows readily from Theorem 6.5, is left as an optional exercise.

If the covariance between two random variables is positive, then $Y$ tends to increase as $X$ increases. If the covariance is negative, then $Y$ tends to decrease as $X$ increases. But what can we say about the numerical value of the covariance? We know that a covariance equal to 0 means that there is no linear relationship between $X$ and $Y$, but when the covariance is nonzero, its absolute value will depend on the units of measurement of $X$ and $Y$. To overcome this difficulty, we define a standardized version of the covariance known as the **coefficient of correlation**.

**Definition 6.11**

The **coefficient of correlation** $\boldsymbol{\rho}$ for two random variables $X$ and $Y$ is

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

where $\boldsymbol{\sigma_x}$ and $\sigma_y$ are the standard deviations of $X$ and $Y$, respectively.

Since $\rho$ is equal to the covariance divided by the product of two positive quantities, $\sigma_x$ and $\sigma_y$, it will have the same sign as the covariance but, in addition, it will (proof omitted) assume a value in the interval $-1 \leq \rho \leq 1$. Values of $\rho = -1$ and $\rho = 1$ imply perfect straight-line relationships between $X$ and $Y$, the former with a negative slope and the latter with a positive slope. A value of $\rho = 0$ implies no linear relationship between $X$ and $Y$.

Property of the Correlation Coefficient

$$-1 \leq \rho \leq 1$$

## Applied Exercises

6.41 *IF-THEN software code.* Find the covariance of the random variables $X$ and $Y$ in Exercise 6.1 (p. 239).

6.42 *Tossing dice.* Find the covariance of the random variables $X$ and $Y$ in Exercise 6.2 (p. 239).

6.43 *Variable speed limit control for freeways.* Find the correlation coefficient $\rho$ for $X$ and $Y$ in Exercise 6.5 (p. 240).

6.44 *Red lights on truck route.* Refer to Exercise 6.7 (p. 240).

a. Find the covariance of the random variables $x$ and $y$.
b. Find the coefficient of correlation $\rho$ for $x$ and $y$.

6.45 *Capacity of tank kerosene.* Commercial kerosene is stocked in a bulk tank at the beginning of each week. Because of limited supplies, the proportion $X$ of the capacity of the tank available for sale and the proportion $Y$ of the capacity of the tank actually sold during the week are continuous random variables. Their joint distribution is given by

$$f(x, y) = \begin{cases} 4x^2 & \text{if } 0 \leq y \leq x; 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the covariance of $X$ and $Y$.

6.46 *Modeling annual rainfall and peaks.* Refer to the *Journal of Hydrological Sciences* (April 2000) study of rainfall in Tokushima, Japan, Exericse 6.34 (p. 249). Recall that the bivariate normal distribution was used to model the joint distribution of annual storm peak $X$ (millimeters/day) and annual rainfall amount $Y$ (millimeters). The article used the following values as estimates of the distribution's parameters: $\mu_x = 147$ mm/day, $\sigma_x = 59$ mm/day, $\mu_y = 223$ mm, $\sigma_y = 117$ mm, and $\rho = .67$. Use this information to find the covariance between $X$ and $Y$.

## Theoretical Exercises

6.47 Refer to Exercise 6.16 (p. 245).

a. Find the covariance of the random variables $X$ and $Y$.
b. Find the coefficient of correlation $\rho$ for $X$ and $Y$.

6.48 Refer to Exercise 6.17 (p. 245).

a. Find the covariance of the random variables $X$ and $Y$.
b. Find the coefficient of correlation $\rho$ for $X$ and $Y$.

6.49 Prove Theorem 6.6 for the discrete case.

6.50 Prove Theorem 6.6 for the continuous case.

6.51 As an illustration of why the converse of Theorem 6.6 is not true, consider the joint distribution of two discrete random variables, $X$ and $Y$, shown in the accompanying table. Show that $\text{Cov}(X, Y) = 0$, but that $X$ and $Y$ are dependent.

| | | $X = x$ | | |
|---|---|---|---|---|
| | | $-1$ | $0$ | $+1$ |
| | $-1$ | $\frac{1}{12}$ | $\frac{2}{12}$ | $\frac{1}{12}$ |
| $Y = y$ | $0$ | $\frac{2}{12}$ | $0$ | $\frac{2}{12}$ |
| | $+1$ | $\frac{1}{12}$ | $\frac{2}{12}$ | $\frac{1}{12}$ |

6.52 Find the covariance of $X$ and $Y$ for the random variables of Exercise 6.18 (p. 245).

## 6.6 Probability Distributions and Expected Values of Functions of Random Variables (*Optional*)

In the three previous sections, we considered functions of two random variables. In this optional section, we consider the more general case, i.e., functions of one or more random variables.

There are essentially three methods for finding the density function for a function of random variables. Two of these—the moment generating function method and the transformation method—are beyond the scope of this text, but a discussion of them can be found in the references at the end of the chapter. The third method, which we will call the **cumulative distribution function method**, will be demonstrated with examples.

Suppose $W$ is a function of one or more random variables. The cumulative distribution function method finds the density function for $W$ by first finding the probability $P(W \le w)$, which is equal to $F(w)$. The density function $f(w)$ is then found by differentiating $F(w)$ with respect to $w$. We will demonstrate the method in Examples 6.12 and 6.13.

**Example 6.12**

Applying the Cumulative Distribution Function Method

Suppose the random variable $Y$ has an exponential density function

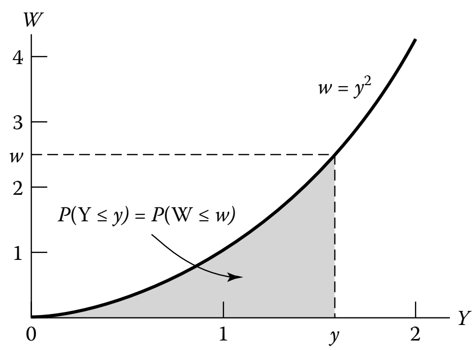
$$f(y) = \begin{cases} \dfrac{e^{-y/\beta}}{\beta} & \text{if } 0 \le y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

and let $W = Y^2$. Find the density function for the random variable $W$.

Solution

A graph of $W = Y^2$ is shown in Figure 6.6. We will denote the cumulative distribution functions of $W$ and $Y$ as $G(w)$ and $F(y)$, respectively. We note from the figure that $W$

**FIGURE 6.6**
A graph of $W = Y^2$



will be less than $w$ whenever $Y$ is less than $y$; it follows that $P(W \le w) = G(w) = F(y)$. Since $W = Y^2$, we have $y = \sqrt{w}$ and

$$F(y) = F(\sqrt{w}) = \int_{-\infty}^{\sqrt{w}} f(y)\, dy = \int_{0}^{\sqrt{w}} \frac{e^{-y/\beta}}{\beta}\, dy = -e^{-y/\beta}\Big]_0^{\sqrt{w}} = 1 - e^{-(\sqrt{w}/\beta)}$$

Therefore, the cumulative distribution function for $W$ is

$$G(w) = 1 - e^{-(\sqrt{w}/\beta)}$$

Differentiating, we obtain the density function for $W$:

$$\frac{dG(w)}{dw} = g(w) = \frac{w^{-1/2}e^{-(\sqrt{w}/\beta)}}{2\beta}$$

**Example 6.13**

**Finding the Distribution of a Sum of Random Variables**

If the random variables $X$ and $Y$ possess a uniform joint density function over the unit square, then $f(x, y) = 1$ for $0 \le X \le 1$ and $0 \le Y \le 1$. Find the density function for the sum $W = X + Y$.

**Solution**

Each value of $W$ corresponds to a series of points on the line $w = x + y$ (see Figure 6.7). Written in the slope–intercept form, $y = w - x$, this is the equation of a line with slope equal to $-1$ and $y$-intercept equal to $w$. The values of $W$ that are less than or equal to $w$ are those corresponding to points $(x, y)$ below the line $w = x + y$. (This area is shaded in Figure 6.7.) Then, for values of the $y$-intercept $w$, $0 \le w \le 1$, the probability that $W$ is less than or equal to $w$ is equal to the volume of a solid over the shaded area shown in the figure. We could find this probability by multiple integration, but it is easier to obtain it with the aid of geometry. Each of the two equal sides of the triangle has length $w$. Therefore, the area of the shaded triangular region is $w^2/2$, the height of the solid over the region is $f(x, y) = 1$, and the volume is

$$P(W \le w) = G(w) = w^2/2 \qquad (0 \le w \le 1)$$

The equation for $G(w)$ is different over the interval $1 \le w \le 2$. The probability $P(W \le w) = G(w)$ is the integral of $f(x, y) = 1$ over the shaded area shown in Figure 6.8. The integral can be found by subtracting from 1 the volume corresponding

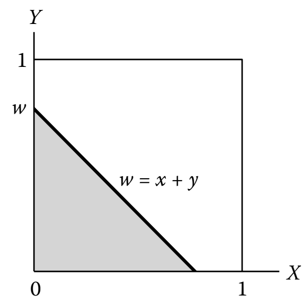



**FIGURE 6.7**
A graph showing the region of integration to find $G(w)$, $0 \le w \le 1$

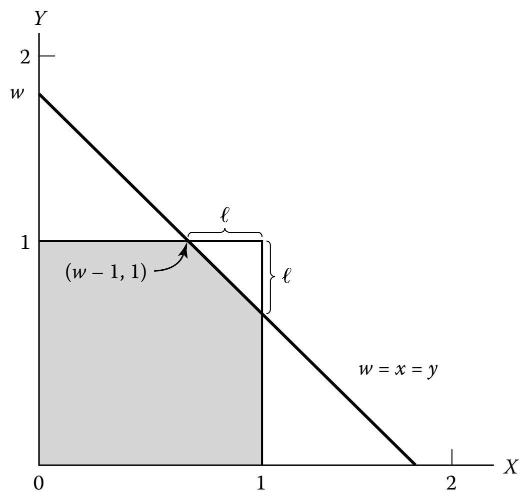



**FIGURE 6.8**
A graph showing the region of integration to find $G(w)$, $1 \le w \le 2$

to the small triangular (nonshaded) area that lies above the line $w = x + y$. To find the length of one side of this triangle, we need to locate the point where the line $w = (x + y)$ intersects the line $y = 1$. Substituting $y = 1$ into the equation of the line, we find

$$w = x + 1 \quad \text{or} \quad x = w - 1$$

The point $(w - 1, 1)$ is shown in Figure 6.8. The two equal sides of the triangle each have length $\ell = 1 - (w - 1) = 2 - w$. The area of the triangle lying above the line $w = x + y$ is then

$$\begin{aligned}\text{Area} &= \frac{1}{2}(\text{Base})(\text{Height})\\ &= \frac{1}{2}(2 - w)(2 - w) = \frac{(2 - w)^2}{2}\end{aligned}$$

Since the height of the solid constructed over the triangle is $f(x, y) = 1$, the probability that $W$ lies above the line $w = x + y$ is $(2 - w)^2/2$. Subtracting this probability from 1, we find the probability that $W$ lies below the line to be

$$\begin{aligned}G(w) = P(W \le w) &= 1 - \frac{(2 - w)^2}{2} = 1 - \frac{4 - 4w + w^2}{2}\\ &= -1 + 2w - w^2/2 \qquad (1 \le w \le 2)\end{aligned}$$

The density function for the sum of the two random variables $X$ and $Y$ is now obtained by differentiating $G(w)$:

$$g(w) = \frac{dG(w)}{dw} = \frac{d(w^2/2)}{dw} = w \qquad (0 \le w \le 1)$$

$$g(w) = \frac{dG(w)}{dw} = \frac{d(-1 + 2w - w^2/2)}{dw} = 2 - w \qquad (1 \le w \le 2)$$

Graphs of the cumulative distribution function and the density function for $W = X + Y$ are shown in Figures 6.9a and 6.9b, respectively. Note that the area under the density function over the interval $0 \le w \le 2$ is equal to 1.

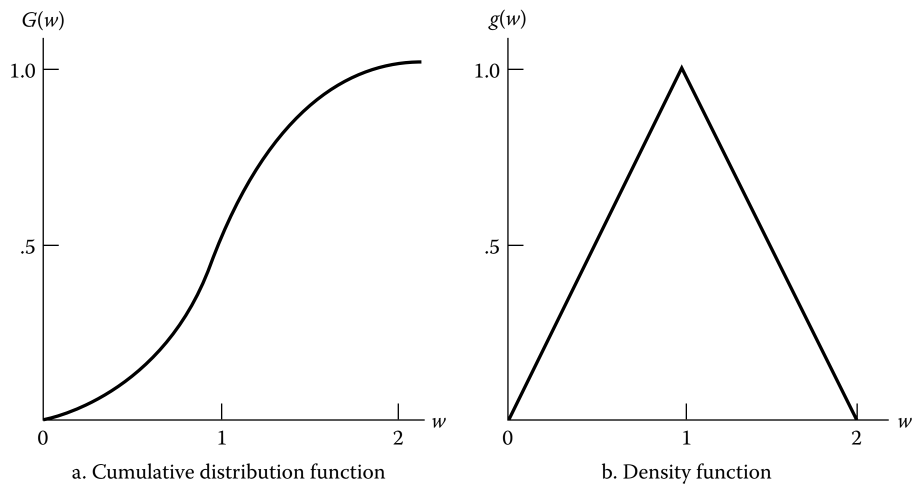



**FIGURE 6.9**
Graphs of the cumulative distribution function and density function for $W = X + Y$

One of the most useful functions of a single continuous random variable is the cumulative distribution function itself. We will show that if $Y$ is a continuous random variable with density function $f(y)$ and cumulative distribution function $F(y)$, then $W = F(y)$ has a uniform probability distribution over the interval $0 \leq w \leq 1$. Using a computer program for generating random numbers, we can generate a random sample of $W$ values. For each value of $W$, we can solve for the corresponding value of $Y$ using the equation $W = F(y)$ and, thereby, obtain a random sample of $Y$ values from a population modeled by the density function $f(y)$. We will present this important transformation as a theorem, prove it, and then demonstrate its use with an example.

**THEOREM 6.7**

Let $Y$ be a continuous random variable with density function $f(y)$ and cumulative distribution $F(y)$. Then the density function of $W = F(y)$ will be a uniform distribution defined over the interval $0 \leq w \leq 1$, i.e.,

$$g(w) = 1 \qquad (0 \leq w \leq 1)$$

**Proof of Theorem 6.7** Figure 6.10 shows the graph of $W = F(y)$ for a continuous random variable $Y$. You can see from the figure that there is a one-to-one correspondence between $y$ values and $w$ values, and that values of $Y$ corresponding to values of $W$ in the interval $0 \leq W \leq w$ will be those in the interval $0 \leq Y \leq y$. Therefore,

$$P(W \leq w) = P(Y \leq y) = F(y)$$

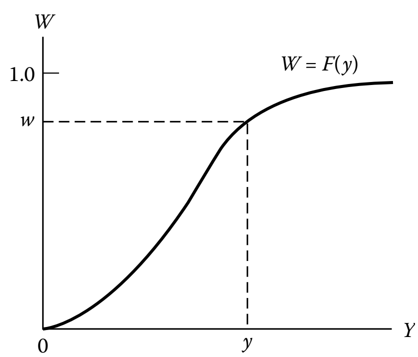
But since $W = F(y)$, we have $F(y) = w$. Therefore, we can write

$$G(w) = P(W \leq w) = F(y) = w$$

Finally, we differentiate over the range $0 \leq w \leq 1$ to obtain the density function:

$$g(w) = \frac{dF(w)}{dw} = 1 \; (0 \leq w \leq 1)$$

**FIGURE 6.10**
Cumulative distribution function $F(y)$




**Example 6.14**
Generating a Random Sample

Use Theorem 6.7 to generate a random sample of $n = 3$ observations from an exponential distribution with $\beta = 2$.

Solution

The density function for the exponential distribution with $\beta = 2$ is

$$f(y) = \begin{cases} \dfrac{e^{-y/2}}{2} & \text{if } 0 \leq y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

and the cumulative distribution function is

$$F(y) = \int_{-\infty}^{y} f(t)\,dt = \int_{0}^{y} \frac{e^{-t/2}}{2}\,dt = -e^{-t/2}\Big]_{0}^{y} = 1 - e^{-y/2}$$

If we let $W = F(y) = 1 - e^{-y/2}$, then Theorem 6.7 tells us that $W$ has a uniform density function over the interval $0 \le W \le 1$.

To draw a random number $Y$ from the exponential distribution, we first randomly draw a value of $W$ from the uniform distribution. This can be done by drawing a random number from Table 1 of Appendix B or using a computer. Suppose, for example, that we draw the random number 10480. This corresponds to the random selection of the value $W_1 = .10480$ from a uniform distribution over the interval $0 \le W \le 1$. Substituting this value of $W_1$ into the formula for $W = F(y)$ and solving for $Y$, we obtain

$$\begin{aligned} W_1 &= F(y) = 1 - e^{-Y_1/2} \\ .10480 &= 1 - e^{-Y_1/2} \\ e^{-Y_1/2} &= .8952 \\ -\frac{Y_1}{2} &= -.111 \\ \text{Then } Y_1 &= .222 \end{aligned}$$

If the next two random numbers selected are 22368 and 24130, then the corresponding values of the uniform random variable are $W_2 = .22368$ and $W_3 = .24130$. By substituting these values into the formula $W = 1 - e^{-Y/2}$, you can verify that $Y_2 = .506$ and $Y_3 = .552$. Thus, $Y_1 = .222$, $Y_2 = .506$, and $Y_3 = .552$ represent three randomly selected observations on an exponential random variable with mean equal to 2.

We conclude this section with a discussion of a very useful function of random variables, called a *linear* function.

**Definition 6.12**

Let $Y_1, Y_2, \ldots, Y_n$ be random variables and let $a_1, a_2, \ldots, a_n$ be constants. Then $\ell$ is a **linear function** of $Y_1, Y_2, \ldots, Y_n$ if

$$\ell = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

The expected value (mean) and variance of a linear function of $Y_1, Y_2, \ldots, Y_n$ may be computed using the formulas presented in Theorem 6.8.

**THEOREM 6.8**

**The Expected Value $E(\ell)$ and Variance $V(\ell)$* of a Linear Function of $Y_1, Y_2, \ldots, Y_n$**

Suppose the means and variances of $Y_1, Y_2, \ldots, Y_n$ are $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \ldots, (\mu_n, \sigma_n^2)$, respectively. If $\ell = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$, then

$$E(\ell) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

and

$$\begin{aligned} \sigma_\ell^2 = V(\ell) &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \\ &\quad + 2a_1a_2\mathrm{Cov}(y_1, y_2) + 2a_1a_3\mathrm{Cov}(y_1, y_3) + \cdots \end{aligned}$$

*In the preceding sections, we have used different subscripts on the symbol $\sigma^2$ to denote the variances of different random variables. This notation is cumbersome if the random variable is a function of several other random variables. Consequently, we will use the notation $\sigma_{()}^2$ or $V()$ interchangeably to denote a variance.

$$+ 2a_1a_n\text{Cov}(y_1, y_n) + 2a_2a_3\text{Cov}(y_2, y_3)$$
$$+ \cdots + 2a_2a_n\text{Cov}(y_2, y_n) + \cdots + 2a_{n-1}a_n\text{Cov}(y_{n-1}, y_n)$$

*Note*: If $Y_1, Y_2, \ldots, Y_n$ are independent, then

$$\sigma_\ell^2 = V(\ell) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

**Proof of Theorem 6.8** By Theorem 6.3, we know

$$E(\ell) = E(a_1Y_1) + E(a_2Y_2) + \cdots + E(a_nY_n)$$

Then, by Theorem 6.2,

$$\begin{aligned} E(\ell) &= a_1E(Y_1) + a_2E(Y_2) + \cdots + a_nE(Y_n) \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \end{aligned}$$

Similarly,

$$\begin{aligned} V(\ell) &= E\{[\ell - E(\ell)]^2\} \\ &= E[(a_1Y_1 + a_2Y_2 + \cdots + a_nY_n - a_1\mu_1 - a_2\mu_2 - \cdots - a_n\mu_n)^2] \\ &= E\{[(a_1(Y_1 - \mu_1) + a_2(Y_2 - \mu_2) + \cdots + a_n(Y_n - \mu_n)]^2\} \\ &= E[a_1^2(Y_1 - \mu_1)^2 + a_2^2(Y_2 - \mu_2)^2 + \cdots + a_n^2(Y_n - \mu_n)^2 \\ &\quad + 2a_1a_2(Y_1 - \mu_1)(Y_2 - \mu_2) + 2a_1a_3(Y_1 - \mu_1)(Y_3 - \mu_3) \\ &\quad + \cdots + 2a_{n-1}a_n(Y_{n-1} - \mu_{n-1})(Y_n - \mu_n)] \\ &= a_1^2E[(Y_1 - \mu_1)^2] + \cdots + a_n^2E[(Y_n - \mu_n)^2] \\ &\quad + 2a_1a_2E[(Y_1 - \mu_1)(Y_2 - \mu_2)] + 2a_1a_3E[(Y_1 - \mu_1)(Y_3 - \mu_3)] \\ &\quad + \cdots + 2a_{n-1}a_nE[(Y_{n-1} - \mu_{n-1})(Y_n - \mu_n)] \end{aligned}$$

By the definitions of variance and covariance, we have

$$E[(Y_i - \mu_j)^2] = \sigma_i^2 \quad \text{and} \quad E[(Y_i - \mu_i)(Y_j - \mu_j)] = \text{Cov}(Y_i, Y_j)$$

Therefore,

$$\begin{aligned} V(\ell) &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 + 2a_1a_2\text{Cov}(Y_1, Y_2) + 2a_1a_3\text{Cov}(Y_1, Y_3) \\ &\quad + \cdots + 2a_2a_3\text{Cov}(Y_2, Y_3) + \cdots + 2a_{n-1}a_n\text{Cov}(Y_{n-1}, Y_n) \end{aligned}$$

**Example 6.15**

**Mean and Variance of a Function of Random Variables**

Suppose $Y_1$, $Y_2$, and $Y_3$ are random variables with $(\mu_1 = 1, \sigma_1^2 = 2)$, $(\mu_2 = 3, \sigma_2^2 = 1)$, $(\mu_3 = 0, \sigma_3^2 = 4)$, $\text{Cov}(Y_1, Y_2) = -1$, $\text{Cov}(Y_1, Y_3) = 2$, and $\text{Cov}(Y_2, Y_3) = 1$. Find the mean and variance of

$$\ell = 2Y_1 + Y_2 - 3Y_3$$

**Solution** The linear function

$$\ell = 2Y_1 + Y_2 - 3Y_3$$

has coefficients $a_1 = 2$, $a_2 = 1$, and $a_3 = -3$. Then by Theorem 6.6,

$$\begin{aligned} \mu_\ell = E(\ell) &= a_1\mu_1 + a_2\mu_2 + a_3\mu_3 \\ &= (2)(1) + (1)(3) + (-3)(0) = 5 \end{aligned}$$

$$\begin{aligned} \sigma_\ell^2 = V(\ell) &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_3^2\sigma_3^2 \\ &\quad + 2a_1a_2\text{Cov}(y_1, y_2) + 2a_1a_3\text{Cov}(y_1, y_3) + 2a_2a_3\text{Cov}(y_2, y_3) \end{aligned}$$

$$= (2)^2(2) + (1)^2(1) + (-3)^2(4)$$
$$+ 2(2)(1)(-1) + 2(2)(-3)(2) + 2(1)(-3)(1)$$
$$= 11$$

These results indicate that the probability distribution of $\ell$ is centered about $E(\ell) = \mu_\ell = 5$ and that its spread is measured by $\sigma_\ell = \sqrt{V(\ell)} = \sqrt{11} = 3.3$. If we were to randomly select values of $Y_1$, $Y_2$, and $Y_3$, we would expect the value of $\ell$ to fall in the interval $\mu_\ell \pm 2\sigma_\ell$, or $-1.6$ to 11.6, according to the Empirical Rule.

**Example 6.16**

Expected Value of the Sample Mean

Let $Y_1, Y_2, \ldots, Y_n$ be a sample of $n$ independent observations selected from a population with mean $\mu$ and variance $\sigma^2$. Find the expected value and variance of the sample mean, $\overline{Y}$.

Solution

The sample measurements, $Y_1, Y_2, \ldots, Y_n$, can be viewed as observations on $n$ independent random variables, where $Y_1$ corresponds to the first observation, $Y_2$ to the second, etc. Therefore, the sample mean $\overline{Y}$ will be a random variable with a probability distribution (or density function).

By writing

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}$$

we see that $\overline{Y}$ is a linear function of $Y_1, Y_2, \ldots, Y_n$, with $a_1 = \frac{1}{n}, a_2 = \frac{1}{n}, \ldots, a_n = \frac{1}{n}$. Since $Y_1, Y_2, \ldots, Y_n$ are independent, it follows from Theorem 6.5 that the covariance of $Y_i$ and $Y_j$, for all pairs with $i \neq j$, will equal 0. Therefore, we can apply Theorem 6.6 to obtain

$$\mu_{\overline{Y}} = E(\overline{Y}) = \left(\frac{1}{n}\right)\mu + \left(\frac{1}{n}\right)\mu + \cdots + \left(\frac{1}{n}\right)\mu = \frac{n\mu}{n} = \mu$$

$$\sigma^2_{\overline{Y}} = V(\overline{Y}) = \left(\frac{1}{n}\right)^2\sigma^2 + \left(\frac{1}{n}\right)^2\sigma^2 + \cdots + \left(\frac{1}{n}\right)^2\sigma^2 = \left(\frac{n}{n^2}\right)\sigma^2 = \frac{\sigma^2}{n}$$

**Example 6.17**

Probability Distribution of the Sample Mean

Suppose that the population of Example 6.16 has mean $\mu = 10$ and variance $\sigma^2 = 4$. Describe the probability distribution for a sample mean based on $n = 25$ observations.

Solution

From Example 6.16, we know that the probability distribution of the sample mean will have mean and variance

$$E(\overline{Y}) = \mu = 10 \quad \text{and} \quad \sigma^2_{\overline{Y}} = V(\overline{Y}) = \frac{\sigma^2}{n} = \frac{4}{25}$$

and thus,

$$\sigma_{\overline{Y}} = \sqrt{V(\overline{Y})} = \sqrt{\frac{4}{25}} = \frac{2}{5} = .4$$

Therefore, the probability distribution of $\overline{Y}$ will be centered about its mean, $\mu = 10$, and most of the distribution will fall in the interval $\mu \pm 2\sigma_{\overline{y}}$, or $10 \pm 2(.4)$ or 9.2 to 10.8. We will learn more about the properties of the probability distribution of $\overline{Y}$ in the remaining sections of this chapter.

## Applied Exercises

6.53 *Tossing dice.* Refer to Exercise 6.2 (p. 239). Find the mean and variance of $(X + Y)$, the sum of the dots showing on the two dice.

6.54 *Red lights on truck route.* Refer to Exercise 6.7 (p. 240). Find the variance of $(X + Y)$. Within what range would you expect $(X + Y)$ to fall?

6.55 *Servicing an automobile.* Refer to Exercise 6.14 (p. 245). Find the variance of $(X - Y)$, the time it takes to actually service the car.

6.56 *Characteristics of a truss subjected to loads.* Refer to the *Journal of Engineering Mechanics* (Dec. 2009) study of the characteristics of a 10-bar truss subjected to loads, Exercise 6.13 (p. 244). Recall that the joint probability distribution of stiffness index $X$ (pounds/sq. in.) and load $Y$ (thousand pounds) is given by the formula,

$$f(x, y) = (1/40)\, e^{-x}, 0 < x < \infty, 80 < y < 120$$

The square root of the load $Y$ is an important variable for determining the level of stress the truss is able to withstand. Let $W = \sqrt{Y}$. Find the probability density function for $W$. Then show that $\int f(w)dw = 1$. [*Hint:* Use the result you obtained in Exercise 6.13c.]

6.57 *Proportion defective in a lot.* A particular manufacturing process yields a proportion $p$ of defective items in each lot. The number $Y$ of defectives in a random sample of $n$ items from the process follows a binomial distribution. Find the expected value and variance of $\hat{p} = Y/n$, the fraction of defectives in the sample. (*Hint*: Write $\hat{p}$ as a linear function of a single random variable $Y$, i.e., $\hat{p} = a_1 Y$, where $a_1 = 1/n$.)

6.58 *Counting microorganisms.* Researchers at the University of Kent (England) developed models for counting the number of colonies of microorganisms in liquid (*Journal of Agricultural, Biological, and Environmental Statistics*, June 2005). The expected number of colonies was derived under the "inhibition" model. Consider a microorganism of species A. Suppose $n$ species A spores are deposited in a Petri dish, and let $Y$ equal the number of species A spores that grow. Then $Y = \Sigma_{i=1}^{n} X_i$ where $X_i = 1$ if a species A spore grows and $X_i = 0$ if a species A spore is inhibited from growing.

a. Show that $E(Y) = np$, where $p = P(X_i = 1)$.

b. In an alternative model, the researchers showed that $p = P(X_i = 1)$ depends on the amount $S$ of soil in the Petri dish, where $p = e^{-\theta S}$ and $\theta$ is a constant. In all their experiments, the amount of soil placed in the Petri dish was essentially the same. For these experiments, explain why $E(Y) = ne^{-\theta S}$.

6.59 *Laser printer paper.* The amount $Y$ of paper used per day by a laser printer at a commercial copy center has an exponential distribution with mean equal to five boxes (i.e., $\beta = 5$). The daily cost of the paper is proportional to $C = (3Y + 2)$. Find the probability density function of the daily cost of paper used by the laser printer.

6.60 *Amount of pollution discharged.* An environmental engineer has determined that the amount $Y$ (in parts per million) of pollutant per water sample collected near the discharge tubes of an island power plant has probability density function

$$f(y) = \begin{cases} \dfrac{1}{10} & \text{if } 0 < y < 10 \\ 0 & \text{elsewhere} \end{cases}$$

A new cleaning device has been developed to help reduce the amount of pollution discharged into the ocean. It is believed that the amount $A$ of pollutant discharged when the device is operating will be related to $Y$ by

$$A = \begin{cases} \dfrac{y}{2} & \text{if } 0 < y < 5 \\ \dfrac{2y - 5}{2} & \text{if } 5 < y < 10 \end{cases}$$

Find the probability density function of $A$.

6.61 *Voltage of circuit.* Researchers at the University of California (Berkeley) have developed a switched-capacitor circuit for generating pseudorandom signals (*International Journal of Circuit Theory and Applications*, May/June 1990). The intensity of the signal (voltage), $Y$, is modeled using the Rayleigh probability distribution with mean $\mu$. This continuous distribution has density function:

$$f(y) = \frac{y}{\mu} \exp^{-y^2/(2\mu)} \quad (y > 0)$$

Find the density function of the random variable $W = Y^2$. Can you name the distribution?

6.62 *Drawing a random sample.* Use Theorem 6.7 to draw a random sample of $n = 5$ observations from a distribution with probability density function

$$f(y) = \begin{cases} e^{y} & \text{if } y < 0 \\ 0 & \text{elsewhere} \end{cases}$$

6.63 *Drawing a random sample.* Use Theorem 6.7 to draw a random sample of $n = 5$ observations from a beta distribution with $\alpha = 2$ and $\beta = 1$.

6.64 *Supercomputer CPU time.* The total time $X$ (in minutes) from the time a supercomputer job is submitted until its run is completed and the time $Y$ the job waits in the job queue before being run have the joint density function

$$f(x, y) = \begin{cases} e^{-x} & \text{if } 0 \le y \le x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

The CPU time for the job (i.e., the length of time the job is in control of the supercomputer's central processing unit) is given by the difference $W = X - Y$. Find the density function of a job's CPU time. [*Hint*: You may use the facts that

$$
\begin{aligned}
P(W \le w) &= P(W \le w, X > w) + P(W \le w, X \le w) \\
&= P(X - w \le Y \le X, w < X < \infty) \\
&\quad + P(0 \le Y \le X, 0 \le X \le w) \\
&= \int_{w_0}^{\infty} \int_{x-w_0}^{x} e^{-x} dy\, dx + \int_0^{w} \int_0^{x} e^{-x} dy\, dx
\end{aligned}
$$

and $\int ye^{-y}\, dy = -ye^{-y} + \int e^{-y}\, dy$ in determining the density function.]

## Theoretical Exercises

6.65 Suppose that $Y_1$, $Y_2$, and $Y_3$ are random variables with $(\mu_1 = 0, \sigma_1^2 = 2)$, $(\mu_2 = -1, \sigma_2^2 = 3)$, $(\mu_3 = 5, \sigma_3^2 = 9)$, $\text{Cov}(Y_1, Y_2) = 1$, $\text{Cov}(Y_1, Y_3) = 4$, and $\text{Cov}(Y_2, Y_3) = -2$. Find the mean and variance of

$$\ell = \frac{1}{2}Y_1 - Y_2 + 2Y_3$$

6.66 Suppose that $Y_1$, $Y_2$, $Y_3$, and $Y_4$ are random variables with

| | | | |
|---|---|---|---|
| $E(Y_1) = 2,$ | $V(Y_1) = 4,$ | $\text{Cov}(Y_1, Y_2) = -1,$ | $\text{Cov}(Y_2, Y_3) = 0$ |
| $E(Y_2) = 4,$ | $V(Y_2) = 8,$ | $\text{Cov}(Y_1, Y_3) = 1,$ | $\text{Cov}(Y_2, Y_4) = 2$ |
| $E(Y_3) = -1,$ | $V(Y_3) = 6,$ | $\text{Cov}(Y_1, Y_4) = \frac{1}{2},$ | $\text{Cov}(Y_3, Y_4) = 0$ |
| $E(Y_4) = 0,$ | $V(Y_4) = 1$ | | |

Find the mean and variance of

$$\ell = -3Y_1 + 2Y_2 + 6Y_3 - Y_4$$

6.67 Let $Y_1, Y_2, \ldots, Y_n$ be a sample of $n$ independent observations selected from a gamma distribution with $\alpha = 1$ and $\beta = 2$. Show that the expected value and variance of the sample mean $\overline{Y}$ are identical to the expected value and variance of a gamma distribution with parameters $\alpha = n$ and $\beta = 2/n$.

6.68 Consider the density function

$$f(y) = \begin{cases} e^{-(y-3)} & \text{if } y > 3 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density function of $W$, where:

a. $W = e^{-Y}$
b. $W = Y - 3$
c. $W = Y/3$

6.69 Consider the density function

$$f(y) = \begin{cases} 2y & \text{if } 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density function of $W$, where:

a. $W = Y^2$
b. $W = 2Y - 1$
c. $W = 1/Y$

# 6.7 Sampling Distributions

Recall that the $n$ measurements in a sample can be viewed as observations on $n$ random variables, $Y_1, Y_2, \ldots, Y_n$. Consequently, the sample mean $\overline{Y}$, the sample variance $s^2$, and other statistics are functions of random variables—functions that we will use in the following chapters to make inferences about population parameters. Thus, a primary reason for presenting the theory of probability and probability distributions in the preceding sections was to enable us to find and evaluate the properties of the probability distribution of a statistic. This probability distribution is often called the **sampling distribution** of the statistic. As is the case for a single random variable, its mean is the expected value of the statistic. Its standard deviation is called the **standard error** of the statistic.

**Definition 6.13**

The **sampling distribution** of a statistic is its probability distribution.

**Definition 6.14**

The **standard error** of a statistic is the standard deviation of its sampling distribution.

Mathematical techniques like those presented in Optional Section 6.6 can be used to find the sampling distribution of a statistic. Except in simple examples, these methods are difficult to apply. An alternative approach is to use a computer to simulate the sampling distribution. (This is the topic of Section 6.8.)

Even if we are unable to find the exact mathematical form of the probability distribution of a statistic and are unable to approximate it using simulation, we can always find its mean and variance using the methods of Chapters 4–6. Then we can obtain an approximate description of the sampling distribution by applying the Empirical Rule.

## 6.8 Approximating a Sampling Distribution by Monte Carlo Simulation

Consider a statistic $W$ that is a function of $n$ sample measurements, $Y_1, Y_2, \ldots, Y_n$. We have shown (in Optional Section 6.6) how we can use probability theory and mathematics to find its sampling distribution. However, the mathematical problem of finding $f(w)$ is often very difficult to solve. When such a situation occurs, we may be able to find an approximation to $f(w)$ by repeatedly generating observations on the statistic $W$ using a random number generator. This method is called **Monte Carlo simulation**. By examining the resulting histogram for $W$, we can approximate $f(w)$.

To illustrate the procedure, we will approximate the sampling distribution for the sum $W = Y_1 + Y_2$ of a sample of $n = 2$ observations from a uniform distribution over the interval $0 \leq Y \leq 1$. Recall that we found an exact expression for this sampling distribution in Example 6.13. Thus, we will be able to compare our simulated sampling distribution with the exact form of the sampling distribution shown in Figure 6.9b.

To begin the Monte Carlo simulation, we used SAS to generate 10,000 pairs of random numbers, with each pair representing a sample $(y_1, y_2)$ from the uniform distribution over the interval $0 \leq Y \leq 1$. We then programmed SAS to calculate the sum $W = Y_1 + Y_2$ for each of the 10,000 pairs. A SAS relative frequency histogram for the 10,000 values of $W$ is shown in Figure 6.11. By comparing Figures 6.9b and 6.11, you can see that the simulated sampling distribution provides a good approximation to the true probability distribution of the sum of a sample of $n = 2$ observations from a uniform distribution.

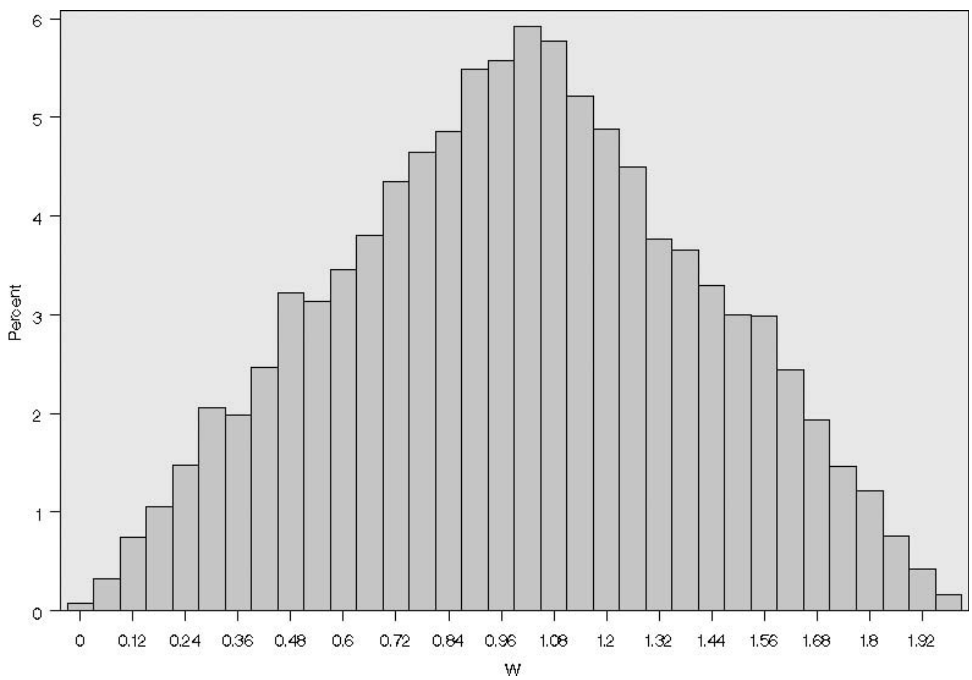



**FIGURE 6.11**
Simulated sampling distribution for sum of two uniform (0, 1) random variables using SAS

**Example 6.18**

**Sampling Distribution Simulation: Uniform**

Simulate the sampling distribution of the sample mean

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

for a sample of $n = 5$ observations drawn from the uniform probability distribution shown in Figure 6.12. Note that the uniform distribution has mean $\mu = .5$. Repeat the procedure for $n = 15, 25, 50,$ and 100. Interpret the results.

**Solution**

We used the SAS RANUNI subroutine to obtain 10,000 random samples of size $n = 5$ from the uniform probability distribution, over the interval (0, 1), and programmed SAS to compute the mean

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

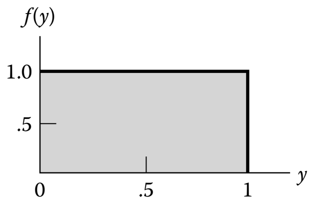



**FIGURE 6.12**
Uniform distribution of Example 6.18

for each sample. The frequency histogram for the 10,000 values of $\overline{Y}$ obtained from the uniform distribution is shown in the top left panel of the MINITAB printout, Figure 6.13. Note its shape for this small value of $n$.

The relative frequency histograms of $\overline{Y}$ based on samples of size $n = 15, 25, 50,$ and 100, also simulated by computer, are shown in the remaining panels of Figure 6.13. Note that the values of $\overline{Y}$ tend to cluster about the mean of the uniform distribution, $\mu = .5$. Furthermore, as $n$ increases, there is less variation in the sampling distribution. You can also see from the figures that as the sample size increases, the shape of the sampling distribution of $\overline{Y}$ tends toward the shape of the normal distribution (symmetric and mound-shaped).

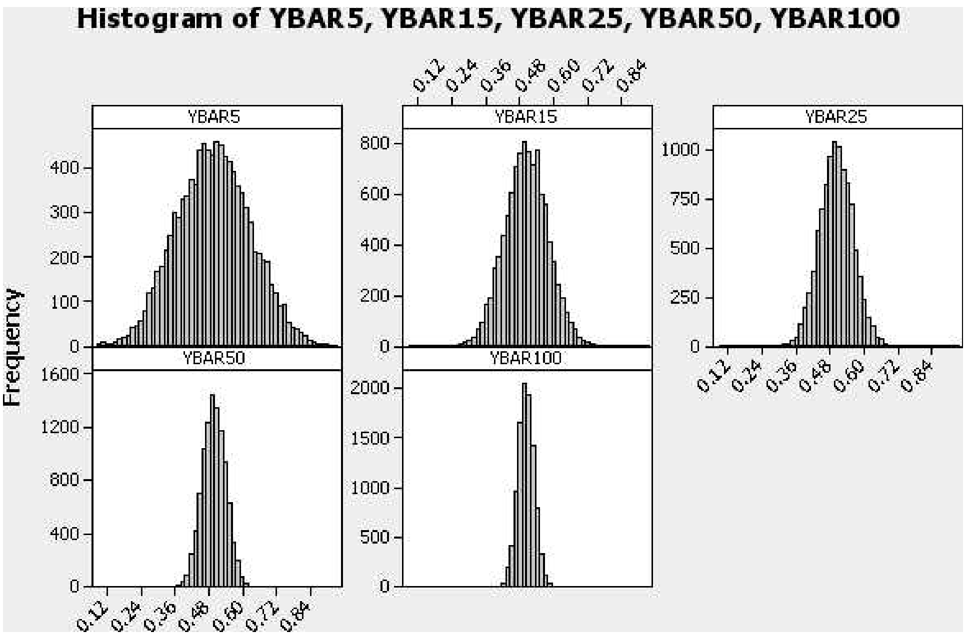



**FIGURE 6.13**
Simulated sampling distributions for means of uniform (0, 1) random variables, $n = 5, 15, 25, 50,$ and 100

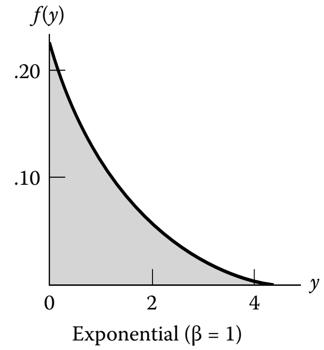
**Example 6.19**

**Sampling Distribution Simulation: Exponential**

Repeat the instructions of Example 6.18, but sample from an exponential probability distribution with mean $\beta = 1$. (See Figure 6.14.)

**Solution** Using the SAS RANEXP function, we simulated the sampling distributions of $\overline{Y}$ for samples of size $n = 5, 15, 25, 50$, and 100 from an exponential distribution. Histograms for these simulated sampling distributions are shown in the MINITAB printout, Figure 6.15. Note the three properties illustrated earlier: (1) values of $\overline{Y}$ tend to cluster about the mean of the exponential probability distribution, $\mu = 1$;(2) the variance of $\overline{Y}$ decreases as $n$ increases; and, (3) the shape of the sampling distribution of $\overline{Y}$ tends toward the shape of the normal distribution as $n$ increases.

**FIGURE 6.14**
Exponential distribution for Example 6.19




In Section 6.9, we generalize the results of Examples 6.18 and 6.19 in the form of a theorem.

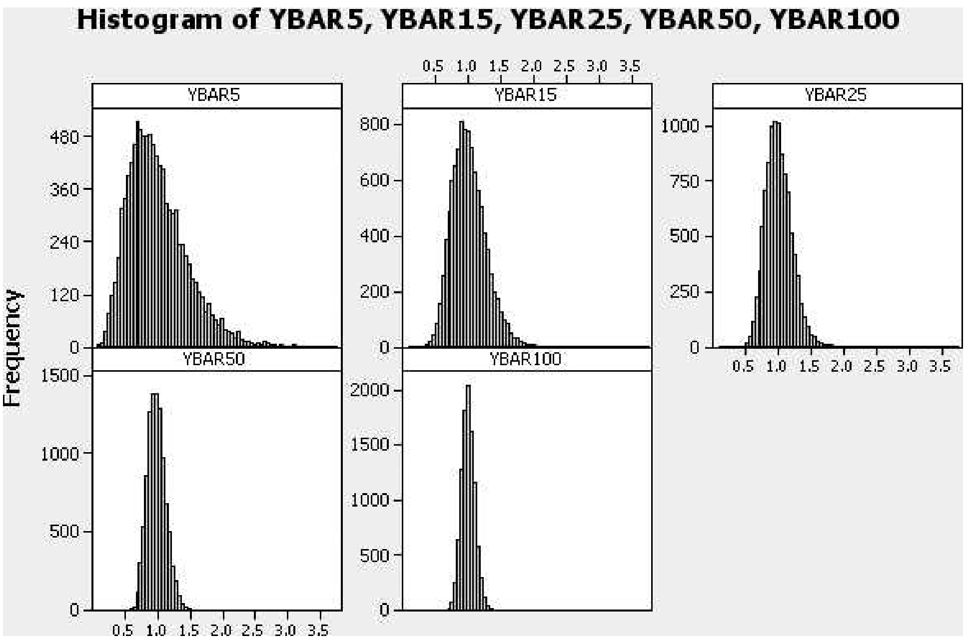



**FIGURE 6.15**
Simulated sampling distributions for means of exponential $(\beta = 1)$ random variables, $n = 5, 15, 25, 50$, and 100

## Applied Exercises

6.70 *Sampling distribution of $s^2$.* Use Monte Carlo simulation to approximate the sampling distribution of $s^2$, the variance of a sample of $n = 100$ observations from a

a. Uniform distribution on the interval (0, 1).
b. Normal distribution, with mean 0 and variance 1.
c. Exponential distribution with mean 1.

6.71 *Sampling distribution of the median.* Use Monte Carlo simulation to approximate the sampling distribution of $M$, the median of a sample of $n = 50$ observations from a uniform distribution on the interval (0, 1).

6.72 *Sampling distribution of the range.* Use Monte Carlo simulation to approximate the sampling distribution of $R$, the range of a sample of $n = 10$ observations from a normal distribution, with mean 0 and variance 1.

# 6.9 The Sampling Distributions of Means and Sums

The simulation of the sampling distribution of the sample mean based on independent random samples from uniform, normal, and exponential distributions in Examples 6.18 and 6.19 illustrates the ideas embodied in one of the most important theorems in statistics. The following version of the theorem applies to the sampling distribution of the sample mean, $\overline{Y}$.

**THEOREM 6.9** The Central Limit Theorem

If a random sample of $n$ observations, $Y_1, Y_2, \ldots, Y_n$, is drawn from a population with finite mean $\mu$ and variance $\sigma^2$, then, when $n$ is sufficiently large, the sampling distribution of the sample mean $\overline{Y}$ can be approximated by a normal density function.

The sampling distribution of $\overline{Y}$, in addition to being approximately normal for large $n$, has other known characteristics, which are given in Definition 6.15.

**Definition 6.15**

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of $n$ observations from a population with finite mean $\mu$ and finite standard deviation $\sigma$. Then, the **mean and standard deviation of the sampling distribution** of $\overline{Y}$, denoted $\mu_{\overline{y}}$ and $\sigma_{\overline{y}}$, respectively, are

$$\mu_{\overline{y}} = \mu, \quad \sigma_{\overline{y}} = \sigma/\sqrt{n}$$

The significance of the central limit theorem and Definition 6.15 is that we can use the normal distribution to approximate the sampling distribution of the sample mean $\overline{y}$ as long as the population possesses a finite mean and variance and the number $n$ of measurements in the sample is sufficiently large. How large the sample size must be will depend on the nature of the sampled population. You can see from our simulated experiments in Examples 6.18 and 6.19 that the sampling distribution of $\overline{Y}$ tends to become very nearly normal for sample sizes as small as $n = 25$ for the uniform and exponential population distributions. When the population distribution is symmetric about its mean, the sampling distribution of $\overline{Y}$ will be mound-shaped and nearly normal for sample sizes as small as $n = 15$. In addition, if the sampled population possesses a normal distribution, then the sampling distribution of $\overline{Y}$ will be a normal density function, regardless of the sample size. In fact, it can be shown that *the sampling distribution of **any linear function** of normally distributed random variables, even those that are correlated and have different means and variances, is a normal distribution*. This important result is presented (without proof) in Theorem 6.10 and illustrated in an example.

**THEOREM 6.10**

Let $a_1, a_2, \ldots, a_n$ be constants and let $Y_1, Y_2, \ldots, Y_n$ be $n$ normally distributed random variables with $E(Y_i) = \mu_i$, $V(Y_i) = \sigma_i^2$, and $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ $(i = 1, 2, \ldots, n)$. Then the sampling distribution of a linear combination of the normal random variables

$$\ell = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

possesses a normal density function with mean and variance*

$$E(\ell) = \mu = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

and

$$\begin{aligned} V(\ell) &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \\ &\quad + 2a_1a_2\sigma_{12} + 2a_1a_3\sigma_{13} + \cdots + 2a_1a_n\sigma_{1n} \\ &\quad + 2a_2a_3\sigma_{23} + \cdots + 2a_2a_n\sigma_{2n} \\ &\quad + \cdots + 2a_{n-1}a_n\sigma_{n-1,\,n} \end{aligned}$$

**Example 6.20**

**Sampling Distribution of $(\bar{Y}_1 - \bar{Y}_2)$**

Suppose you select independent random samples from two normal populations, $n_1$ observations from population 1 and $n_2$ observations from population 2. If the means and variances for populations 1 and 2 are $(\mu_1, \sigma_1^2)$ and $(\mu_2, \sigma_2^2)$, respectively, and if $\bar{Y}_1$ and $\bar{Y}_2$ are the corresponding sample means, find the distribution of the difference $(\bar{Y}_1 - \bar{Y}_2)$.

**Solution** Since $\bar{y}_1$ and $\bar{y}_2$ are both linear functions of normally distributed random variables, they will be normally distributed by Theorem 6.10. The means and variances of the sample means (see Example 6.16) are

$$E(\bar{Y}_i) = \mu_i \quad \text{and} \quad V(\bar{Y}_i) = \frac{\sigma_i^2}{n_i} \quad (i = 1, 2)$$

Then, $\ell = \bar{Y}_1 - \bar{Y}_2$ is a linear function of two normally distributed random variables, $\bar{y}_1$ and $\bar{y}_2$. According to Theorem 6.10, $\ell$ will be normally distributed with

$$E(\ell) = \mu_\ell = E(\bar{Y}_1) - E(\bar{Y}_2) = \mu_1 - \mu_2$$

$$V(\ell) = \sigma_\ell^2 = (1)^2V(\bar{Y}_1) + (-1)^2V(\bar{Y}_2) + 2(1)(-1)\text{Cov}(\bar{Y}_1, \bar{Y}_2)$$

But, since the samples were independently selected, $\bar{Y}_1$ and $\bar{Y}_2$ are independent and $\text{Cov}(\bar{Y}_1, \bar{Y}_2) = 0$. Therefore,

$$V(\ell) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

We have shown that $(\bar{Y}_1 - \bar{Y}_2)$ is a normally distributed random variable with mean $(\mu_1 - \mu_2)$ and variance $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Typical applications of the central limit theorem, however, involve samples selected from nonnormal or unknown populations, as illustrated in Examples 6.21 and 6.22.

**Example 6.21**

**Sampling Distribution of $\bar{Y}$: Inference**

Engineers responsible for the design and maintenance of aircraft pavements traditionally use pavement-quality concrete. A study was conducted at Luton Airport (United Kingdom) to assess the suitability of concrete blocks as a surface for aircraft pavements (*Proceedings of the Institute of Civil Engineers*, Apr. 1986). The original pavement-quality concrete of the western end of the runway was overlaid with 80-mm-thick concrete blocks. A series of plate-bearing tests was carried out to

*The formulas for the mean and variance of a linear function of any random variables, $Y_1, Y_2, \ldots, Y_n$, were given in Theorem 6.8.

determine the load classification number (LCN)—a measure of breaking strength—of the surface. Let $\bar{y}$ represent the mean LCN of a sample of 25 concrete block sections on the western end of the runway.

a. Prior to resurfacing, the mean LCN of the original pavement-quality concrete of the western end of the runway was known to be $\mu = 60$, and the standard deviation was $\sigma = 10$. If the mean strength of the new concrete block surface is no different from that of the original surface, describe the sampling distribution of $\bar{Y}$.

b. If the mean strength of the new concrete block surface is no different from that of the original surface, find the probability that $\bar{Y}$, the sample mean LCN of the 25 concrete block sections, exceeds 65.

c. The plate-bearing tests on the new concrete block surface resulted in $\bar{Y} = 73$. Based on this result, what can you infer about the true mean LCN of the new surface?

Solution

a. Although we have no information about the shape of the relative frequency distribution of the breaking strengths (LCNs) for sections of the new surface, we can apply Theorem 6.9 to conclude that the sampling distribution of $\bar{Y}$, the mean LCN of the sample, is approximately normally distributed. In addition, if $\mu = 60$ and $\sigma = 10$, the mean, $\mu_{\bar{y}}$, and the standard deviation, $\sigma_{\bar{y}}$, of the sampling distribution are given by

$$\mu_{\bar{y}} = \mu = 60$$

and

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

b. We want to calculate $P(\bar{Y} > 65)$. Since $\bar{Y}$ has an approximate normal distribution, we have

$$\begin{aligned} P(\bar{Y} > 65) &= P\left(\frac{\bar{Y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} > \frac{65 - \mu_{\bar{y}}}{\sigma_{\bar{y}}}\right) \\ &\approx P\left(Z > \frac{65 - 60}{2}\right) = P(Z > 2.5) \end{aligned}$$

where $Z$ is a standard normal random variable. Using Table 5 of Appendix B, we obtain

$$P(Z > 2.5) = .5 - .4938 = .0062$$

Therefore, $P(\bar{Y} > 65) = .0062$.

c. If there is no difference between the true mean strengths of the new and original surfaces (i.e., $\mu = 60$ for both surfaces), the probability that we would obtain a sample mean LCN for concrete block of 65 or greater is only .0062. Observing $\bar{Y} = 73$ provides strong evidence that the true mean breaking strength of the new surface exceeds $\mu = 60$. Our reasoning stems from the rare event philosophy of Chapter 3, which states that such a large sample mean $(\bar{Y} = 73)$ is very unlikely to occur if $\mu = 60$.

## Example 6.22

**Sampling Distribution of a Proportion**

Consider a binomial experiment with $n$ Bernoulli trials and probability of success $p$ on each trial. The number $Y$ of successes divided by the number $n$ of trials is called the **sample proportion of successes** and is denoted by the symbol $\hat{p} = Y/n$. Explain why the random variable

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

has approximately a standard normal distribution for large values of $n$.

**Solution** If we denote the outcome of the $i$th Bernoulli trial as $Y_i\ (i = 1, 2, \ldots, n)$, where

$$Y_i = \begin{cases} 1 & \text{if outcome is a success} \\ 0 & \text{if outcome is a failure} \end{cases}$$

then the number $Y$ of successes in $n$ trials is equal to the sum of $n$ independent Bernoulli random variables:

$$\sum_{i=1}^{n} Y_i$$

Therefore, $\hat{p} = Y/n$ is a sample mean and, according to Theorem 6.9, $\hat{p}$ will be approximately normally distributed when the sample size $n$ is large. To find the expected value and variance of $\hat{p}$, we can view $\hat{p}$ as a linear function of a single random variable $Y$:

$$\hat{p} = \ell = a_1 Y_1 = \left(\frac{1}{n}\right)Y \qquad \text{where } a_1 = \frac{1}{n} \quad \text{and} \quad Y_1 = Y$$

We now apply Theorem 6.8 to obtain $E(\ell)$ and $V(\ell)$:

$$E(\hat{p}) = \frac{1}{n}E(Y) = \frac{1}{n}(np) = p$$

$$V(\hat{p}) = \left(\frac{1}{n}\right)^2 V(Y) = \frac{1}{n^2}(npq) = \frac{pq}{n}$$

Therefore,

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$$

is equal to the deviation between a normally distributed random variable $\hat{p}$ and its mean $p$, expressed in units of its standard deviation, $\sqrt{pq/n}$. This satisfies the definition of a standard normal random variable given in Section 5.5.

The central limit theorem also applies to the sum of a sample of $n$ measurements subject to the conditions stated in Theorem 6.9. The only difference is that the approximating normal distribution will have mean $n\mu$ and variance $n\sigma^2$.

### The Sampling Distribution of a Sum of Random Variables

If a random sample of $n$ observations, $Y_1, Y_2, \ldots, Y_n$, is drawn from a population with finite mean $\mu$ and variance $\sigma^2$, then, when $n$ is sufficiently large, the sampling distribution of the sum

$$\sum_{i=1}^{n} Y_i$$

can be approximated by a normal density function with mean $E(\Sigma Y_i) = n\mu$ and $V(\Sigma Y_i) = n\sigma^2$.

In Section 6.10, we apply the central limit theorem for sums to show that the normal density function can be used to approximate the binomial probability distribution when the number $n$ of trials is large.

## Applied Exercises

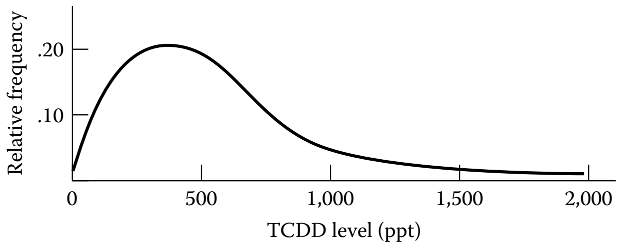
6.73 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010), study of the trend in the design of social robots, Exercise 2.1 (p. 26). The researchers obtained a random sample of 106 social robots obtained through a web search and determined the number that were designed with legs, but no wheels. Let $\hat{p}$ represent the sample proportion of social robots designed with legs, but no wheels. Assume that in the population of all social robots, 40% are designed with legs, but no wheels.

a. Give the mean and standard deviation of the sampling distribution of $\hat{p}$.
b. Describe the shape of the sampling distribution of $\hat{p}$.
c. Find $P(\hat{p} > .59)$.
d. Recall that the researchers found that 63 of the 106 robots were built with legs only. Does this result cast doubt on the assumption that 40% of all social robots are designed with legs, but no wheels? Explain.

6.74 *Uranium in the Earth's crust.* Refer to the *American Mineralogist* (October 2009) study of the evolution of uranium minerals in the Earth's crust, Exercise 5.17 (p. 199). Recall that researchers estimate that the trace amount of uranium $Y$ in reservoirs follows a uniform distribution ranging between 1 and 3 parts per million. In a random sample of $n = 60$ reservoirs, let $\overline{Y}$ represent the sample mean amount of uranium.

a. Find $E(\overline{Y})$ and interpret its value.
b. Find Var $(\overline{Y})$.
c. Describe the shape of the sampling distribution of $\overline{Y}$.
d. Find the probability that $\overline{Y}$ is between 1.5 ppm and 2.5 ppm.
e. Find the probability that $\overline{Y}$ exceeds 2.2 ppm.

6.75 *Chemical dioxin exposure.* The National Institute for Occupational Safety and Health (NIOSH) evaluated the level of exposure of workers to the chemical dioxin, 2,3,7,8-TCDD. The distribution of TCDD levels in parts per trillion (ppt) of production workers at a Newark, New Jersey, chemical plant had a mean of 293 ppt and a standard deviation of 847 ppt (*Chemosphere*, Vol. 20, 1990). A graph of the distribution is shown here.




In a random sample of $n = 50$ workers selected at the New Jersey plant, let $\overline{Y}$ represent the sample mean TCDD level.

a. Find the mean and standard deviation of the sampling distribution of $\overline{Y}$.
b. Draw a sketch of the sampling distribution of $\overline{Y}$. Locate the mean on the graph.
c. Find the probability that $\overline{Y}$ exceeds 550 ppt.

6.76 *Levelness of concrete slabs.* Geotechnical engineers use water-level "manometer" surveys to assess the levelness of newly constructed concrete slabs. Elevations are typically measured at eight points on the slab; of interest is the maximum differential between elevations. The *Journal of Performance of Constructed Facilities* (Feb. 2005) published an article on the levelness of slabs in California residential developments. Elevation data collected for over 1,300 concrete slabs *before tensioning* revealed that maximum differential, $Y$, has a mean of $\mu = .53$ inch and a standard deviation of $\sigma = .193$ inch. Consider a sample of $n = 50$ slabs selected from those surveyed and let $\overline{Y}$ represent the mean of the sample.

a. Fully describe the sampling distribution of $\overline{Y}$.
b. Find $P(\overline{Y} > .58)$.
c. The study also revealed that the mean maximum differential of concrete slabs measured *after tensioning and loading* is $\mu = .58$ inch. Suppose the sample data yields $\overline{Y} = .59$ inch. Comment on whether the sample measurements were obtained before tensioning or after tensioning and loading.

6.77 *Surface roughness of pipe.* Refer to the *Anti-Corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of oil field pipes, Exercise 2.20 (p. 37). Recall that a scanning probe instrument was used to measure the surface roughness $Y$ (in micrometers) of 20 sampled sections of coated interior pipe. Consider the sample mean, $\overline{Y}$.

a. Assume that the surface roughness distribution has a mean of $\mu = 1.8$ micrometers and a standard deviation of $\sigma = .5$ micrometer. Use this information to find the probability that $\overline{Y}$ exceeds 1.85 micrometers.
b. The sample data is reproduced in the table. Compute $\bar{y}$.
c. Based on the result, part **b**, comment on the validity of the assumptions made in part **a**.

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-Corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

6.78 *Handwashing versus handrubbing.* The *British Medical Journal* (August 17, 2002) published a study to compare the effectiveness of handwashing with soap and handrubbing with alcohol. Health care workers who used handrubbing had a mean bacterial count of 35 per hand with a standard deviation of 59. Health care workers who used handwashing had a mean bacterial count of 69 per hand with a standard deviation of 106. In a random sample of 50 health care workers, all using the same method of cleaning their hands, the mean bacterial count per hand $\overline{Y}$ is less than 30. Give your opinion on whether this sample of workers used handrubbing with alcohol or handwashing with soap.

6.79 *Tomato as a taste modifier.* Miraculin is a protein naturally produced in a rare tropical fruit that can convert a sour taste into a sweet taste. Refer to the *Plant Science* (May 2010) investigation of the ability of a hybrid tomato plant to produce miraculin, Exercise 5.29 (p. 204). Recall that the amount $Y$ of miraculin produced in the plant had a mean of 105.3 micro-grams per gram of fresh weight with a standard deviation of 8.0. Consider a random sample of $n = 64$ hybrid tomato plants and let $\overline{Y}$ represent the sample mean amount of miraculin produced. Would you expect to observe a value of $\overline{Y}$ less than 103 micro-grams per gram of fresh weight? Explain.

**PHISHING**

6.80 *Phishing attacks to email accounts.* In Exercise 2.24 (p. 38), you learned that *phishing* describes an attempt to extract personal/financial information from unsuspecting people through fraudulent email. Data from an actual phishing attack against an organization were presented in *Chance* (Summer 2007). The interarrival times, i.e., the time differences (in seconds), for 267 fraud box email notifications were recorded and are saved in the **PHISHING** file. For this exercise, consider these interarrival times to represent the population of interest.

a. In Exercise 2.24 you constructed a histogram for the interarrival times. Describe the shape of the population of interarrival times.
b. Find the mean and standard deviation of the population of interarrival times.
c. Now consider a random sample of $n = 40$ interarrival times selected from the population. Describe the shape of the sampling distribution of $\overline{Y}$, the sample mean. Theoretically, what are $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$?
d. Find $P(\overline{Y} < 90)$.
e. Use a random number generator to select a random sample of $n = 40$ interarrival times from the population, and calculate the value of $\overline{Y}$. (Every student in the class should do this.)
f. Refer to part e. Obtain the values of $\overline{Y}$ computed by the students and combine them into a single data set. Form a histogram for these values of $\overline{Y}$. Is the shape approximately normal?
g. Refer to part f. Find the mean and standard deviation of the $\overline{Y}$-values. Do these values approximate $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ respectively?

6.81 *Modeling machine downtime.* An article in *Industrial Engineering* (Aug. 1990) discussed the importance of modeling machine downtime correctly in simulation studies. As an illustration, the researcher considered a single-machine-tool system with repair times (in minutes) that can be modeled by a gamma distribution with parameters $\alpha = 1$ and $\beta = 60$. Of interest is the mean repair time, $\overline{Y}$, of a sample of 100 machine breakdowns.

a. Find $E(\overline{Y})$ and $\text{Var}(\overline{Y})$.
b. What probability distribution provides the best model of the sampling distribution of $\overline{Y}$? Why?
c. Calculate the probability that the mean repair time, $\overline{Y}$, is no longer than 30 minutes.

6.82 *Diesel system maintenance.* The U.S. Army Engineering and Housing Support Center recently sponsored a study of the reliability, availability, and maintainability (RAM) characteristics of small diesel and gas-powered systems at commercial and military facilities (*IEEE Transactions on Industry Applications*, July/Aug. 1990). The study revealed that the time, $Y$, to perform corrective maintenance on continuous diesel auxiliary systems has an approximate exponential distribution with an estimated mean of 1,700 hours.

a. Assuming $\mu = 1{,}700$, find the probability that the mean time to perform corrective maintenance for a sample of 70 continuous diesel auxiliary systems exceeds 2,500 hours.
b. If you observe $\overline{Y} > 2{,}500$, what inference would you make about the value of $\mu$?

6.83 *Freight elevator maximum load.* A large freight elevator can transport a maximum of 10,000 pounds (5 tons). Suppose a load of cargo containing 45 boxes must be transported via the elevator. Experience has shown that the weight $Y$ of a box of this type of cargo follows a probability distribution with mean $\mu = 200$ pounds and standard deviation $\sigma = 55$ pounds. What is the probability that all 45 boxes can be loaded onto the freight elevator and transported simultaneously? (*Hint*: Find $P(\sum_{i=1}^{45} y_i \leq 10{,}000)$).

## Theoretical Exercises

6.84 If $Y$ has a $\chi^2$ distribution with $n$ degrees of freedom (see Section 5.7), then $Y$ could be represented by $Y = \sum_{i=1}^{n} X_i$, where the $X_i$'s are independent $\chi^2$ distributions, each with 1 degree of freedom.

a. Show that $Z = (Y - n)/\sqrt{2n}$ has approximately a standard normal distribution for large values of $n$.
b. If $y$ has a $\chi^2$ distribution with 30 degrees of freedom, find the approximate probability that $Y$ falls within 2 standard deviations of its mean, i.e., find $P(\mu - 2\sigma < Y < \mu + 2\sigma)$.

6.85 Let $\hat{p}_1$ be the sample proportion of successes in a binomial experiment with $n_1$ trials and let $\hat{p}_2$ be the sample proportion of successes in a binomial experiment with $n_2$ trials, conducted independently of the first. Let $p_1$ and $p_2$ be the corresponding population parameters. Show that

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}}$$

has approximately a standard normal distribution for large values of $n_1$ and $n_2$.

## 6.10 Normal Approximation to the Binomial Distribution

Consider the binomial random variable $Y$ with parameters $n$ and $p$. Recall that $Y$ has mean $\mu = np$ and variance $\sigma^2 = npq$. We showed in Example 6.22 that the number $Y$ of successes in $n$ trials can be regarded as a sum consisting of $n$ values of 0 and 1, with each 0 and 1 representing the outcome (failure or success, respectively) of a particular trial, i.e.,

$$Y = \sum_{i=1}^{n} Y_i \qquad \text{where } Y_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$$

Then, according to the central limit theorem for sums, the binomial probability distribution $p(y)$ should become more nearly normal as $n$ becomes larger. The normal approximation to a binomial probability distribution is reasonably good even for small samples—say, $n$ as small as 10—when $p = .5$, and the distribution of $Y$ is therefore symmetric about its mean $\mu = np$. When $p$ is near 0 (or 1), the binomial probability distribution will tend to be skewed to the right (or left), but this skewness will disappear as $n$ becomes large. In general, the approximation will be good when $n$ is large enough so that $\mu - 2\sigma = np - 2\sqrt{npq}$ and $\mu + 2\sigma = np + 2\sqrt{npq}$ both lie between 0 and $n$. It can be shown (proof omitted) that for both $\mu - 2\sigma$ and $\mu + 2\sigma$ to fall between 0 and $n$, both $np$ and $nq$ must be greater than or equal to 4.

### Condition Required to Apply a Normal Approximation to a Binomial Probability Distribution

The approximation will be good if both $\mu - 2\sigma = np - 2\sqrt{npq}$ and $\mu + 2\sigma = np + 2\sqrt{npq}$ lie between 0 and $n$. This condition will be satisfied if both $np \geq 4$ and $nq \geq 4$.

**Example 6.23**

**Finding a Binomial Probability Using Normal Approximation**

Let $Y$ be a binomial probability distribution with $n = 10$ and $p = .5$.

a. Graph $p(y)$ and superimpose on the graph a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$.

b. Use Table 2 of Appendix B to find $P(Y \leq 4)$.

c. Use the normal approximation to the binomial probability distribution to find an approximation to $P(Y \leq 4)$.

**Solution**

a. The graphs of $p(y)$ and a normal distribution with

$$\mu = np = (10)(.5) = 5$$

and

$$\sigma = \sqrt{npq} = \sqrt{(10)(.5)(.5)} = 1.58$$

are shown in Figure 6.16. Note that both $np = 5$ and $nq = 5$ both exceed 4. Thus, the normal density function with $\mu = 5$ and $\sigma = 1.58$ provides a good approximation to $p(y)$.

b. From Table 2 of Appendix B, we obtain

$$\sum_{y=0}^{4} p(y) = .377$$

c. By examining Figure 6.16, you can see that $P(Y \leq 4)$ is the area under the normal curve to the left of $Y = 4.5$. Note that the area to the left of $Y = 4$ would *not* be appropriate because it would omit half the probability rectangle corresponding to

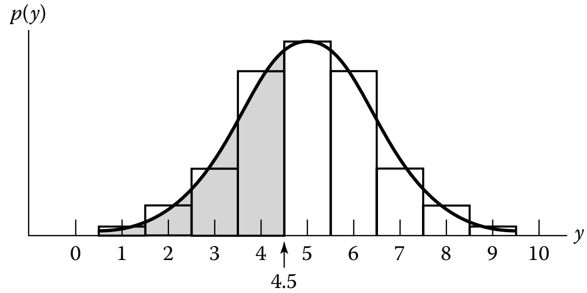



**FIGURE 6.16**
A binomial probability distribution $(n = 10, p = .5)$ and the approximating normal distribution $(\mu = np = 5$ and $\sigma = \sqrt{npq} = 1.58)$

$Y = 4$. We need to add .5 to 4 before calculating the probability to correct for the fact that we are using a continuous probability distribution to approximate a discrete probability distribution. The value .5 is called the **continuity correction factor** for the normal approximation to the binomial probability (see the box). The $Z$ value corresponding to the corrected value $Y = 4.5$ is

$$Z = \frac{Y - \mu}{\sigma} = \frac{4.5 - 5}{1.58} = \frac{-.5}{1.58} = -.32$$

The area between $Z = 0$ and $Z = .32$, given in Table 5 of Appendix B, is $A = .1255$. Therefore,

$$P(Y \leq 4) \approx .5 - A = .5 - .1255 = .3745$$

Thus, the normal approximation to $P(Y \leq 4) = .377$ is quite good, although $n$ is as small as 10. The sample size would have to be larger to apply the approximation if $p$ were not equal to .5.

### Continuity Correction for the Normal Approximation to a Binomial Probability

Let $Y$ be a binomial random variable with parameters $n$ and $p$, and let $Z$ be a standard random variable. Then,

$$P(Y \leq a) \approx P\left(Z < \frac{(a + .5) - np}{\sqrt{npq}}\right)$$

$$P(Y \geq a) \approx P\left(Z > \frac{(a - .5) - np}{\sqrt{npq}}\right)$$

$$P(a \leq Y \leq b) \approx P\left(\frac{(a - .5) - np}{\sqrt{npq}} < Z < \frac{(b + .5) - np}{\sqrt{npq}}\right)$$

## Applied Exercises

6.86 *Female fire fighters.* According to the International Association of Women in Fire and Protection Services, 4% of all fire fighters in the world are female.

a. Approximate the probability that more than 100 of a random sample of 500 fire fighters are female.
b. Approximate the probability that 5 or fewer of a random sample of 500 fire fighters are female.

6.87 *Defects in semiconductor wafers.* The computer chips in notebook and laptop computers are produced from semiconductor wafers. Certain semiconductor wafers are exposed to an environment that generates up to 100 possible defects per wafer. The number of defects per wafer, $Y$, was found to follow a binomial distribution if the manufacturing process is stable and generates defects that are randomly distributed on the wafers. (*IEEE Transactions on Semiconductor Manufacturing*, May 1995.) Let $p$ represent the probability that a defect occurs at any one of the 100 points of the wafer. For each of the following cases, determine whether the normal approximation can be used to characterize $Y$.

a. $p = .01$
b. $p = .50$
c. $p = .90$

6.88 *Chemical signals of mice.* Refer to the *Cell* (May 14, 2010) study of the ability of a mouse to recognize the odor of a potential predator, Exercise 4.27 (p. 153). You learned that 40% of lab mice cells exposed to chemically produced major urinary proteins (Mups) from a cat responded positively (i.e., recognized the danger of the lurking predator). Again, consider a sample of 100 lab mice cells, each exposed to chemically produced cat Mups, and let Y represent the number of cells that respond positively. How likely is it that less than half of the cells respond positively to cat Mups?

6.89 *Ecotoxicological survival study.* Refer to the *Journal of Agricultural, Biological and Environmental Sciences* (Sep. 2000) evaluation of the risk posed by hazardous pollutants, Exercise 4.28 (p. 153). In the experiment, guppies (all the same age and size) were released into a tank of natural seawater polluted with the pesticide dieldrin and the number of guppies surviving after 5 days was determined. Recall that the researchers estimated that the probability of any single guppy surviving was .60. If 300 guppies are released into the polluted tank, estimate the probability that fewer than 100 guppies survive after 5 days.

6.90 *Mercury contamination of swordfish. Consumer Reports* found widespread contamination of seafood in New York and Chicago supermarkets. For example, 40% of the swordfish pieces available for sale have a level of mercury above the Food and Drug Administration (FDA) limit. Consider a random sample of 20 swordfish pieces from New York and Chicago supermarkets.

a. Use the normal approximation to the binomial to calculate the probability that fewer than 2 of the 20 swordfish pieces have mercury levels exceeding the FDA limit.
b. Use the normal approximation to the binomial to calculate the probability that more than half of the 20 swordfish pieces have mercury levels exceeding the FDA limit.
c. Use the binomial tables to calculate the exact probabilities in parts **a** and **b**. Does the normal distribution provide a good approximation to the binomial distribution?

6.91 *Analysis of bottled water.* Refer to the *Scientific American* (July 2003) report on whether bottled water is really purified water, Exercise 4.29 (p. 153). Recall that the Natural Resources Defense Council found that 25% of bottled water brands fill their bottles with just tap water. In a random sample of 65 bottled water brands, is it likely that 20 or more brands will contain tap water? Explain.

6.92 *Bridge inspection ratings.* Refer to the *Journal of Performance of Constructed Facilities* (Feb. 2005) study of inspection ratings of all major Denver bridges, Exercise 4.30 (p. 153). Recall that the National Bridge Inspection Standard (NBIS) rating scale ranges from 0 (poorest rating) to 9 (highest rating). Engineers forecast that 9% of all major Denver bridges will have ratings of 4 or below in the year 2020.

a. Use the forecast to approximate the probability that in a random sample of 70 major Denver bridges, at least half will have an inspection rating of 4 or below in 2020.
b. Suppose that you actually observe at least 35 of the sample of 70 bridges with inspection ratings of 4 or below in 2020. What inference can you make? Why?

6.93 *Fingerprint expertise.* Refer to the *Psychological Science* (August 2011) study of fingerprint identification, Exercise 4.32 (p. 153). Recall that when presented with prints from the same individual, a fingerprint expert will correctly identify the match 92% of the time. Consider a forensic data base of 1,000 different pairs of fingerprints, where each pair is a match.

a. What proportion of the 1,000 pairs would you expect an expert to correctly identify as a match?
b. What is the probability that an expert will correctly identify less than 900 of the fingerprint matches?

6.94 *Airport luggage inspection. New Jersey Business* reports that Newark International Airport's terminal handles an average of 3,000 international passengers an hour but is capable of handling twice that number. Also, after scanning all luggage, 20% of arriving international passengers are detained for intrusive luggage inspection. The inspection facility can handle 600 passengers an hour without unreasonable delays for the travelers.

a. When international passengers arrive at the rate of 1,500 per hour, what is the expected number of passengers who will be detained for luggage inspection?
b. In the future, it is expected that as many as 4,000 international passengers will arrive per hour. When that occurs, what is the expected number of passengers who will be detained for luggage inspection?
c. Refer to part **b**. Find the approximate probability that more than 600 international passengers will be detained for luggage inspection. (This is also the probability that travelers will experience unreasonable luggage inspection delays.)

## 6.11 Sampling Distributions Related to the Normal Distribution

In this section, we present the sampling distributions of several well-known statistics that are based on random samples of observations from a normal population. These statistics are the $\chi^2$, $t$, and $F$ statistics. In Chapter 7, we show how to use these statistics to estimate the values of certain population parameters. The following results are stated without proof. Proofs using the methodology of Section 6.2 can be found in the references at the end of this chapter.

**THEOREM 6.11**

If a random sample of $n$ observations, $Y_1, Y_2, \dots, Y_n$, is selected from a normal distribution with mean $\mu$ and variance $\sigma^2$, then the sampling distribution of

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a **chi-square density function** (see Section 5.7) with $\nu = (n-1)$ degrees of freedom. *Note*: The random variable $s^2$ represents the sample variance.

**THEOREM 6.12**

If $\chi_1^2$ and $\chi_2^2$ are independent chi-square random variables with $\nu_1$ and $\nu_2$ degrees of freedom, respectively, then the sum $(\chi_1^2 + \chi_2^2)$ has a **chi-square distribution** with $(\nu_1 + \nu_2)$ degrees of freedom.

**Definition 6.16**

Let $Z$ be a standard normal random variable and $\chi^2$ be a chi-square random variable with $\nu$ degrees of freedom. If $Z$ and $\chi^2$ are independent, then

$$T = \frac{Z}{\sqrt{\chi^2/\nu}}$$

is said to possess a **Student's *T* distribution** (or, simply, ***T* distribution**) with $\nu$ degrees of freedom.

**Definition 6.17**

Let $\chi_1^2$ and $\chi_2^2$ be chi-square random variables with $\nu_1$ and $\nu_2$ degrees of freedom, respectively. If $\chi_1^2$ and $\chi_2^2$ are independent, then

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

is said to have an ***F* distribution** with $\nu_1$ numerator degrees of freedom and $\nu_2$ denominator degrees of freedom.

*Note*: The sampling distributions for the $T$ and $F$ statistics can also be derived using the methods of Optional Section 6.6. Both sampling distributions are related to the density function for a beta-type random variable (see Section 5.9). It can be shown (proof omitted) that a $T$ distribution with $\nu$ degrees of freedom is actually a special case of an $F$ distribution with $\nu_1 = 1$ and $\nu_2 = \nu$ degrees of freedom. Neither of the cumulative distribution functions can be obtained in closed form. Consequently, we dispense with the equations of the density functions and present useful values of the statistics and corresponding areas in tabular form in Appendix B as well as using statistical software to find probabilities.

The following examples illustrate how these statistics can be used to make probability statements about population parameters.

**Example 6.24**

Chi-square Distribution Application

Consider a cannery that produces 8-ounce cans of processed corn. Quality control engineers have determined that the process is operating properly when the true variation $\sigma^2$ of the fill amount per can is less than .0025. A random sample of $n = 10$ cans is selected from a day's production, and the fill amount (in ounces) recorded for each. Of interest is the sample variance, $S^2$. If, in fact, $\sigma^2 = .001$, find the probability that $S^2$ exceeds .0025. Assume that the fill amounts are normally distributed.

Solution

We want to calculate $P(S^2 > .0025)$. Assume the sample of 10 fill amounts is selected from a normal distribution. Theorem 6.11 states that the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-square probability distribution with $\nu = (n-1)$ degrees of freedom. Consequently, the probability we seek can be written

$$P(S^2 > .0025) = P\left[\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)(.0025)}{\sigma^2}\right]$$

$$= P\left[\chi^2 > \frac{(n-1)(.0025)}{\sigma^2}\right]$$

Substituting $n = 10$ and $\sigma^2 = .001$, we have

$$P(S^2 > .0025) = P\left[\chi^2 > \frac{9(.0025)}{.001}\right] = P(\chi^2 > 22.5)$$

Upper-tail areas of the chi-square distribution have been tabulated and are given in Table 8 of Appendix B, a portion of which is reproduced in Table 6.2. The table gives the values of $\chi^2$, denoted $\chi^2_a$, that locate an area (probability) $a$ in the upper tail of the distribution, i.e., $P(\chi^2 > \chi^2_a) = a$. In our example, we want to find the probability $a$ such that $\chi^2_a > 22.5$.

Now, for $n = 10$, we have $\nu = n - 1 = 9$ degrees of freedom. Searching Table 6.2 in the row corresponding to $\nu = 9$, we find that $\chi^2_{.01} = 21.666$ and $\chi^2_{.005} = 23.5893$. (These values are shaded in Table 6.2.) Consequently, the probability that we seek falls between $a = .01$ and $a = .005$, i.e.,

$$.005 < P(\chi^2 > 22.5) < .01 \qquad \text{(see Figure 6.17)}$$

Thus, the probability that the variance of the sample fill amounts exceeds .0025 is small (between .005 and .01) when the true population variance $\sigma^2$ equals .001.

**FIGURE 6.17**
Finding $P(\chi^2 > 22.5)$ in Example 6.24

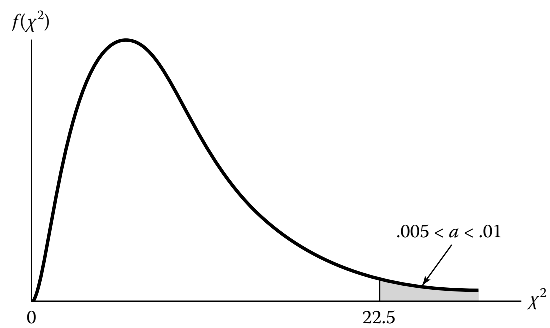
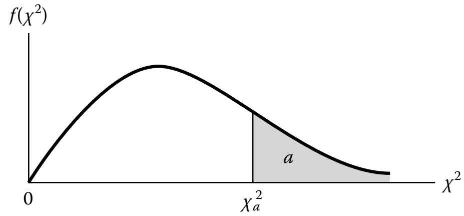


**TABLE 6.2 Abbreviated Version of Table 8 of Appendix B: Tabulated Values of $\chi^2$**




| Degrees of Freedom | $\chi^2_{.100}$ | $\chi^2_{.050}$ | $\chi^2_{.025}$ | $\chi^2_{.010}$ | $\chi^2_{.005}$ |
|---|---|---|---|---|---|
| 1 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| 2 | 4.60517 | 5.99147 | 7.37776 | 9.21034 | 10.5966 |
| 3 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8381 |
| 4 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8602 |
| 5 | 9.23635 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 7 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 13.3616 | 15.5073 | 17.5346 | 20.0902 | 21.9550 |
| 9 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 10 | 15.9871 | 18.3070 | 20.4831 | 23.2093 | 25.1882 |
| 11 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7569 |
| 12 | 18.5494 | 21.0261 | 23.3367 | 26.2170 | 28.2995 |
| 13 | 19.8119 | 22.3621 | 24.7356 | 27.6883 | 29.8194 |
| 14 | 21.0642 | 23.6848 | 26.1190 | 29.1413 | 31.3193 |
| 15 | 22.3072 | 24.9958 | 27.4884 | 30.5779 | 32.8013 |
| 16 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 |
| 17 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 |
| 18 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1564 |
| 19 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5822 |

The exact probability in Example 6.24 can be found using statistical software. Figure 6.18 is a MINITAB printout showing the probability for a chi-square distribution with 9 degrees of freedom. Note that (by default), MINITAB computes the cumulative probability $P(\chi^2 < 22.5) = .99278$. Consequently, the exact probability we need is

$$P(\chi^2 > 22.5) = 1 - P(\chi^2 < 22.5) = 1 - .99278 = .00722$$

**FIGURE 6.18**
MINITAB Chi-square Probability

```
Cumulative Distribution Function

Chi-Square with 9 DF

   x  P( X <= x )
22.5     0.992578
```

**Example 6.25**

Derivation of Student's T-distribution

Suppose the random variables $\overline{Y}$ and $S^2$ are the mean and variance of a random sample of $n$ observations from a normally distributed population with mean $\mu$ and variance $\sigma^2$. It can be shown (proof omitted) that $\overline{Y}$ and $S^2$ are statistically independent when the sampled population has a normal distribution. Use this result to show that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

possesses a $T$ distribution with $\nu = (n - 1)$ degrees of freedom.*

Solution

We know from Theorem 6.10 that $\overline{Y}$ is normally distributed with mean $\mu$ and variance $\sigma^2/n$. Therefore,

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable. We also know from Theorem 6.11 that

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a $\chi^2$ random variable with $\nu = (n - 1)$ degrees of freedom. Then, using Definition 6.15 and the information that $\overline{Y}$ and $S^2$ are independent, we conclude that

$$T = \frac{Z}{\sqrt{\chi^2/\nu}} = \frac{\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}} = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

has a Student's $T$ distribution with $\nu = (n - 1)$ degrees of freedom. As we will learn in Chapter 7, the $T$ distribution is useful for making inferences about the population mean $\mu$ when the population standard deviation $\sigma$ is unknown (and must be estimated by $S^2$).

Theorem 6.11 and Examples 6.24 and 6.25 identify the sampling distributions of two statistics that will play important roles in statistical inference. Others are presented without proof in Tables 6.3a and 6.3b. All are based on random sampling from normally distributed populations. These results will be needed in Chapter 7.

*The result was first published in 1908 by W. S. Gosset, who wrote under the pen name of Student. Thereafter, this statistic became known as Student's $T$.

**TABLE 6.3a Sampling Distributions of Statistics Based on Independent Random Samples of $n_1$ and $n_2$ Observations, Respectively, from Normally Distributed Populations with Parameters $(\mu_1, \sigma_1^2)$ and $(\mu_2, \sigma_2^2)$**

| Statistic | Sampling Distribution | Additional Assumptions | Basis of Derivation of Sampling Distribution |
|---|---|---|---|
| $\chi^2 = \dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$ where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ | Chi-square with $\nu = (n_1 + n_2 - 2)$ degrees of freedom | $\sigma_1^2 = \sigma_2^2 = \sigma^2$ | Theorems 6.11–6.12 |
| $T = \dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ | Student's $T$ with $\nu = (n_1 + n_2 - 2)$ degrees of freedom | $\sigma_1^2 = \sigma_2^2 = \sigma^2$ | Theorems 6.10–6.11 and Definition 6.15 |
| $F = \left(\dfrac{S_1^2}{S_2^2}\right)\left(\dfrac{\sigma_2^2}{\sigma_1^2}\right)$ | $F$ distribution with $\nu_1 = (n_1 - 1)$ numerator degrees of freedom and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom | None | Theorem 6.11 and Definition 6.17 |

**TABLE 6.3b Sampling Distributions of Statistics Based on a Random Sample from a Single Normally Distributed Population with Mean $\mu$ and Variance $\sigma^2$**

| Statistic | Sampling Distribution | Additional Assumptions | Basis of Derivation of Sampling Distribution |
|---|---|---|---|
| $\chi^2 = \dfrac{(n - 1)S^2}{\sigma^2}$ | Chi-square with $\nu = (n - 1)$ degrees of freedom | None | Methods of Section 6.7 |
| $t = \dfrac{\bar{y} - \mu}{S/\sqrt{n}}$ | Student's $T$ with $\nu = (n - 1)$ degrees of freedom | None | Theorems 6.10–6.11 and Definition 6.15 |

## Applied Exercises

6.95 *Natural gas consumption and temperature.* Refer to the *Transactions of the ASME* (June 2004) study on predicting daily natural gas consumption using temperature, Exercise 5.32 (p. 205). Recall that the researchers showed that the daily July temperature in Buenos Aires, Argentina, is normally distributed with $\mu = 11°C$ and $\sigma = 3°C$. Consider a random sample of $n$ daily July temperatures from the population and let $S^2$ represent the sample variance. Use Table 8 of Appendix B to estimate the following probabilities:

a. $P(S^2 > 14.4)$ when $n = 10$
b. $P(S^2 > 33.3)$ when $n = 5$
c. $P(S^2 > 16.7)$ when $n = 22$

6.96 Refer to Exercise 6.95. Find the exact probabilities using statistical software.

6.97 *Monitoring impedance to leg movements.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) study of impedance to leg movements, Exercise 2.46 (p. 51). Recall that engineers attached electrodes to the ankles and knees of volunteers and measured the signal-to-noise ratio (SNR) of impedance changes. For a particular ankle-knee electrode pair, the SNR values were measured for a sample of $n = 10$ volunteers. Assume the distribution of SNR values in the population is normal with $\mu = 20$ and $\sigma = 5$.

a. Describe the sampling distribution of $T = \sqrt{n}(\overline{Y} - \mu)/S$.

b. Describe the sampling distribution of $\chi^2 = (n - 1)S^2/\sigma^2$.

6.98 *Bearing strength of concrete FRP strips.* Refer to the *Composites Fabrication Magazine* (Sept. 2004) evaluation of a new method of fastening fiber-reinforced polymer (FRP) strips to concrete, Exercise 2.47 (p. 51). Recall that a sample of 10 FRP strips mechanically fastened to highway bridges were tested for bearing strength. The strength measurement $Y$ (in mega Pascal units, MPa) was recorded for each strip. Assume that $Y$ is normally distributed with variance $\sigma^2 = 100$.

a. Describe the sampling distribution of $S^2$, the sample variance.

b. Find the approximate probability that $S^2$ is less than 16.92.

c. The data for the experiment are reproduced in the table. Do these data tend to contradict or support the assumption that $\sigma^2 = 100$?

**FRP**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 240.9 | 248.8 | 215.7 | 233.6 | 231.4 | 230.9 | 225.3 | 247.3 | 235.5 | 238.0 |

Data are simulated from summary information provided in *Composites Fabrication Magazine*, Sept. 2004, p. 32 (Table 1).

6.99 Seismic ground noise. Refer to the *Earthquake Engineering and Engineering Vibration* (March 2013) study of the structural damage to a three-story building caused by seismic ground noise, Exercise 5.33 (p. 205). Recall that the acceleration $Y$ (in meters per second-squared) of the seismic ground noise was modeled using a normal probability distribution. Assume $\mu = .5$ and unknown $\sigma$. Consider a random sample of $n = 16$ acceleration measurements from this population, and let $\overline{Y}$ represent the sample mean and $S$ the sample standard deviation.

a. Describe the sampling distribution of the statistic, $T = 4(\overline{Y} - .5)/S$.

b. Suppose the sample standard deviation is $S = .015$. Use this value, the result, part a, and statistical software to find the exact probability that the sample mean acceleration is less than .52 meters per sec$^2$.

6.100 *Flicker in an electrical power system.* Refer to the *Electrical Engineering* (March 2013) assessment of the quality of electrical power, Exercise 5.37 (p. 205). Recall that a measure of quality is the degree to which voltage fluctuations cause light flicker in the system. The perception of light flicker $Y$ in a system (measured periodically over 10-minute intervals) follows (approximately) a normal distribution with $\mu = 2.2\%$ and $\sigma = .5\%$. Consider a random sample of 35 intervals.

a. Use statistical software to find the exact probability that the sample standard deviation, $S$, is less than .75%.

b. If $S = .4\%$, use statistical software to find the exact probability that the sample mean perception of light flicker over the 35 sample intervals exceeds 2%.

## Theoretical Exercises

6.101 Let $Y_1, Y_2, \ldots, Y_{n_1}$ be a random sample of $n_1$ observations from a normal distribution with mean $\mu_1$ and variance $\sigma_1^2$. Let $X_1, X_2, \ldots, X_{n_2}$ be a random sample of $n_2$ observations from a normal distribution with mean $\mu_2$ and variance $\sigma_2^2$. Assuming the samples were independently selected, show that

$$F = \left(\frac{S_1^2}{S_2^2}\right)\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$

has an $F$ distribution with $\nu_1 = (n_1 - 1)$ numerator degrees of freedom and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom.

6.102 Let $S_1^2$ and $S_2^2$ be the variances of independent random samples of sizes $n_1$ and $n_2$ selected from normally distributed populations with parameters $(\mu_1, \sigma^2)$ and $(\mu_2, \sigma^2)$, respectively. Thus, the populations have different means, but a common variance $\sigma^2$. To estimate the common variance, we can combine information from both samples and use the **pooled estimator**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Use Theorems 6.11 and 6.12 to show that $(n_1 + n_2 - 2)S_p^2/\sigma^2$ has a chi-square distribution with $\nu = (n_1 + n_2 - 2)$ degrees of freedom.

6.103 Let $\overline{Y}_1$ and $\overline{Y}_2$ be the means of independent random samples of sizes $n_1$ and $n_2$ selected from normally distributed populations with parameters $(\mu_1, \sigma^2)$ and $(\mu_2, \sigma^2)$, respectively. If

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

show that

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

has a Student's $T$ distribution with $\nu = (n_1 + n_2 - 2)$ degrees of freedom.

## *STATISTICS IN ACTION REVISITED*

### Availability of an Up/Down Maintained System

Recall (from SIA, p. 235) that "cycle availability" is the probability that a system is functioning at any point in time during the maintenance cycle. If the random variable $X$ represents the time between failures of the system (i.e., the "up" time) and the random variable Y represents the time to repair the system during a maintenance cycle (i.e., the "down" time), then $(X + Y)$ represents the total cycle time. Further, the random variable cycle availability, A, is defined as follows:

$$A = X/(X + Y), X > 0 \text{ and } Y > 0$$

As an example, the Department of Defense used the exponential distribution (Section 5.7) to model both failure time, $X$, and repair time, $Y$. Given this assumption, what are the properties of the probability distribution of availability, $A$? The Department of Defense derived this distribution both theoretically (using the method of transformations) and with Monte Carlo simulation. An outline of the method of transformations (details of which are beyond the scope of this text) follows.

Assume that failure time $X$ has an exponential distribution with mean 1 hour, repair time $Y$ has an exponential distribution with mean 1 hour, and that $X$ and $Y$ are independent. Then,

$$f_1(x) = e^{-x}, \quad x > 0$$

$$f_2(y) = e^{-y}, \quad y > 0$$

and, since $X$ and $Y$ are independent,

$$f(x, y) = f_1(x) \cdot f_2(y) = e^{-x} e^{-y} = e^{-(x+y)}$$

Now $A = X/(X + Y)$, and let $B = (X + Y)$. Then, using some algebra, we can show that $X = A \cdot B$ and $Y = (1 - A) \cdot B$.

To find the density function for $A$, $g(a)$, we first need to find the joint probability density function for $A$ and $B$, $g(a, b)$. This density is derived by substituting $x = ab$ and $y = (1 - a)b$ into the density, $f(x, y)$, and multiplying by an appropriate derivative. (This is the transformation method.) Given the preceding assumptions, it can be shown (proof omitted) that

$$g(a, b) = be^{-b}, \quad 0 < a < 1 \quad \text{and} \quad b > 0$$

To find $g(a)$, we integrate $g(a, b)$ over the range of $b$, as demonstrated in Section 5.4.

$$g(a) = \int_0^{\infty} g(a, b)\, db = \int_0^{\infty} (be^{-b})\, db = -e^{-b}\Big]_0^{\infty} = [-e^{-\infty} - (-e^0)] = 1$$

Since $A$ ranges between 0 and 1, you can see that $g(a)$ is simply the probability density function for a uniform random variable between 0 and 1 (see Section 5.4).

Theoretical performance measures for the system can now be obtained using the parameters of the uniform distribution. For example, the expected value of availability is $E(A) = .5$ and the variance is $Var(A) = 1/12 = .083$. Also, the 10th percentile, i.e., the value of $a$ so that $P(A \leq a) = .1$, is $a = .1$. Similarly, the lower and upper quartiles are $Q_1 = .25$ and $Q_3 = .75$.

**AVAILEXP**

The Department of Defense also used Monte Carlo simulation to derive the distribution of cycle availability. A total of $n = 5{,}000$ random values of $X$ and $Y$ were obtained (independently) from an exponential random number generator. The value of $A = X/(X + Y)$ was calculated for each of these random values. We performed this simulation and saved the results in the **AVAILEXP** file. Summary statistics for the 5,000 values of $A$ are shown in the MINITAB printout, Figure SIA6.1. From the printout, the mean, variance, and lower and upper quartiles are .503, .084, .247, and .753, respectively. These simulated values agree very closely with the derived theoretical uniform distribution.

**Descriptive Statistics: A**

| Variable | N | Mean | Variance | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| A | 5000 | 0.50251 | 0.08418 | 0.000590 | 0.24749 | 0.50667 | 0.75266 | 0.99999 |

**FIGURE SIA6.1**
MINITAB summary statistics for simulated availability using the exponential distribution for failure time and repair time

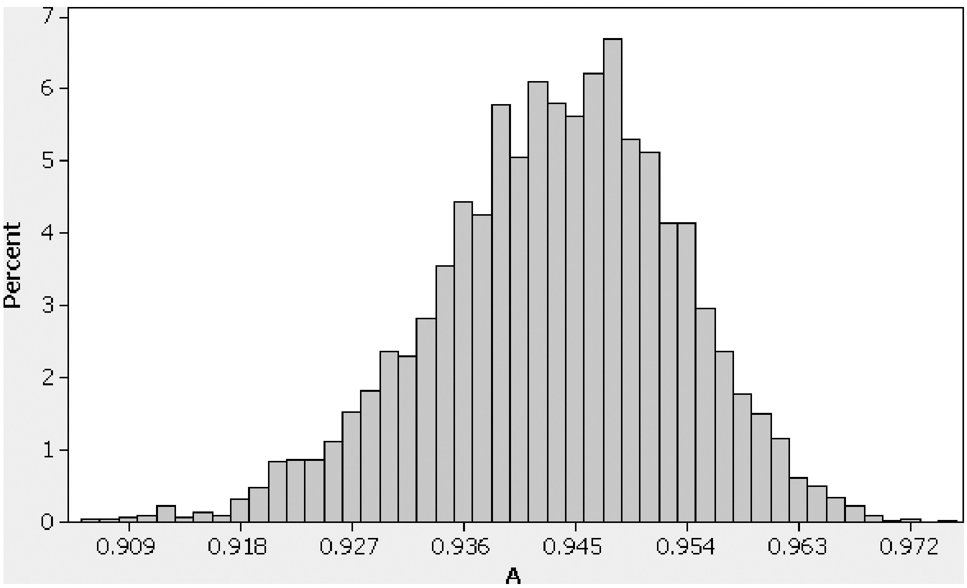
The Monte Carlo simulation approach is especially valuable for failure time and repair time distributions that are not modeled well by exponential distributions. Engineers have found that cycle availability generally can be modeled by a beta distribution (see Section 5.9), where the parameter $\alpha$ represents the mean time between failures (MTBF) and the parameter $\beta$ represents the mean time to repair (MTTR). As a more realistic example, the Department of Defense considered a system with $\alpha =$ MTBF $=$ 500 hours and $\beta =$ MTTR $=$ 30 hours.

**AVAILBETA**

We simulated $n = 5{,}000$ values of $A$ from this beta distribution using MINITAB and saved the results in the **AVAILBETA** file. A histogram, as well as summary statistics, for these 5,000 values of $A$ are shown in the MINITAB printout, Figure SIA6.2. Recalling that the availability, $A$, represents the probability that the system is "up" and available, these summary statistics provide performance measures of the system. For example, the mean of .943 implies that the system is "up", on average, 94.3% of the time. The mean plus or minus 2 standard deviations gives the interval:

$$.943 \pm 2\sqrt{.0000988} = .943 \pm .0199 = (.923, .963)$$

Thus (applying the Empirical Rule), on approximately 95% of the cycles the probability that the system will be "up" will range anywhere between .923 and .963. The lower-quartile value of .937 implies that the probability of the system being "up" will fall below .937 on only 25% of the cycles. Other important performance measures, such as the 10th percentile and $P(A > .95)$, are easily obtained from the simulated histogram.




**Descriptive Statistics: A**

| Variable | N | Mean | Variance | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| A | 5000 | 0.94339 | 0.0000988 | 0.90555 | 0.93705 | 0.94401 | 0.95026 | 0.97437 |

**FIGURE SIA6.2**
MINITAB descriptive statistics for simulated availability using the beta distribution with MTBF $=$ 500 and MTTR $=$ 30

# Quick Review

## Key Terms

[*Note: Items marked with an asterisk (*) are from the optional section in this chapter*.]



## Key Formulas

| | | |
|---|---|---|
| Conditional probability distribution for discrete random variable: | $p(x \mid y) = p(x, y)/p(y)$ if $X$ and $Y$ are dependent<br>$= p(x)$ if $X$ and $Y$ are independent | 238<br>238 |
| Conditional density function for continuous random variable: | $f(x \mid y) = f(x, y)/f(y)$ if $X$ and $Y$ are dependent<br>$= f(x)$ if $X$ and $Y$ are independent | 242<br>242 |
| Expected values: | $E(c) = c$<br>$E[c \cdot g(X, Y)] = c \cdot E[g(X, Y)]$<br>$E[g_1(X, Y) + g_2(X, Y)] = E[g_1(X, Y)] + E[g_2(X, Y)]$<br>$E(XY) = E(X) \cdot E(Y)$ if $X$ and $Y$ are independent | 246 |
| Covariance: | $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$ if $X$ and $Y$ dependent<br>$= 0$ if $X$ and $Y$ independent | 251<br>251 |
| Correlation: | $\rho = \dfrac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$ if $X$ and $Y$ dependent<br>$= 0$ if $X$ and $Y$ independent | 283 |
| Normal approximation to binomial: | $P(a < Y < b) = P\left\{\dfrac{(a - .5) - np}{\sqrt{npq}} < Z < \dfrac{(b + .5) - np}{\sqrt{npq}}\right\}$ | 272 |
| Sampling distribution of $\overline{Y}$: | Mean $= \mu$ Standard deviation $= \sigma/\sqrt{n}$ | 263 |
| Sampling distribution of $\Sigma Y$: | Mean $= n\mu$ Standard deviation $= \sqrt{n}\,\sigma$ | 268 |

## LANGUAGE LAB

| *Symbol* | *Pronunciation* | *Description* |
|---|---|---|
| $p(x \mid y)$ | $p$ of $x$ given $y$ | Conditional probability distribution for $X$ given $Y$ |
| $f(x \mid y)$ | $f$ of $x$ given $y$ | Conditional density function for $X$ given $Y$ |
| $\text{Cov}(X, Y)$ | Covariance | Covariance of $X$ and $Y$ |
| $\rho$ | rho | Correlation coefficient for $X$ and $Y$ |
| $\mu_{\bar{y}}$ | mu of $\overline{Y}$ | Mean of sampling distribution of $\overline{Y}$ |
| $\sigma_{\bar{y}}$ | sigma of $\overline{Y}$ | Standard deviation of sampling distribution of $\overline{Y}$ |

## Chapter Summary

- The **joint probability distribution** for two random variables is called a **bivariate distribution**.
- The **conditional probability distribution** for a random variable $X$, given $Y$, is the joint probability distribution for $X$ and $Y$ divided by the marginal probability distribution for $Y$.
- The **covariance** of $X$ and $Y$: $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$.
- The **correlation**: of $X$ and $Y$: $\rho = \text{Cov}(X, Y)/(\sigma_x \sigma_y)$
- For two **independent** random variables, (1) the joint probability distribution is the product of the two respective marginal probability distributions, (2) $E(XY) = E(X) \cdot E(y)$, (3) covariance equals 0, and (4) correlation equals 0.
- The **sampling distribution** of a statistic is the theoretical probability distribution of the statistic in repeated sampling.
- The **standard error** of a statistic is the standard deviation of the sampling distribution.
- **Monte Carlo simulation** involves repeatedly generating observations on a statistic in order to approximate the sampling distribution.
- The **central limit theorem** states that the sampling distribution of $\overline{Y}$ is approximately normal for large $n$.
- Two properties of the sampling distribution of $\overline{Y}$: mean $= \mu$, standard deviation $= \sigma/\sqrt{n}$
- Normal distribution can be used to approximate a binomial probability when $\mu \pm 2\sigma$ falls within the interval $(0, n)$. This will be true when $np \geq 4$ **and** $nq \geq 4$.
- Some sampling distributions related to the normal distribution: **chi-square distribution, Student's *T* distribution**, and ***F* distribution**.

## Supplementary Exercises

6.104 *Automated drilling machine.* Refer to the *Journal of Engineering for Industry* (Aug. 1993) study of an automated drilling machine, Exercise 3.67 (p. 120). The eight machining conditions used in the study are reproduced here.

| *Experiment* | *Workpiece Material* | *Dril Size (in.)* | *Drill Speed (rpm)* | *Feed Rate (ipr)* |
|---|---|---|---|---|
| 1 | Cast iron | .25 | 1,250 | .011 |
| 2 | Cast iron | .25 | 1,800 | .005 |
| 3 | Steel | .25 | 3,750 | .003 |
| 4 | Steel | .25 | 2,500 | .003 |
| 5 | Steel | .25 | 2,500 | .008 |
| 6 | Steel | .125 | 4,000 | .0065 |
| 7 | Steel | .125 | 4,000 | .009 |
| 8 | Steel | .125 | 3,000 | .010 |

Suppose that two of the machining conditions listed will detect a flaw in the automated system. Define $X$ as the number of these two conditions with steel material, and define $Y$ as the number of these two conditions with .25-inch drill size.

a. Find the bivariate probability distribution $p(x, y)$.
b. Find the marginal probability distribution $p_2(y)$.
c. Find the conditional probability distribution $p_1(x \mid y)$.

**DDT**

6.105 *Contaminated fish.* Refer to the U.S. Army Corps of Engineers data on contaminated fish saved in the **DDT** file. (See *Statistics in Action*, Chapters 1 and 2.) Recall that the length (in centimeters), weight (in grams), and DDT level (in parts per million) were measured for each of 144 fish caught from the polluted Tennessee River in Alabama.

a. An analysis of the data reveals that the distribution of fish lengths is skewed to the left. Assume that this is true of the population of fish lengths and that the population has mean $\mu = 43$ centimeters and $\sigma = 7$ centimeters. Use this information to describe the sampling distribution of $\bar{y}$, the mean length of a sample of $n = 40$ fish caught from the Tennessee River.
b. An analysis reveals that the distribution of fish weights is approximately normal. Assume that this is true of the population of fish weights and that the population has mean $\mu = 1{,}050$ grams and $\sigma = 376$ grams. Use this information to describe the sampling distribution of $\overline{Y}$, the mean weight of a sample of $n = 40$ fish caught from the Tennessee River.
c. An analysis reveals that the distribution of fish DDT levels is highly skewed to the right. Assume that this is true of the population of fish DDT levels and that the population has mean $\mu = 24$ ppm and $\sigma = 98$ ppm. Use this information to describe the sampling distribution of $\overline{Y}$, the mean DDT level of a sample of $n = 40$ fish caught from the Tennessee River.

Table for Exercise 6.106

| | | $X=x$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| | 1 | .001 | .002 | .002 | .025 | .040 | .025 | .005 | .005 | 0 | 0 |
| | 2 | .005 | .005 | .010 | .075 | .100 | .075 | .050 | .030 | .030 | .025 |
| $Y=y$ | 3 | 0 | 0 | 0 | .025 | .050 | .080 | .050 | .080 | .040 | .030 |
| | 4 | 0 | .001 | .002 | .005 | .010 | .025 | .010 | .003 | .001 | .001 |
| | 5 | 0 | .002 | .005 | .005 | .020 | .030 | .015 | 0 | 0 | 0 |

6.106 *Decision-support system.* The management of a bank must decide whether to install a commercial loan decision-support system (an online management information system) to aid its analysts in making commercial loan decisions. Past experience shows that $X$, the additional number (per year) of correct loan decisions—accepting good loan applications and rejecting those that would eventually be defaulted—attributable to the decision-support system, and $Y$, the lifetime (in years) of the decision-support system, have the joint probability distribution shown in the table at the top of the page.

a. Find the marginal probability distributions, $p_1(x)$ and $p_2(y)$.
b. Find the conditional probability distribution, $p_1(x \mid y)$.
c. Given that the decision-support system is in its third year of operation, find the probability that at least 40 additional correct loan decisions will be made.
d. Find the expected lifetime of the decision-support system, i.e., find $E(Y)$.
e. Are $X$ and $Y$ correlated? Are $X$ and $Y$ independent?
f. Each correct loan decision contributes approximately \$25,000 to the bank's profit. Compute the mean and standard deviation of the additional profit attributable to the decision-support system. [*Hint*: Use the marginal distribution $p_1(x)$.]

6.107 *Quality control inspectors.* Suppose that $X$ and $Y$, the proportions of an 8-hour workday that two quality control inspectors actually spend on performing their assigned duties, have joint probability density

$$f(x, y) = \begin{cases} x + y & \text{if } 0 \le x \le 1;\ 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the marginal probability distributions, $f_1(x)$ and $f_2(y)$.
b. Verify that

$$\int_{-\infty}^{\infty} f_1(x)\, dx = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} f_2(y)\, dy = 1$$

c. Find the conditional probability distributions, $f_1(x \mid y)$ and $f_2(y \mid x)$.
d. Verify that

$$\int_{-\infty}^{\infty} f_1(x|y)\, dx = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} f_2(y \mid x)\, dy = 1$$

e. Are $X$ and $Y$ correlated? Are $X$ and $Y$ independent?
f. The proportion $d$ of "dead" time (i.e., time when no assigned duties are performed) for the two attendants is given by the relation $D = 1 - (X + Y)/2$. Find $E(D)$ and $V(D)$. Within what limits would you expect $D$ to fall?

6.108 *Sleep-inducing hormone.* Studies by neuroscientists at the Massachusetts Institute of Technology (MIT) reveal that melatonin, which is secreted by the pineal gland in the brain, functions naturally as a sleep-inducing hormone. Male volunteers were given various doses of melatonin or placebos and then placed in a dark room at midday and told to close their eyes and fall asleep on demand. Of interest to the MIT researchers is the time $Y$ (in minutes) required for each volunteer to fall asleep. With the placebo (i.e., no hormone), the researchers found that the mean time to fall asleep was 15 minutes. Assume that with the placebo treatment $\mu = 15$ and $\sigma = 5$.

a. Consider a random sample of $n = 20$ men who are given the sleep-inducing hormone, melatonin. Let $\overline{Y}$ represent the mean time to fall asleep for this sample. If the hormone is *not* effective in inducing sleep, describe the sampling distribution of $\overline{Y}$.
b. Refer to part a. Find $P(\overline{Y} \le 6)$.
c. In the actual study, the mean time to fall asleep for the 20 volunteers was $\overline{Y} = 5$. Use this result to make an inference about the true value of $\mu$ for those taking the melatonin.

6.109 *Merging into traffic.* The merging process from an acceleration lane to the through lane of a freeway constitutes an important aspect of traffic operation at interchanges. A study of parallel interchange ramps in Israel revealed that many drivers do not use the entire length of parallel lanes for acceleration, but seek as soon as possible an appropriate gap in the major stream of traffic for merging (*Transportation Engineering*, Nov. 1985). At one site (Yavneh), 54% of the drivers use less than half the lane length available before merging. Suppose we plan to monitor the merging patterns of a random sample of 330 drivers at the Yavneh site.

a. What is the approximate probability that fewer than 100 of the drivers will use less than half the acceleration lane length before merging?
b. What is the approximate probability that 200 or more of the drivers will use less than half the acceleration lane length before merging?

6.110 *Creep in concrete.* Concrete experiences a characteristic marked increase in "creep" when it is heated for the first time under load. An experiment was conducted to investigate the transient thermal strain behavior of concrete (*Magazine of Concrete Research*, Dec. 1985). Two variables thought to affect thermal strain are $X$, rate of heating (degrees centigrade per minute), and $Y$, level of load (percentage of initial strength). Concrete specimens are prepared and tested under various combinations of heating rate and load, and the thermal strain is determined for each. Suppose the joint probability distribution for $X$ and $Y$ for those specimens that yielded acceptable results is as given in the table. Suppose a concrete specimen is randomly selected from among those in the experiment that yielded acceptable thermal strain behavior.

| | | $X$(°C/minute) | | | | |
|---|---|---|---|---|---|---|
| | | .1 | .2 | .3 | .4 | .5 |
| | 0 | .17 | .11 | .07 | .05 | .05 |
| $Y$ | 10 | .10 | .06 | .05 | .02 | .01 |
| | 20 | .09 | .04 | .03 | .01 | 0 |
| | 30 | .08 | .04 | .02 | 0 | 0 |

a. Find the probability that the concrete specimen was heated at a rate of .3°C/minute.
b. Given that the concrete specimen was heated at .3°C/minute, find the probability that the specimen had a load of 20%.
c. Are rate of heating $X$ and level of load $Y$ correlated?
d. Are rate of heating $X$ and level of load $Y$ independent?

6.111 *Demand for home heating oil.* A supplier of home heating oil has a 250-gallon tank that is filled at the beginning of each week. Since the weekly demand for the oil increases steadily up to 100 gallons and then levels off between 100 and 250 gallons, the probability distribution of the weekly demand $Y$ (in hundreds of gallons) can be represented by

$$f(y) = \begin{cases} \dfrac{y}{2} & \text{if } 0 \leq y \leq 1 \\ \dfrac{1}{2} & \text{if } 1 \leq y \leq 2.5 \\ 0 & \text{elsewhere} \end{cases}$$

If the supplier's profit is given by $W = 10Y - 2$, find the probability density function of $W$.

6.112 *Dioxin study.* Dioxin, often described as the most toxic chemical known, is created as a by-product in the manufacture of herbicides such as Agent Orange. Scientists have found that .000005 gram (five-millionths of a gram) of dioxin—a dot barely visible to the human eye—is a lethal dose for experimental guinea pigs in more than half the animals tested, making dioxin 2,000 times more toxic than strychnine. Assume that the amount of dioxin required to kill a guinea pig has a relative frequency distribution with mean $\mu = .000005$ gram and standard deviation $\sigma = .000002$ gram. Consider an experiment in which the amount of dioxin required to kill each of $n = 50$ guinea pigs is measured, and the sample mean $\overline{Y}$ is computed.

a. Calculate $\mu_{\overline{y}}$ and $\sigma_{\overline{y}}$.
b. Find the probability that the mean amount of dioxin required to kill the 50 guinea pigs is larger than .0000053 gram.

6.113 *Forest canopy closure.* The determination of the percent canopy closure of a forest is essential for wildlife habitat assessment, watershed runoff estimation, erosion control, and other forest management activities. One way in which geoscientists estimate percent forest canopy closure is through the use of a satellite sensor called the Landsat Thematic Mapper. A study of the percent canopy closure in the San Juan National Forest (Colorado) was conducted by examining Thematic Mapper Simulator (TMS) data collected by aircraft at various forest sites (*IEEE Transactions on Geoscience and Remote Sensing*, Jan. 1986). The mean and standard deviation of the readings obtained from TMS Channel 5 were found to be 121.74 and 27.52, respectively.

a. Let $\overline{Y}$ be the mean TMS reading for a sample of 32 forest sites. Assuming the figures given are population values, describe the sampling distribution of $\overline{Y}$.
b. Use the sampling distribution of part **a** to find the probability that $\overline{Y}$ falls between 118 and 130.

6.114 *Canopy closure variance.* Refer to Exercise 6.113. Let $S^2$ be the variance of the TMS readings for the 32 sampled forest sites. Assuming the sample is from **a** normal population, estimate the probability that $S^2$ exceeds 1,311.

6.115 *Monitoring the filling process.* University of Louisville researchers examined the process of filling plastic pouches of dry blended biscuit mix (*Quality Engineering*, Vol. 91, 1996). The current fill mean of the process is set at $\mu = 406$ grams and the process fill standard deviation is $\sigma = 10.1$ grams. (According to the researchers, "the high level of variation is due to the fact that the product has poor flow properties and is, therefore, difficult to fill consistently from pouch to pouch.") Operators monitor the process by randomly sampling 36 pouches each day and measuring the amount of biscuit mix in each. Consider $\overline{Y}$, the main fill amount of the sample of 36 products. Suppose that on one particular day, the operators observe $\overline{Y} = 400.8$. One of the operators believes that this indicates that the true process fill mean $\mu$ for that day is less than 406 grams. Another operator argues that

$\mu = 406$ and the small value of $\overline{Y}$ observed is due to random variation in the fill process. Which operator do you agree with? Why?

6.116 *Lot acceptance sampling.* Quality control is a problem with items that are mass-produced. The production process must be monitored to ensure that the rate of defective items is kept at an acceptably low level. One method of dealing with this problem is **lot acceptance sampling**, in which a random sample of items produced is selected and each item in the sample is carefully tested. The entire lot of items is then accepted or rejected, based on the number of defectives observed in the sample. Suppose a manufacturer of pocket calculators randomly chooses 200 stamped circuits from a day's production and determines $Y$, the number of defective circuits in the sample. If a sample defective rate of 6% or less is considered acceptable and, unknown to the manufacturer, 8% of the entire day's production of circuits is defective, find the approximate probability that the lot of stamped circuits will be rejected.

6.117 *Reflection of neutral particles.* Refer to the problem of transporting neutral particles in a nuclear fusion reactor, described in Exercise 3.36 (p. 103). Recall that particles released into a certain type of evacuated duct collide with the inner duct wall and are either scattered (reflected) with probability .16 or absorbed with probability .84 (*Nuclear Science and Engineering*, May 1986). Suppose 2,000 neutral particles are released into an unknown type of evacuated duct in a nuclear fusion reactor. Of these, 280 are reflected. What is the approximate probability that as few as 280 (i.e., 280 or fewer) of the 2,000 neutral particles would be reflected off the inner duct wall if the reflection probability of the evacuated duct is .16?

6.118 *Math programming problem. IEEE Transactions* (June 1990) presented a hybrid algorithm for solving polynomial 0–1 mathematical programming problems. The solution time (in seconds) for a randomly selected problem solved using the hybrid algorithm has a normal probability distribution with mean $\mu = .8$ second and $\sigma = 1.5$ seconds. Consider a random sample of $n = 30$ problems solved with the hybrid algorithm.

a. Describe the sampling distribution of $S^2$, the variance of the solution times for the 30 problems.
b. Find the approximate probability that $S^2$ will exceed 3.30.

6.119 *Flaws in aluminum siding.* A building contractor has decided to purchase a load of factory-reject aluminum siding as long as the average number of flaws per piece of siding in a sample of size 35 from the factory's reject pile is 2.1 or less. If it is known that the number of flaws per piece of siding in the factory's reject pile has a Poisson probability distribution with a mean of 2.5, find the approximate probability that the contractor will not purchase a load of siding. (*Hint*: If $Y$ is a Poisson random variable with mean $\lambda$, then $\sigma_y^2$ also equals $\lambda$.)

6.120 *Machine repair time.* An article in *Industrial Engineering* (August 1990) discussed the importance of modeling machine downtime correctly in simulation studies. As an illustration, the researcher considered a single-machine-tool system with repair times (in minutes) that can be modeled by an exponential distribution with $\beta = 60$. Of interest is the mean repair time, $\overline{Y}$, of a sample of 100 machine breakdowns.

a. Find $E(\overline{Y})$ and the variance of $\overline{Y}$.
b. What probability distribution provides the best model of the sampling distribution of $\overline{Y}$? Why?
c. Calculate the probability that the mean repair time, $\overline{Y}$, is no longer than 30 minutes.

6.121 *Fecal pollution at Huntington Beach.* The State of California mandates fecal indicator bacteria monitoring at all public beaches. When the concentration of fecal bacteria in a single sample of water exceeds 400 colony-forming units of fecal coliform per 100 milliliters, local health officials must post a sign (called surf zone posting) warning beachgoers of potential health risks from entering the water. Engineers at the University of California–Irvine conducted a study of the surf water quality at Huntington Beach in California and reported the results in *Environmental Science & Technology* (Sept. 2004). The researchers found that beach closings were occurring despite low pollution levels in some instances and in others, signs were not posted when the fecal limit was exceeded. They attributed these "surf zone posting errors" to the variable nature of water quality in the surf zone (for example, fecal bacteria concentration tends to be higher during ebb tide and at night) and the inherent time delay between when a water sample is collected and when a sign is posted or removed. In order to prevent posting errors, the researchers recommend using an averaging method rather than a single sample to determine unsafe water quality. (For example, one simple averaging method is to take a random sample of multiple water specimens and compare the average fecal bacteria level of the sample to the limit of 400 cpu/100 mL in order to determine whether the water is safe.) Discuss the pros and cons of using the single-sample standard versus the averaging method. Part of your discussion should address the probability of posting a sign when, in fact, the water is safe, and the probability of posting a sign, when, in fact, the water is unsafe. (Assume the fecal bacteria concentrations of water specimens at Huntington Beach follow an approximate normal distribution.)

6.122 The waiting time $Y$ until delivery of a new component for a data-processing unit is uniformly distributed over the interval from 1 to 5 days. The cost $C$ (in hundreds of dollars) of this delay to the purchaser is given by $C = (2Y^2 + 3)$. Find the probability that the cost of delay is at least \$2,000, i.e., compute $P(C \geq 20)$.

## Theoretical Exercises

6.123 Let $X$ and $Y$ be two continuous random variables with joint density $f(x, y)$. Show that

$$f_2(y \mid x)f_1(x) = f_1(x \mid y)f_2(y)$$

6.124 Let $X$ and $Y$ be uncorrelated random variables. Verify each of the following:

a. $V(X + Y) = V(X - Y)$
b. $\text{Cov}[(X + Y),(X - Y)] = V(X) - V(Y)$

6.125 Suppose three continuous random variables $Y_1$, $Y_2$, $Y_3$ have the joint distribution

$$f(y_1, y_2, y_3) = \begin{cases} c(y_1 + y_2)e^{-y_3} & \text{if } 0 \leq y_1 \leq 1; 0 \leq y_2 \leq 2; y_3 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of $c$ that makes $f(y_1, y_2, y_3)$ a probability density.
b. Are the three variables independent? [*Hint*: If $f(y_1, y_2, y_3) = f_1(y_1)f_2(y_2)f_3(y_3)$ then $Y_1$, $Y_2$ and $Y_3$ are independent.]

6.126 Consider the density function

$$f(y) = \begin{cases} 3y^2 & \text{if } 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density function of $W$, where:

a. $W = \sqrt{Y}$
b. $W = 3 - Y$
c. $W = -\ln(Y)$

6.127 Let $Y_1$ and $Y_2$ be a sample of $n = 2$ observations from a gamma random variable with parameters $\alpha = 1$ and arbitrary β, and corresponding density function

$$f(y_i) = \begin{cases} \dfrac{1}{\beta}e^{-y_i/\beta} & \text{if } y_i > 0 \ (i = 1, 2) \\ 0 & \text{elsewhere} \end{cases}$$

Show that the sum $W = (Y_1 + Y_2)$ is also a gamma random variable with parameters $\alpha = 2$ and $\beta$. [*Hint*: You may use the result

$$P(W \leq w) = P(0 < Y_2 \leq w - Y_1, 0 \leq Y_1 < w) = \int_0^w \int_0^{w-y_1} f(y_1, y_2)\, dy_2\, dy_1$$

Then use the fact that

$$f(y_1, y_2) = f(y_1)f(y_2)$$

since $Y_1$ and $Y_2$ are independent.]

6.128 Let $Y$ have an exponential density with mean $\beta$. Show that $W = 2Y/\beta$ has a $\chi^2$ density with $\nu = 2$ degrees of freedom.

6.129 The lifetime $Y$ of an electronic component of a laptop computer has a *Rayleigh density*, given by

$$f(y) = \begin{cases} \left(\dfrac{2y}{\beta}\right)e^{-y^2/\beta} & \text{if } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability density function for $W = Y^2$, and identify the type of density function. (*Hint:* You may use the result

$$\int \frac{2y}{\beta}e^{-y^2/\beta}\, dy = -e^{-y^2/\beta}$$

in determining the density function for $W$.)

6.130 Let $Y_1$ and $Y_2$ be a random sample of $n = 2$ observations from a normal distribution with mean $\mu$ and variance $\sigma^2$.

a. Show that

$$Z = \frac{Y_1 - Y_2}{\sqrt{2}\sigma}$$

has a standard normal distribution.

b. Given the result in part a, show that $Z^2$ possesses a $\chi^2$ distribution with 1 degree of freedom. [*Hint:* First show that $S^2 = (Y_1 - Y_2)^2/2$; then apply Theorem 6.11.]

6.131 Refer to Exercise 6.62 (p. 260). Use the computer to generate a random sample of $n = 100$ observations from a distribution with probability density

$$f(y) = \begin{cases} e^y & \text{if } y < 0 \\ 0 & \text{elsewhere} \end{cases}$$

Repeat the procedure 1,000 times and compute the sample mean $\overline{Y}$ for each of the 1,000 samples of size $n = 100$. Then generate (by computer) a relative frequency histogram for the 1,000 sample means. Does your result agree with the theoretical sampling distribution described by the central limit theorem?

6.132 Use Theorem 6.7 to draw a random sample of $n = 5$ observations from a population with probability density function given by

$$f(y) = \begin{cases} 2(y - 1) & \text{if } 1 \leq y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

6.133 Use Theorem 6.7 to draw a random sample of $n = 5$ observations from a population with probability density function given by

$$f(y) = \begin{cases} 2ye^{-y^2} & \text{if } 0 < y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

6.134 The continuous random variable $Y$ is said to have a lognormal distribution with parameters $\mu$ and $\sigma$ if its probability density function, $f(y)$, satisfies

$$f(y) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left\{-\frac{(\ln y - \mu)^2}{2\sigma^2}\right\} \qquad (y > 0)$$

Show that $X = \ln(Y)$ has a normal distribution with mean $\mu$ and variance $\sigma^2$.

CHAPTER 7

# Estimation Using Confidence Intervals

## OBJECTIVE

*To explain the basic concepts of statistical estimation; to present some estimators and to illustrate their use in practical sampling situations involving one or two samples*

## CONTENTS

## *STATISTICS IN ACTION*
## Bursting Strength of PET Beverage Bottles

Polyethylene terephthalate (PET) bottles are used for carbonated beverages. At a certain facility, PET bottles are produced by inserting injection molded pre-forms into a stretch blow machine with 24 cavities. Each machine can produce 440 bottles per minute. A critical property of PET bottles is their bursting strength — the pressure at which bottles filled with water burst when pressurized.

In the *Journal of Data Science* (May 2003), researchers measured and analyzed the bursting strength of PET bottles made from two different mold designs – an old design and a new design. The new mold design reduces the time to change molds in the blow machine, thus reducing the downtime of the machine. This advantage, however, will be negated if the new design has problems with low bursting strength. Consequently, an analysis was performed to compare bursting strengths of PET bottles produced from the two mold designs.

The data for the analysis was obtained by testing one bottle per cavity per day produced over a period of 32 days for each design. Since the machine has 24 cavities, there were a total of $32 \times 24 = 768$ PET bottles tested for each design. Each bottle was filled with water and pressurized until it burst and the resulting pressure (in pounds per square inch) was recorded. These bursting strengths are saved in the **PETBOTTLE** file, described in Table SIA7.1. The researchers showed that there were no significant trends in bursting strength over time (days) and that there were no "cavity effects" (i.e., no significant bursting strength differences among the 24 cavities within each design). Thus, the data for all cavities and all days were pooled in order to compare the two mold designs.

**PETBOTTLE**

**TABLE SIA7.1:**

| Variable Name | Description | Data Type |
|---|---|---|
| DESIGN | Mold design (OLD or NEW) | Qualitative |
| DAY | Day number | Quantitative |
| CAVITY | Cavity number | Quantitative |
| STRENGTH | Bursting strength (psi) | Quantitative |

In the Statistics in *Action Revisited* at the end of this chapter, we demonstrate how the methods outlined in this chapter can be used to compare bursting strengths of PET bottles produced from the two mold designs.

## 7.1 Point Estimators and Their Properties

An inference about a population parameter can be made in either of two ways—we can estimate the unknown parameter value or we can make a decision about a hypothesized value of the parameter. To illustrate, we can estimate the mean time $\mu$ for an industrial robot to complete a task, or we might want to decide whether the mean $\mu$ exceeds some value—say, 3 minutes. The method for making a decision about one or more population parameters, called a **statistical test of a hypothesis**, is the topic of Chapter 8. This chapter will be concerned with **estimation**.

Suppose we want to estimate some population parameter, which we denote by $\theta$. For example, $\theta$ could be a population mean $\mu$, a population variance $\sigma^2$, or the probability $F(a)$ that an observation selected from the population is less than or equal to the value $a$. A **point estimator**, designated by the symbol $\hat{\theta}$ (i.e., we place a "hat" over the symbol of a parameter to denote its estimator), is a rule or formula that tells us how to use the observations in a sample to compute a single number (a point) that serves as an **estimate** of the value of $\theta$. For example, let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observed values of the random variable, $Y$. Then the sample mean, $\bar{y}$, is a point estimator of the population mean, $E(Y) = \mu$—i.e., $\hat{\mu} = \bar{y}$. Similarly, the sample variance $s^2$ is a point estimator of $\sigma^2$—i.e., $\hat{\sigma}^2 = s^2$.*

**Definition 7.1**

A **point estimator** is a random variable that represents a numerical estimate of a population parameter based on the measurements contained in a sample. The single number that results from the calculation is called a **point estimate**.

Another way to estimate the value of a population parameter $\theta$ is to use an interval estimator. An **interval estimator** is a rule, usually expressed as a formula, for calculating two points from the sample data. The objective is to form an interval that contains $\theta$ with a high degree of confidence. For example, if we estimate the mean time $\mu$ for a robot to complete a task to be between 2.7 and 3.1 minutes, then the interval 2.7 to 3.1 is an interval estimate of $\mu$.

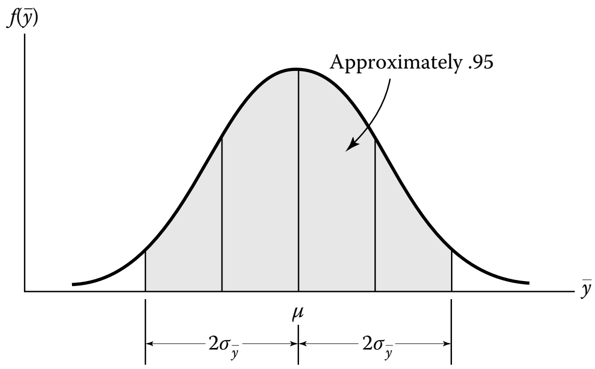
**Definition 7.2**

An **interval estimator** is a pair of random variables obtained from the sample data used to form an interval that estimates a population parameter.

Since a point estimator is calculated from a sample, it possesses a sampling distribution. The sampling distribution of a point estimator completely describes its properties. For example, according to the central limit theorem, the sampling distribution for a sample mean will be approximately normally distributed for large sample sizes, say, $n = 30$ or more, with mean $\mu$ and standard error $\sigma/\sqrt{n}$ (see Figure 7.1). The figure shows that a sample mean $\bar{y}$ is equally likely to fall above or below $\mu$ and that the probability is approximately .95 that it will not deviate from $\mu$ by more than $2\sigma_{\bar{y}} = 2\sigma/\sqrt{n}$.

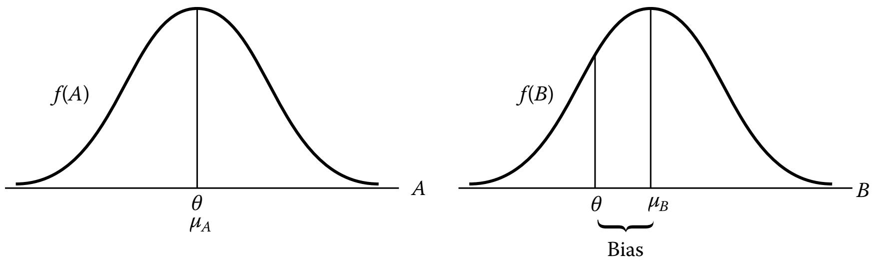
The characteristics exhibited in Figure 7.1 identify the two most desirable properties of estimators. First, we would like the sampling distribution of an estimator to be centered over the parameter being estimated. If the mean of the sampling distribution of an estimator $\hat{\theta}$ is equal to the estimated parameter $\theta$, then the estimator is said to be **unbiased**. If not, the estimator is said to be **biased**. The sample mean is an unbiased estimator of the population mean $\mu$. Sampling distributions for unbiased and biased estimators are shown in Figures 7.2a and 7.2b, respectively.

**FIGURE 7.1**
Sampling distribution of a sample mean for large samples




*In this chapter (and throughout the remainder of the text), we simplify notation and use lowercase letters (e.g., $\bar{y}$ and $s^2$) to represent both a function of random variables and the values the function takes on.

**FIGURE 7.2**
Sampling distributions for unbiased and biased estimators of $\theta$




a. Estimator $A$ is unbiased.
b. Estimator $B$ is biased.

**Definition 7.3**
An estimator $\hat{\theta}$ of a parameter $\theta$ is **unbiased** if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, the estimator is said to be **biased**.

**Definition 7.4**
The **bias** $b(\theta)$ of an estimator $\hat{\theta}$ is equal to the difference between the mean $E(\hat{\theta})$ of the sampling distribution of $\hat{\theta}$ and $\theta$, i.e.,

$$b(\theta) = E(\hat{\theta}) - \theta$$

Unbiased estimators are generally preferred over biased estimators. In addition, given a set of unbiased estimators, we prefer the estimator with *minimum variance*. That is, we want the spread of the sampling distribution of the estimator to be as small as possible so that estimates will tend to fall close to $\theta$.

Figure 7.3 portrays the sampling distribution of two unbiased estimators, $A$ and $B$, with $A$ having smaller variance than $B$. An unbiased estimator that has the minimum variance among all unbiased estimators is called the **minimum variance unbiased estimator (MVUE)**. For example, $\bar{y}$ is the MVUE for $\mu$. That is, $\text{Var}(\bar{y}) = \sigma^2/n$ is the smallest variance among all unbiased estimators of $\mu$. (Proof omitted.)

**Definition 7.5**
The **minimum variance unbiased estimator (MVUE)** of a parameter $\theta$ is the estimator $\hat{\theta}$ that has the smallest variance of all unbiased estimators.

Sometimes we cannot achieve both unbiasedness and minimum variance in the same estimator. For example, Figure 7.4 shows a biased estimator $A$ with a slight bias but with a smaller variance than the MVUE $B$. In such a case, we prefer the estimator that minimizes the **mean squared error**, the mean of the squared deviations between $\hat{\theta}$ and $\theta$:

$$\text{Mean squared error for } \hat{\theta}\text{: } E[(\hat{\theta} - \theta)^2]$$

**FIGURE 7.3**
Sampling distributions for two unbiased estimators of $\theta$ with different variances

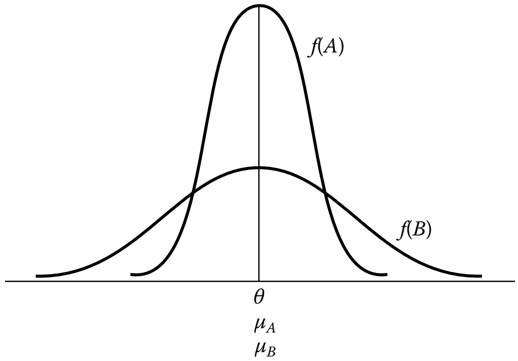
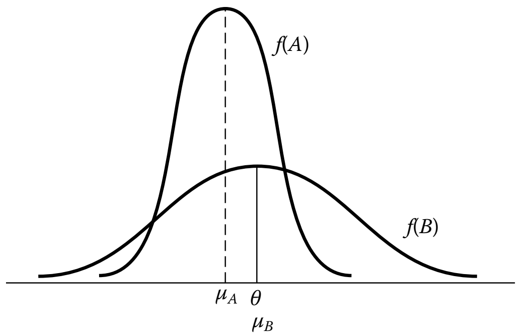


**FIGURE 7.4**
Sampling distributions of biased estimator $A$ and MVUE $B$




It can be shown (proof omitted) that

$$E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + b^2(\theta)$$

Therefore, if $\hat{\theta}$ is unbiased, i.e., if $b(\theta) = 0$, then the mean squared error is equal to $V(\hat{\theta})$. Furthermore, when the bias $b(\theta) = 0$, the estimator $\hat{\theta}$ that yields the smallest mean squared error is also the MVUE for $\theta$.

**Example 7.1**

Unbiased Estimator of $\sigma^2$

Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations on a random variable $Y$ mean $\mu$ and variance $\sigma^2$. Show that the sample variance $s^2$ is an unbiased estimator of the population variance $\sigma^2$ when:

a. The sampled population has a normal distribution.
b. The distribution of the sampled population is unknown.

Solution

a. From Theorem 6.11, we know that when sampling from a normal distribution,

$$\frac{(n-1)s^2}{\sigma^2} = \chi^2$$

where $\chi^2$ is a chi-square random variable with $\nu = (n-1)$ degrees of freedom. Rearranging terms yields

$$s^2 = \frac{\sigma^2}{(n-1)}\chi^2$$

from which it follows that

$$E(s^2) = E\left[\frac{\sigma^2}{(n-1)}\chi^2\right]$$

Applying Theorem 5.2, we obtain

$$E(s^2) = \frac{\sigma^2}{(n-1)}E(\chi^2)$$

We know from Section 5.7 that $E(\chi^2) = \nu$; thus

$$E(s^2) = \frac{\sigma^2}{(n-1)}\nu = \frac{\sigma^2}{(n-1)}(n-1) = \sigma^2$$

Therefore, by Definition 7.3, we conclude that $s^2$ is an unbiased estimator of $\sigma^2$.

b. By the definition of sample variance, we have

$$s^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{n} y_i^2 - \frac{(\sum y_i)^2}{n}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right]$$

From Theorem 4.4, $\sigma^2 = E(Y^2) - \mu^2$. Consequently, $E(Y^2) = \sigma^2 + \mu^2$ for a random variable $Y$. Since each $Y$ value, $y_1, y_2, \ldots, y_n$, was randomly selected from a population with mean $\mu$ and variance $\sigma^2$, it follows that

$$E(y_i^2) = \sigma^2 + \mu^2 (i = 1, 2, \ldots, n)$$

and

$$E(\bar{y}_i^2) = \sigma_{\bar{y}}^2 + (\mu_{\bar{y}})^2 = \sigma^2/n + \mu^2$$

Taking the expected value of $s^2$ and substituting these expressions, we obtain

$$\begin{aligned} E(s^2) &= E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right]\right\} \\ &= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n} y_i^2\right] - E[n(\bar{y})^2]\right\} \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[y_i^2] - nE[(\bar{y})^2]\right\} \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} \\ &= \frac{1}{n-1}[(n\sigma^2 + n\mu^2) - \sigma^2 - n\mu^2] \\ &= \frac{1}{n-1}[n\sigma^2 - \sigma^2] \\ &= \left(\frac{n-1}{n-1}\right)\sigma^2 = \sigma^2 \end{aligned}$$

This shows that, regardless of the nature of the sampled population, $s^2$ is an unbiased estimator of $\sigma^2$.

## Theoretical Exercises

7.1 Let $y_1, y_2, y_3$ be a random sample from an exponential distribution with mean $\theta$, i.e., $E(y_i) = \theta$, $i = 1, 2, 3$. Consider three estimators of $\theta$:

$$\hat{\theta}_1 = \bar{y} \qquad \hat{\theta}_2 = y_1 \qquad \hat{\theta}_3 = \frac{y_1 + y_2}{2}$$

a. Show that all three estimators are unbiased.

b. Which of the estimators has the smallest variance? [*Hint:* Recall that, for an exponential distribution, $V(y_i) = \theta^2$.]

7.2 Let $y_1, y_2, y_3, \ldots, y_n$ be a random sample from a Poisson distribution with mean $\lambda$, i.e., $E(y_i) = \lambda$, $i = 1, 2, \ldots, n$. Consider four estimators of $\lambda$:

$$\hat{\lambda}_1 = \bar{y} \qquad \hat{\lambda}_2 = n(y_1 + y_2 + \cdots + y_n)$$

$$\hat{\lambda}_3 = \frac{y_1 + y_2}{2} \qquad \hat{\lambda}_4 = \frac{y_1}{n}$$

a. Which of the four estimators are unbiased?

b. Of the unbiased estimators, which has the smallest variance? [*Hint:* Recall that, for a Poisson distribution, $V(y_i) = \lambda$.]

7.3 Suppose the random variable $Y$ has a binomial distribution with parameters $n$ and $p$.

a. Show that $\hat{p} = y/n$ is an unbiased estimator of $p$.

b. Find the variance of $\hat{p}$.

7.4 Let $y_1, y_2, \ldots, y_n$ be a random sample from a gamma distribution with parameters $\alpha = 2$ and $\beta$ unknown.

a. Show that $\bar{y}$ is a biased estimator of $\beta$. Compute the bias.

b. Show that $\hat{\beta} = \bar{y}/2$ is an unbiased estimator of $\beta$.

c. Find the variance of $\hat{\beta} = \bar{y}/2$. [*Hint:* Recall that, for a gamma distribution, $E(y_i) = 2\beta$ and $V(y_i) = 2\beta^2$.]

7.5 Show that $E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + b^2(\theta)$ where the bias $b(\theta) = E(\hat{\theta}) - \theta$. [*Hint:* Write $(\hat{\theta} - \theta) = [\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]$.]

7.6 Let $y_1$ be a sample of size 1 from a uniform distribution over the interval from 2 to $\theta$.

a. Show that $y_1$ is a biased estimator of $\theta$ and compute the bias.

b. Show that $2(y_1 - 1)$ is an unbiased estimator of $\theta$.

c. Find the variance of $2(y_1 - 1)$.

7.7 Let $y_1, y_2, \ldots, y_n$ be a random sample from a normal distribution, with mean $\mu$ and variance $\sigma^2$. Show that the variance of the sampling distribution of $s^2$ is $2\sigma^4/(n - 1)$.

## 7.2 Finding Point Estimators: Classical Methods of Estimation

There are a number of different methods for finding point estimators of parameters. Two classical methods, the **method of moments** and the **method of maximum likelihood**, are the main topics of this section. These techniques produce the estimators of the population parameters encountered in Sections 7.4–7.10. A brief discussion of other methods for finding point estimators is given at the end of this section. Two of these alternative methods are the topic of optional Section 7.12.

### Method of Moments

The method of estimation that we have employed thus far is to use sample numerical descriptive measures to estimate their population parameters. For example, we used the sample mean $\bar{y}$ to estimate the population mean $\mu$. From Definition 4.7, we know that the parameter $E(Y) = \mu$ is the first moment about the origin or, as it is sometimes called, the **first population moment**. Similarly, we define the **first sample moment** as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The general technique of using sample moments to estimate their corresponding population moments is called the **method of moments**. For the parameters discussed in this chapter, the method of moments yields estimators that have the two desired properties mentioned earlier, i.e., unbiased estimators and estimators with minimum variance.

**Definition 7.6**

Let $y_1, y_2, \ldots, y_n$ represent a random sample of $n$ observed values on a random variable $Y$ with some probability distribution (discrete or continuous). The ***k*th population moment** and ***k*th sample moment** are defined as follows:

$$k\text{th population moment: } E(Y^k)$$

$$k\text{th sample moment: } m_k = \frac{\sum_{i=1}^{n} y_i^k}{n}$$

For the case $k = 1$, the first population moment is $E(Y) = \mu$ and the first sample moment is $m_1 = \bar{y}$.

**Definition 7.7**

Let $y_1, y_2, \ldots, y_n$ represent a random sample of $n$ observations on a random variable $Y$ with a probability distribution (discrete or continuous) with parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the **moment estimators, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$,** are obtained by equating the first $m$ sample moments to the corresponding first $m$ population moments:

$$E(Y) = \frac{1}{n}\sum y_i$$

$$E(Y^2) = \frac{1}{n}\sum y_i^2$$

$$\cdots$$

$$E(Y^k) = \frac{1}{n}\sum y_i^k$$

and solving for $\theta_1, \theta_2, \ldots, \theta_k$. (Note that the first $m$ population moments will be functions of $\theta_1, \theta_2, \ldots, \theta_k$.)

*Note*: For the special case $m = 1$, the moment estimator of $\theta$ is some function of the sample mean $\bar{y}$.

**Example 7.2**

**Point Estimate of a Mean: Auditory Nerve Response Rate**

Research in the *Journal of the Acoustic Society of America* found that the response rate $Y$ of auditory nerve fibers in cats has an approximate Poisson distribution with unknown mean $\lambda$. Suppose the auditory nerve fiber response rate (recorded as number of spikes per 200 milliseconds of noise burst) was measured in each of a random sample of 10 cats. The data follow:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15.1 | 14.6 | 12.0 | 19.2 | 16.1 | 15.5 | 11.3 | 18.7 | 17.1 | 17.2 |

Calculate a point estimate for the mean response rate $\lambda$ using the method of moments.

Solution

We have only one parameter, $\lambda$, to estimate; therefore, the moment estimator is found by setting the first population moment, $E(Y)$, equal to the first sample moment, $\bar{y}$. For the Poisson distribution, $E(Y) = \lambda$; hence, the moment estimator is

$$\hat{\lambda} = \bar{y}$$

For this example,

$$\bar{y} = \frac{15.1 + 14.6 + \cdots + 17.2}{10} = 15.68$$

Thus, our estimate of the mean auditory nerve fiber response rate $\lambda$ is 15.68 spikes per 200 milliseconds of noise burst.

**Example 7.3** *(optional)*

**Moment Estimators of a Gamma Distribution**

Research in *IEEE Transactions on Energy Conversion* found that the time $Y$ until failure from fatigue cracks for underground cable possesses an approximate gamma probability distribution with parameters $\alpha$ and $\beta$. Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations on the random variable $Y$. Find the moment estimators of $\alpha$ and $\beta$.

Solution

Since we must estimate two parameters, $\alpha$ and $\beta$, the method of moments requires that we set the first two population moments equal to their corresponding sample moments. From Section 5.6, we know that for the gamma distribution

$$\mu = E(Y) = \alpha\beta$$
$$\sigma^2 = \alpha\beta^2$$

Also, from Theorem 4.4, $\sigma^2 = E(Y^2) - \mu^2$. Thus, $E(Y^2) = \sigma^2 + \mu^2$. Then for the gamma distribution, the first two population moments are

$$E(Y) = \alpha\beta$$
$$E(Y^2) = \sigma^2 + \mu^2 = \alpha\beta^2 + (\alpha\beta)^2$$

Setting these equal to their respective sample moments, we have

$$\hat{\alpha}\hat{\beta} = \bar{y}$$
$$\hat{\alpha}\hat{\beta}^2 + (\hat{\alpha}\hat{\beta})^2 = \frac{\sum y_i^2}{n}$$

Substituting $\bar{y}$ for $\hat{\alpha}\hat{\beta}$ in the second equation, we obtain

$$\bar{y}\hat{\beta} + (\bar{y})^2 = \frac{\sum y_i^2}{n}$$

or,

$$\bar{y}\hat{\beta} = \frac{\sum y_i^2}{n} - (\bar{y})^2$$

$$= \frac{\sum y_i^2 - n(\bar{y})^2}{n} = \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n}$$

$$= \frac{(n-1)s^2}{n}$$

Our two equations are now reduced to

$$\hat{\alpha}\hat{\beta} = \bar{y}$$

$$\bar{y}\hat{\beta} = \left(\frac{n-1}{n}\right)s^2$$

Solving these equations simultaneously, we obtain the moment estimators

$$\hat{\beta} = \left(\frac{n-1}{n}\right)\frac{s^2}{\bar{y}} \quad \text{and} \quad \hat{\alpha} = \left(\frac{n}{n-1}\right)\frac{\bar{y}^2}{s^2} = \left(\frac{n}{n-1}\right)\left(\frac{\bar{y}}{s}\right)^2$$

**Method of Maximum Likelihood** The method of maximum likelihood and an exposition of the properties of maximum likelihood estimators are the results of work by Sir Ronald A. Fisher (1890–1962). Fisher's logic can be seen by considering the following example: If we randomly select a sample of $n$ independent observations $y_1$, $y_2, \ldots, y_n$, of a discrete random variable $Y$ and if the probability distribution $p(y)$ is a function of a single parameter $\theta$, then the probability of observing these $n$ independent values of $Y$ is

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2)\cdots p(y_n)$$

Fisher called this joint probability of the sample values, $y_1, y_2, \ldots, y_n$, the **likelihood function** $L$ of the sample, and suggested that one should choose as an estimate of $\theta$ the value of $\theta$ that maximizes $L$. If the likelihood $L$ of the sample is a function of two parameters, say, $\theta_1$ and $\theta_2$, then the maximum likelihood estimates of $\theta_1$ and $\theta_2$ are the values that maximize $L$. The notion is easily extended to the situation in which $L$ is a function of more than two parameters.

**Definition 7.8**

a. The **likelihood function** $L$ of a sample of $n$ observations $y_1, y_2, \ldots, y_n$, is the joint probability function $p(y_1, y_2, \ldots, y_n)$ when $Y_1, Y_2, \ldots, Y_n$ are discrete random variables.
b. The **likelihood function** $L$ of a sample of $n$ observations, $y_1, y_2, \ldots, y_n$, is the joint density function $f(y_1, y_2, \ldots, y_n)$ when $Y_1, Y_2, \ldots, Y_n$ are continuous random variables.

*Note:* For fixed values of $y_1, y_2, \ldots, y_n$, $L$ will be a function of $\theta$.

Theorem 7.1 follows directly from the definition of independence and Definitions 6.8 and 6.9.

**THEOREM 7.1**

a. Let $y_1, y_2, \ldots, y_n$ represent a random sample of $n$ independent observations on a random variable $Y$. Then $L = p(y_1)p(y_2)\cdots p(y_n)$ when $Y$ is a discrete random variable with probability distribution $p(y)$.

b. Let $y_1, y_2, \ldots, y_n$ represent a random sample of $n$ independent observations on a random variable $Y$. Then $L = f(y_1)f(y_2)\cdots f(y_n)$ when $Y$ is a continuous random variable with density function $f(y)$.

**Definition 7.9**

Let $L$ be the likelihood of a sample, where $L$ is a function of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the **maximum likelihood estimators** of $\theta_1, \theta_2, \ldots, \theta_k$ are the values of $\theta_1, \theta_2, \ldots, \theta_k$ that maximize $L$.

Fisher showed that maximum likelihood estimators of population means and proportions possess some very desirable properties. As the sample size $n$ becomes larger and larger, the sampling distribution of a maximum likelihood estimator $\hat{\theta}$ tends to become more and more nearly normal, with mean equal to $\theta$ and a variance that is equal to or less than the variance of *any other* estimator. Although these properties of maximum likelihood estimators pertain only to estimates based on large samples, they tend to provide support for the maximum likelihood method of estimation. The properties of maximum likelihood estimators based on small samples can be acquired by using the methods of Chapters 4, 5, and 6 to derive their sampling distributions or, at the very least, to acquire their means and variances.

To simplify our explanation of how to find a maximum likelihood estimator, we will assume that $L$ is a function of a single parameter $\theta$. Then, from differential calculus, we know that the value of $\theta$ that maximizes (or minimizes) $L$ is the value for which $\frac{dL}{d\theta} = 0$. Obtaining this solution, which always yields a maximum (proof omitted), can be difficult because $L$ is usually the product of a number of quantities involving $\theta$. Differentiating a sum is easier than differentiating a product, so we attempt to maximize the logarithm of $L$ rather than $L$ itself. Since the logarithm of $L$ is a monotonically increasing function of $L$, $L$ will be maximized by the same value of $\theta$ that maximizes its logarithm. We illustrate the procedure in Examples 7.4 and 7.5.

**Example 7.4**

Finding a Maximum Likelihood Estimator

Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations on a random variable $Y$ with the exponential density function

$$f(y) = \begin{cases} \dfrac{e^{-y/\beta}}{\beta} & \text{if } 0 \le y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Determine the maximum likelihood estimator of $\beta$.

Solution

Since $y_1, y_2, \ldots, y_n$ are independent random variables, we have

$$\begin{aligned} L &= f(y_1)f(y_2)\cdots f(y_n) \\ &= \left(\frac{e^{-y_1/\beta}}{\beta}\right)\left(\frac{e^{-y_2/\beta}}{\beta}\right)\cdots\left(\frac{e^{-y_n/\beta}}{\beta}\right) \\ &= \frac{e^{-\Sigma_{i=1}^{n} y_i/\beta}}{\beta^n} \end{aligned}$$

Taking the natural logarithm of $L$ yields

$$\ln(L) = \ln(e^{-\Sigma_{i=1}^{n} y_i/\beta}) - n\ln(\beta) = -\frac{\sum_{i=1}^{n} y_i}{\beta} - n\ln(\beta)$$

Then,

$$\frac{d\ln(L)}{d\beta} = \frac{\sum_{i=1}^{n} y_i}{\beta^2} - \frac{n}{\beta}$$

Setting this derivative equal to 0 and solving for $\hat{\beta}$, we obtain

$$\frac{\sum_{i=1}^{n} y_i}{\hat{\beta}^2} - \frac{n}{\hat{\beta}} = 0 \quad \text{or} \quad n\hat{\beta} = \sum_{i=1}^{n} y_i$$

This yields

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Therefore, the maximum likelihood estimator (MLE) of $\beta$ is the sample mean $\bar{y}$, i.e., $\hat{\beta} = \bar{y}$.

**Example 7.5** *(optional)*

Maximum Likelihood Estimators of $\mu$ and $\sigma^2$.

Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations on the random variable $Y$, where $f(y)$ is a normal density function with mean $\mu$ and variance $\sigma^2$. Find the maximum likelihood estimators of $\mu$ and $\sigma^2$.

Solution

Since $y_1, y_2, \ldots, y_n$ are independent random variables, it follows that

$$\begin{aligned} L &= f(y_1)f(y_2)\cdots f(y_n) \\ &= \left(\frac{e^{-(y_1-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\right)\left(\frac{e^{-(y_2-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\right)\cdots\left(\frac{e^{-(y_n-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\right) \\ &= \frac{e^{-\sum_{i=1}^{n}(y_i-\mu)^2/(2\sigma^2)}}{\sigma^n(2\pi)^{n/2}} \end{aligned}$$

and

$$\ln(L) = -\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi)$$

Taking partial derivatives of $\ln(L)$ with respect to $\mu$ and $\sigma$ and setting them equal to 0 yields

$$\frac{\partial \ln(L)}{\partial \mu} = \frac{\sum_{i=1}^{n} 2(y_i - \hat{\mu})}{2\hat{\sigma}^2} - 0 - 0 = 0$$

and

$$\frac{\partial \ln(L)}{\partial \sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{2\hat{\sigma}^4} - \frac{n}{2}\left(\frac{1}{\hat{\sigma}^2}\right) - 0 = 0$$

The values of $\mu$ and $\sigma^2$ that maximize $L$ [and hence $\ln(L)$] will be the simultaneous solution of these two equations. The first equation reduces to

$$\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \quad \text{or} \quad \sum_{i=1}^{n} y_i - n\hat{\mu} = 0$$

and it follows that

$$n\hat{\mu} = \sum_{i=1}^{n} y_i \quad \text{and} \quad \hat{\mu} = \bar{y}$$

Substituting $\hat{\mu} = \bar{y}$ into the second equation and multiplying by $2\hat{\sigma}^2$, we obtain

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\hat{\sigma}^2} = n \quad \text{or} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Therefore, the maximum likelihood estimators of $\mu$ and $\sigma^2$ are

$$\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Note: The maximum likelihood estimator of $\sigma^2$ is equal to the sum of squares of deviations $\sum_{i=1}^{n}(y_i - \bar{y})^2$ divided by $n$, whereas the sample variance $s^2$ uses a divisor of $(n - 1)$. We showed in Example 7.1 that $s^2$ is an unbiased estimator of $\sigma^2$. Therefore, the maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{(n - 1)}{n}s^2$$

is a biased estimator of $\sigma^2$. However,

$$\text{Var}(\sigma^2) = \text{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}(S^2) < \text{Var}(S^2)$$

Thus, although unbiased, the maximum likelihood estimator of $\sigma^2$ has a smaller variance than the sample variance, $s^2$.

**Method of Least Squares** Another useful technique for finding point estimators is the **method of least squares.** This method finds the estimate of $\theta$ that minimizes the **mean squared error (MSE)**:

$$\text{MSE} = E[(\hat{\theta} - \theta)^2]$$

The method of least squares—a widely used estimation technique—is discussed in detail in Chapter 10. Several other estimation methods are briefly described here; consult the references at the end of this chapter if you want to learn more about their use.

**Jackknife Estimators** Tukey (1958) developed a "leave-one-out-at-a-time" approach to estimation, called the **jackknife**,* that is gaining increasing acceptance among practitioners. Let $y_1, y_2, \ldots, y_n$ be a sample of size $n$ from a population with parameter $\theta$. An estimate $\hat{\theta}_{(i)}$ is obtained by omitting the $i$th observation (i.e., $y_i$) and computing the estimate based on the remaining $(n - 1)$ observations. This calculation is performed for each observation in the data set, and the procedure results in $n$ estimates of $\theta$: $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(n)}$. The **jackknife estimator** of $\theta$ is then some suitably chosen

*The procedure derives its name from the Boy Scout jackknife; like the jackknife, the procedure serves as a handy tool in a variety of situations when specialized techniques may not be available.

linear combination (e.g., a weighted average) of the $n$ estimates. Application of the jackknife is suggested for situations where we are likely to have outliers or biased samples, or find it difficult to assess the variability of the more traditional estimators.

**Robust Estimators** Many of the estimators discussed in Sections 7.4–7.10 are based on the assumption that the sampled population is approximately normal. When the distribution of the sampled population deviates greatly from normality, such estimators do not have desirable properties (e.g., unbiasedness and minimum variance). An estimator that performs well for a very wide range of probability distributions is called a **robust estimator**. For example, a robust estimate of the population mean $\mu$, called the ***M*-estimator**, compares favorably to the sample mean $\bar{y}$ when the sampled population is normal and is considerably better than $\bar{y}$ when the population is heavy-tailed or skewed. A type of robust estimator derived from "bootstrapping" is discussed in optional Section 7.12. See Mosteller and Tukey (1977) and Devore (1987) for a good practical discussion of robust estimation techniques.

**Bayes Estimators** The classical approach to estimation is based on the concept that the unknown parameter $\theta$ is a constant. All the information available to us about $\theta$ is contained in the random sample $y_1, y_2, \ldots, y_n$ selected from the relevant population. In contrast, the **Bayesian** approach to estimation regards $\theta$ as a random variable with some known (**prior**) probability distribution $g(\theta)$. The sample information is used to modify the prior distribution on $\theta$ to obtain the **posterior** distribution, $f(\theta \mid y_1, y_2, \ldots, y_n)$. The **Bayes estimator** of $\theta$ is then the mean of the posterior probability distribution [see Wackerly, Mendenhall, and Scheaffer (2008)]. A brief discussion of Bayes estimators is given in optional Section 7.12.

## Theoretical Exercises

7.8 A binomial experiment consisting of $n$ trials resulted in Bernoulli observations $y_1, y_2, \ldots, y_n$, where

$$y_i = \begin{cases} 1 & \text{if the } i\text{th trial was a success} \\ 0 & \text{if not} \end{cases}$$

and $P(y_i = 1) = p$, $P(y_i = 0) = 1 - p$. Let $Y = \sum_{i=1}^{n} y_i$ be the number of successes in $n$ trials.

a. Find the moment estimator of $p$.
b. Is the moment estimator unbiased?
c. Find the maximum likelihood estimator of $p$. [*Hint:* $L = p^y(1 - p)^{n-y}$.]
d. Is the maximum likelihood estimator unbiased?

7.9 Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations from a Poisson distribution with probability function

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad (y = 0, 1, 2, \ldots)$$

a. Find the maximum likelihood estimator of $\lambda$.
b. Is the maximum likelihood estimator unbiased?

7.10 Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations on a random variable $Y$, where $f(y)$ is a gamma density function with $\alpha = 2$ and unknown $\beta$:

$$f(y) = \begin{cases} \dfrac{ye^{-y/\beta}}{\beta^2} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the maximum likelihood estimator of $\beta$.
b. Find $E(\hat{\beta})$ and $V(\hat{\beta})$.

7.11 Refer to Exercise 7.10.

a. Find the moment estimator of $\beta$.
b. Find $E(\hat{\beta})$ and $V(\hat{\beta})$.

7.12 Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations from a normal distribution with mean 0 and unknown variance $\sigma^2$. Find the maximum likelihood estimator of $\sigma^2$.

7.13 Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations from an exponential distribution with density

$$f(y) = \begin{cases} \dfrac{1}{\beta}e^{-y/\beta} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the moment estimator of $\beta$.
b. Is the moment estimator unbiased?
c. Find $V(\hat{\beta})$.

## 7.3 Finding Interval Estimators: The Pivotal Method

In Section 7.1, we defined an interval estimator as a rule that tells how to use the sample observations to calculate two numbers that define an interval that will enclose the estimated parameter with a high degree of confidence. The resulting random interval (random, because the sample observations used to calculate the endpoints of the interval are random variables) is called a **confidence interval**, and the probability (prior to sampling) that it contains the estimated parameter is called its **confidence coefficient**. If a confidence interval has a confidence coefficient equal to .95, we call it a 95% confidence interval. If the confidence coefficient is .99, the interval is said to be a 99% confidence interval, etc. A more practical interpretation of the confidence coefficient for a confidence interval is given later in this section.

**Definition 7.10**

The **confidence coefficient** for a confidence interval is equal to the probability that the random interval, prior to sampling, will contain the estimated parameter.

One way to find a confidence interval for a parameter $\theta$ is to acquire a **pivotal statistic**, a statistic that is a function of the sample values and the single parameter $\theta$. Because many statistics are approximately normally distributed when the sample size $n$ is large (central limit theorem), we can construct confidence intervals for their expected values using the standard normal random variable $Z$ as a pivotal statistic.

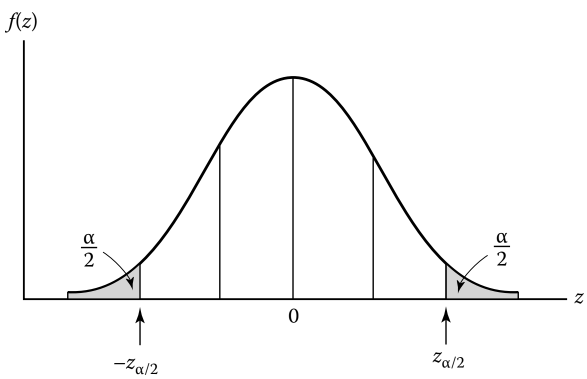
To illustrate, let $\hat{\theta}$ be a statistic with a sampling distribution that is approximately normally distributed for large samples with mean $E(\hat{\theta}) = \theta$ and standard error $\sigma_{\hat{\theta}}$. Then,

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is a standard normal random variable. Since $Z$ is also a function of only the sample statistic $\hat{\theta}$ and the parameter $\theta$, we will use it as a pivotal statistic. To derive a confidence interval for $\theta$, we first make a probability statement about the pivotal statistic. To do this, we locate values $z_{\alpha/2}$ and $-z_{\alpha/2}$ that place a probability of $\alpha/2$ in each tail of the $Z$ distribution (see Figure 7.5), i.e., $P(z > z_{\alpha/2}) = \alpha/2$. It can be seen from Figure 7.5 that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

**FIGURE 7.5**
Locating $z_{\alpha/2}$ for a confidence interval



Substituting the expression for $z$ into the probability statement and using some simple algebraic operations on the inequality, we obtain

$$\begin{aligned} P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) &= P\left(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right) \\ &= P(-z_{\alpha/2}\sigma_{\hat{\theta}} \le \hat{\theta} - \theta \le z_{\alpha/2}\sigma_{\hat{\theta}}) \\ &= P(-\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \le -\theta \le -\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) \\ &= P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha \end{aligned}$$

Therefore, the probability that the interval formed by

$$\text{LCL} = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \quad \text{to} \quad \text{UCL} = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$$

will enclose $\theta$ is equal to $(1 - \alpha)$. The quantities LCL and UCL are called the **lower** and **upper confidence limits**, respectively, for the confidence interval. The confidence coefficient for the interval will be $(1 - \alpha)$.

The derivation of a large-sample $(1 - \alpha)100\%$ confidence interval for $\theta$ is summarized in Theorem 7.2.

**THEOREM 7.2**

Let $\hat{\theta}$ be normally distributed for large samples with $E(\hat{\theta}) = \theta$ and standard error $\sigma_{\hat{\theta}}$. Then a $(1 - \alpha)100\%$ confidence interval for $\theta$ is

$$\hat{\theta} \pm (z_{\alpha/2})\sigma_{\hat{\theta}}$$

The large-sample confidence interval can also be acquired intuitively by examining Figure 7.6. The $z$ value corresponding to an area $A = .475$—i.e., the $z$ value that places area $\alpha/2 = .025$ in the upper tail of the $Z$ distribution—is (see Table 5 of Appendix B) $z_{.025} = 1.96$. Therefore, the probability that $\hat{\theta}$ will lie within $1.96\sigma_{\hat{\theta}}$ of $\theta$ is .95. You can see from Figure 7.6 that whenever $\hat{\theta}$ falls within the interval $\theta \pm 1.96\sigma_{\hat{\theta}}$, then the interval $\hat{\theta} \pm 1.96\sigma_{\hat{\theta}}$ will enclose $\theta$. Therefore, $\hat{\theta} \pm 1.96\sigma_{\hat{\theta}}$ yields a 95% confidence interval for $\theta$.

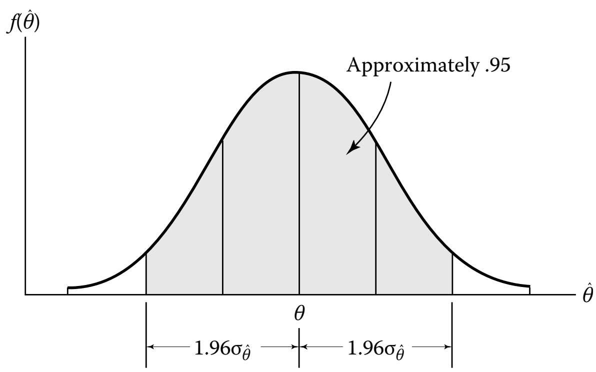



**FIGURE 7.6**
The sampling distribution of $\hat{\theta}$ for large samples

We may encounter one slight difficulty when we attempt to apply this confidence interval in practice. It is often the case that $\sigma_{\hat{\theta}}$ is a function of the parameter $\theta$ that we are attempting to estimate. However, when the sample size $n$ is large (which we have assumed throughout the derivation), we can substitute the estimate $\hat{\theta}$ for the parameter $\theta$ to obtain an approximate value for $\sigma_{\hat{\theta}}$.

In Example 7.6 we will use a pivotal statistic to find a confidence interval for $\mu$ when the sample size is small, say, $n < 30$.

**Example 7.6**

Finding a 95% confidence Interval for $\mu$: Pivotal Method

Let $\bar{y}$ and $s$ be the sample mean and standard deviation based on a random sample of $n$ observations ($n < 30$) from a normal distribution with mean $\mu$ and variance $\sigma^2$

a. Derive an expression for a $(1 - \alpha) \times 100\%$ confidence interval for $\mu$.

b. Find a 95% confidence interval for $\mu$ if $\bar{y} = 9.1$, $s = 1.1$, and $n = 10$.

Solution

a. A pivotal statistic for $\mu$ can be constructed using the $T$ statistic of Chapter 6. By Definition 6.16,

$$T = \frac{Z}{\sqrt{\chi^2/\nu}}$$

where $Z$ and $\chi^2$ are independent random variables and $\chi^2$ is based on $\nu$ degrees of freedom. We know that $\bar{y}$ is normally distributed and that

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable. From Theorem 6.11, it follows that

$$\frac{(n - 1)s^2}{\sigma^2} = \chi^2$$

is a chi-square random variable with $\nu = (n - 1)$ degrees of freedom. We state (without proof) that $\bar{y}$ and $s^2$ are independent when they are based on a random sample selected from a normal distribution. Therefore, $z$ and $\chi^2$ will be independent random variables. Substituting the expressions for $z$ and $\chi^2$ into the formula for $T$, we obtain

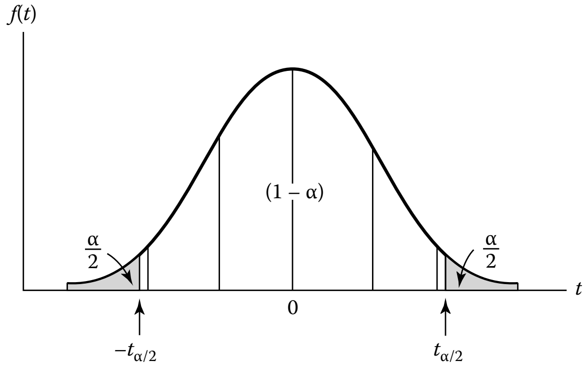
$$T = \frac{Z}{\sqrt{\chi^2/\nu}} = \frac{\dfrac{\bar{y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n - 1)s^2}{\sigma^2}\Big/(n - 1)}} = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

Note that the pivotal statistic is a function only of $\mu$ and the sample statistics $\bar{y}$ and $s^2$.

The next step in finding a confidence interval for $\mu$ is to make a probability statement about the pivotal statistic $T$. We will select two values of $T$—call them $t_{\alpha/2}$ and $-t_{\alpha/2}$—that correspond to probabilities of $\alpha/2$ in the upper and lower tails, respectively, of the $T$ distribution (see Figure 7.7). From Figure 7.7, it can be seen that

$$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) = 1 - \alpha$$

**FIGURE 7.7**
The location of $t_{\alpha/2}$ and $-t_{\alpha/2}$ for a Student's $T$ distribution




Substituting the expression for $t$ into the probability statement, we obtain

$$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) = P\left(-t_{\alpha/2} \le \frac{\bar{y} - \mu}{s/\sqrt{n}} \le t_{\alpha/2}\right) = 1 - \alpha$$

Multiplying the inequality within the brackets by $s/\sqrt{n}$, we obtain

$$P\left[-t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \bar{y} - \mu \le t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Subtracting $\bar{y}$ from each part of the inequality yields

$$P\left[-\bar{y} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le -\mu \le -\bar{y} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Finally, we multiply each term of the inequality by $-1$, thereby reversing the inequality signs. The result is

$$P\left[\bar{y} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{y} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Therefore, a $(1 - \alpha)100\%$ confidence interval for $\mu$ when $n$ is small is

$$\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

b. For a 95% confidence interval, $\alpha = .05$ and $\alpha/2 = .025$. Consequently, we need to find $t_{.025}$. A portion of Table 7, Appendix B, is shown in Table 7.1. When $\nu = n - 1 = 9$ degrees of freedom, $t_{.025} = 2.262$ (shaded on Table 7.1).

Substituting $\bar{y} = 9.7$, $s = 1.1$, $n = 10$, and $t_{.025} = 2.262$ into the confidence interval formula, part **a**, we obtain

$$9.7 \pm 2.262\left(\frac{1.1}{\sqrt{10}}\right) = 9.7 \pm .79 = (8.91, 10.49).$$

Since the confidence coefficient is $1 - \alpha = .95$, we say that we are 95% confident that the interval 8.91 to 10.49 contains the true mean $\mu$.

**TABLE 7.1 An Abbreviated Version of Table 7 of Appendix B**

| Degrees of Freedom | $t_{.100}$ | $t_{.050}$ | $t_{.025}$ | $t_{.010}$ | $t_{.005}$ |
|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |

### Practical Interpretation of a Confidence Interval

If a $(1 - \alpha)100\%$ confidence interval for a parameter $\theta$ is (LCL, UCL), then we are $(1 - \alpha)100\%$ "confident" that $\theta$ falls between LCL and UCL.

The phrase "95% confident" in the solution to Example 7.6b has a very special meaning in interval estimation. To illustrate this, we used Monte Carlo simulation to draw 100 samples of size $n = 10$ from a normal distribution with known mean $\mu = 10$ and variance $\sigma^2 = 1$. A 95% confidence interval for $\mu$ was computed for each of the 100 samples. These are shown in the EXCEL workbook, Figure 7.8. Only the 5 intervals that are highlighted fail to enclose the mean $\mu = 10$. The proportion that encloses $\mu$, .95, is exactly equal to the confidence coefficient. This explains why we are highly confident that the interval calculated in Example 7.6b (8.91, 10.49), encloses the true value of $\mu$. *If we were to employ our interval estimator on repeated occasions*, 95% *of the intervals constructed would contain* $\mu$.

### Theoretical Interpretation of the Confidence Coefficient $(1 - \alpha)$

If we were to repeatedly collect a sample of size $n$ from the population and construct a $(1 - \alpha)100\%$ confidence interval for each sample, then we expect $(1 - \alpha)100\%$ of the intervals to enclose the true parameter value.

Confidence intervals for population parameters other than the population mean can be derived using the pivotal method outlined in this section. The estimators and pivotal statistics for many of these parameters are well known. In Sections 7.4–7.10, we give the confidence interval formulas for several population parameters that are commonly encountered in practice and illustrate each with a practical example.

| | A | B | C |
|---|---|---|---|
| 1 | SAMPLE | LOWER95 | UPPER95 |
| 2 | 1 | 9.35 | 10.51 |
| 3 | 2 | 9.65 | 11.46 |
| 4 | 3 | 9.32 | 11.13 |
| 5 | 4 | 9.04 | 10.17 |
| 6 | 5 | 9.56 | 10.88 |
| 7 | 6 | 9.42 | 10.74 |
| 8 | 7 | 9.5 | 10.33 |
| 9 | 8 | 9.92 | 10.89 |
| 10 | 9 | 9.64 | 10.55 |
| 11 | 10 | 9.51 | 10.95 |
| 12 | 11 | 9.5 | 10.77 |
| 13 | 12 | 9.74 | 10.98 |
| 14 | 13 | 9.71 | 10.46 |
| 15 | 14 | 9.19 | 10.98 |
| 16 | 15 | 8.92 | 10.67 |
| 17 | 16 | 9.35 | 10.72 |
| 18 | 17 | 9.15 | 10.51 |
| 19 | 18 | 9.62 | 10.42 |
| 20 | 19 | 9.2 | 10.34 |
| 21 | 20 | 9.02 | 10.67 |
| 22 | 21 | 9.89 | 10.87 |
| 23 | 22 | 8.67 | 10.16 |
| 24 | 23 | 9.85 | 11.42 |
| 25 | 24 | 9.62 | 10.95 |
| 26 | 25 | 9.34 | 10.91 |
| 27 | 26 | 9.15 | 10.69 |
| 28 | 27 | 8.98 | 10.6 |
| 29 | 28 | 10.02 | 11.32 |
| 30 | 29 | 8.72 | 10.69 |
| 31 | 30 | 9.03 | 10.56 |
| 32 | 31 | 9.24 | 10.67 |
| 33 | 32 | 9.28 | 10.66 |
| 34 | 33 | 9.29 | 10.7 |
| 35 | 34 | 9.61 | 10.8 |
| 36 | 35 | 9.48 | 10.4 |
| 37 | 36 | 8.85 | 11.27 |
| 38 | 37 | 9.15 | 10.35 |
| 39 | 38 | 9.49 | 10.92 |
| 40 | 39 | 9.38 | 10.61 |
| 41 | 40 | 9.28 | 10.44 |
| 42 | 41 | 9.45 | 10.62 |
| 43 | 42 | 9.17 | 11.01 |
| 44 | 43 | 9.91 | 10.74 |
| 45 | 44 | 9.15 | 10.65 |
| 46 | 45 | 9.25 | 10.52 |
| 47 | 46 | 9.14 | 10.41 |
| 48 | 47 | 9.15 | 10.95 |
| 49 | 48 | 9.63 | 10.78 |
| 50 | 49 | 8.85 | 10.55 |
| 51 | 50 | 9.28 | 10.79 |
| 52 | 51 | 8.81 | 10.2 |
| 53 | 52 | 8.9 | 10.56 |
| 54 | 53 | 8.74 | 10.85 |
| 55 | 54 | 9.39 | 10.65 |
| 56 | 55 | 9.2 | 11.31 |
| 57 | 56 | 9.06 | 10.81 |
| 58 | 57 | 8.96 | 10.52 |
| 59 | 58 | 10 | 11.1 |
| 60 | 59 | 9.39 | 11.37 |
| 61 | 60 | 8.83 | 10.71 |
| 62 | 61 | 8.89 | 10.19 |
| 63 | 62 | 9.02 | 10.32 |
| 64 | 63 | 9.38 | 10.47 |
| 65 | 64 | 9.2 | 10.46 |
| 66 | 65 | 10.12 | 11.09 |
| 67 | 66 | 9.84 | 11.49 |
| 68 | 67 | 9.43 | 10.62 |
| 69 | 68 | 9.34 | 10.5 |
| 70 | 69 | 8.44 | 9.86 |
| 71 | 70 | 9.58 | 10.45 |
| 72 | 71 | 9.13 | 10.11 |
| 73 | 72 | 9.22 | 10.57 |
| 74 | 73 | 9.49 | 10.41 |
| 75 | 74 | 9.48 | 10.77 |
| 76 | 75 | 9.67 | 10.93 |
| 77 | 76 | 9.19 | 10.79 |
| 78 | 77 | 9.1 | 10.05 |
| 79 | 78 | 10.19 | 10.96 |
| 80 | 79 | 8.76 | 10.54 |
| 81 | 80 | 9.47 | 10.67 |
| 82 | 81 | 8.84 | 10.32 |
| 83 | 82 | 8.69 | 9.9 |
| 84 | 83 | 9.38 | 10.62 |
| 85 | 84 | 8.54 | 10.42 |
| 86 | 85 | 9.65 | 11.02 |
| 87 | 86 | 9.65 | 10.77 |
| 88 | 87 | 9.46 | 11.5 |
| 89 | 88 | 9.3 | 10.69 |
| 90 | 89 | 9.25 | 10.89 |
| 91 | 90 | 9.51 | 10.79 |
| 92 | 91 | 9.26 | 10.78 |
| 93 | 92 | 9.16 | 10.7 |
| 94 | 93 | 9 | 10.36 |
| 95 | 94 | 8.97 | 10.86 |
| 96 | 95 | 9.44 | 10.09 |
| 97 | 96 | 9.46 | 10.87 |
| 98 | 97 | 8.74 | 10.48 |
| 99 | 98 | 9.56 | 10.95 |
| 100 | 99 | 9.17 | 10.67 |
| 101 | 100 | 8.98 | 10.85 |

**FIGURE 7.8**
EXCEL workbook showing 100 95% confidence intervals for the mean of a normal distribution ($\mu = 10, \sigma = 1$)

## Applied Exercises

7.14 *Finding t-values.* Use Table 7 of Appendix B to determine the values of $t_{\alpha/2}$ that would be used in the construction of a confidence interval for a population mean for each of the following combinations of confidence coefficient and sample size:

a. Confidence coefficient .99, $n = 18$
b. Confidence coefficient .95, $n = 10$
c. Confidence coefficient .90, $n = 15$

7.15 *Comparing z and t-values.* It can be shown (proof omitted) that as the sample size $n$ increases, the $T$ distribution tends to normality and the value $t_\alpha$, such that $P(T > t_\alpha) = \alpha$, approaches the value $z_\alpha$, such that $P(Z > z_\alpha) = \alpha$. Use Table 7 of Appendix B to verify that as the sample size $n$ gets infinitely large, $t_{.05} = z_{.05}$, $t_{.025} = z_{.025}$, and $t_{.01} = z_{.01}$.

## Theoretical Exercises

7.16 Let $Y$ be the number of successes in a binomial experiment with $n$ trials and probability of success $p$. Assuming that $n$ is large, use the sample proportion of successes $\hat{p} = Y/n$ to form a confidence interval for $p$ with confidence coefficient $(1 - \alpha)$. [*Hint:* Start with the pivotal statistic

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}\hat{q}}{n}}}$$

and use the fact (proof omitted) that for large $n$, $Z$ is approximately a standard normal random variable.]

7.17 Let $y_1, y_2, \ldots, y_n$ be a random sample from a Poisson distribution with mean $\lambda$. Suppose we use $\bar{y}$ as an estimator of $\lambda$. Derive a $(1 - \alpha)100\%$ confidence interval for $\lambda$. [*Hint:* Start with the pivotal statistic

$$Z = \frac{\bar{y} - \lambda}{\sqrt{\lambda/n}}$$

and show that for large samples, $Z$ is approximately a standard normal random variable. Then substitute $\bar{y}$ for $\lambda$ in the denominator (why can you do this?) and follow the pivotal method of Example 7.6.]

7.18 Let $y_1, y_2, \ldots, y_n$ be a random sample of $n$ observations from an exponential distribution with mean $\beta$. Derive a large-sample confidence interval for $\beta$. [*Hint:* Start with the pivotal statistic

$$Z = \frac{\bar{y} - \beta}{\beta/\sqrt{n}}$$

and show that for large samples, $Z$ is approximately a standard normal random variable. Then substitute $\bar{y}$ for $\beta$ in the denominator (why can you do this?) and follow the pivotal method of Example 7.6.]

7.19 Let $\bar{y}_1$ and $s_1^2$ be the sample mean and sample variance, respectively, of $n_1$ observations randomly selected from a population with mean $\mu_1$ and variance $\sigma_1^2$. Similarly, define $\bar{y}_2$ and $s_2^2$ for an independent random sample of $n_2$ observations from a population with mean $\mu_2$ and $\sigma_2^2$. Derive a large-sample confidence interval for $(\mu_1 - \mu_2)$. [*Hint:* Start with the pivotal statistic

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

and show that for large samples, $Z$ is approximately a standard normal random variable. Substitute $s_1^2$ for $\sigma_1^2$ and $s_2^2$ for $\sigma_2^2$ (why can you do this?) and follow the pivotal method of Example 7.6.]

7.20 Let $(\bar{y}_1, s_1^2)$ and $(\bar{y}_2, s_2^2)$ be the means and variances of two independent random samples of sizes $n_1$ and $n_2$, respectively, selected from normal populations with different means, $\mu_1$ and $\mu_2$, but with a common variance, $\sigma^2$.

a. Show that $E(\bar{y}_1 - \bar{y}_2) = \mu_1 - \mu_2$.
b. Show that

$$V(\bar{y}_1 - \bar{y}_2) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

c. Explain why

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

is a standard normal random variable.

7.21 Refer to Exercise 7.20. According to Theorem 6.11,

$$\chi_1^2 = \frac{(n_1 - 1)s_1^2}{\sigma^2} \quad \text{and} \quad \chi_2^2 = \frac{(n_2 - 1)s_2^2}{\sigma^2}$$

are independent chi-square random variables with $(n_1 - 1)$ and $(n_2 - 1)$ df, respectively. Show that

$$\chi^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2}$$

is a chi-square random variable with $(n_1 + n_2 - 2)$ df.

7.22 Refer to Exercises 7.20 and 7.21. The pooled estimator of the common variance $\sigma^2$ is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Show that

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

has a Student's $T$ distribution with $(n_1 + n_2 - 2)$ df. [*Hint:* Recall that $T = Z/\sqrt{\chi^2/\nu}$ has a Student's $T$ distribution with $\nu$ df and use the results of Exercises 7.20c and 7.21.]

7.23 Use the pivotal statistic $T$ given in Exercise 7.22 to derive a $(1 - \alpha)100\%$ small-sample confidence interval for $(\mu_1 - \mu_2)$.

## 7.4 Estimation of a Population Mean

From our discussions in Section 7.2, we already know that a useful point estimate of the population mean $\mu$ is $\bar{y}$, the sample mean. According to the central limit theorem (Theorem 6.9), we also know that for sufficiently large $n$, the sampling distribution of the sample mean $\bar{y}$ is approximately normal with $E(\bar{y}) = \mu$ and $V(\bar{y}) = \sigma^2/n$. The fact that $E(\bar{y}) = \mu$ implies that $\bar{y}$ is an unbiased estimator of $\mu$. Furthermore, it can be shown (proof omitted) that $\bar{y}$ has the smallest variance among all unbiased estimators of $\mu$. Hence, $\bar{y}$ is the MVUE for $\mu$. Therefore, it is not surprising that $\bar{y}$ is considered the best estimator of $\mu$.

Since $\bar{y}$ is approximately normal for large $n$, we can apply Theorem 7.2 to construct a large-sample $(1 - \alpha)100\%$ confidence interval for $\mu$. Substituting $\hat{\theta} = \bar{y}$ and $\sigma_{\hat{\theta}} = \sigma/\sqrt{n}$ into the confidence interval formula given in Theorem 7.2, we obtain the formula given in the following box.

### Large-Sample $(1 - \alpha)100\%$ Confidence Interval for the Population Mean, $\mu$

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} = \bar{y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the $z$ value that locates an area of $\alpha/2$ to its right, $\sigma$ is the standard deviation of the population from which the sample was selected, $n$ is the sample size, and $\bar{y}$ is the value of the sample mean.

[*Note:* When the value of $\sigma$ is unknown (as will usually be the case), the sample standard deviation $s$ may be used to approximate $\sigma$ in the formula for the confidence interval. The approximation is generally quite satisfactory when $n \geq 30$.]

*Assumptions:* None (since the Central Limit Theorem guarantees that $\bar{y}$ is approximately normal regardless of the distribution of the sampled population)

*Note:* The value of the sample size $n$ required for the sampling distribution of $\bar{y}$ to be approximately normal will vary depending on the shape (distribution) of the target population (see Examples 6.18 and 6.19). As a general rule of thumb, a sample size $n$ of 30 or more will be considered *sufficiently large* for the central limit theorem to apply.

**Example 7.7**

Large-Sample Estimation of $\mu$: Mean Failure Time

Suppose a PC manufacturer wants to evaluate the performance of its hard disk memory system. One measure of performance is the average time between failures of the disk drive. To estimate this value, a quality control engineer recorded the time between failures for a random sample of 45 disk-drive failures. The following sample statistics were computed:

$$\bar{y} = 1{,}762 \text{ hours} \qquad s = 215 \text{ hours}$$

a. Estimate the true mean time between failures with a 90% confidence interval.

b. If the hard disk memory system is running properly, the true mean time between failures will exceed 1,700 hours. Based on the interval, part **a**, what can you infer about the disk memory system?

Solution

a. For a confidence coefficient of $1 - \alpha = .90$, we have $\alpha = .10$ and $\alpha/2 = .05$; therefore, a 90% confidence interval for $\mu$ is given by

$$\begin{aligned}\bar{y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &= \bar{y} \pm z_{.05}\left(\frac{\sigma}{\sqrt{n}}\right)\\ &\approx \bar{y} \pm z_{0.5}\left(\frac{s}{\sqrt{n}}\right)\\ &= 1{,}762 \pm z_{.05}\left(\frac{215}{\sqrt{45}}\right)\end{aligned}$$

where $z_{.05}$ is the $z$ value corresponding to an upper-tail area of .05. From Table 5 of Appendix B, $z_{.05} = 1.645$. Then the desired interval is

$$\begin{aligned}1{,}762 \pm z_{.05}\left(\frac{215}{\sqrt{45}}\right) &= 1{,}762 \pm 1.645\left(\frac{215}{\sqrt{45}}\right)\\ &= 1{,}762 \pm 52.7\end{aligned}$$

or 1,709.3 to 1,814.7 hours. We are 90% confident that the interval (1,709.3, 1,814.7) encloses $\mu$, the true mean time between disk failures.

b. Since all values within the 90% confidence interval exceed 1,700 hours, we can infer (with 90% confidence) that the hard disk memory system is running properly.

Sometimes, time or cost limitations may restrict the number of sample observations that may be obtained for estimating $\mu$. In the case of small samples, (say, $n < 30$), the following two problems arise:

1. Since the central limit theorem applies only to large samples, we are not able to assume that the sampling distribution of $\bar{y}$ is approximately normal. Therefore, we cannot apply Theorem 7.2. For small samples, the sampling distribution of $\bar{y}$ depends on the particular form of the relative frequency distribution of the population being sampled.
2. The sample standard deviation $s$ may not be a satisfactory approximation to the population standard deviation $\sigma$ if the sample size is small.

Fortunately, we may proceed with estimation techniques based on small samples if we can assume that the population from which the sample is selected has an approximate normal distribution. If this assumption is valid, then we can use the procedure of Example 7.6 to construct a confidence interval for $\mu$. The general form of a small-sample confidence interval for $\mu$, based on the Student's $T$ distribution, is as shown in the next box.

**Small-Sample $(1 - \alpha)100\%$ Confidence Interval for the Population Mean, $\mu$**

$$\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ is obtained from the Student's $T$ distribution with $(n - 1)$ degrees of freedom.

*Assumption:* The population from which the sample is selected has an approximately normal distribution.

**Example 7.8**

Small-Sample Estimation of $\mu$: Mean Silica Content

The Geothermal Loop Experimental Facility, located in the Salton Sea in southern California, is a U.S. Department of Energy operation for studying the feasibility of generating electricity from the hot, highly saline water of the Salton Sea. Operating experience has shown that these brines leave silica scale deposits on metallic plant piping, causing excessive plant outages. Research published in the *Journal of Testing and Evaluation* found that scaling can be reduced somewhat by adding chemical solutions to the brine. In one screening experiment, each of five antiscalants was added to an aliquot of brine, and the solutions were filtered. A silica determination (parts per million of silicon dioxide) was made on each filtered sample after a holding time of 24 hours, with the results shown in Table 7.2. Estimate the mean amount of silicon dioxide present in the five antiscalant solutions. Use a 99% confidence interval.

**TABLE 7.2 Silica measurements for Example 7.8**

BRINE

| | | | | |
|---|---|---|---|---|
| 229 | 255 | 280 | 203 | 229 |

Solution

The first step in constructing the confidence interval is to compute the mean, $\bar{y}$, and standard deviation, $s$, of the sample of five silicon dioxide amounts. These values, $\bar{y} = 239.2$ and $s = 29.3$, are shaded in the MINITAB printout, Figure 7.9.

For a confidence coefficient of $1 - \alpha = .99$, we have $\alpha = .01$ and $\alpha/2 = .005$. Since the sample size is small $(n = 5)$, our estimation technique requires the assumption that the amount of silicon dioxide present in an antiscalant solution has an approximately normal distribution (i.e., the sample of 5 silicon amounts is selected from a normal population).

Substituting the values for $\bar{y}$, $s$, and $n$ into the formula for a small-sample confidence interval for $\mu$, we obtain

$$\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = \bar{y} \pm t_{.005}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 239.2 \pm t_{.005}\left(\frac{29.3}{\sqrt{5}}\right)$$

where $t_{.005}$ is the value corresponding to an upper-tail area of .005 in the Student's $T$ distribution based on $(n - 1) = 4$ degrees of freedom. From Table 7 of Appendix B, the required $t_{.005} = 4.604$. Substitution of this value yields

$$239.2 \pm t_{.005}\left(\frac{29.3}{\sqrt{5}}\right) = 239.2 \pm (4.604)\left(\frac{29.3}{\sqrt{5}}\right)$$

$$= 239.2 \pm 60.3$$

or, 178.9 to 299.5 ppm. (*Note:* This interval is highlighted on the MINITAB printout, Figure 7.9.) Thus, if the distribution of silicon dioxide amounts is approximately normal, we can be 99% confident that the interval (178.9, 299.5) encloses $\mu$, the true mean amount of silicon dioxide present in an antiscalant solution.

**FIGURE 7.9**
MINITAB descriptive statistics and confidence interval for silica data

**One-Sample T: PPM**

```
Variable  N     Mean   StDev  SE Mean          99% CI
PPM       5  239.200  29.295   13.101  (178.881, 299.519)
```

Before we conclude this section, two comments are necessary. The first concerns the assumption that the sampled population is normally distributed. In the real world, we rarely know whether a sampled population has an exact normal distribution. However, empirical studies indicate that moderate departures from this assumption do not seriously affect the confidence coefficients for small-sample confidence intervals. For example, if the population of silicon dioxide amounts for the antiscalant solutions of Example 7.8 has a distribution that is mound-shaped but nonnormal, it is likely that the actual confidence coefficient for the 99% confidence interval will be close to

.99—at least close enough to be of practical use. As a consequence, the small-sample confidence interval is frequently used by experimenters when estimating the population mean of a nonnormal distribution as long as the distribution is mound-shaped and only moderately skewed. For populations that depart greatly from normality, other estimation techniques (such as robust estimation) or methods that are distribution-free (called **nonparametrics**) are recommended. Nonparametric statistics are the topic of Chapter 15.

The second comment focuses on whether $\sigma$ is known or unknown. We have previously stated (Chapter 6) that when $\sigma$ is known and the sampled population is normally distributed, the sampling distribution of $\bar{y}$ is normal regardless of the size of the sample. That is, if you know the value of $\sigma$ and you know that the sample comes from a normal population, then you can use the $z$ distribution rather than the $t$ distribution to form confidence intervals. In reality, however, $\sigma$ is rarely (if ever) known. Consequently, you will always be using $s$ in place of $\sigma$ in the confidence interval formulas, and the sampling distribution of $\bar{y}$ will be a Student's $T$ distribution. This is why the formula for a large-sample confidence interval given earlier in this section is only approximate; in the large-sample case, $t \approx z$. Many statistical software packages give the results for *exact* confidence intervals when $\sigma$ is unknown; thus, these results are based on the $T$ distribution. For practical reasons, however, we will continue to distinguish between $Z$ and $T$ confidence intervals based on whether the sample size is large or small.

## Applied Exercises

7.24 *Characteristics of DNA in antigen-produced protein.* *Ascaridia galli* is a parasitic roundworm that attacks the intestines of birds, especially chickens and turkeys. Scientists are working on a synthetic vaccine (antigen) for the parasite. The success of the vaccine hinges on the characteristics of DNA in peptide (protein) produced by the antigen. In the journal, *Gene Therapy and Molecular Biology* (June, 2009), scientists tested alleles of antigen-produced protein for level of peptide. For a sample of 4 alleles, the mean peptide score was 1.43 and the standard deviation was .13.

a. Use this information to construct a 90% confidence interval for the true mean peptide score in alleles of the antigen-produced protein.
b. Interpret the interval for the scientists.
c. What is meant by the phrase "90% confidence"?

7.25 *Laser scanning for fish volume estimation.* Engineers design tanks for rearing commercial fish to minimize both the use of natural resources (water) and the rearing volume necessary to ensure fish welfare. One key to a well-designed tank is obtaining a reliable estimate of the volume (biomass) of fish reared in the tank. The feasibility of a laser scanning technique for estimating fish biomass was investigated in the *Journal of Aquacultural Engineering* (Nov. 2012). Fifty turbot fish were reared in a tank for experimental purposes. A laser scan was executed in four randomly selected locations in the tank and the volume (in kilograms) of fish layer at each location was measured. The four laser scans yielded a mean volume of 240 kg with a standard deviation of 15 kg. From this information, estimate the true mean volume of fish layer in the tank with 99% confidence. Interpret the result, practically. What assumption about the data is necessary for the inference derived from the analysis to be valid?

7.26 *Surface roughness of pipe.* Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated interior pipe used in oil fields, Exercise 2.20 (p. 37). The data (in micrometers) for 20 sampled pipe sections are reproduced in the table on p. 312. A MINITAB analysis of the data is shown below.

a. Locate a 95% confidence interval for the mean surface roughness of coated interior pipe on the accompanying MINITAB printout.
b. Would you expect the average surface roughness to be as high as 2.5 micrometers? Explain.

**MINITAB Output for Exercise 7.26**

**One-Sample T: ROUGH**

| Variable | N | Mean | StDev | SE Mean | 95% CI |
|---|---|---|---|---|---|
| ROUGH | 20 | 1.88100 | 0.52391 | 0.11715 | (1.63580, 2.12620) |

## Data for Exercise 7.26

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

7.27 *Evaporation from swimming pools.* A new formula for estimating the water evaporation from occupied swimming pools was proposed and analyzed in the journal, *Heating/Piping/Air Conditioning Engineering* (Apr. 2013). The key components of the new formula are number of pool occupants, area of pool's water surface, and the density difference between room air temperature and the air at the pool's surface. Data were collected from a wide range of pools where the evaporation level was known. The new formula was applied to each pool in the sample, yielding an estimated evaporation level. The absolute value of the deviation between the actual and estimated evaporation level was then recorded as a percentage. The researchers reported the following summary statistics for absolute deviation percentage: $\bar{y} = 18$, $s = 20$. Assume that the sample contained $n = 500$ swimming pools.

a. Estimate the true mean absolute deviation percentage for the new formula with a 90% confidence interval.
b. The American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) handbook also provides a formula for estimating pool evaporation. Suppose the ASHRAE mean absolute deviation percentage is $\mu = 34\%$. (This value was reported in the article.) On average, is the new formula "better" than the ASHRAE formula? Explain.

7.28 *Radon exposure in Egyptian tombs.* Many ancient Egyptian tombs were cut from limestone rock that contained uranium. Since most tombs are not well-ventilated, guards, tour guides, and visitors may be exposed to deadly radon gas. In *Radiation Protection Dosimetry* (December 2010), a study of radon exposure in tombs in the Valley of Kings, Luxor, Egypt (recently opened for public tours) was conducted. The radon levels — measured in becquerels per cubic meter ($Bq/m^3$) — in the inner chambers of a sample of 12 tombs were determined. Summary statistics follow: $\bar{y} = 3{,}643$ $Bq/m^3$ and $s = 4{,}487$ $Bq/m^3$. Use this information to estimate, with 95% confidence, the true mean level of radon exposure in tombs in the Valley of Kings. Interpret the resulting interval.

7.29 *Contamination of New Jersey wells.* Methyl *t*-butyl ether (MTBE) is an organic water contaminant that often results from gasoline spills. The level of MTBE (in parts per billion) was measured for a sample of 12 well sites located near a gasoline service station in New Jersey. (*Environmental Science & Technology*, Jan. 2005.) The data are listed in the accompanying table.

**NJGAS**

| | | | | | |
|---|---|---|---|---|---|
| 150 | 367 | 38 | 12 | 11 | 134 |
| 12 | 251 | 63 | 8 | 13 | 107 |

*Source:* Kuder, T., et al. "Enrichment of stable carbon and hydrogen isotopes during anaerobic biodegradation of MTBE: Microcosm and field evidence." *Environmental Science & Technology*, Vol. 39, No. 1, Jan. 2005 (Table 1).

a. Give a point estimate for $\mu$, the true mean MTBE level for all well sites located near the New Jersey gasoline service station.
b. Calculate and interpret a 99% confidence interval for $\mu$.
c. What assumptions are required for the interval, part **b**, to be valid? Are these assumptions reasonably satisfied?

7.30 *Intellectual development of engineering students.* Refer to the *Journal of Engineering Education* (Jan. 2005) study of the intellectual development of undergraduate engineering students, Exercise 1.27 (p. 20). Intellectual development (Perry) scores were determined for 21 students in a first-year, project-based design course. (Recall that a Perry score of 1 indicates the lowest level of intellectual development, and a Perry score of 5 indicates the highest level.) The average Perry score for the 21 students was 3.27 and the standard deviation was .40. Apply the confidence interval method of this section to estimate the true mean Perry score of all undergraduate engineering students with 99% confidence. Interpret the results.

7.31 *Radioactive lichen.* Refer to the Lichen Radionuclide Baseline Research project, Exercise 2.15 (p. 36). University of Alaska researchers determined the amount of the radioactive element cesium-137 for nine lichen specimens sampled from various Alaskan locations. The data values (measured in microcuries per milliliter) are given in the table.

**LICHEN**

| | | | | |
|---|---|---|---|---|
| 0.0040868 | 0.0157644 | 0.0023579 | 0.0067379 | |
| 0.0165727 | 0.0067379 | 0.0078284 | 0.0111090 | 0.0100518 |

*Source:* Lichen Radionuclide Baseline Research project, 2003.

a. Describe the population of interest to the researchers.
b. Estimate the mean of the population using a 95% confidence interval.
c. Make an inference about the magnitude of the population mean.
d. For the inference to be valid, how must the population data be distributed?

7.32 *Crude oil biodegradation.* Refer to the *Journal of Petroleum Geology* (April 2010) study of the environmental factors associated with biodegradation in crude oil reservoirs, Exercise 2.18 (p. 37). One indicator of biodegradation is the level of dioxide in the water. Recall that 16 water specimens were randomly selected from various locations in a reservoir on the floor of a mine and the amount of dioxide (milligrams/liter) as well as presence of oil was determined for each specimen. These data are reproduced in the table on p. 313.

a. Estimate the true mean amount of dioxide present in water specimens that contain oil using a 95% confidence interval. Give a practical interpretation of the interval.
b. Repeat part a for water specimens that do not contain oil.
c. Based on the results, parts a and b, make an inference about biodegradation at the mine reservoir.

**BIODEG**

| Dioxide Amount | Crude Oil Present |
|---|---|
| 3.3 | No |
| 0.5 | Yes |
| 1.3 | Yes |
| 0.4 | Yes |
| 0.1 | No |
| 4.0 | No |
| 0.3 | No |
| 0.2 | Yes |
| 2.4 | No |
| 2.4 | No |
| 1.4 | No |
| 0.5 | Yes |
| 0.2 | Yes |
| 4.0 | No |
| 4.0 | No |
| 4.0 | No |

*Source:* Permanyer, A., et al. "Crude oil biodegradation and environmental factors at the Riutort oil shale mine, SE Pyrenees", *Journal of Petroleum Geology*, Vol. 33, No. 2, April 2010 (Table 1).

7.33 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study on the current trend in the design of social robots, Exercise 2.37 (p. 49). Recall that in a random sample of social robots obtained through a web search, 28 were built with wheels. The number of wheels on each of the 28 robots are reproduced in the accompanying table.

a. Estimate $\mu$, the average number of wheels used on all social robots built with wheels, with 99% confidence.
b. Practically interpret the interval, part a.
c. Refer to part a. In repeated sampling, what proportion of all similarly constructed confidence intervals will contain the true mean, $\mu$?

**ROBOTS**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 3 | 3 | 3 | 6 | 4 | 2 | 2 | 2 | 1 | 3 | 3 | 3 |
| 3 | 4 | 4 | 3 | 2 | 8 | 2 | 2 | 3 | 4 | 3 | 3 | 4 | 2 |

*Source:* Chew, S., et al. "Do social robots walk or roll?", *International Conference on Social Robotics*, Vol. 6414, 2010 (adapted from Figure 2).

7.34 *Monitoring impedance to leg movements.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) experiment to monitor the impedance to leg movements, Exercise 2.46 (p. 51). Recall that engineers attached electrodes to the ankles and knees of volunteers and measured the signal-to-noise ratio (SNR) of impedance changes. For a particular ankle–knee electrode pair, a sample of 10 volunteers had SNR values with a mean of 19.5 and a standard deviation of 4.7.

a. Form a 95% confidence interval for the true mean SNR impedance change. Interpret the result.
b. In Exercise 2.36, you found an interval that contains about 95% of all SNR values in the population. Compare this interval to the confidence interval, part **a.** Explain why these two intervals differ.

7.35 *Oven cooking study.* A group of Harvard University School of Public Health researchers studied the impact of cooking on the size of indoor air particles. (*Environmental Science & Technology*, September 1, 2000.) The decay rate (measured as $\mu$m/hour) for fine particles produced from oven cooking or toasting was recorded on six randomly selected days. These six measurements are:

**DECAY**

| | | | | | |
|---|---|---|---|---|---|
| .95 | .83 | 1.20 | .89 | 1.45 | 1.12 |

*Source:* Abt, E., et al., "Relative contribution of outdoor and indoor particle sources to indoor concentrations." *Environmental Science & Technology*, Vol. 34, No. 17, Sept. 1, 2000 (Table 3).

a. Find and interpret a 95% confidence interval for the true average decay rate of fine particles produced from oven cooking or toasting.
b. Explain what the phrase "95% confident" implies in the interpretation of part **a**.
c. What must be true about the distribution of the population of decay rates for the inference to be valid?

**PONDICE**

7.36 *Albedo of ice meltponds.* Refer to the National Snow and Ice Data Center (NSIDC) collection of data on the albedo, depth, and physical characteristics of ice meltponds in the Canadian Arctic, first presented in Example 2.1 (p. 17). Albedo is the ratio of the light reflected by the ice to that received by it. (High albedo values give a white appearance to the ice.) Visible albedo values were recorded for a sample of 504 ice meltponds located in the Barrow Strait in the Canadian Arctic; these data are saved in the **PONDICE** file.

a. Find a 90% confidence interval for the true mean visible albedo value of all Canadian Arctic ice ponds.
b. Give both a practical and theoretical interpretation of the interval.
c. Recall from Example 2.1 that type of ice for each pond was classified as first-year ice, multiyear ice, or landfast ice. Find 90% confidence intervals for mean visible albedo for each of the three ice types. Interpret the intervals.

7.37 *Oxygen bubbles in molten salt.* Molten salt is used in an electro-refiner to treat nuclear fuel waste. Eventually, the salt needs to be purified (for reuse) or disposed of. A promising method of purification involves oxidation. Such a method was investigated in *Chemical Engineering Research and Design* (Mar. 2013). An important aspect of the purification process is the rising velocity of oxygen bubbles in the molten salt. An experiment was conducted in which oxygen was inserted (at a designated sparging rate) into molten salt and photographic images of the bubbles taken. A random sample of 25 images yielded the data on bubble velocity (measured in meters per second) shown in the table. (Note: These data are simulated based on information provided in the article.)

a. Use statistical software to find a 95% confidence interval for the mean bubble rising velocity of the population. Interpret the result.
b. The researchers discovered that the mean bubble rising velocity is $\mu = .338$ when the sparging rate of oxygen is $3.33 \times 10^{-6}$. Do you believe that the data in the table were generated at this sparging rate? Explain.

**BUBBLE**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.275 | 0.261 | 0.209 | 0.266 | 0.265 | 0.312 | 0.285 | 0.317 | 0.229 |
| 0.251 | 0.256 | 0.339 | 0.213 | 0.178 | 0.217 | 0.307 | 0.264 | 0.319 |
| 0.298 | 0.169 | 0.342 | 0.270 | 0.262 | 0.228 | 0.220 | | |

## 7.5 Estimation of the Difference Between Two Population Means: Independent Samples

In Section 7.4, we learned how to estimate the parameter $\mu$ from a single population. We now proceed to a technique for using the information in two samples to estimate the difference between two population means, $(\mu_1 - \mu_2)$, when the samples are collected independently. For example, we may want to compare the mean starting salaries for college graduates with mechanical engineering and civil engineering degrees, or the mean operating costs of automobiles with rotary engines and standard engines, or the mean failure times of two electronic components. The technique to be presented is a straightforward extension of that used for estimation of a single population mean.

Suppose we select independent random samples of sizes $n_1$ and $n_2$ from populations with means $\mu_1$ and $\mu_2$, respectively. Intuitively, we want to use the difference between the sample means, $(\bar{y}_1 - \bar{y}_2)$, to estimate $(\mu_1 - \mu_2)$. In Example 6.20, we showed that

$$E(\bar{y}_1 - \bar{y}_2) = \mu_1 - \mu_2$$
$$V(\bar{y}_1 - \bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

You can see that $(\bar{y}_1 - \bar{y}_2)$ is an unbiased estimator for $(\mu_1 - \mu_2)$. Further, it can be shown (proof omitted) that $V(\bar{y}_1 - \bar{y}_2)$ is smallest among all unbiased estimators, i.e., $(\bar{y}_1 - \bar{y}_2)$ is the MVUE for $(\mu_1 - \mu_2)$.

According to the central limit theorem, $(\bar{y}_1 - \bar{y}_2)$ will also be approximately normal for large $n_1$ and $n_2$ regardless of the distributions of the sampled populations. Thus, we can apply Theorem 7.2 to construct a large-sample confidence interval for $(\mu_1 - \mu_2)$. The procedure for forming a large-sample confidence interval for $(\mu_1 - \mu_2)$ appears in the box.

**Large-Sample $(1 - \alpha)100\%$ Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples**

$$(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sigma_{(\bar{y}_1 - \bar{y}_2)} = (\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
$$\approx (\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(*Note:* We have used the sample variances $s_1^2$ and $s_2^2$ as approximations to the corresponding population parameters.)

*Assumptions:* 1. The two random samples are selected in an independent manner from the target populations. That is, the choice of elements in one sample does not affect, and is not affected by, the choice of elements in the other sample.

2. The sample sizes $n_1$ and $n_2$ are sufficiently large for the central limit theorem to apply. (We recommend $n_1 \geq 30$ and $n_2 \geq 30$.)

**Example 7.9**

Large-Sample Confidence Interval for $(\mu_1 - \mu_2)$: Comparing Mean Salaries of Engineers

We want to estimate the difference between the mean starting salaries for recent graduates with mechanical engineering and electrical engineering bachelor's degrees from the University of Michigan (UM). The following information is available:*

1. A random sample of 48 starting salaries for UM mechanical engineering graduates produced a sample mean of \$64,650 and a standard deviation of \$7,000.
2. A random sample of 32 starting salaries for UM aerospace engineering graduates produced a sample mean of \$58,420 and a standard deviation of \$6,830.

Solution

We will let the subscript 1 refer to the mechanical engineering graduates and the subscript 2 to the aerospace engineering graduates. We will also define the following notation:

$\mu_1$ = Population mean starting salary of all recent UM mechanical engineering graduates

$\mu_2$ = Population mean starting salary of all recent UM aerospace engineering graduates

Similarly, let $\bar{y}_1$ and $\bar{y}_2$ denote the respective sample means; $s_1$ and $s_2$, the respective sample standard deviations; and $n_1$ and $n_2$, the respective sample sizes. The given information is summarized in Table 7.3.

**TABLE 7.3 Summary of Information for Example 7.9**

| | Mechanical Engineers | Aerospace Engineers |
|---|---|---|
| *Sample Size* | $n_1 = 48$ | $n_2 = 32$ |
| *Sample Mean* | $\bar{y}_1 = 64{,}650$ | $\bar{y}_2 = 58{,}420$ |
| *Sample Standard Deviation* | $s_1 = 7{,}000$ | $s_2 = 6{,}830$ |

*Source:* Engineering Career Resource Center, University of Michigan.

The general form of a 95% confidence interval for $(\mu_1 - \mu_2)$, based on large, independent samples from the target populations, is given by

$$(\bar{y}_1 - \bar{y}_2) \pm z_{.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Recall that $z_{.025} = 1.96$ and use the information in Table 7.3 to make the following substitutions to obtain the desired confidence interval:

$$\begin{aligned}
&(64{,}650 - 58{,}420) \pm 1.96\sqrt{\sigma_1^2/48 + \sigma_2^2/32} \\
&\approx (64{,}650 - 58{,}420) \pm 1.96\sqrt{(7{,}000)^2/48 + (6{,}830)^2/32} \\
&\approx 6{,}230 \pm 3{,}085
\end{aligned}$$

or (\$3,145, \$9,315).

If we were to use this method of estimation repeatedly to produce confidence intervals for $(\mu_1 - \mu_2)$, the difference between population means, we would expect

*The information for this example was based on a 2011–2012 survey of graduates conducted by the Engineering Career Resource Center, University of Michigan.

95% of the intervals to enclose $(\mu_1 - \mu_2)$. Since the interval includes only positive differences, we can be reasonably confident that the mean starting salary of mechanical engineering graduates of UM is between \$3,145 and \$9,315 higher than the mean starting salary of aerospace engineering graduates.

### Practical Interpretation of a Confidence Interval for $(\theta_1 - \theta_2)$

Let (LCL, UCL) represent a (1-$\alpha$) 100% confidence interval for $(\theta_1 - \theta_2)$.

- If LCL > 0 and UCL > 0, conclude $\theta_1 > \theta_2$
- If LCL < 0 and UCL < 0, conclude $\theta_1 < \theta_2$
- If LCL < 0 and UCL > 0 (i.e., the interval includes 0), conclude no evidence of a difference between $\theta_1$ and $\theta_2$

A confidence interval for $(\mu_1 - \mu_2)$, based on small samples from each population, is derived using Student's *T* distribution. As was the case when estimating a single population mean from information in a small sample, we must make specific assumptions about the relative frequency distributions of the two populations, as indicated in the box. These assumptions are required if either sample is small (i.e., if either $n_1 < 30$ or $n_2 < 30$).

### Small-Sample $(1 - \alpha)100\%$ Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples and $\sigma_1^2 = \sigma_2^2$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and the value of $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ degrees of freedom.

*Assumptions:*
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The variances $\sigma_1^2$ and $\sigma_2^2$ of the two populations are equal.
3. The random samples are selected in an independent manner from the two populations.

Note that this procedure requires that the samples be selected from two normal populations that have equal variances (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Since we are assuming the variances are equal, we construct an estimate of $\sigma^2$ based on the information contained in *both* samples. This **pooled estimate** is denoted by $s_p^2$ and is computed as shown in the previous box. You will notice that $s_p^2$ is a weighted average of the two sample variances, $s_1^2$ and $s_2^2$, with the weights proportional to the respective sample sizes.

**Example 7.10**

Small-Sample Confidence Interval for $\mu_1 - \mu_2$: Permeability of Concrete

The *Journal of Testing and Evaluation* reported on the results of laboratory tests conducted to investigate the stability and permeability of open-graded asphalt concrete. In one part of the experiment, four concrete specimens were prepared for asphalt contents of 3% and 7% by total weight of mix. The water permeability of each concrete specimen was determined by flowing deaerated water across the specimen and measuring the amount of water loss. The permeability measurements (recorded in inches per hour) for the eight concrete specimens are shown in Table 7.4. Find a 95% confidence interval for the difference between the mean permeabilities of concrete made with asphalt contents of 3% and 7%. Interpret the interval.

Solution

First, we calculate the means and variances of the two samples, using the computer. A SAS printout giving descriptive statistics for the two samples is shown in Figure 7.10.

CONCRETE

**TABLE 7.4 Permeability Measurements for 3% and 7% Asphalt Concrete, Example 7.10**

| | | | | | |
|---|---|---|---|---|---|
| *Asphalt Content* | 3% | 1,189 | 840 | 1,020 | 980 |
| | 7% | 853 | 900 | 733 | 785 |

*Source:* Woelfl, G., Wei, I., Faulstich, C., and Litwack, H. "Laboratory testing of asphalt concrete for porous pavements." *Journal of Testing and Evaluation*, Vol. 9, No. 4, July 1981, pp. 175–181. Copyright American Society for Testing and Materials.

For the 3% asphalt, $\bar{y}_1 = 1{,}007.25$ and $s_1 = 143.66$; for the 7% asphalt, $\bar{y}_2 = 817.75$ and $s_2 = 73.63$.

Since both samples are small $(n_1 = n_2 = 4)$, the procedure requires the assumption that the two samples of permeability measurements are independently and randomly selected from normal populations with equal variances. The 95% small-sample confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{.025}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= (1{,}007.25 - 817.75) \pm t_{.025}\sqrt{s_p^2\left(\frac{1}{4} + \frac{1}{4}\right)}$$

where $t_{.025} = 2.447$ is obtained from the $T$ distribution (Table 7 of Appendix B) based on $n_1 + n_2 - 2 = 4 + 4 - 2 = 6$ degrees of freedom, and

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_s^2}{n_1 + n_2 - 2} = \frac{3(143.66)^2 + 3(73.63)^2}{6}$$
$$= 13{,}028.92$$

```
Sample Statistics

     Group            N        Mean     Std. Dev.   Std. Error
     ---------------------------------------------------------
     3%               4     1007.25       143.66        71.828
     7%               4      817.75       73.627        36.813

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are    t statistic      Df        Pr > t
     ---------------------------------------------------------
     Equal                     2.348       6        0.0572
     Not Equal                 2.348    4.47        0.0718

95% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                  -8.00          387.00
```

**FIGURE 7.10**
SAS descriptive statistics and confidence interval for concrete data

is the pooled sample variance. Substitution yields the interval

$$(1{,}007.25 - 817.75) \pm 2.447\sqrt{13{,}028.92\left(\frac{1}{4} + \frac{1}{4}\right)}$$
$$= 189.5 \pm 197.50$$

or, $-8.00$ to 387.00. (*Note:* This interval is also displayed in Figure 7.10.) The interval is interpreted as follows: We are 95% confident that the interval $(-8, 387)$ encloses the true difference between the mean permeabilities of the two types of concrete. Since the interval includes 0, we are unable to conclude that the two population means differ.

As with the one-sample case, the assumptions required for estimating $(\mu_1 - \mu_2)$ with small samples do not have to be satisfied exactly for the interval estimate to be useful in practice. Slight departures from these assumptions do not seriously affect the level of confidence in the procedure. For example, when the variances $\sigma_1^2$ and $\sigma_2^2$ of the sampled populations are unequal, researchers have found that the formula for the small-sample confidence interval for $(\mu_1 - \mu_2)$ still yields valid results in practice as long as the two populations are normal and the sample sizes are equal, i.e., $n_1 = n_2$.

This situation occurs in Example 7.10. The sample standard deviations given in Figure 7.10 are $s_1 = 143.66$ and $s_2 = 73.63$. Thus, it is very likely that the population variances, $\sigma_1^2$ and $\sigma_2^2$, are unequal.* However, since $n_1 = n_2 = 4$, the inference derived from this interval is still valid if we use $s_1^2$ and $s_2^2$ as estimates for the population variances (rather than using the pooled sample variance, $s_p^2$).

In the case where $\sigma_1^2 \neq \sigma_2^2$ and $n_1 \neq n_2$, an approximate confidence interval for $(\mu_1 - \mu_2)$ can be constructed by modifying the degrees of freedom associated with the $t$ distribution, and, again, substituting $s_1^2$ for $\sigma_1^2$ and $s_2^2$ for $\sigma_2^2$. These modifications are shown in the box.

### Approximate Small-Sample Inferences for $(\mu_1 - \mu_2)$ when $\sigma_1^2 \neq \sigma_2^2$

To obtain approximate confidence intervals and tests for $(\mu_1 - \mu_2)$ when $\sigma_1^2 \neq \sigma_2^2$, make the following modifications to the degrees of freedom, $\nu$, used in the $T$ distribution and the estimated standard error:

$$n_1 = n_2 = n: \quad \nu = n_1 + n_2 - 2 = 2(n-1) \qquad \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{1}{n}(s_1^2 + s_2^2)}$$

$$n_1 \neq n_2: \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

[*Note:* In the case of $n_1 \neq n_2$, the value of $\nu$ will not generally be an integer. Round $\nu$ down to the nearest integer to use the $t$ table.**]

*Assumptions:*
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The random samples are selected in an independent manner from the two populations.

*A method for comparing two population variances is presented in Section 7.10.

**Rounding the value of $\nu$ down will produce wider, more conservative confidence intervals.

## Applied Exercises

7.38 *Muscle activity of harvesting foresters.* Research in the *International Journal of Foresting Engineering* (Vol. 19, 2008) investigated the muscle activity patterns in the neck and upper extremities among forestry vehicle operators. Two types of harvesting vehicles — Timberjack and Valmet — were compared since they differed dramatically in design of the control levers. Independent random samples of 7 Timberjack and 6 Valmet harvester operators participated in the study. Muscle rest (seconds per minute) in the right trapezius (a neck muscle) was determined for each operator. The 7 Timberjack operators had a mean muscle rest of 10.35 seconds/minute, while the 6 Valmet operators had a mean of 3.30 seconds/minute.

a. Explain why one cannot make any reliable inferences about $(\mu_T - \mu_V)$, the true mean difference in muscle rest in the right trapezius of Timberjack and Valmet harvestor operators based on the information provided.

b. Suppose the standard deviations for the samples of Timberjack and Valmet operators are 4.0 and 2.5 seconds/minute, respectively. Use this additional information to construct a 99% confidence interval for $(\mu_T - \mu_V)$. Practically interpret the resulting interval.

c. What assumptions about the data must be made in order for the inference, part b, to be valid?

7.39 *Drug content assessment.* Refer to Exercise 5.45 (p. 210) and the *Analytical Chemistry* (Dec. 15, 2009) study in which scientists used high-performance liquid chromatography to determine the amount of drug in a tablet. Twenty-five tablets were produced at each of two different, independent sites. Drug concentrations (measured as a percentage) for the tablets produced at the two sites are listed in the accompanying table. The scientists want to know whether there is any difference between the mean drug concentration in tablets produced at Site 1 and the corresponding mean at Site 2. Use the MINITAB printout below to help the scientists draw a conclusion.

**SANDSTONE**

7.40 *Permeability of sandstone during weathering.* Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercise 5.51 (p. 211). Recall that blocks of sandstone were cut into 300 equal-sized slices and the slices randomly divided into three groups of 100 slices each. Slices in group A were not exposed to any type of weathering; slices in group B were repeatedly sprayed with a 10% salt solution (to simulate wetting by driven rain) under temperate conditions; and, slices in group C were soaked in a

**DRUGCON**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Site 1* | | | | | | | | |
| 91.28 | 92.83 | 89.35 | 91.90 | 82.85 | 94.83 | 89.83 | 89.00 | 84.62 |
| 86.96 | 88.32 | 91.17 | 83.86 | 89.74 | 92.24 | 92.59 | 84.21 | 89.36 |
| 90.96 | 92.85 | 89.39 | 89.82 | 89.91 | 92.16 | 88.67 | | |
| *Site 2* | | | | | | | | |
| 89.35 | 86.51 | 89.04 | 91.82 | 93.02 | 88.32 | 88.76 | 89.26 | 90.36 |
| 87.16 | 91.74 | 86.12 | 92.10 | 83.33 | 87.61 | 88.20 | 92.78 | 86.35 |
| 93.84 | 91.20 | 93.44 | 86.77 | 83.77 | 93.19 | 81.79 | | |

*Source:* Borman, P.J., Marion, J.C., Damjanov,I., & Jackson, P. "Design and analysis of method equivalence studies", *Analytical Chemistry*, Vol. 81, No. 24, December 15, 2009 (Table 3).

**MINITAB Output for Exercise 7.39**

**Two-Sample T-Test and CI: Content, Site**

```
Two-sample T for Content

Site   N   Mean  StDev  SE Mean
1     25  89.55   3.07     0.61
2     25  89.03   3.34     0.67


Difference = mu (1) - mu (2)
Estimate for difference:  0.515
95% CI for difference:  (-1.308, 2.338)
T-Test of difference = 0 (vs not =): T-Value = 0.57  P-Value = 0.573  DF = 48
Both use Pooled StDev = 3.2057
```

10% salt solution and then dried (to simulate blocks of sandstone exposed during a wet winter and dried during a hot summer). All sandstone slices were then tested for permeability, measured in milliDarcies (mD). The data for the study (simulated) are saved in the **SANDSTONE** file. Let $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$ represent the sample mean permeability measurements for slices in group A, B, and C, respectively.

a. In Exercise 5.51 you determined that the permeability measurements in any one of the three experimental groups are not approximately normally distributed. How does this impact the shape of the sampling distributions of $(\bar{y}_A - \bar{y}_B)$ and $(\bar{y}_B - \bar{y}_C)$?
b. Find a 95% confidence interval for $(\mu_B - \mu_C)$, the true mean difference in the mean permeability of sandstone slices in groups B and C. From this interval, what do you conclude about which group has the larger mean permeability?
c. Find a 95% confidence interval for $(\mu_A - \mu_B)$, the true mean difference in the mean permeability of sandstone slices in groups A and B. From this interval, what do you conclude about which group has the larger mean permeability?

7.41 *Hippo grazing patterns in Kenya.* In Kenya, human-induced land-use changes and excessive resource extraction has threatened the jungle ecosystem by reducing animal grazing areas and disrupting access to water sources. In *Landscape & Ecology Engineering* (Jan. 2013), researchers compared hippopotamus grazing patterns in two Kenyan areas — a national reserve and a community pastoral ranch. Each area was subdivided into plots of land. The plots were sampled (406 plots in the national reserve and 230 plots in the pastoral ranch) and the number of hippo trails from a water source was determined for each plot. Sample statistics are provided in the table. The researchers concluded that the mean number of hippo trails was higher in the national reserve than in the pastoral ranch. Do you agree? Support your answer with a 95% confidence interval.

| | National Reserve | Pastoral Ranch |
|---|---|---|
| Sample size: | 406 | 230 |
| Mean number of trails: | .31 | .13 |
| Standard deviation: | .4 | .3 |

*Source:* Kanga, E.M., et al. "Hippopotamus and livestock grazing: influences on riparian vegetation and facilitation of other herbivores in the Mara Region of Kenya", *Landscape & Ecology Engineering*, Vol. 9, No. 1, January 2013.

7.42 *Index of Biotic Integrity.* The Ohio Environmental Protection Agency used the Index of Biotic Integrity (IBI) to measure the biological condition or health of an aquatic region. The IBI is the sum of metrics which measure the presence, abundance, and health of fish in the region. (Higher values of the IBI correspond to healthier fish populations.) Researchers collected IBI measurements for sites located in different Ohio river basins. (*Journal of Agricultural, Biological, and Environmental Sciences*, June 2005.) Summary data for two river basins, Muskingum and Hocking, are given in the table. Use a 90% confidence interval to compare the mean IBI values of the two river basins.

| River Basin | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| Muskingum | 53 | .035 | 1.046 |
| Hocking | 51 | .340 | .960 |

*Source:* Boone, E. L., Keying, Y., and Smith, E. P. "Evaluating the relationship between ecological and habitat conditions using hierarchical models." *Journal of Agricultural, Biological, and Environmental Sciences*, Vol. 10, No. 2, June 2005 (Table 1).

7.43 *High-strength aluminum alloys.* Mechanical engineers have developed a new high-strength aluminum alloy for use in antisubmarine aircraft, tankers, and long-range bombers. (*JOM*, Jan. 2003.) The new alloy is obtained by applying a retrogression and reaging (RAA) heat treatment to the current strongest aluminum alloy. A series of strength tests were conducted to compare the new RAA alloy to the current strongest alloy. Three specimens of each type of aluminum alloy were produced and the yield strength (measured in mega-pascals, MPa) of each specimen determined. The results are summarized in the table.

| | Alloy Type | |
|---|---|---|
| | RAA | Current |
| *Number of specimens* | 3 | 3 |
| *Mean yield strength (MPa)* | 641.0 | 592.7 |
| *Standard deviation* | 19.3 | 12.4 |

a. Estimate the difference between the mean yield strengths of the two alloys using a 95% confidence interval.
b. The researchers concluded that the RAA-processed aluminum alloy is superior to the current strongest aluminum alloy with respect to yield strength. Do you agree?

7.44 *Patent infringement case. Chance* (Fall 2002) described a lawsuit where Intel Corp. was charged with infringing on a patent for an invention used in the automatic manufacture of computer chips. In response, Intel accused the inventor of adding material to his patent notebook after the patent was witnessed and granted. The case rested on whether a patent witness's signature was written on top of key text in the notebook or under the key text. Intel hired a physicist who used an X-ray beam to measure the relative concentration of certain elements (e.g., nickel, zinc, potassium) at several spots on the notebook page. The zinc measurements for three notebook locations—on a text line, on a witness line, and on the intersection of the witness and text lines—are provided in the table.

**PATENT**

| | | | | | | |
|---|---|---|---|---|---|---|
| *Text line*: | .335 | .374 | .440 | | | |
| *Witness line*: | .210 | .262 | .188 | .329 | .439 | .397 |
| *Intersection*: | .393 | .353 | .285 | .295 | .319 | |

a. Use a 95% confidence interval to compare the mean zinc measurement for the text line with the mean for the intersection.
b. Use a 95% confidence interval to compare the mean zinc measurement for the witness line with the mean for the intersection.
c. From the results, parts **a** and **b**, what can you infer about the mean zinc measurements at the three notebook locations?
d. What assumptions are required for the inferences to be valid? Are they reasonably satisfied?

7.45 *Producer willingness to supply biomass.* The conversion of biomass to energy is critical for producing transportation fuels. How willing are producers to supply biomass products such as cereal straw, corn stover and surplus hay? To answer this question, researchers conducted a survey of producers in both mid-Missouri and southern Illinois. (*Biomass and Energy*, Vol. 36, 2012.) Independent samples of 431 Missouri producers and 508 Illinois producers participated in the survey. Each producer was asked to give the maximum proportion of hay produced that they would be willing to sell to the biomass market. Summary statistics for the two groups of producers are listed in the table. Does the mean amount of surplus hay producers willing to sell to the biomass market differ for the two areas, Missouri and Illinois? Use a 95% confidence interval to make the comparison.

| | Missouri producers | Illinois producers |
|---|---|---|
| Sample size: | 431 | 508 |
| Mean amount of hay (%): | 21.5 | 22.2 |
| Standard deviation (%): | 33.4 | 34.9 |

*Source:* Altman, I. & Sanders, D. "Producer willingness and ability to supply biomass: Evidence from the U.S. Midwest", *Biomass and Energy*, Vol. 36, No. 8, 2012 (Tables 3 and 7).

7.46 *Process voltage readings.* Refer to the Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally, Exercise 2.72 (p. 70). Test devices (pilots) were set up at both the old and new locations and voltage readings on 30 production runs at each location were obtained. The data are reproduced in the table. Descriptive statistics are displayed in the SAS printout at the bottom of the page. (*Note:* Larger voltage readings are better than smaller voltage readings.)

a. Compare the mean voltage readings at the two locations using a 90% confidence interval.
b. Based on the interval, part **a**, does it appear that the manufacturing process can be established locally?

**VOLTAGE**

| Old Location | | | New Location | | |
|---|---|---|---|---|---|
| 9.98 | 10.12 | 9.84 | 9.19 | 10.01 | 8.82 |
| 10.26 | 10.05 | 10.15 | 9.63 | 8.82 | 8.65 |
| 10.05 | 9.80 | 10.02 | 10.10 | 9.43 | 8.51 |
| 10.29 | 10.15 | 9.80 | 9.70 | 10.03 | 9.14 |
| 10.03 | 10.00 | 9.73 | 10.09 | 9.85 | 9.75 |
| 8.05 | 9.87 | 10.01 | 9.60 | 9.27 | 8.78 |
| 10.55 | 9.55 | 9.98 | 10.05 | 8.83 | 9.35 |
| 10.26 | 9.95 | 8.72 | 10.12 | 9.39 | 9.54 |
| 9.97 | 9.70 | 8.80 | 9.49 | 9.48 | 9.36 |
| 9.87 | 8.72 | 9.84 | 9.37 | 9.64 | 8.68 |

*Source:* Harris Corporation, Melbourne, Fla.

7.47 *Converting powders to solids. Sintering*, one of the most important techniques of materials science, is used to convert powdered material into a porous solid body. The following two measures characterize the final product:

**SAS Output for Exercise 7.46**

```
Sample Statistics

     Group          N       Mean     Std. Dev.    Std. Error
     -------------------------------------------------------
     NEW           30   9.422333       0.4789        0.0874
     OLD           30   9.803667       0.5409        0.0988


Hypothesis Test

     Null hypothesis:     Mean 1 - Mean 2 =  0
     Alternative:         Mean 1 - Mean 2 ^= 0

     If Variances Are     t statistic      Df        Pr > t
     -------------------------------------------------------
     Equal                   -2.891         58        0.0054
     Not Equal               -2.891      57.16        0.0054


90% Confidence Interval for the Difference between Two Means

                 Lower Limit    Upper Limit
                 -----------    -----------
                    -0.60          -0.16
```

$V_V$ = Percentage of total volume of final product that is solid

$$= \left( \frac{\text{Solid volume}}{\text{Porous volume} + \text{Solid volume}} \right) \cdot 100$$

$S_V$ = Solid–pore interface area per unit volume of the product

When $V_V = 100\%$, the product is completely solid—i.e., it contains no pores. Both $V_V$ and $S_V$ are estimated by a microscopic examination of polished cross sections of sintered material. The accompanying table gives the mean and standard deviation of the values of $S_V$ (in squared centimeters per cubic centimeter) and $V_V$ (percentage) for $n = 100$ specimens of sintered nickel for two different sintering times.

| | $S_V$ | | $V_V$ | |
|---|---|---|---|---|
| Time | $\bar{y}$ | $s$ | $\bar{y}$ | $s$ |
| 10 minutes | 736.0 | 181.9 | 96.73 | 2.1 |
| 150 minutes | 299.5 | 161.0 | 97.82 | 1.5 |

Data and experimental information provided by Guoquan Liu while visiting at the University of Florida.

a. Find a 95% confidence interval for the mean change in $S_V$ between sintering times of 10 minutes and 150 minutes. What inference would you make concerning the difference in mean sintering times?
b. Repeat part **a** for $V_V$.

## 7.6 Estimation of the Difference Between Two Population Means: Matched Pairs

The large- and small-sample procedures for estimating the difference between two population means presented in Section 7.5 were based on the assumption that the samples were randomly and independently selected from the target populations. Sometimes we can obtain more information about the difference between population means, $(\mu_1 - \mu_2)$, by selecting **paired observations**.

For example, suppose you want to compare two methods for drying concrete using samples of five cement mixes with each method. One method of sampling would be to randomly select 10 mixes (say, A, B, C, D, . . . , J) from among all available mixes and then randomly assign 5 to drying method 1 and 5 to drying method 2 (see Table 7.5). The strength measurements obtained after conducting a series of strength tests would represent independent random samples of strengths attained by concrete specimens dried by the two different methods. The difference between the mean strength measurements, $(\mu_1 - \mu_2)$, could be estimated using the confidence interval procedure described in Section 7.5.

**TABLE 7.5 Independent Random Samples of Cement Mixes Assigned to Each Method**

| Method 1 | Method 2 |
|---|---|
| Mix A | Mix B |
| Mix E | Mix C |
| Mix F | Mix D |
| Mix H | Mix G |
| Mix J | Mix I |

A better method of sampling would be to match the concrete specimens in pairs according to type of mix. From each mix pair, one specimen would be randomly selected to be dried by method 1; the other specimen would be assigned to be dried by method 2, as shown in Table 7.6. Then the differences between **matched pairs** of strength measurements should provide a clearer picture of the difference in strengths for the two drying methods because the matching would tend to cancel the effects of the factors that formed the basis of the matching (i.e., the effects of the different cement mixes).

**TABLE 7.6 Setup of the Matched-Pairs Design for Comparing Two Methods of Drying Concrete**

| Type of Mix | Method 1 | Method 2 |
|---|---|---|
| A | Specimen 2 | Specimen 1 |
| B | Specimen 2 | Specimen 1 |
| C | Specimen 1 | Specimen 2 |
| D | Specimen 2 | Specimen 1 |
| E | Specimen 1 | Specimen 2 |

In a matched-pairs experiment, the symbol $\mu_d$ is commonly used to denote the mean difference between matched pairs of measurements, where $\mu_d = (\mu_1 - \mu_2)$. Once the differences in the sample are calculated, a confidence interval for $\mu_d$ is identical to the confidence interval for the mean of a single population given in Section 7.4.

The procedure for estimating the difference between two population means based on matched-pairs data for both large and small samples is given in the box.

**$(1 - \alpha)100\%$ Confidence Interval for $\mu_d = (\mu_1 - \mu_2)$: Matched Pairs**

Let $d_1, d_2, \ldots, d_n$ represent the differences between the pairwise observations in a *random sample* of $n$ matched pairs, $\bar{d}$ = mean of the $n$ sample differences, and $s_d$ = standard deviation of the $n$ sample differences.

| Large Sample | Small Sample |
|---|---|
| $\bar{d} \pm z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)$ | $\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$ |
| where $\sigma_d$ is the population deviation of differences. | where $t_{\alpha/2}$ is based on $(n - 1)$ degrees of freedom. |
| *Assumption*: $n \geq 30$ | *Assumption*: The population of paired differences is normally distributed. |

[*Note:* When $\sigma_d$ is unknown (as is usually the case), use $s_d$ to approximate $\sigma_d$.]

**Example 7.11**

Matched Pairs Confidence Interval — Driver Reaction Time

A federal traffic safety engineer wants to ascertain the effect of wearing safety devices (shoulder harnesses, seat belts) on reaction times to peripheral stimuli. A study was designed as follows: A random sample of 15 student drivers was selected from students enrolled in a driver-education program. Each driver performed a simulated driving task that allowed reaction times to be recorded under two conditions, wearing a safety device (restrained condition) and wearing no safety device (unrestrained condition). Thus, each student driver received two reaction-time scores, one for the restrained condition and one for the unrestrained condition. The data (in hundredths of a second) are shown in Table 7.7 and saved in the **SAFETY** file. Find and interpret a

**SAFETY**

**TABLE 7.7 Reaction time data for Example 7.11**

| | Condition | | |
|---|---|---|---|
| Driver | Restrained | Unrestrained | Difference |
| 1 | 36.7 | 36.1 | 0.6 |
| 2 | 37.5 | 35.8 | 1.7 |
| 3 | 39.3 | 38.4 | 0.9 |
| 4 | 44.0 | 41.7 | 2.3 |
| 5 | 38.4 | 38.3 | 0.1 |
| 6 | 43.1 | 42.6 | 0.5 |
| 7 | 36.2 | 33.6 | 2.6 |
| 8 | 40.6 | 40.9 | -0.3 |
| 9 | 34.9 | 32.5 | 2.4 |
| 10 | 31.7 | 30.7 | 1.0 |
| 11 | 37.5 | 37.4 | 0.1 |
| 12 | 42.8 | 40.2 | 2.6 |
| 13 | 32.6 | 33.1 | -0.5 |
| 14 | 36.8 | 33.6 | 3.2 |
| 15 | 38.0 | 37.5 | 0.5 |

95% confidence interval for the difference between the mean reaction time scores of restrained and unrestrained drivers.

Solution: Since each of the student drivers performed the simulated driving task under both conditions, the data are collected as matched pairs. Each student in the sample represents one of the 15 matched pairs. We want to estimate $\mu_d = (\mu_1 - \mu_2)$, where

$\mu_1 =$ mean reaction time of all drivers in the restrained condition

$\mu_2 =$ mean reaction time of all drivers in the unrestrained condition

The differences between pairs of reaction times are computed as

$d =$ (Restrained reaction time) – (Unrestrained reaction time)

and are also shown in Table 7.7. Now the number of differences, $n = 15$, is small; consequently, we must assume that these differences are from an approximately normal distribution in order to proceed.

The mean and standard deviation of these sample differences are shown (highlighted) on the MINITAB printout, Figure 7.11. From the printout, $\bar{d} = 1.18$ and $s_d = 1.19$. The value of $t_{.025}$, based on $(n - 1) = 14$ degrees of freedom, is given in Table 7 of Appendix B as $t_{.025} = 2.145$. Substituting these values into the formula for the small-sample confidence interval, we obtain

$$\bar{d} \pm t_{.025}(s_d/\sqrt{n}) = 1.18 \pm 2.145(1.19/\sqrt{15}) = 1.18 \pm .66 = (.52, 1.84)$$

Note that this interval is also shown (highlighted) on the printout, Figure 7.11.

We estimate with 95% confidence that $\mu_d = (\mu_1 - \mu_2)$, the difference between the mean reaction times of students in the restrained and unrestrained conditions, falls between .52 and 1.84 hundredths of a second. Since all values in the interval are positive, we can infer that the mean reaction time ($\mu_1$) of students in the restrained condition is anywhere from .52 to 1.84 hundredths of a second higher than the mean reaction time ($\mu_2$) of students in the unrestrained condition.

**Paired T-Test and CI: REAC-R, REAC-U**

```
Paired T for REAC-R - REAC-U

             N    Mean  StDev  SE Mean
REAC-R      15  38.007  3.576    0.923
REAC-U      15  36.827  3.616    0.934
Difference  15   1.180  1.191    0.307


95% CI for mean difference: (0.521, 1.839)
T-Test of mean difference = 0 (vs not = 0): T-Value = 3.84  P-Value = 0.002
```

**FIGURE 7.11**
MINITAB Printout of Matched-Paired Analysis, Example 7.11

In an analysis of matched-pair observations, it is important to stress that the pairing of the experimental units (the objects upon which the measurements are paired) must be performed *before* the data are collected. By using the matched pairs of units that have similar characteristics, we are able to cancel out the effects of the variables used to match the pairs. On the other hand, if you collect the data as matched pairs but employ a statistical method of analysis that does not account for the matching (e.g., a confidence interval for $\mu_1 - \mu_2$ based on *independent* samples), the characteristics of the matched pairs will not be cancelled out. This will typically result in a wider confidence interval, and, consequently, a potentially invalid inference. We illustrate this last point in the next example.

**Example 7.12**

**Independent Samples Confidence Interval Applied to Match-Paired Data**

Refer to the driver reaction time study, Example 7.11. Although the data was collected as matched pairs, suppose a researcher mistakenly analyzes the data using an independent (small) samples, 95% confidence interval for $\mu_1 - \mu_2$ . This confidence interval is shown on the MINITAB printout, Figure 7.12. Locate and interpret the interval. Explain why the results are misleading.

Solution: The formula for a small-sample 95% confidence interval for $(\mu_1 - \mu_2)$ using independent samples is (from Section 7.5):

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This interval, highlighted on Figure 7.12, is $(-1.51, 3.87)$. Note that the value 0 falls within the interval. This implies that there is insufficient evidence of a difference between the mean reaction times of restrained and unrestrained drivers — an inference we know to be invalid. The problem results from a comparison of the standard errors used in the matched-paired analysis (Example 7.11) and the independent samples analysis (Example 7.12). Using values shown on the two printouts, Figures 7.11 and 7.12, the standard errors are calculated as follows:

Standard error, matched pairs: $s_d/\sqrt{n} = 1.19/\sqrt{15} = .307$

Standard error, independent samples:

$$\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{14(3.58)^2 + 14(3.62)^2}{28}\left(\frac{1}{15} + \frac{1}{15}\right)} = 1.315$$

You can see that the independent samples standard error is much larger than the corresponding value for matched pairs. The greater variation in the independent samples analysis results from failing to account for the driver-to-driver variation in the matched pairs. Because this variation is not cancelled out, the consequence is a wider independent samples confidence interval — one which ultimately leads to a potentially erroneous conclusion.

**Two-Sample T-Test and CI: REAC-R, REAC-U**

```
Two-sample T for REAC-R vs REAC-U

        N   Mean  StDev  SE Mean
REAC-R  15  38.01   3.58     0.92
REAC-U  15  36.83   3.62     0.93


Difference = mu (REAC-R) - mu (REAC-U)
Estimate for difference:  1.18
95% CI for difference:  (-1.51, 3.87)
T-Test of difference = 0 (vs not =): T-Value = 0.90  P-Value = 0.377  DF = 27
```

**FIGURE 7.12**
MINITAB Printout of Independent-Samples Analysis of Match-Paired Data, Example 7.12

## Applied Exercises

7.48 *Device designed for obstetric delivery.* A team of biomedical engineers designed a prototype device for assisting in obstetric delivery (*International Journal for Service Learning in Engineering*, Fall 2012). To determine the effectiveness of the device, five medical students were recruited to participate in the study. Initially, each student was asked to deliver a model "baby" from a birthing simulator using the prototype device without any training on its use. A pre-training score was assigned based on the time taken to deliver the model, with penalties incurred for errors during the procedure. A higher score (the maximum score was 900 points) indicated a more proficient use of the prototype. Next, each student participated in a 30-minute training workshop on the use of the prototype device. Following the workshop, the students repeated the birthing test and a post-training delivery score was assigned to each. The team of biomedical engineers want to estimate the average increase in delivery score following training.

a. What is the parameter of interest to the engineers?
b. Give details on the design used to collect the data.
c. The following information was provided in the article: $\bar{y}_{\text{Pre}} = 481.8$, $\bar{y}_{\text{Post}} = 712.4$. Is this sufficient information to find the estimate of the parameter of interest? If so, give the estimate?
d. The following additional information was provided in the article: $s_{\text{Pre}} = 99.1$, $s_{\text{Post}} = 31.0$. Is this sufficient information (along with the information in part c) to find a 95% confidence interval for the parameter of interest? If so, give the interval estimate? If not, what further information do you need?

7.49 *Twinned drill holes.* A traditional method of verifying mineralization grades in mining is to drill twinned holes, i.e., the drilling of a new hole, or "twin", next to an earlier drillhole. The use of twinned drill holes was investigated in *Exploration and Mining Geology* (Vol. 18, 2009). Geologists use data collected at both holes to estimate the total amount of heavy minerals (THM) present at the drilling site. The data in the next table (based on information provided in the journal article) represent THM percentages for a sample of 15 twinned holes drilled at a diamond mine in Africa. The geologists want to know if there is any evidence of a difference in the true THM means of all original holes and their twin holes drilled at the mine.

a. Explain why the data should be analyzed as paired differences.
b. Compute the difference between the "1st hole" and "2nd hole" measurements for each drilling location.
c. Find the mean and standard deviation of the differences, part b.
d. Use the summary statistics, part c, to find a 90% confidence interval for the true mean difference ("1st hole" minus "2nd hole") in THM measurements.
e. Interpret the interval, part d. Can the geologists conclude that there is no evidence of a difference in the true THM means of all original holes and their twin holes drilled at the mine?

**TWINHOLE**

| Location | 1st Hole | 2nd Hole |
|---|---|---|
| 1 | 5.5 | 5.7 |
| 2 | 11.0 | 11.2 |
| 3 | 5.9 | 6.0 |
| 4 | 8.2 | 5.6 |
| 5 | 10.0 | 9.3 |
| 6 | 7.9 | 7.0 |
| 7 | 10.1 | 8.4 |
| 8 | 7.4 | 9.0 |
| 9 | 7.0 | 6.0 |
| 10 | 9.2 | 8.1 |
| 11 | 8.3 | 10.0 |
| 12 | 8.6 | 8.1 |
| 13 | 10.5 | 10.4 |
| 14 | 5.5 | 7.0 |
| 15 | 10.0 | 11.2 |

7.50 *Settlement of shallow foundations.* Structures built on a shallow foundation (e.g., a concrete slab-on-grade foundation) are susceptible to settlement. Consequently, accurate settlement prediction is essential in the design of the foundation. Several methods for predicting settlement of shallow foundations on cohesive soil were com-

**SHALLOW**

| Structure | Actual | Predicted |
|---|---|---|
| 1 | 11 | 11 |
| 2 | 11 | 11 |
| 3 | 10 | 12 |
| 4 | 8 | 6 |
| 5 | 11 | 9 |
| 6 | 9 | 10 |
| 7 | 9 | 9 |
| 8 | 39 | 51 |
| 9 | 23 | 24 |
| 10 | 269 | 252 |
| 11 | 4 | 3 |
| 12 | 82 | 68 |
| 13 | 250 | 264 |

*Source:* Ozur, M. "Comparing Methods for Predicting Immediate Settlement of Shallow Foundations on Cohesive Soils Based on Hypothetical and Real Cases", *Environmental & Engineering Geoscience*, Vol. 18, No. 4, November 2012 (from Table 4).

pared in *Environmental & Engineering Geoscience* (Nov., 2012). Settlement data for a sample of 13 structures built on a shallow foundation were collected. (These structures included office buildings, bridge piers, and concrete test plates.) The actual settlement values (measured in millimeters) for each structure were compared to settlement predictions made using a formula that accounts for dimension, rigidity, and embedment depth of the foundation. The data are listed in the table on p. 326.

a. What type of design was employed to collect the data?
b. Construct a 99% confidence interval for the mean difference between actual and predicted settlement value. Give a practical interpretation of the interval.
c. Explain the meaning of "99% confidence" for this application.

7.51 *Acidity of mouthwash.* Acid has been found to be a primary cause of dental caries (cavities). It is theorized that oral mouthwashes contribute to the development of caries due to the antiseptic agent oxidizing into acid over time. This theory was tested in the *Journal of Dentistry, Oral Medicine and Dental Education* (Vol. 3, 2009). Three bottles of mouthwash, each of a different brand, were randomly selected from a drug store. The pH level (where lower pH levels indicate higher acidity) of each bottle was measured on the date of purchase and after 30 days. The data are shown in the table. Use a 95% confidence interval to determine if the mean initial pH level of mouthwash differs significantly from the mean pH level after 30 days.

**MOUTHWASH**

| Mouthwash Brand | Initial pH | Final pH |
|---|---|---|
| LMW | 4.56 | 4.27 |
| SMW | 6.71 | 6.51 |
| RMW | 5.65 | 5.58 |

*Source:* Chunhye, K.L. & Schmitz, B.C. "Determination of pH, total acid, and total ethanol in oral health products: Oxidation of ethanol and recommendations to mitigate its association with dental caries", *Journal of Dentistry, Oral Medicine and Dental Education*, Vol. 3, No. 1, 2009 (Table 1).

7.52 *Testing electronic circuits.* Japanese researchers have developed a compression/depression method of testing electronic circuits based on Huffman coding. (*IEICE Transactions on Information & Systems*, Jan. 2005.) The new method is designed to reduce the time required for input decompression and output compression—called the compression ratio. Experimental results were obtained by testing a sample of 11 benchmark circuits (all of different sizes) from a SUN Blade 1000 workstation. Each circuit was tested using the standard compression/depression method and the new Huffman-based coding method, and the compression ratio was recorded. The data are given in the next table. Compare the two methods with a 95% confidence interval. Which method has the smaller mean compression ratio?

**CIRCUITS**

| Circuit | Standard Method | Huffman Coding Method |
|---|---|---|
| 1 | .80 | .78 |
| 2 | .80 | .80 |
| 3 | .83 | .86 |
| 4 | .53 | .53 |
| 5 | .50 | .51 |
| 6 | .96 | .68 |
| 7 | .99 | .82 |
| 8 | .98 | .72 |
| 9 | .81 | .45 |
| 10 | .95 | .79 |
| 11 | .99 | .77 |

*Source:* Ichihara, H., Shintani, M., and Inoue, T. "Huffman-based test response coding." *IEICE Transactions on Information & Systems*, Vol. E88-D, No. 1, Jan. 2005 (Table 3).

7.53 *Exposure to low-frequency sound. Infrasound* refers to sound waves or vibrations with a frequency below the audibility range of the human ear. Even though infrasound cannot be heard, it can produce physiological effects in humans. Mechanical science engineers in China studied the impact of infrasound on a person's blood pressure and heart rate (*Journal of Low Frequency Noise, Vibration and Active Control*, Mar. 2004). Six university students were exposed to infrasound for 1 hour. The table gives the blood pressure and heart rate for each student, both before and after exposure.

**INFRASOUND**

| | Systolic Pressure (mm Hg) | | Diastolic Pressure (mm Hg) | | Heart Rate (beats/min) | |
|---|---|---|---|---|---|---|
| Student | Before | After | Before | After | Before | After |
| 1 | 105 | 118 | 60 | 73 | 70 | 70 |
| 2 | 113 | 129 | 60 | 73 | 69 | 80 |
| 3 | 106 | 117 | 60 | 79 | 76 | 84 |
| 4 | 126 | 134 | 79 | 86 | 77 | 86 |
| 5 | 113 | 115 | 73 | 66 | 64 | 76 |

*Source:* Qibai, C. Y. H., and Shi, H. "An investigation on the physiological and psychological effects of infrasound on persons." *Journal of Low Frequency Noise, Vibration and Active Control*, Vol. 23, No. 1, Mar. 2004 (Table V).

a. Compare the before and after mean systolic blood pressure readings using a 99% confidence interval. Interpret the result.
b. Compare the before and after mean diastolic blood pressure readings using a 99% confidence interval. Interpret the result.
c. Compare the before and after mean heart rate readings using a 99% confidence interval. Interpret the result.

**CRASH**

7.54 *NHTSA new-car crash tests.* Each year the National Highway Traffic Safety Administration (NHTSA) conducts crash tests for new cars. Crash-test dummies are placed in the driver's seat and front passenger's seat of a new-car model, and the car is steered by remote control into a head-on collision with a fixed barrier while traveling at 35 miles per hour. The results for 98 new cars are saved in the **CRASH** file. Two of the variables measured for each car in the data set are (1) the severity of the driver's chest injury and (2) the severity of the passenger's chest injury. (The more points assigned to the chest injury rating, the more severe the injury.) Suppose the NHTSA wants to determine whether the true mean driver chest injury rating exceeds the true mean passenger chest injury rating, and if so, by how much.

a. State the parameter of interest to the NHTSA.
b. Explain why the data should be analyzed as matched pairs.
c. Find a 99% confidence interval for the true difference between the mean chest injury ratings of drivers and front-seat passengers.
d. Interpret the interval, part **c**. Does the true mean driver chest injury rating exceed the true mean passenger chest injury rating? If so, by how much?
e. What conditions are required for the analysis to be valid? Do these conditions hold for this data?

7.55 *Alcoholic fermentation in wines.* Determining alcoholic fermentation in wine is critical to the wine-making process. Must/wine density is a good indicator of the fermentation point since the density value decreases as sugars are converted into alcohol. For decades, winemakers have measured must/wine density with a hydrometer. Although accurate, the hydrometer employs a manual process that is very time-consuming. Consequently, large wineries are searching for more rapid measures of density measurement. An alternative method utilizes the hydrostatic balance instrument (similar to the hydrometer, but digital). A winery in Portugal collected the must/wine density measurements for white wine samples randomly selected from the fermentation process for a recent harvest. For each sample, the density of the wine at 20°C was measured with both the hydrometer and the hydrostatic balance. The densities for 40 wine samples are saved in the **WINE40** file. (The first five and last five observations are shown in the accompanying table.) The winery will use the alternative method of measuring wine density only if it can be demonstrated that the mean difference between the density measurements of the two methods does not exceed .002. Perform the analysis for the winery and give your recommendation.

**WINE40**

| Sample | Hydrometer | Hydrostatic |
|---|---|---|
| 1 | 1.08655 | 1.09103 |
| 2 | 1.00270 | 1.00272 |
| 3 | 1.01393 | 1.01274 |
| 4 | 1.09467 | 1.09634 |
| 5 | 1.10263 | 1.10518 |
| ⋮ | ⋮ | ⋮ |
| 36 | 1.08084 | 1.08097 |
| 37 | 1.09452 | 1.09431 |
| 38 | 0.99479 | 0.99498 |
| 39 | 1.00968 | 1.01063 |
| 40 | 1.00684 | 1.00526 |

*Source:* Cooperative Cellar of Borba (*Adega Cooperative de Borba*), Portugal.

7.56 *Impact of red light cameras on car crashes.* To combat red-light-running crashes — the phenomenon of a motorist entering an intersection after the traffic signal turns red and causing a crash — many states are adopting photo-red enforcement programs. In these programs, red light cameras installed at dangerous intersections photograph the license plates of vehicles that run the red light. How effective are photo-red enforcement programs in reducing red-light-running crash incidents at intersections? The Virginia Department of Transportation (VDOT) conducted a comprehensive study of its newly adopted photo-red enforcement program and published the results in a report. In one portion of the study, the VDOT provided crash data both before and after installation of red light cameras at several intersections. The data (measured as the number of crashes caused by red light running per intersection per year) for 13 intersections in Fairfax County, VA are given in the table. Analyze the data for the VDOT. What do you conclude?

**REDLIGHT**

| Intersection | Before Camera | After Camera |
|---|---|---|
| 1 | 3.60 | 1.36 |
| 2 | 0.27 | 0 |
| 3 | 0.29 | 0 |
| 4 | 4.55 | 1.79 |
| 5 | 2.60 | 2.04 |
| 6 | 2.29 | 3.14 |
| 7 | 2.40 | 2.72 |
| 8 | 0.73 | 0.24 |
| 9 | 3.15 | 1.57 |
| 10 | 3.21 | 0.43 |
| 11 | 0.88 | 0.28 |
| 12 | 1.35 | 1.09 |
| 13 | 7.35 | 4.92 |

*Source:* Virginia Transportation Research Council, "Research Report: The Impact of Red Light Cameras (Photo-Red Enforcement) on Crashes in Virginia", June 2007.

## 7.7 Estimation of a Population Proportion

We will now consider the method for estimating the binomial proportion $p$ of successes—that is, the proportion of elements in a population that have a certain characteristic. For example, a quality control inspector may be interested in the proportion of defective items produced on an assembly line, or a supplier of heating oil may be interested in the proportion of homes in its service area that are heated by natural gas.

A logical candidate for a point estimate of the population proportion $p$ is the sample proportion $\hat{p} = y/n$, where $y$ is the number of observations in a sample of size $n$ that have the characteristic of interest (i.e., the random variable $Y$ is the number of "successes" in a binomial experiment). In Example 6.23, we showed that for large $n$, $\hat{p}$ is approximately normal with mean

$$E(\hat{p}) = p$$

and variance

$$V(\hat{p}) = \frac{pq}{n}$$

Therefore, $\hat{p}$ is an unbiased estimator of $p$ and (proof omitted) has the smallest variance among all unbiased estimators; that is, $\hat{p}$ is the MVUE for $p$. Since $\hat{p}$ is approximately normal, we can use it as a pivotal statistic and apply Theorem 7.2 to derive the formula for a large-sample confidence interval for $p$ shown in the box.

**Large-Sample $(1 - \alpha)100\%$ Confidence Interval for a Population Proportion, $p$**

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p}$ is the sample proportion of observations with the characteristic of interest, and $\hat{q} = 1 - \hat{p}$.
(*Note:* The interval is approximate since we must substitute the sample $\hat{p}$ and $\hat{q}$ for the corresponding population values for $\sigma_{\hat{p}}$ .)

*Assumption:* The sample size $n$ is sufficiently large so that the approximation is valid. As a rule of thumb, the condition of a "sufficiently large" sample size will be satisfied if $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$.

Note that we must substitute $\hat{p}$ and $\hat{q}$ into the formula for $\sigma_{\hat{p}} = \sqrt{pq/n}$ to construct the interval. This approximation will be valid as long as the sample size $n$ is sufficiently large. Many researchers adopt the rule of thumb that $n$ is "sufficiently large" if the interval $\hat{p} \pm 2\sqrt{\hat{p}\hat{q}/n}$ does not contain 0 or 1. Recall (Section 6.10) that this rule is satisfied if $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$.

**Example 7.13**

Confidence Interval for a proportion: Alloy Failure Rate

Stainless steels are frequently used in chemical plants to handle corrosive fluids. However, these steels are especially susceptible to stress corrosion cracking in certain environments. In a sample of 295 steel alloy failures that occurred in oil refineries and petrochemical plants in Japan, 118 were caused by stress corrosion cracking and corrosion fatigue. Construct a 95% confidence interval for the true proportion of alloy failures caused by stress corrosion cracking.

Solution The sample proportion of alloy failures caused by corrosion is

$$\hat{p} = \frac{\text{Number of alloy failures in sample caused by corrosion}}{\text{Number of alloy failures in sample}}$$

$$= \frac{118}{295} = .4$$

Thus, $\hat{q} = 1 - .4 = .6$. The approximate 95% confidence interval is then

$$\hat{p} \pm z_{.025}\sqrt{\frac{\hat{p}\hat{q}}{n}} = .4 \pm 1.96\sqrt{\frac{(.4)(.6)}{295}} = .4 \pm .056$$

or (.344, .456). (Note that the approximation is valid since $n\hat{p} = 118$ and $n\hat{q} = 177$ both exceed 4.)

We are 95% confident that the interval from .344 to .456 encloses the true proportion of alloy failures that were caused by corrosion. If we repeatedly selected random samples of $n = 295$ alloy failures and constructed a 95% confidence interval based on each sample, then we would expect 95% of the confidence intervals constructed to contain $p$.

Small-sample procedures are available for the estimation of a population proportion $p$. These techniques are similar to those small-sample procedures for estimating a population mean $\mu$. (Recall that $\hat{p} = y/n$ can be thought of as a mean of a sample of 0–1 Bernoulli outcomes.) The details are not included in our discussion, however, because most surveys in actual practice use samples that are large enough to employ the procedure of this section.

## Applied Exercises

7.57 *Cell phone use by drivers.* Studies have shown that drivers who use cell phones while operating a motor passenger vehicle increase their risk of an accident. Nevertheless, drivers continue to make cell phone calls while driving. A 2011 *Harris Poll* of 2,163 adults found that 60% (1,298 adults) use cell phones while driving.

a. Give a point estimate of $p$, the true driver cell phone use rate (i.e., the proportion of all drivers who are using a cell phone while operating a motor passenger vehicle).
b. Find a 95% confidence interval for $p$.
c. Give a practical interpretation of the interval, part b.

7.58 *Microsoft program security issues.* Refer to the *Computers & Security* (July 2013) study of security issues with Microsoft products, Exercise 2.4 (p. 27). Recall that Microsoft periodically issues a Security Bulletin that reports the software affected by the vulnerability. In a sample of 50 bulletins issued in a recent year, 32 reported a security problem with Microsoft Windows.

a. Find a point estimate of the proportion of security bulletins issued that reported a problem with Windows during the year.
b. Find an interval estimate for the proportion, part a. Use a 90% confidence interval.
c. Practically interpret the confidence interval, part b. Your answer should begin with, "We are 90% confident ... ".
d. Give a theoretical interpretation of the phrase "90% confident".

**ASWELLS**

7.59 *Arsenic in groundwater. Environmental Science & Technology* (Jan. 2005) reported on a study of the reliability of a commercial kit to test for arsenic in groundwater. The field kit was used to test a sample of 328 groundwater wells in Bangladesh. If the color indicator on the field kit registers red, the level of arsenic in the water is estimated to be at least 50 micrograms per liter; if the color registers green, the arsenic level is estimated to be below 50 micrograms per liter. The data for the study is saved in the **ASWELLS** file. A summary of the results of the arsenic tests is displayed in the MINITAB printout below. Use the information to find a 90% confidence interval for the true proportion of all groundwater wells in Bangladesh that have an estimated arsenic level below 50 micrograms per liter. Give a practical interpretation of the interval.

**Tally for Discrete Variables: KIT-COLOR**

| KIT-COLOR | Count | Percent |
|---|---|---|
| Green | 178 | 54.27 |
| Red | 150 | 45.73 |
| N= | 328 | |

7.60 *Annual survey of computer crimes.* Refer to the *CSI Computer Crime and Security Survey*, first presented in Exercise 1.19 (p. 15). Recall that of the 351 organizations that responded to the survey, 144 (or 41%) admitted unauthorized use of computer systems at their firms during the year. Estimate the probability of unauthorized use of computer systems at an organization with a 90% confidence interval. Explain how 90% is used as a measure of reliability for the confidence interval.

7.61 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010), study of the trend in the design of social robots, Exercise 7.33 (p. 313). The researchers obtained a random sample of 106 social robots through a web search and determined that 63 were designed with legs, but no wheels.

a. Find a 99% confidence interval for the proportion of all social robots designed with legs, but no wheels. Interpret the result.
b. Is it valid to assume that in the population of all social robots, 40% are designed with legs, but no wheels? Explain.

7.62 *Material safety data sheets.* The Occupational Safety & Health Administration requires companies that handle hazardous chemicals to complete material safety data sheets (MSDS). These MSDS have been criticized for being too hard to understand and complete by workers. A study of 150 MSDS revealed that only 11% were satisfactorily completed. (*Chemical & Engineering News*, Feb. 7, 2005.) Give a 95% confidence interval for the true percentage of MSDS that are satisfactorily completed.

7.63 *Study of aircraft bird-strikes.* As world-wide air traffic volume has grown over the years, the problem of airplanes striking birds and other flying wildlife has increased dramatically. The *International Journal for Traffic and Transport Engineering* (Vol. 3, 2013) reported on a study of aircraft bird strikes at Aminu Kano International Airport in Nigeria. During the survey period, a sample of 44 aircraft bird strikes were analyzed. The researchers found that 36 of the 44 bird strikes at the airport occurred above 100 feet. Suppose an airport air traffic controller estimates that less than 70% of aircraft bird strikes occur above 100 feet. Comment on the accuracy of this estimate. Use a 95% confidence interval to support your inference.

7.64 *Estimating the age of glacial drifts.* Refer to the *American Journal of Science* (Jan. 2005) study of the chemical makeup of buried glacial drifts (or tills) in Wisconsin, Exercise 2.22 (p. 38). The ratio of aluminum (Al) to beryllium (Be) in sediment for each of a sample of 26 buried till specimens is given in the table at the bottom of the page.

a. Recall that the researchers desired an estimate of the proportion of till specimens in Wisconsin with an Al/Be ratio that exceeds 4.5. Compute this estimate from the sample data.
b. Form a 95% confidence interval around the estimate, part **a**. Interpret the interval.

7.65 *Orientation cues for astronauts.* Astronauts often report episodes of disorientation as they move around the zero-gravity spacecraft. To compensate, crew members rely heavily on visual information to establish a top-down orientation. An empirical study was conducted to assess the potential of using color brightness as a body orientation cue (*Human Factors*, Dec. 1988). Ninety college students, reclining on their backs in the dark, were disoriented when positioned on a rotating platform under a slowly rotating disk that filled their entire field of vision. Half the disk was painted with a brighter level of color than the other half. The students were asked to say "stop" when they believed they were right-side-up, and the brightness level of the disk was recorded. Of the 90 students, 58 selected the brighter color level.

a. Use this information to estimate the true proportion of subjects who use the bright color level as a cue to being right-side-up. Construct a 95% confidence interval for the true proportion.
b. Can you infer from the result, part **a**, that a majority of subjects would select bright color levels over dark color levels as a cue to being right-side-up? Explain.

**TILLRATIO**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.75 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 |
| 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 | 2.73 | 2.95 | 2.25 | 2.73 | 2.55 | 3.06 |

*Source:* Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

## 7.8 Estimation of the Difference Between Two Population Proportions

This section extends the method of Section 7.7 to the case in which we want to estimate the difference between two binomial proportions. For example, we may be interested in comparing the proportion $p_1$ of defective items produced by machine 1 to the proportion $p_2$ of defective items produced by machine 2.

Let $y_1$ and $y_2$ represent the numbers of successes in two independent binomial experiments with samples of size $n_1$ and $n_2$, respectively. To estimate the difference $(p_1 - p_2)$, where $p_1$ and $p_2$ are binomial parameters—i.e., the probabilities of success in the two independent binomial experiments—consider the proportion of successes in each of the samples:

$$\hat{p}_1 = \frac{y_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{y_2}{n_2}$$

Intuitively, we would expect $(\hat{p}_1 - \hat{p}_2)$ to provide a reasonable estimate of $(p_1 - p_2)$. Since $(\hat{p}_1 - \hat{p}_2)$ is a linear function of the binomial random variables $Y_1$ and $Y_2$, where $E(y_i) = n_i p_i$ and $V(y_i) = n_i p_i q_i$, we have

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= E(\hat{p}_1) - E(\hat{p}_2) = E\left(\frac{y_1}{n_1}\right) - E\left(\frac{y_2}{n_2}\right) \\ &= \frac{1}{n_1}E(y_1) - \frac{1}{n_2}E(y_2) = \frac{1}{n_1}(n_1 p_1) - \frac{1}{n_2}(n_2 p_2) \\ &= p_1 - p_2 \end{aligned}$$

and

$$\begin{aligned} V(\hat{p}_1 - \hat{p}_2) &= V(\hat{p}_1) + V(\hat{p}_2) - 2\,\mathrm{Cov}(\hat{p}_1, \hat{p}_2) \\ &= V\left(\frac{y_1}{n_1}\right) + V\left(\frac{y_2}{n_2}\right) - 0 \qquad \text{since } y_1 \text{ and } y_2 \text{ are independent} \\ &= \frac{1}{n_1^2}V(y_1) + \frac{1}{n_2^2}V(y_2) \\ &= \frac{1}{n_1^2}(n_1 p_1 q_1) + \frac{1}{n_2^2}(n_2 p_2 q_2) \\ &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \end{aligned}$$

Thus, $(\hat{p}_1 - \hat{p}_2)$ is an unbiased estimator of $(p_1 - p_2)$ and, in addition, it has minimum variance (proof omitted).

The central limit theorem also guarantees that, for sufficiently large sample sizes $n_1$ and $n_2$, the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ will be approximately normal. It follows (Theorem 7.2) that a large-sample confidence interval for $(p_1 - p_2)$ may be obtained as shown in the following box.

Note that we must substitute the values of $\hat{p}_1$ and $\hat{p}_2$ for $p_1$ and $p_2$, respectively, to obtain an estimate of $\sigma_{(\hat{p}_1 - \hat{p}_2)}$. As in the one-sample case, this approximation is reasonably accurate when both $n_1$ and $n_2$ are sufficiently large, i.e., if the intervals

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} \quad \text{and} \quad \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

do not contain 0 or the sample size ($n_1$ or $n_2$). This will be true if $n_1\hat{p}_1$, $n_2\hat{p}_2$, $n_1\hat{q}_1$, and $n_2\hat{q}_2$ are all greater than or equal to 4.

**Large-Sample $(1 - \alpha)100\%$ Confidence Interval for $(p_1 - p_2)$**

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observations with the characteristic of interest.

[*Note:* We have followed the usual procedure of substituting the sample values $\hat{p}_1$, $\hat{q}_1$, $\hat{p}_2$, and $\hat{q}_2$ for the corresponding population values required for $\sigma_{(\hat{p}_1 - \hat{p}_2)}$.

*Assumption:* The samples are sufficiently large that the approximation is valid. As a general rule of thumb, we will require that $n_1\hat{p}_1 \geq 4$, $n_1\hat{q}_1 \geq 4$, $n_2\hat{p}_2 \geq 4$, and $n_2\hat{q}_2 \geq 4$.

## Example 7.14

Confidence Interval for $(p_1 - p_2)$: Speed Limit Violations

A traffic engineer conducted a study of vehicular speeds on a segment of street that had the posted speed limit changed several times. When the posted speed limit on the street was 30 miles per hour, the engineer monitored the speeds of 100 randomly selected vehicles traversing the street and observed 49 violations of the speed limit. After the speed limit was raised to 35 miles per hour, the engineer again monitored the speeds of 100 randomly selected vehicles and observed 19 vehicles in violation of the speed limit. Find a 99% confidence interval for $(p_1 - p_2)$, where $p_1$ is the true proportion of vehicles that (under similar driving conditions) exceed the lower speed limit (30 miles per hour) and $p_2$ is the true proportion of vehicles that (under similar driving conditions) exceed the higher speed limit (35 miles per hour). Interpret the interval.

**Solution** In this example,

$$\hat{p}_1 = \frac{49}{100} = .49 \quad \text{and} \quad \hat{p}_2 = \frac{19}{100} = .19$$

Note that

$$n_1\hat{p}_1 = 49 \qquad n_1\hat{q}_1 = 51$$
$$n_2\hat{p}_2 = 19 \qquad n_2\hat{q}_2 = 81$$

all exceed 4. Thus, we can apply the approximation for a large-sample confidence interval for $(p_1 - p_2)$.

For a confidence interval of $(1 - \alpha) = .99$, we have $\alpha = .01$ and $z_{\alpha/2} = z_{.005} = 2.58$ (from Table 5 of Appendix B). Substitution into the confidence interval formula yields:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_1}}$$
$$= (.49 - .19) \pm 2.58\sqrt{\frac{(.49)(.51)}{100} + \frac{(.19)(.81)}{100}}$$
$$= .30 \pm .164 = (.136, .464)$$

This interval is also shown (highlighted) on the MINITAB Printout, Figure 7.13.

```
Test and CI for Two Proportions

Sample   X    N  Sample p
1       49  100  0.490000
2       19  100  0.190000


Difference = p (1) - p (2)
Estimate for difference:  0.3
99% CI for difference:  (0.136318, 0.463682)
Test for difference = 0 (vs not = 0):  Z = 4.72  P-Value = 0.000

Fisher's exact test: P-Value = 0.000
```

**FIGURE 7.13**
MINITAB Printout of Comparison of Two Proportions, Example 7.13

Our interpretation is that the true difference, $(p_1 - p_2)$, falls between .136 and .464 with 99% confidence. Since the lower bound on our estimate is positive (.136), we are fairly confident that the proportion of all vehicles in violation of the lower speed limit (30 miles per hour) exceeds the corresponding proportion in violation of the higher speed limit (35 miles per hour) by at least .136.

Small-sample estimation procedures for $(p_1 - p_2)$ will not be discussed here for the reasons outlined at the end of Section 7.7.

## Applied Exercises

**MTBE**

7.66 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in New Hampshire wells, Exercise 2.12 (p. 29). Data collected for a sample of 223 wells are saved in the **MTBE** file. Recall that each well was classified according to well class (public or private) and detectable level of MTBE (below limit or detect). The accompanying SPSS printout gives the number of wells in the sample with a detectable level of MTBE for both the 120 public wells and the 103 private wells.

**DETECT * CLASS Crosstabulation**

Count

| | | CLASS | | |
|---|---|---|---|---|
| | | Private | Public | Total |
| DETECT | Below Limit | 81 | 72 | 153 |
| | Detect | 22 | 48 | 70 |
| Total | | 103 | 120 | 223 |

a. Estimate the true proportion of public wells with a detectable level of MTBE.
b. Estimate the true proportion of private wells with a detectable level of MTBE.
c. Compare the two proportions, parts **a** and **b**, with a 95% confidence interval.
d. Give a practical interpretation of the confidence interval, part **c**. Which type of well has the higher proportion of wells with a detectable level of MTBE?

7.67 *Producer willingness to supply biomass.* Refer to the *Biomass and Energy* (Vol. 36, 2012) study of the willingness of producers to supply biomass products such as surplus hay, Exercise 7.45 (p. 321). Recall that independent samples of Missouri producers and Illinois producers were surveyed. Another aspect of the study focused on the services producers are willing to supply. One key service involves windrowing (mowing and piling) of hay. Of the 558 Missouri producers surveyed, 187 were willing to offer windrowing services; of the 940 Illinois producers surveyed, 380 were willing to offer windrowing services. The researchers want to know if the proportion of producers who are willing to offer windrowing services to the biomass market differ for the two areas, Missouri and Illinois.

a. Specify the parameter of interest to the researchers.
b. A MINITAB printout of the analysis is provided below. Locate a 99% confidence interval for the difference between the proportions of producers who are willing to offer windrowing services in Missouri and Illinois.
c. What inference can you make about the two proportions based on the confidence interval, part b?

7.68 *Study of armyworm pheromones.* A study was conducted to determine the effectiveness of pheromones produced by two different strains of fall armyworms — the corn-strain and the rice-strain (*Journal of Chemical Ecology*, March 2013). Both corn-strain and rice-strain male armyworms were released into a field containing a synthetic pheromone made from a corn-strain blend. A count of the number of males trapped by the pheromone was then determined. The experiment was conducted once in a corn field, then again in a grass field. The results are provided in the table on p. 335.

a. Consider the corn field results. Construct a 90% confidence interval for the difference between the proportions of corn-strain and rice-strain males trapped by the pheromone.
b. Consider the grass field results. Construct a 90% confidence interval for the difference between the proportions of corn-strain and rice-strain males trapped by the pheromone.

**MINITAB Output for Exercise 7.67**

**Test and CI for Two Proportions**

```
Sample     X     N  Sample p
1        187   558  0.335125
2        380   940  0.404255


Difference = p (1) - p (2)
Estimate for difference:  -0.0691299
99% CI for difference:  (-0.135079, -0.00318070)
Test for difference = 0 (vs not = 0):  Z = -2.67  P-Value = 0.008

Fisher's exact test: P-Value = 0.008
```

c. Based on the confidence intervals, parts a and b, what can you conclude about the effectiveness of a corn-blend synthetic pheromone placed in a corn field? A grass field?

| | Corn Field | Grass Field |
|---|---|---|
| Number of corn-strain males released | 112 | 215 |
| Number trapped | 86 | 164 |
| Number of rice-strain males released | 150 | 669 |
| Number trapped | 92 | 375 |

7.69 *The winner's curse in road contract bidding.* In sealed bidding on state road construction contracts, the "winner's curse" is the phenomenon of the winning (or highest) bid price being above the expected price of the contract (called the Department of Transportation engineer's estimate). *The Review of Economics and Statistics* (Aug. 2001) published a study on whether bid experience impacts the likelihood of the winner's curse occurring. Two groups of bidders in a sealed-bid auction were compared: (1) super-experienced bidders and (2) less-experienced bidders. In the super-experienced group, 29 of 189 winning bids were above the item's expected price; in the less-experienced group, 32 of 149 winning bids were above the item's expected price.

a. Find an estimate of $p_1$, the true proportion of super-experienced bidders who fall prey to the winner's curse.
b. Find an estimate of $p_2$, the true proportion of less-experienced bidders who fall prey to the winner's curse.
c. Construct a 90% confidence interval for $p_1 - p_2$.
d. Give a practical interpretation of the confidence interval, part **c**. Make a statement about whether bid experience impacts the likelihood of the winner's curse occurring.

7.70 *Effectiveness of drug tests of Olympic athletes.* Erythropoietin (EPO) is a banned drug used by athletes to increase the oxygen-carrying capacity of their blood. New tests for EPO were first introduced prior to the 2000 Olympic Games held in Sydney, Australia. *Chance* (Spring 2004) reported that of a sample of 830 world-class athletes, 159 did not compete in the 1999 World Championships (a year prior to the new EPO test). Similarly, 133 of 825 potential athletes did not compete in the 2000 Olympic Games. Was the new test effective in deterring an athlete's participation in the 2000 Olympics? If so, then the proportion of nonparticipating athletes in 2000 will be greater than the proportion of nonparticipating athletes in 1999. Use a 90% confidence interval to compare the two proportions and make the proper conclusion.

7.71 *Teeth defects and stress in prehistoric Japan.* Linear enamel hypoplasia (LEH) defects are pits or grooves on the tooth surface that are typically caused by malnutrition, chronic infection, stress and trauma. A study of LEH defects in prehistoric Japanese cultures was published in the *American Journal of Physical Anthropology* (May 2010). Three groups of Japanese people were studied: Yayoi farmers (early agriculturists), eastern Jomon foragers (broad-based economy), and western Jomon foragers (wet rice economy). LEH defect prevalence was determined from skulls of individuals obtained from each of the three cultures. The results (percentage of individuals with at least one LEH defect) are provided in the accompanying table. Two theories were tested. Theory 1 states that foragers with a broad-based economy will have a lower LEH defect prevalence than early agriculturists. Theory 2 states that foragers with a wet rice economy will not differ in LEH defect prevalence from early agriculturists.

| Group | Number of individuals | Percent LEH |
|---|---|---|
| Yayoi | 182 | 63.1 |
| Eastern Jomon | 164 | 48.2 |
| Western Jomon | 122 | 64.8 |

*Source:* Temple, D.H. "Patterns of systemic stress during the agricultural transition in prehistoric Japan", *American Journal of Physical Anthropology*, Vol. 142, No. 1, May 2010 (Table 3).

a. Use a 99% confidence interval to determine whether there is evidence to support Theory 1.
b. Use a 99% confidence interval to determine whether there is evidence to support Theory 2.

**SWDEFECTS**

7.72 *Predicting software defects.* Refer to the PROMISE Software Engineering Repository data on 498 modules of software code written in C language for a NASA spacecraft instrument, saved in the **SWDEFECTS** file. (See *Statistics in Action*, Chapter 3.) Recall that the software code in each module was evaluated for defects; 49 were classified as "true" (i.e., module has defective code) and 449 were classified as "false" (i.e., module has correct code). Consider these to be independent random samples of software code modules. Researchers predicted the defect status of each module using the simple algorithm "if number of lines of code in the module exceeds 50, predict the module to have a defect." The SPSS printout below shows the number of modules in each of the two samples that were predicted to have defects (PRED_LOC = "yes") and predicted to have no defects (PRED_LOC = "no"). Now, define the *accuracy rate* of the algorithm as the proportion of modules that were correctly predicted. Compare the accuracy rate of the algorithm when applied to modules with defective code to the accuracy rate of the algorithm when applied to modules with correct code. Use a 99% confidence interval.

**DEFECT * PRED_LOC Crosstabulation**

Count

| | | PRED_LOC | | |
|---|---|---|---|---|
| | | no | yes | Total |
| DEFECT | false | 400 | 49 | 449 |
| | true | 29 | 20 | 49 |
| Total | | 429 | 69 | 498 |

## 7.9 Estimation of a Population Variance

In the previous sections, we considered interval estimates for population means and proportions. In this section, we discuss confidence intervals for a population variance $\sigma^2$, and in Section 7.10, confidence intervals for the ratio of two variances, $\sigma_1^2/\sigma_2^2$. Unlike means and proportions, the pivotal statistics for variances do not possess a normal ($z$) distribution or a $t$ distribution. In addition, certain assumptions are required regardless of the sample size.

Let $y_1, y_2, \ldots, y_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$. From Theorem 6.11, we know that

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

possesses a chi-square distribution with $(n - 1)$ degrees of freedom. Confidence intervals for $\sigma^2$ are based on the pivotal statistic, $\chi^2$.

Recall that upper-tail areas of the chi-square distribution have been tabulated and are given in Table 8 of Appendix B. Unlike the $Z$ and $T$ distributions, the chi-square distribution is not symmetric about 0. To find values of $\chi^2$ that locate an area $\alpha$ in the lower tail of the distribution, we must find $\chi^2_{1-\alpha}$, where $P(\chi^2 > \chi^2_{1-\alpha}) = 1 - \alpha$. For example, the value of $\chi^2$ that places an area $\alpha = .05$ in the lower tail of the distribution when df $= 9$ is $\chi^2_{1-\alpha} = \chi^2_{.95} = 3.32511$ (see Table 8 of Appendix B). We use this fact to write a probability statement for the pivotal statistic $\chi^2$:

$$P(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}) = 1 - \alpha$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are tabulated values of $\chi^2$ that place a probability of $\alpha/2$ in each tail of the chi-square distribution (see Figure 7.14).

Substituting $[(n-1)s^2]/\sigma^2$ for $\chi^2$ in the probability statement and performing some simple algebraic manipulations, we obtain

$$P\left(\chi^2_{(1-\alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right)$$

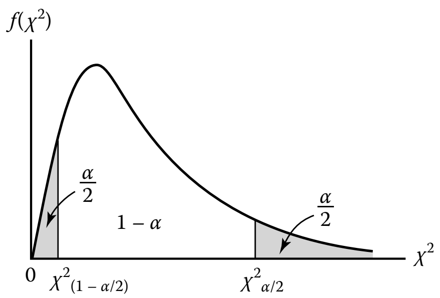
$$= P\left(\frac{\chi^2_{(1-\alpha/2)}}{(n-1)s^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{\alpha/2}}{(n-1)s^2}\right)$$

$$= P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}\right) = 1 - \alpha$$

Thus, a $(1 - \alpha)100\%$ confidence interval for $\sigma^2$ is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

**FIGURE 7.14**
The location of $\chi^2_{(1-\alpha/2)}$ and $\chi^2_{\alpha/2}$ for a chi-square distribution



### A $(1 - \alpha)100\%$ Confidence Interval for a Population Variance, $\sigma^2$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are values of $\chi^2$ that locate an area of $\alpha/2$ to the right and $\alpha/2$ to the left, respectively, of a chi-square distribution based on $(n - 1)$ degrees of freedom.

*Assumption:* The population from which the sample is selected has an approximate normal distribution.

Note that the estimation technique applies to either large or small $n$ and that the assumption of normality is required in either case.

**Example 7.15**

Confidence Interval for $\sigma^2$: Can Fill Weights

A quality control supervisor in a cannery knows that the exact amount each can contains will vary, since there are certain uncontrollable factors that affect the amount of fill. The mean fill per can is important, but equally important is the variation, $\sigma^2$, of the amount of fill. If $\sigma^2$ is large, some cans will contain too little and others too much. To estimate the variation of fill at the cannery, the supervisor randomly selects 10 cans and weighs the contents of each. The weights (in ounces) are listed in Table 7.8. Construct a 90% confidence interval for the true variation in fill of cans at the cannery.

FILLWTS

**TABLE 7.8 Fill Weights of Cans**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7.96 | 7.90 | 7.98 | 8.01 | 7.97 | 7.96 | 8.03 | 8.02 | 8.04 | 8.02 |

Solution

The supervisor wishes to estimate $\sigma^2$, the population variance of the amount of fill. A $(1 - \alpha)100\%$ confidence interval for $\sigma^2$ is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

For the confidence interval to be valid, we must assume that the sample of observations (amounts of fill) is selected from a normal population.

To compute the interval, we need to calculate either the sample variance $s^2$ or the sample standard deviation $s$. Descriptive statistics for the sample data are provided in the SAS printout shown in Figure 7.15 (p. 338). The value of $s$, shaded in Figure 7.15, is $s = .043$.

Now, $(1 - \alpha) = .90$ and $\alpha/2 = .10/2 = .05$. Therefore, the tabulated values $\chi^2_{.05}$ and $\chi^2_{.95}$ for $(n - 1) = 9$ df (obtained from Table 8, Appendix B) are

$$\chi^2_{.05} = 16.9190 \quad \text{and} \quad \chi^2_{.95} = 3.32511$$

Substituting these values into the formula, we obtain

$$\frac{(10-1)(.043)^2}{16.9190} \le \sigma^2 \le \frac{(10-1)(.043)^2}{3.32511}$$

$$.00098 \le \sigma^2 \le .00500$$

We are 90% confident that the true variance in amount of fill of cans at the cannery falls between .00098 and .00500. The quality control supervisor could use this interval to check whether the variation of fill at the cannery is too large and in violation of government regulatory specifications.

The TTEST Procedure

Variable: WEIGHT (WEIGHT)

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 10 | 7.9890 | 0.0431 | 0.0136 | 7.9000 | 8.0400 |

| Mean | 90% CL Mean | | Std Dev | 90% CL Std Dev | |
|---|---|---|---|---|---|
| 7.9890 | 7.9640 | 8.0140 | 0.0431 | 0.0314 | 0.0708 |

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 9 | 586.66 | <.0001 |

**FIGURE 7.15**
SAS Descriptive Statistics and Confidence Interval for Fill Weights

**Example 7.16**

Confidence Interval for $\sigma$: Can Fill Weights

Refer to Example 7.15. Find a 90% confidence interval for $\sigma$, the true standard deviation of the can weights.

Solution

A confidence interval for $\sigma$ is obtained by taking the square roots of the lower and upper endpoints of a confidence interval for $\sigma^2$. Thus, the 90% confidence interval is

$$\sqrt{.00098} \leq \sigma \leq \sqrt{.00500}$$

$$.031 \leq \sigma \leq .071$$

This interval is also shown (shaded) on Figure 7.15. We are 90% confident that the true standard deviation of can weights is between .031 and .071 ounce.

## Applied Exercises

7.73 *Finding chi-square values.* For each of the following combinations of $\alpha$ and degrees of freedom (df), find the value of chi-square, $\chi^2_\alpha$, that places an area $\alpha$ in the upper tail of the chi-square distribution:

a. $\alpha = .05$, df = 7 b. $\alpha = .10$, df = 16
c. $\alpha = .01$, df = 10 d. $\alpha = .025$, df = 8
e. $\alpha = .005$, df = 5

7.74 *Characteristics of a rock fall.* Refer to the *Environmental Geology* (Vol. 58, 2009) simulation study of how far a block from a collapsing rock wall will bounce down a soil slope, Exercise 2.29 (p. 43). Rebound lengths (in meters) were estimated for 13 rock bounces. The data are repeated in the table below. A MINITAB analysis of the data is shown in the printout on p. 339.

a. Locate a 95% confidence interval for $\sigma^2$ on the printout. Interpret the result.
b. Locate a 95% confidence interval for $\sigma$ on the printout. Interpret the result.
c. What conditions are required for the intervals, parts a and b, to be valid?

7.75 *Oil content of fried sweet potato chips.* The characteristics of sweet potato chips fried at different temperatures were investigated in the *Journal of Food Engineering* (Sep., 2013). A sample of 6 sweet potato slices were fried at 130º using a vacuum fryer. One characteristic of interest to the researchers was internal oil content (measured in gigagrams). The results were: $\bar{y}$=.178 g/g and $s$ = .011 g/g. Use this information to construct a 95% confidence interval for the true standard deviation of the internal oil content distribution for the sweet potato chips. Interpret the result, practically.

**ROCKFALL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10.94 | 13.71 | 11.38 | 7.26 | 17.83 | 11.92 | 11.87 | 5.44 | 13.35 | 4.90 | 5.85 | 5.10 | 6.77 |

*Source:* Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy", *Environmental Geology,* Vol. 58, 2009. (Table 2.)

**MINITAB Output for Exercise 7.74**

```
Test and CI for One Variance: REB-LENGTH

Method

The chi-square method is only for the normal distribution.
The Bonett method is for any continuous distribution.


Statistics

Variable     N  StDev  Variance
REB-LENGTH  13   4.09      16.8


95% Confidence Intervals

                              CI for        CI for
Variable    Method            StDev        Variance
REB-LENGTH  Chi-Square  (2.94, 6.76)  (8.6, 45.7)
            Bonett      (2.97, 6.64)  (8.8, 44.1)
```

7.76 *DNA in antigen-produced protein.* Refer to the *Gene Therapy and Molecular Biology* (June 2009) study of DNA in peptide (protein) produced by antigens for a parasitic roundworm in birds, Exercise 7.24 (p. 311). Recall that scientists tested each in a sample of 4 alleles of antigen-produced protein for level of peptide. The results were: $\bar{y}$= 1.43 and $s = 13$. Use this information to construct a 90% confidence interval for the true variation in peptide scores for alleles of the antigen-produced protein. Interpret the interval for the scientists.

7.77 *Radon exposure in Egyptian tombs.* Refer to the *Radiation Protection Dosimetry* (December 2010) study of radon exposure in tombs carved from limestone in the Egyptian Valley of Kings, Exercise 7.28 (p. 312). The radon levels in the inner chambers of a sample of 12 tombs were determined, yielding the following summary statistics: $\bar{y} = 3{,}643$ Bq/m$^3$ and $s = 4{,}487$ Bq/m$^3$. Use this information to estimate, with 95% confidence, the true standard deviation of radon levels in tombs in the Valley of Kings. Interpret the resulting interval.

7.78 *Monitoring impedance to leg movements.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) experiment to monitor the impedance to leg movements, Exercise 7.34 (p. 313). Engineers attached electrodes to the ankles and knees of volunteers and measured the signal-to-noise ratio (SNR) of impedance changes. Recall that for a particular ankle–knee electrode pair, a sample of 10 volunteers had SNR values with a mean of 19.5 and a standard deviation of 4.7. Form a 95% confidence interval for the true standard deviation of the SNR impedance changes. Interpret the result.

7.79 *Drug content assessment.* Refer to the *Analytical Chemistry* (Dec. 15, 2009) study of a new method used by GlaxoSmithKline Medicines Research Center to determine the amount of drug in a tablet, Exercise 5.45 (p. 210). Drug concentrations (measured as a percentage) for 50 randomly selected tablets are repeated in the table below. For comparisons against a standard method, the scientists at GlaxoSmithKline desire an estimate of the variability in

**DRUGCON**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 91.28 | 92.83 | 89.35 | 91.90 | 82.85 | 94.83 | 89.83 | 89.00 | 84.62 |
| 86.96 | 88.32 | 91.17 | 83.86 | 89.74 | 92.24 | 92.59 | 84.21 | 89.36 |
| 90.96 | 92.85 | 89.39 | 89.82 | 89.91 | 92.16 | 88.67 | 89.35 | 86.51 |
| 89.04 | 91.82 | 93.02 | 88.32 | 88.76 | 89.26 | 90.36 | 87.16 | 91.74 |
| 86.12 | 92.10 | 83.33 | 87.61 | 88.20 | 92.78 | 86.35 | 93.84 | 91.20 |
| 93.44 | 86.77 | 83.77 | 93.19 | 81.79 | | | | |

*Source:* Borman, P.J., Marion, J.C., Damjanov,I., & Jackson, P. "Design and analysis of method equivalence studies", *Analytical Chemistry*, Vol. 81, No. 24, December 15, 2009 (Table 3).

drug concentrations for the new method. Obtain the estimate for the scientists using a 99% confidence interval. Interpret the interval.

**PONDICE**

7.80 *Albedo of ice meltponds.* Refer to the National Snow and Ice Data Center (NSIDC) collection of data on the albedo of ice meltponds, Exercise 7.36 (p. 313). The visible albedo values for a sample of 504 ice meltponds located in the Canadian Arctic are saved in the **PONDICE** file. Find a 90% confidence interval for the true variance of the visible albedo values of all Canadian Arctic ice ponds. Give both a practical and theoretical interpretation of the interval.

**PHISHING**

7.81 *Phishing attacks to email accounts.* Refer to the *Chance* (Summer 2007) study of an actual phishing attack against an organization, Exercise 2.24 (p. 38). Recall that *phishing* describes an attempt to extract personal/financial information from unsuspecting people through fraudulent email. The interarrival times (in seconds) for 267 fraud box email notifications are saved in the **PHISHING** file. Like with Exercise 2.24, consider these interarrival times to represent the population of interest.

a. Obtain a random sample of $n = 10$ interarrival times from the population.
b. Use the sample, part b, to obtain an interval estimate of the population variance of the interarrival times. What is the measure of reliability for your estimate?
c. Find the true population variance for the data. Does the interval, part b, contain the true variance? Give one reason why it may not.

## 7.10 Estimation of the Ratio of Two Population Variances

The common statistical procedure for comparing two population variances, $\sigma_1^2$ and $\sigma_2^2$, makes an inference about the ratio $\sigma_1^2/\sigma_2^2$. This is because the sampling distribution of the estimator of $\sigma_1^2/\sigma_2^2$ is well known when the samples are randomly and independently selected from two normal populations. Under these assumptions, a confidence interval for $\sigma_1^2/\sigma_2^2$ is based on the pivotal statistic

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

where $\chi_1^2$ and $\chi_2^2$ are independent chi-square random variables with $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$ degrees of freedom, respectively. Substituting $(n-1)s^2/\sigma^2$ for $\chi^2$ (see Theorem 6.11), we may write

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = \frac{\dfrac{(n_1 - 1)s_1^2}{\sigma_1^2}\Big/(n_1 - 1)}{\dfrac{(n_2 - 1)s_2^2}{\sigma_2^2}\Big/(n_2 - 1)}$$

$$= \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

$$= \left(\frac{s_1^2}{s_2^2}\right)\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$

From Definition 6.17 we know that $F$ has an $F$ distribution with $\nu_1 = (n_1 - 1)$ numerator degrees of freedom and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom. An $F$ distribution can be symmetric about its mean, skewed to the left, or skewed to the right; its exact shape depends on the degrees of freedom associated with $s_1^2$ and $s_2^2$, i.e., $(n_1 - 1)$ and $(n_2 - 1)$.

To establish lower and upper confidence limits for $\sigma_1^2/\sigma_2^2$, we need to be able to find tabulated $F$ values corresponding to the tail areas of the distribution. The *upper-tail F* values can be found in Tables 9–12 of Appendix B for $a = .10, .05, .025$, and $.01$, respectively. Table 10 of Appendix B is partially reproduced in Table 7.9. The

**TABLE 7.9 Abbreviated Version of Table 10 of Appendix B: Tabulated Values of the *F* Distribution, $\alpha = .05$**

| $\nu_2$ \ $\nu_1$ (Denominator Degrees of Freedom) | Numerator Degrees of Freedom | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | | 8 | 9 |
| **1** | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 |
| **2** | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 7 | 19.37 | 19.38 |
| **3** | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| **4** | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| **5** | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| **6** | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| **7** | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| **8** | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| **9** | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | | 3.23 | 3.18 |
| **10** | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 |
| **11** | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.29 | 2.95 | 2.90 |
| **12** | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 |
| **13** | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 |
| **14** | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 |

columns of the table correspond to various degrees of freedom for the numerator sample variance, $s_1^2$, in the pivotal statistic, whereas the rows correspond to the degrees of freedom for the denominator sample variance, $s_2^2$. For example, with numerator degrees of freedom $\nu_1 = 7$ and denominator degrees of freedom $\nu_2 = 9$, we have $F_{.05} = 3.29$ (shaded in Table 7.9). Thus, $\alpha = .05$ is the tail area to the right of 3.29 in the $F$ distribution with 7 numerator df and 9 denominator df, i.e., $P(F > F_{.05}) = .05$.

*Lower-tail* values of the $F$ distribution are not given in Tables 9–12 of Appendix B. However, it can be shown (proof omitted) that

$$F_{1-a(\nu_1,\,\nu_2)} = \frac{1}{F_{a(\nu_2,\,\nu_1)}}$$

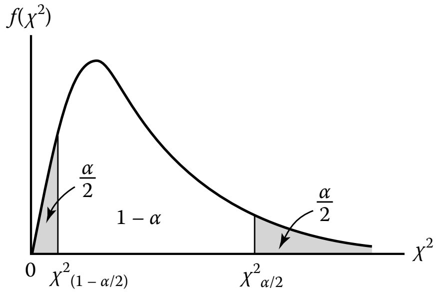
where $F_{1-a(\nu_1,\nu_2)}$ is the $F$ value that cuts off an area $a$ in the *lower* tail of an $F$ distribution based on $\nu_1$ numerator and $\nu_2$ denominator degrees of freedom, and $F_{a(\nu_2,\nu_1)}$ is the $F$ value that cuts off an area $a$ in the *upper* tail of an $F$ distribution based on $\nu_2$ numerator and $\nu_1$ denominator degrees of freedom. For example, suppose we want to find the value that locates an area $a = .05$ in the *lower* tail of an $F$ distribution with $\nu_1 = 7$ and $\nu_2 = 9$. That is, we want to find $F_{1-a(\nu_1,\nu_2)} = F_{.95(7,9)}$. First, we find the upper-tail values, $F_{.05(9,7)} = 3.68$, from Table 7.9. (Note that we must switch the numerator and denominator degrees of freedom to obtain this value.) Then, we calculate

$$F_{.95(7,9)} = \frac{1}{F_{.05(9,7)}} = \frac{1}{3.68} = .272$$

Using the notation established previously, we can write a probability statement for the pivotal statistic $F$ (see Figure 7.16):

$$P(F_{1-\alpha/2(\nu_1,\nu_2)} \leq F \leq F_{\alpha/2(\nu_1,\nu_2)}) = 1 - \alpha$$

**FIGURE 7.16**
$F$ distribution with $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$




Letting $F_L = F_{1-\alpha/2}$ and $F_U = F_{\alpha/2}$, and substituting $(s_1^2/s_2^2)(\sigma_2^2/\sigma_1^2)$ for $F$, we obtain:

$$P(F_L \le F \le F_U) = P\left[F_L \le \left(\frac{s_1^2}{s_2^2}\right)\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \le F_U\right]$$

$$= P\left(\frac{s_2^2}{s_1^2}F_L \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{s_2^2}{s_1^2}F_U\right)$$

$$= P\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_U} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_L}\right) = 1 - \alpha$$

or

$$P\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2(\nu_1,\nu_2)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2(\nu_1,\nu_2)}}\right) = 1 - \alpha$$

Replacing $F_{1-\alpha/2(\nu_1,\nu_2)}$ with $1/F_{\alpha/2(\nu_2,\nu_1)}$, we obtain the final form of the confidence interval:

$$P\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2(\nu_1,\nu_2)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2(\nu_2,\nu_1)}\right) = 1 - \alpha$$

### A $(1 - \alpha)100\%$ Confidence Interval for the Ratio of Two Population Variances, $\sigma_1^2/\sigma_2^2$

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2(\nu_1,\nu_2)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}F_{\alpha/2(\nu_2,\nu_1)}$$

where $F_{\alpha/2(\nu_1,\nu_2)}$ is the value of $F$ that locates an area $\alpha/2$ in the upper tail of the $F$ distribution with $\nu_1 = (n_1 - 1)$ numerator and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom, and $F_{\alpha/2(\nu_2,\nu_1)}$ is the value of $F$ that locates an area $\alpha/2$ in the upper tail of the $F$ distribution with $\nu_2 = (n_2 - 1)$ numerator and $\nu_1 = (n_1 - 1)$ denominator degrees of freedom.

*Assumptions:*
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The random samples are selected in an independent manner from the two populations.

As in the one-sample case, normal populations must be assumed regardless of the sizes of the two samples.

**Example 7.17**

**Confidence Interval for $\sigma_1^2/\sigma_2^2$: Comparing Two Assembly Lines**

A firm has been experimenting with two different physical arrangements of its assembly line. It has been determined that both arrangements yield approximately the same average number of finished units per day. To obtain an arrangement that produces greater process control, you suggest that the arrangement with the smaller variance in the number of finished units produced per day be permanently adopted. Two independent random samples yield the results shown in Table 7.10. Construct a 95% confidence interval for $\sigma_1^2/\sigma_2^2$, the ratio of the variance of the number of finished units for the two assembly line arrangements. Based on the result, which of the two arrangements would you recommend?

ASSEMBLY

**TABLE 7.10 Number of Finished Units Produced per Day by Two Assembly Lines**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Line 1* | 448 | 523 | 506 | 500 | 533 | 447 | 524 | 469 | 470 | 494 | 536 |
| | 481 | 492 | 567 | 492 | 457 | 497 | 483 | 533 | 408 | 453 | |
| *Line 2* | 372 | 446 | 537 | 592 | 536 | 487 | 592 | 605 | 550 | 489 | 461 |
| | 500 | 430 | 543 | 459 | 429 | 494 | 538 | 540 | 481 | 484 | 374 |
| | 495 | 503 | 547 | | | | | | | | |

```
Sample Statistics

     LINE
     Group              N      Mean     Std. Dev.    Variance
     ----------------------------------------------------------
     1                 21  491.0952       37.522      1407.89
     2                 25    499.36       61.069     3729.407

Hypothesis Test

     Null hypothesis:    Variance 1 / Variance 2 =  1
     Alternative:        Variance 1 / Variance 2 ^= 1

                     - Degrees of Freedom -
          F             Numer.    Denom.         Pr > F
     ----------------------------------------------------
        0.38              20        24           0.0304

95% Confidence Interval of the Ratio of Two Variances

                    Lower Limit    Upper Limit
                    -----------    -----------
                       0.1622         0.9089
```

**FIGURE 7.17**
SAS descriptive statistics and confidence interval for assembly line data

Solution

Summary statistics for the assembly line data are shown (highlighted) on the SAS printout, Figure 7.17. Note that $s_1^2 = 1407.89$ and $s_2^2 = 3729.41$.

To construct the confidence interval, we must assume that the distributions of the numbers of finished units for the two assembly lines are both approximately normal. Since we want a 95% confidence interval, the value of $\alpha/2$ is .025, and we need to find $F_{.025(\nu_1,\nu_2)}$ and $F_{.025(\nu_2,\nu_1)}$. The sample sizes are $n_1 = 21$ and $n_2 = 25$; thus, $F_{.025(\nu_1,\nu_2)}$ is based on $\nu_1 = (n_1 - 1) = 20$ numerator df and $\nu_2 = (n_2 - 1) = 24$ denominator df. Consulting Table 11 of Appendix B, we obtain $F_{.025(20,24)} = 2.33$. In contrast, $F_{.025(\nu_2,\nu_1)}$ is based on $\nu_2 = (n_2 - 1) = 24$ numerator df and $\nu_1 = (n_1 - 1) = 20$ denominator df; hence (from Table 11 of Appendix B), $F_{.025(24,\,20)} = 2.41$.

Substituting the values for $s_1^2$, $s_2^2$, $F_{.025(\nu_1,\nu_2)}$ and $F_{.025(\nu_2,\nu_1)}$ into the confidence interval formula, we have

$$\frac{(1407.89)}{(3729.41)}\left(\frac{1}{2.33}\right) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{1407.89}{3729.41}(2.41)$$

$$.162 \le \frac{\sigma_1^2}{\sigma_2^2} \le .909$$

(*Note:* This interval is shown at the bottom of Figure 7.17.)

We estimate with 95% confidence that the ratio $\sigma_1^2/\sigma_2^2$ of the true population variances will fall between .162 and .909. Since all the values within the interval are less than 1.0, we can be confident that the variance in the number of units finished on line 1 (as measured by $\sigma_1^2$) is less than the corresponding variance for line 2 (as measured by $\sigma_2^2$).

## Applied Exercises

7.82 *Finding F values.* Find $F_\alpha$ for an $F$ distribution with 15 numerator df and 12 denominator df for the following values of $\alpha$:

a. $\alpha = .025$ b. $\alpha = .05$ c. $\alpha = .10$

7.83 *Finding F values.* Find $F_{.05}$ for an $F$ distribution with:

a. Numerator df = 7, denominator df = 25
b. Numerator df = 10, denominator df = 8
c. Numerator df = 30, denominator df = 60
d. Numerator df = 15, denominator df = 4

**DRUGCON**

7.84 *Drug content assessment.* Refer to Exercise 7.39 (p. 319) and the *Analytical Chemistry* (Dec. 15, 2009) study in which scientists used high-performance liquid chromatography to determine the amount of drug in a tablet. Recall that 25 tablets were produced at each of two different, independent sites. The researchers want to determine if the two sites produce drug concentrations with different variances. A MINITAB printout of the analysis is provided on p. 345. Locate a 95% confidence interval for $\sigma_1^2/\sigma_2^2$ on the printout. Based on this interval, what inference can you draw concerning the variances in drug concentrations at the two sites?

7.85 *Hippo grazing patterns in Kenya.* Refer to the *Landscape & Ecology Engineering* (Jan. 2013) study of hippopotamus grazing patterns in Kenya, Exercise 7.41 (p. 320). Recall that plots of land were sampled in two areas — a national reserve and a pastoral ranch — and the number of hippo trails from a water source was determined for each plot. Sample statistics are reproduced in the next table.

a. Find an interval estimate of $\sigma_1^2/\sigma_2^2$, the ratio of the variances associated with the two areas. Use a 90% confidence level.
b. Can the researchers reliably conclude that the variability in number of hippo trails from a water source in the national reserve differs from the variability in number of hippo trails from a water source in the pastoral ranch? Explain.

| | National Reserve | Pastoral Ranch |
|---|---|---|
| Sample size: | 406 | 230 |
| Mean number of trails: | 0.31 | 0.13 |
| Standard deviation: | 0.4 | 0.3 |

*Source:* Kanga, E.M., et al. "Hippopotamus and livestock grazing: influences on riparian vegetation and facilitation of other herbivores in the Mara Region of Kenya", *Landscape & Ecology Engineering*, Vol. 9, No. 1, January 2013.

7.86 *Oil content of fried sweet potato chips.* Refer to the *Journal of Food Engineering* (Sep. 2013) study of the characteristics of fried sweet potato chips, Exercise 7.78 (p. 339). Recall that a sample of 6 sweet potato slices fried at 130º using a vacuum fryer yielded the following statistics on internal oil content (measured in gigagrams): $\bar{y}_1 = .178$ g/g and $s_1 = .011$ g/g. A second sample of 6 sweet potato slices was obtained, only these were subjected to a two-stage frying process (again, at 130º) in an attempt to improve texture and appearance. Summary statistics on internal oil content for this second sample follows: $\bar{y}_2 = .140$ g/g and $s_2 = .002$ g/g. The researchers want to compare the mean internal oil contents of sweet potato chips fried with the two methods; however, they recognize that the sample sizes are small.

a. What assumption about the data is required in order for the comparison of means to be valid?

**MINITAB Output for Exercise 7.84**

**Test and CI for Two Variances: Content vs Site**

```
Method

Null hypothesis          Variance(1) / Variance(2) = 1
Alternative hypothesis   Variance(1) / Variance(2) not = 1
Significance level       Alpha = 0.05


Statistics

Site   N  StDev  Variance
1     25  3.067     9.406
2     25  3.339    11.147

Ratio of standard deviations = 0.919
Ratio of variances = 0.844


95% Confidence Intervals

                                        CI for
Distribution    CI for StDev       Variance
of Data             Ratio              Ratio
Normal         (0.610, 1.384)  (0.372, 1.915)
Continuous     (0.497, 1.315)  (0.247, 1.729)


Tests

                                                  Test
Method                           DF1  DF2  Statistic  P-Value
F Test (normal)                   24   24       0.84    0.681
Levene's Test (any continuous)     1   48       0.64    0.427
```

b. Construct a 95% confidence interval for the ratio of the two population variances of interest.

c. Based on the interval, part b, is there a violation of the assumption, part a? Explain.

7.87 *Sensor motion of a robot.* Refer to *The International Journal of Robotics Research* (Dec. 2004) algorithm for estimating the sensor motion of a robotic arm, Exercise 2.64 (p. 60). A key variable is the error of estimating arm rotation (measured in radians). Data on rotation error for 11 experiments with different combinations of perturbed intrinsics and perturbed projections are reproduced in the table. Suppose the rotation error variance is important and the researchers want to compare the variances for different intrinsics and projections. In particular, they want an estimate of the ratio of the variance for trials with perturbed intrinsics but no perturbed projections, to the variance for trials with no perturbed intrinsics but perturbed projections. Use a 90% confidence interval to estimate the desired parameter. (*Hint:* Delete the data for the first trial.)

**SENSOR**

| Trial | Perturbed Intrinsics | Perturbed Projections | Rotation Error (radians) |
|---|---|---|---|
| 1 | No | No | .0000034 |
| 2 | Yes | No | .032 |
| 3 | Yes | No | .030 |
| 4 | Yes | No | .094 |
| 5 | Yes | No | .046 |
| 6 | Yes | No | .028 |
| 7 | No | Yes | .27 |
| 8 | No | Yes | .19 |
| 9 | No | Yes | .42 |
| 10 | No | Yes | .57 |
| 11 | No | Yes | .32 |

*Source:* Strelow, D., and Singh, S. "Motion estimation form image and inertial measurements." *The International Journal of Robotics Research*, Vol. 23, No. 12, Dec. 2004 (Table 4).

7.88 *Atmospheric transport of pollutants.* In *Environmental Science & Technology* (Oct. 1993), scientists reported on a study of the transport and transformation of PCDD, a pollutant emitted from solid waste incineration, motor vehicles, steel mills, and metal production. Ambient air specimens were collected over several different days at two locations in Sweden: Rörvik (11 days) and Gothenburg (3 days). The level of PCDD (measured in pg/m$^3$) detected in each specimen is recorded here. Use interval estimation to compare the variation in PCDD levels at the two locations. Draw an inference from the analysis.

**PCDDAIR**

| Rörvik | | | | Gothenburg | | |
|---|---|---|---|---|---|---|
| 2.38 | 3.03 | 1.44 | .47 | .50 | .61 | .90 |
| .50 | .22 | .26 | .31 | | | |
| .46 | 1.09 | 2.14 | | | | |

*Source:* Tysklind, M., et al. "Atmospheric transport and transformation of polychlorinated dibenzo-*p*-dioxins and dibenzofurans." *Environmental Science & Technology*, Vol. 27, No. 10, Oct. 1993, p. 2193 (Table III).

## 7.11 Choosing the Sample Size

One of the first problems encountered when applying statistics in a practical situation is to decide on the number of measurements to include in the sample(s). The solution to this problem depends on the answers to the following questions: Approximately how wide do you want your confidence interval to be? What confidence coefficient do you require?

You have probably noticed that the half-widths of many of the confidence intervals presented in Sections 7.4–7.10 are functions of the sample size and the estimated standard error of the point estimator involved. For example, the half-width $H$ of the small-sample confidence interval for $\mu$ is

$$H = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ depends on the sample size $n$ and $s$ is a statistic computed from the sample data. Since we will not know $s$ before selecting the sample and we have no control over its value, the easiest way to decrease the width of the confidence interval is to increase the sample size $n$. Generally speaking, the larger the sample size, the more information you will acquire and the smaller will be the width of the confidence interval. We illustrate the procedure for selecting the sample size in the next two examples.

**Example 7.18**

Choosing $n$ to Estimate $\mu$: Mean Expenditure on Heating Fuel

As part of a Department of Energy (DOE) survey, American families will be randomly selected and questioned about the amount of money they spent last year on home heating oil or gas. Of particular interest to the DOE is the average amount $\mu$ spent last year on heating fuel. If the DOE wants the estimate of $\mu$ to be correct to within \$10 with a confidence coefficient of .95, how many families should be included in the sample?

Solution

The DOE wants to obtain an interval estimate of $\mu$, with confidence coefficient equal to $(1 - \alpha) = .95$ and half-width of the interval equal to 10. The half-width of a large-sample confidence interval for $\mu$ is

$$H = z_{\alpha/2}\sigma_{\bar{y}} = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

In this example, we have $H = 10$ and $z_{\alpha/2} = z_{.025} = 1.96$. To solve the equation for $n$, we need to know $\sigma$. But, as will usually be the case in practice, $\sigma$ is unknown. Suppose, however, that the DOE knows from past records that the yearly amounts spent on heating fuel have a range of approximately \$520. Then we could approximate $\sigma$ by letting the range equal $4\sigma$.* Thus,

$$4\sigma \approx 520 \quad \text{or} \quad \sigma \approx 130$$

Solving for $n$, we have

$$H = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad 10 = 1.96\left(\frac{130}{\sqrt{n}}\right)$$

or

$$n = \frac{(1.96)^2(130)^2}{(10)^2} \approx 650$$

Consequently, the DOE will need to elicit responses from 650 American families to estimate the mean amount spent on home heating fuel last year to within \$10 with 95% confidence. Since this would require an extensive and costly survey, the DOE might decide to allow a larger half-width (say, $H = 15$ or $H = 20$) to reduce the sample size, or the DOE might decrease the desired confidence coefficient. The important point is that the experimenter can obtain an idea of the sampling effort necessary to achieve a specified precision in the final estimate by determining the approximate sample size *before* the experiment is begun.

**Example 7.19**

**Choosing the Sample to Estimate $(p_1 - p_2)$: Defective Items**

A production supervisor suspects a difference exists between the proportions $p_1$ and $p_2$ of defective items produced by two different machines. Experience has shown that the proportion defective for each of the two machines is in the neighborhood of .03. If the supervisor wants to estimate the difference in the proportions correct to within .005 with probability .95, how many items must be randomly sampled from the production of each machine? (Assume that you want $n_1 = n_2 = n$.)

**Solution**

Since we want to estimate $(p_1 - p_2)$ with a 95% confidence interval, we will use $z_{\alpha/2} = z_{.025} = 1.96$. For the estimate to be correct to within .005, the half-width of the confidence interval must equal .005. Then, letting $p_1 = p_2 = .03$ and $n_1 = n_2 = n$, we find the required sample size per machine by solving the following equation for $n$:

$$H = z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} \quad \text{or} \quad H = z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$.005 = 1.96\sqrt{\frac{(.03)(.97)}{n} + \frac{(.03)(.97)}{n}}$$

$$.005 = 1.96\sqrt{\frac{2(.03)(.97)}{n}}$$

$$n = \frac{(1.96)^2(2)(.03)(.97)}{(.005)^2} \approx 8{,}943$$

You can see that this may be a tedious sampling procedure. If the supervisor insists on estimating $(p_1 - p_2)$ correct to within .005 with probability equal to .95, approximately 9,000 items will have to be inspected for each machine.

*From the Empirical Rule, we expect about 95% of the observations to fall between $\mu - 2\sigma$ and $\mu + 2\sigma$. Thus,

$$\text{Range} \approx (\mu + 2\sigma) - (\mu - 2\sigma) = 4\sigma$$

You can see from the calculations in Example 7.18 that $\sigma_{(\hat{p}_1-\hat{p}_2)}$ (and hence the solution, $n_1 = n_2 = n$) depends on the actual (but unknown) values of $p_1$ and $p_2$. In fact, the required sample size $n_1 = n_2 = n$ is largest when $p_1 = p_2 = .5$. Therefore, if you have no prior information on the approximate values of $p_1$ and $p_2$, use $p_1 = p_2 = .5$ in the formula for $\sigma_{(\hat{p}_1-\hat{p}_2)}$. If $p_1$ and $p_2$ are in fact close to .5, then the resulting values of $n_1$ and $n_2$ will be correct. If $p_1$ and $p_2$ differ substantially from .5, then your solutions for $n_1$ and $n_2$ will be larger than needed. Consequently, using $p_1 = p_2 = .5$ when solving for $n_1$ and $n_2$ is a conservative procedure because the sample sizes $n_1$ and $n_2$ will be at least as large as (and probably larger than) needed.

The formulas for calculating the sample size(s) required for estimating the parameters $\mu$, $(\mu_1 - \mu_2)$, $p$, and $(p_1 - p_2)$ are summarized in the following boxes. Sample size calculations for variances require more sophisticated techniques and are beyond the scope of this text.

### Choosing the Sample Size for Estimating a Population Mean $\mu$ to Within $H$ Units with Probability $(1-\alpha)$

$$n = \left(\frac{z_{\alpha/2}\sigma}{H}\right)^2$$

(*Note:* The population standard deviation $\sigma$ will usually have to be approximated.)

### Choosing the Sample Sizes for Estimating the Difference $(\mu_1 - \mu_2)$ Between a Pair of Population Means Correct to Within $H$ Units with Probability $(1-\alpha)$

$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{H}\right)^2 (\sigma_1^2 + \sigma_2^2)$$

where $n_1$ and $n_2$ are the numbers of observations sampled from each of the two populations, and $\sigma_1^2$ and $\sigma_2^2$ are the variances of the two populations.

### Choosing the Sample Size for Estimating a Population Proportion $p$ to Within $H$ Units with Probability $(1-\alpha)$

$$n = \left(\frac{z_{\alpha/2}}{H}\right)^2 pq$$

where $p$ is the value of the population proportion that you are attempting to estimate, and $q = 1 - p$.
(*Note:* This technique requires previous estimates of $p$ and $q$. If none are available, use $p = q = .5$ for a conservative choice of $n$.)

### Choosing the Sample Sizes for Estimating the Difference $(p_1 - p_2)$ Between Two Population Proportions to Within $H$ Units with Probability $(1-\alpha)$

$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{H}\right)^2 (p_1q_1 + p_2q_2)$$

where $p_1$ and $p_2$ are the proportions for populations 1 and 2, respectively, and $n_1$ and $n_2$ are the numbers of observations to be sampled from each population.

## Applied Exercises

7.89 *Radioactive lichen.* Refer to the Alaskan Lichen Radionuclide Baseline Research study, Exercise 7.31 (p. 312). In a sample of $n = 9$ lichen specimens, the researchers found the mean and standard deviation of the amount of the radioactive element, cesium-137, present to be .009 and .005 microcuries per milliliter, respectively. Suppose the researchers want to increase the sample size in order to estimate the mean, $\mu$, to within .001 microcuries per milliliter of its true value using a 95% confidence interval.

a. What is the confidence level desired by the researchers?
b. What is the sampling error desired by the researchers?
c. Compute the sample size necessary to obtain the desired estimate.

7.90 *Aluminum cans contaminated by fire.* A gigantic warehouse located in Tampa, Florida, stores approximately 60 million empty aluminum beer and soda cans. Recently, a fire occurred at the warehouse. The smoke from the fire contaminated many of the cans with blackspot, rendering them unusable. A University of South Florida statistician was hired by the insurance company to estimate $p$, the true proportion of cans in the warehouse that were contaminated by the fire. How many aluminum cans should be randomly sampled to estimate the true proportion to within .02 with 90% confidence?

7.91 *Laser scanning for fish volume estimation.* Refer to the *Journal of Aquacultural Engineering* (Nov. 2012) study of the feasibility of laser scanning to estimate the volume of fish in a tank, Exercise 7.25 (p. 311). Recall that turbot fish were reared in a tank for experimental purposes and laser scans were executed in randomly selected locations in the tank, each scan providing an estimate of the volume (in kilograms) of fish layer. Determine the number of laser scans that must be executed in order to estimate the true mean volume of fish layer in the tank to within 5 kg with 95% confidence. (*Hint*: Use the sample standard deviation obtained in the original study, 15 kg, as an estimate of the true volume standard deviation.)

7.92 *Muscle activity of harvesting foresters.* Refer to the *International Journal of Foresting Engineering* (Vol. 19, 2008) investigation of the muscle activity patterns in the neck and upper extremities among forestry vehicle operators, Exercise 7.38 (p. 319). Recall that two types of harvesting vehicles — Timberjack and Valmet — were compared using an independent-samples design. How many Timberjack and Valmet harvester operators should participate in the study in order to estimate $(\mu_T - \mu_V)$, the true mean difference in muscle rest (seconds per minute) for the two harvesting vehicles, to within 1.5 seconds/minute with 90% confidence? Assume equal sample sizes.

7.93 *Settlement of shallow foundations.* Refer to the *Environmental & Engineering Geoscience* (Nov. 2012) study of settlement of shallow foundations on cohesive soil, Exercise 7.50 (p. 326). Recall that the researchers sampled 13 structures built on a shallow foundation and measured both the actual settlement value (measured in millimeters) and predicted settlement value (based on a formula) for each structure. How many more structures must be sampled in order to estimate the true mean difference between actual and predicted settlement value to within 2 millimeters with 99% confidence?

7.94 *Microsoft program security issues.* Refer to the *Computers & Security* (July 2013) study of security issues with Microsoft products, Exercise 7.58 (p. 330). Determine the number of security bulletins to be sampled in order for Microsoft to estimate the proportion of bulletins that report a problem with Windows to within .075 with 90% confidence.

7.95 *Study of armyworm pheromones.* Refer to the *Journal of Chemical Ecology* (March 2013) study of the effectiveness of pheromones to attract two different strains of fall armyworms, Exercise 7.68 (p. 334). Recall that both corn-strain and rice-strain male armyworms were released into a corn field containing the pheromone and the percentage of males trapped by the pheromone for the two different strains was compared. If the researchers want to estimate the difference in percentages to within 5% with a 90% confidence interval, how many armyworms of each strain need to be released into the field? Assume an equal number of corn-strain and rice-strain males will be released.

7.96 *Oven cooking study.* Refer to Exercise 7.35 (p. 313). Suppose that we want to estimate the average decay rate of fine particles produced from oven cooking or toasting to within .04 with 95% confidence. How large a sample should be selected?

7.97 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in New Hampshire wells, Exercise 7.66 (p. 334). How many public and how many private wells must be sampled in order to estimate the difference between the proportions of wells with a detectable level of MTBE to within .06 with 95% confidence?

7.98 *High-strength aluminum alloys.* Refer to the *JOM* (Jan. 2003) comparison of a new high-strength RAA aluminum alloy to the current strongest aluminum alloy, Exercise 7.43 (p. 320). Suppose the researchers want to estimate the difference between the mean yield strengths of the two alloys to within 15 MPa using a 95% confidence interval. How many alloy specimens of each type must be tested in order to obtain the desired estimate?

## Theoretical Exercise

7.99 When determining the sample size required to estimate $p$, show that the sample size $n$ is largest when $p = .5$.

## 7.12 Alternative Interval Estimation Methods: Bootstrapping and Bayesian Methods (*Optional*)

In Sections 7.4–7.10, the classical statistics approach to estimation was employed to find confidence intervals for population parameters. The key to this methodology is finding the *pivotal statistic* (defined in Section 7.3) for the population parameter of interest. Since the probability distribution of the pivotal statistic is known (either by applying the Central Limit Theorem for large samples or by making assumptions about the population data for small samples), a probability statement about the pivotal statistic is used to find the lower and upper endpoints of a confidence interval.

In this optional section, we present two very different approaches to interval estimation: the *bootstrapping* method and a *Bayesian* method. In certain sampling situations, one or both of these methods may yield narrower and/or more valid confidence intervals for the target parameter.

### Bootstrap Confidence Intervals

The **bootstrap procedure**, developed by Bradley Efron (1979), is a Monte Carlo method that involves ***resampling***—that is, taking repeated samples of size $n$ (*with replacement*) from the original sample data set. Efron's use of the term *bootstrap* was derived from the phrase "to pull oneself up by one's bootstrap;" thus, colorfully describing a computer-based technique that can be used to obtain reliable inferences even in sampling situations where the data do not adhere to the underlying assumptions.

Let $y_1, y_2, y_3, \ldots, y_n$, represent a random sample of size $n$ selected from a population with unknown mean $E(Y)=\mu$. The steps required to obtain a bootstrap confidence interval for $\mu$ are as follows.

**Step 1** Select $j$, where $j$ is the number of times you will resample. (Usually, $j$ is a very large number, say $j = 1{,}000$ or $j = 3{,}000$.)

**Step 2** Randomly sample, with replacement, $n$ values of $Y$ from the original sample data set, $y_1, y_2, y_3, \ldots, y_n$. This is called resampling. (*Note*: Since we are sampling with replacement, it is likely that a single sample will have multiple values of the same $Y$ value.)

**Step 3** Repeat step 2 a total of $j$ times, computing the sample mean, $\bar{y}$, each time.

**Step 4** Let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_j$ represent the $j$ sample means from resampling. The (simulated) distribution of these sample means approximates the true sampling distribution of $\bar{y}$.

**Step 5** Find the (approximate) $(\alpha/2)$ 100% and $(1 - \alpha/2)100\%$ percentiles of the simulated sampling distribution of $\bar{y}$ from step 4. These two percentiles represent the lower and upper endpoints, respectively, of an approximate $(1 - \alpha)100\%$ confidence interval for $\mu$.

Upon initial exposure to this methodology, it may not be immediately clear why resampling the sample will yield reliable estimates and a valid confidence interval. The key is understanding that our only information on the sampling variability of the sample mean, $\bar{y}$, is within the sample itself. Consequently, by resampling the sample we simulate the actual variability around $\bar{y}$. Of course, we need a computer to carry out the thousands of resamplings necessary to obtain a good approximation to the sampling distribution of $\bar{y}$.

We illustrate the bootstrap procedure in the next example.

**Example 7.20**

Bootstrap Estimate of Poisson $\lambda$

Refer to Example 7.2 and the study of auditory nerve fiber response rates in cats. The data (number of spikes per 200 milliseconds on noise burst) for a random sample of 10 cats is reproduced in Table 7.11. Recall that we want to use the sample data to estimate the true mean response rate, $\lambda$. Since the sample size is small ($n = 10$), the sampled population must be approximately normally distributed for the small-sample confidence interval methodology of Section 7.3 to be valid. However, we know that the response rate, $Y$, has an approximate Poisson distribution; consequently, the normality assumption is violated in this sampling situation. Apply the bootstrap procedure to obtain a valid 90% confidence interval for $\lambda$.

CATNERVE

**TABLE 7.11 Auditory Nerve Fiber Response Rates**

| | | | | |
|---|---|---|---|---|
| 15.1 | 14.6 | 12.0 | 19.2 | 16.1 |
| 15.5 | 11.3 | 18.7 | 17.1 | 17.2 |

Solution

To obtain the bootstrap confidence interval, we follow the steps outlined above.

*Step 1* We chose $j = 3{,}000$ for resampling.

*Steps 2–3* SAS was programmed to generate 3,000 random samples of size $n = 10$ (selecting observations with replacement) from the sample data in Table 7.11. The response rates for the first five resamples are shown in Table 7.12.

*Steps 3–4* Next, we programmed SAS to compute the sample mean $\bar{y}$ for each sample. Summary statistics for these 3,000 sample means are displayed in the SAS printout, Figure 7.18.

**TABLE 7.12 Bootstrap Resampling from Data in Table 7.11 (First 5 Samples)**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Sample 1* | 12.0 | 11.3 | 18.7 | 17.2 | 17.2 | 11.3 | 17.1 | 17.2 | 17.1 | 14.6 |
| *Sample 2* | 15.1 | 19.2 | 18.7 | 14.6 | 11.3 | 17.1 | 17.2 | 14.6 | 12.0 | 16.1 |
| *Sample 3* | 17.2 | 18.7 | 15.1 | 14.6 | 11.3 | 15.1 | 17.1 | 16.1 | 17.1 | 15.1 |
| *Sample 4* | 15.1 | 15.1 | 18.7 | 15.5 | 16.1 | 17.1 | 14.6 | 19.2 | 19.2 | 17.2 |
| *Sample 5* | 17.1 | 15.1 | 15.1 | 17.2 | 19.2 | 18.7 | 17.1 | 16.1 | 19.2 | 19.2 |

```
                          The UNIVARIATE Procedure
                              Variable:  YBAR

                                  Moments

N                          3000    Sum Weights                3000
Mean                 15.6907033    Sum Observations       47072.11
Std Deviation        0.79486784    Variance             0.63181488
Skewness             -0.0969683    Kurtosis             -0.0683272
Uncorrected SS       740489.326    Corrected SS         1894.81282
Coeff Variation      5.06585218    Std Error Mean       0.01451223

                        Basic Statistical Measures

            Location                       Variability

        Mean     15.69070     Std Deviation              0.79487
        Median   15.70000     Variance                   0.63181
        Mode     15.37000     Range                      4.89000
                              Interquartile Range        1.05000


                         Quantiles (Definition 5)

                          Quantile       Estimate

                          100% Max         18.110
                          99%              17.505
                          95%              16.980
                          90%              16.690
                          75% Q3           16.230
                          50% Median       15.700
                          25% Q1           15.180
                          10%              14.645
                          5%               14.300
                          1%               13.750
                          0% Min           13.220
```

**FIGURE 7.18**
SAS summary statistics for 3,000 bootstrap values of the sample mean, $\bar{y}$

*Step 5* For a 90% confidence interval, $\alpha = .10$, $\alpha/2 = .05$, and $(1 - \alpha/2) = .95$. Consequently, the bootstrap 90% confidence interval for the population mean, $\lambda$, requires that we obtain the 5th and 95th percentiles of the sampling distribution of $\bar{y}$. These two percentiles are highlighted on Figure 7.16. The 5th percentile is 14.30 and the 95th percentile is 16.98. Thus, the bootstrap 90% confidence interval for $\lambda$ is (14.30, 16.98).

The procedure for obtaining bootstrap confidence intervals for any population parameter, $\theta$, (e.g., $\theta = \sigma^2$, $\theta = \mu_1 - \mu_2$, etc.) follows the same logic as the procedure for estimating $\mu$. The steps are outlined in the box.

### Bootstrap Confidence Interval for a Population Parameter, $\theta$

Let $y_1, y_2, y_3, \ldots, y_n$ represent a random sample of $n$ observations on a random variable $Y$ selected from a population probability distribution with unknown parameter $\theta$. Let $\hat{\theta}$ represent the sample estimate of $\theta$. The bootstrap confidence interval for $\theta$ is obtained by following several steps:

*Step 1* Select $j$, where $j$ is the number of times you will resample.

*Step 2* Randomly sample, with replacement, $n$ values of $Y$ from the original sample data set, $y_1, y_2, y_3, \ldots, y_n$.

*Step 3* Repeat step 2 a total of $j$ times, computing $\hat{\theta}$ each time.

*Step 4* Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_j$ represent the $j$ sample estimates of $\theta$ from resampling.

*Step 5* Find the (approximate) $(\alpha/2)100\%$ and $(1 - \alpha)100\%$ percentiles of the simulated sampling distribution of $\hat{\theta}$ using the sample values from step 4. These two percentiles represent the lower and upper endpoints, respectively, of an approximate $(1 - \alpha)100\%$ confidence interval for $\theta$.

The confidence intervals generated using the procedure in the box are called **percentile bootstrap confidence intervals**. There are several alternative approaches to generating confidence intervals using the bootstrap procedure, such as *bootstrap t intervals, $BC_\alpha$ bootstrap intervals*, and *ABC bootstrap intervals*. These are beyond the scope of this text. Consult the references for details on how to use these procedures.

## Bayesian Estimation Methods

In optional Section 3.7, we introduced *Bayesian statistical methods*—procedures that employ the logic of Bayes' rule (p. 91) to make statistical inferences. For the problem of estimating some population parameter $\theta$, the Bayesian approach is to consider $\theta$ as a random variable with some known probability distribution, $h(\theta)$. In other words, *prior to selecting the sample*, Bayesians assess the likelihood that $\theta$ will take on certain values by choosing an appropriate probability distribution for $\theta$. For this reason, $h(\theta)$ is called the **prior distribution for $\boldsymbol{\theta}$**.

Let $y_1, y_2, y_3, \ldots, y_n$ represent a random sample of size $n$ selected from a population with unknown parameter $\theta$. Also, let $f(y_1, y_2, y_3, \ldots, y_n \mid \theta)$ represent the joint conditional probability distribution of the sample values given $\theta$, and let $h(\theta)$ represent the prior distribution for $\theta$. The key to finding a Bayesian estimator of $\theta$ is the conditional distribution, $g(\theta \mid y_1, y_2, y_3, \ldots, y_n)$—called the **posterior distribution for $\boldsymbol{\theta}$**. Applying Bayes' rule, we can show (proof omitted) that the posterior distribution is

$$g(\theta \mid y_1, y_2, y_3, \ldots, y_n) = \frac{f(y_1, y_2, y_3, \ldots, y_n \mid \theta) \cdot h(\theta)}{f(y_1, y_2, y_3, \ldots, y_n)}$$

where $f(y_1, y_2, y_3, \ldots, y_n) = \int f(y_1, y_2, y_3, \ldots, y_n \mid \theta) \cdot h(\theta)\, d\theta$.

Once the posterior distribution $g(\theta \mid y_1, y_2, y_3, \ldots, y_n)$ is determined, the Bayesian estimate of $\theta$, denoted $\hat{\theta}_B$, is suitably chosen. For example, $\hat{\theta}_B$ may be the mean of the posterior distribution or the median of the posterior distribution.

Bayesians usually assess a penalty for selecting a value of $\hat{\theta}_B$ that differs greatly from the true value of $\theta$. Typically, a **squared error loss function** is used, where the loss is $L = [\theta - \hat{\theta}_B]^2$. Then the estimate of $\theta$ will be the one that minimizes $L$. It can be shown (proof omitted), that the value of $\hat{\theta}_B$ that minimizes squared error loss is the mean of the posterior distribution, denoted $E(\theta \mid Y_1, Y_2, Y_3, \ldots, Y_n)$.

We illustrate the Bayesian estimation method in an example.

**Example 7.21**

**Bayes Estimate of Binomial $p$**

Let $y_1, y_2, y_3, \ldots, y_n$ represent a random sample of size $n$ selected from a Bernoulli probability distribution with unknown probability of success, $p$. Let $X$ represent the sum of the Bernoulli values, $X = \Sigma y_i$. We know (from Section 4.6) that $X$ has a binomial probability distribution with parameters $n$ and $p$. Using a squared error loss function, find the Bayesian estimate of $p$. Assume that the prior distribution for $p$ is a beta probability distribution with parameters $\alpha = 1$ and $\beta = 2$.

Solution

Since we know that $X$ has a discrete binomial distribution (conditional on $n$ and $p$) and $p$ has a continuous prior beta distribution (with $\alpha = 1$ and $\beta = 2$), we have

$$p(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad 0 \le x \le n$$

$$h(p) = \frac{\Gamma(\alpha+\beta)p^{\alpha-1}(1-p)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\Gamma(3)p^0(1-p)^1}{\Gamma(1)\Gamma(2)}$$

$$= 2(1-p), \qquad 0 < p < 1$$

The posterior distribution of $p$ (conditional on the sum, $X = x$) is

$$g(p \mid x) = p(x \mid p) \cdot h(p)/p(x)$$

The denominator of the posterior distribution is found as follows:

$$p(x) = \int_0^1 p(x \mid p) \cdot h(p)dp = \int_0^1 \binom{n}{x} p^x(1-p)^{n-x} 2(1-p)dp$$

$$= 2\binom{n}{x}\int_0^1 p^x(1-p)^{n-x+1}dp$$

Note that the expression being integrated takes the form of a beta distribution with $\alpha = (x+1)$ and $\beta = (n-x+2)$. Consequently, it is equal to $\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta) = \Gamma(x+1)\Gamma(n-x+2)/\Gamma(n+3)$. Thus, we have

$$p(x) = 2\binom{n}{x}\int_0^1 p^x(1-p)^{n-x+1}dp$$

$$= 2\binom{n}{x}\Gamma(x+1)\,\Gamma(n-x+2)/\Gamma(n+3)$$

Now, we can find the posterior distribution of $p$:

$$g(p|x) = p(x|n,p) \cdot h(p)/p(x) = \frac{\binom{n}{x} p^x(1-p)^{n-x} 2(1-p)\Gamma(n+3)}{2\binom{n}{x}\Gamma(x+1)\Gamma(n-x+2)}$$

$$= p^x(1-p)^{n-x+1}\Gamma(n+3)/[\Gamma(x+1)\Gamma(n-x+2)]$$

You can see that $g(p|x)$ has the form of a beta distribution with $\alpha = (x+1)$ and $\beta = (n-x+2)$. With squared error loss, the Bayesian estimate of $p$ will be the mean of this conditional distribution. The mean of a beta distribution is $\alpha/(\alpha+\beta)$. Thus, the Bayesian estimate of $p$ is

$$\hat{p}_B = \alpha/(\alpha+\beta) = (x+1)/[x+1+n-x+2] = (x+1)/(n+3)$$

With some algebra, the expression can be written as follows:

$$\hat{p}_B = (x+1)/(n+3) = \left(\frac{n}{n+3}\right)\left(\frac{x}{n}\right) + \left(\frac{1}{3}\right)\left(\frac{3}{n+3}\right)$$

$$= \left(\frac{n}{n+3}\right)\bar{x} + \left(\frac{1}{3}\right)\left(\frac{3}{n+3}\right)$$

Note that $[3/(n+3)]$ is the mean of a beta distribution with $\alpha = 1$ and $\beta = 2$. Thus, the Bayesian estimate of $p$ is simply a weighted average of the sample mean, $\bar{x}$, and the mean of the prior probability distribution for $p$.

Bayesian interval estimates for a parameter $\theta$ can also be derived from the conditional posterior probability distribution for $\theta$. These intervals, called **credible intervals** (or **probability intervals**), take the form

$$P(L < \theta < U) = \int_{L}^{U} g(\theta \mid y_1, y_2, y_3, \ldots, y_n)\, d\theta$$

The lower and upper bounds of the interval, $L$ and $U$, both functions of the sample $Y$ values, are chosen so that the probability that $\theta$ is in the interval is, say, .95. For example, suppose the posterior probability distribution of $\theta$ is determined to have a $\chi^2$ distribution with degrees of freedom $v = 2\,\Sigma y_i$. To guarantee that $P(L < \theta < U) = .95$, $L$ and $U$ will be the 2.5th and 97.5th percentiles of this $\chi^2$ distribution, respectively. Consult the references for more details on how to obtain these Bayesian probability intervals.

## Applied Exercises

7.100 *Bearing strength of concrete FRP strips.* Refer to the *Composites Fabrication Magazine* (Sept. 2004) study of the strength of fiber-reinforced polymer (FRP) composite materials, Exercise 2.47 (p. 51). Recall that 10 specimens of pultruded FRP strips were mechanically fastened to highway bridges and tested for bearing strength. The strength measurements (recorded in megapascal units, MPa) are reproduced in the table. Use the bootstrap procedure to estimate true mean strength of mechanically fastened FRP strips with a 90% confidence interval. Interpret the result.

**FRP**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 240.9 | 248.8 | 215.7 | 233.6 | 231.4 | 230.9 | 225.3 | 247.3 | 235.5 | 238.0 |

*Source:* Data are simulated from summary information provided in *Composites Fabrication Magazine*, Sept. 2004, p. 32 (Table 1).

7.101 *Contamination of New Jersey wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of contaminated well sites located near a New Jersey gasoline service station, Exercise 7.29 (p. 312). The data (parts per billion) on the level of methyl *tert*-butyl ether (MTBE) at 12 sampled sites are reproduced in the table. Use the bootstrap procedure to estimate the mean MTBE level for all well sites located near the New Jersey gasoline service station with a 99% confidence interval. Compare the bootstrap interval with the interval you obtained in Exercise 7.29, part **b**.

**NJGAS**

| | | | | | |
|---|---|---|---|---|---|
| 150 | 367 | 38 | 12 | 11 | 134 |
| 12 | 251 | 63 | 8 | 13 | 107 |

7.102 *Estimating the age of glacial drifts.* Refer to the *American Journal of Science* (Jan. 2005) study of the chemical makeup of buried glacial drifts (or tills) in Wisconsin, Exercise 7.64 (p. 331). The data on the ratios of aluminum (Al) to beryllium (Be) in sediment for a sample of 26 buried till specimens is reproduced in the table.

a. Use the bootstrap procedure to estimate the true proportion of till specimens with an Al/Be ratio that exceeds 4.5 using a 95% confidence interval.

b. Compare the bootstrap interval with the interval you obtained in Exercise 7.56, part **b**. Why might the bootstrap interval be more appropriate?

**TILLRATIO**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.76 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 |
| 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 | 2.73 | 2.95 | 2.25 | 2.73 | 2.55 | 3.06 |

*Source:* Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

7.103 *White light interferometry.* In fields that require high-precision surface height maps, white light interferometry (WLI) has become the standard inspection tool. Because WLI generates two-dimensional height profiles, and standard mechanical devices used with WLI generate only one-dimensional profiles, engineers must estimate the mean height of pixels generated in WLI surface maps. In *Optical Engineering* (Jan. 2005), German researchers applied Bayesian estimation to solve the problem. A simplified version of the research is stated as follows: Let $Y$ represent the height for a pixel generated by WLI. Assume that $Y$ takes on the value 1 with probability $p$ and the value 0 with probability $(1 - p)$. Also, assume that $p$ has a beta distribution with parameters $\alpha = 1$ and $\beta = 2$. Now let $y_1, y_2, y_3, \ldots, y_n$ represent the heights for a sample of $n$ pixels. Using a squared-error loss function, find the Bayesian estimate of $p$ if $\bar{y} = .80$. [*Hint*: Use the result in Example 7.20.]

## Theoretical Exercises

7.104 Let $y_1, y_2, y_3, \ldots, y_n$, represent a random sample of size $n$ selected from a Poisson probability distribution with unknown mean $\lambda$. Let $X$ represent the sum of the Poisson values, $X = \Sigma Y_i$. Then $X$ has a Poisson distribution with mean $n\lambda$. Assume that the prior distribution for $\lambda$ is an exponential probability distribution with parameter $\beta$.

a. Find the posterior distribution, $g(\lambda | x)$.

b. Using a squared error loss function, find the Bayesian estimate of $\lambda$.

7.105 Let $y_1, y_2, y_3, \ldots, y_n$, represent a random sample of size $n$ selected from a normal probability distribution with unknown mean $\mu$ and variance $\sigma^2 = 1$. Then the sample mean, $\bar{y}$, has a normal distribution with mean $\mu$ and variance $\sigma^2 = 1/n$. Assume that the prior distribution for $\mu$ is a normal distribution with a mean of 5 and a variance of 1.

a. Find the posterior distribution, $g(\mu|\bar{y})$.

b. Using a squared error loss function, show that the Bayesian estimate of $\mu$ is a weighted average of $\bar{y}$ and the mean of the prior distribution, 5.

## *STATISTICS IN ACTION REVISITED*
### Bursting Strength of PET Beverage Bottles

We now return to the *Journal of Data Science* study of the bursting strength of PET bottles made from two different mold designs (SIA, p. 289). The researchers want to determine if the new mold design (which reduces the downtime of the manufacturing process) is comparable to the old design in terms of bursting strength. Recall that experimental data was obtained by testing 768 PET bottles of each design. The bursting strengths (pounds per square inch) are saved in the **PETBOTTLE** file.

Initially, the researchers compared the mean bursting strengths of the two designs. A large-sample 95% confidence interval for the difference, $(\mu_{NEW} - \mu_{OLD})$, is shown at the bottom of the SAS printout, Figure SIA7.1. The resulting interval, (-19.91, -14.16), implies that the mean bursting strength of the old mold design exceeds the mean bursting strength of the new design, and this difference is anywhere from 14.61 to 19.91 psi. Based on this result, it appears that the new design is inferior to the old design. Before recommending that the facility continue with the old mold design, the researchers compared the bursting strength distributions for the two designs.

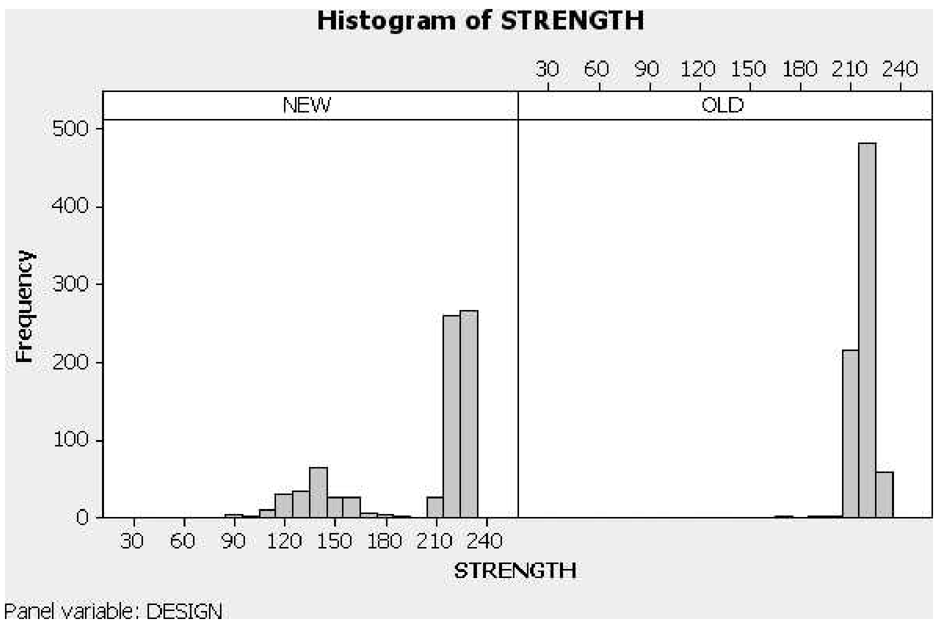
Histograms for the bursting strengths of the old and new designs are shown in the MINITAB printout, Figure SIA7.2. Note that neither distribution appears to be normally distributed. This fact does not compromise

**FIGURE SIA7.1**
SAS 95% Confidence Interval for Difference Between Mean Bursting Strengths of the Two Mold Designs

```
Sample Statistics

     Group            N       Mean     Std. Dev.    Std. Error
     ----------------------------------------------------------
     NEW            768   199.3958       39.897        1.4396
     OLD            768   216.4323       7.5173        0.2713

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are    t statistic       Df        Pr > t
     ----------------------------------------------------------
     Equal                 -11.629       1534        <.0001
     Not Equal             -11.629     821.39        <.0001

95% Confidence Interval for the Difference between Two Means

                   Lower Limit    Upper Limit
                   -----------    -----------
                     -19.91         -14.16
```

the inference made from the large-sample confidence interval above, since the Central Limit Theorem guarantees that the distribution of the difference between the sample means, $(\bar{y}_{\text{NEW}} - \bar{y}_{\text{OLD}})$, is approximately normally distributed. However, the histograms give some insight into the nature of the problem with the new design. Note the bi-modal distribution for the new design in Figure SIA7.2. The new mold design appears to be experiencing a phenomenon called "early or infant mortality" – that is, some bottles produced from the new design are bursting at unusually low pressures. This could be caused by unfamiliarity with the new mold or inadequacies in the blow machine's air flow to the new mold. The researchers also noticed that at the upper end of the bursting strength distribution, the new design tends to have higher bursting strengths than the old design. Consequently, they suggest that if the early mortality problem can be identified and removed, the new design may actually be more reliable (i.e., have a larger bursting strength mean) than the old design.

**FIGURE SIA7.2**
MINITAB Histograms of Bursting Strengths for the Two Mold Designs



To investigate this phenomenon, the researchers removed all observations from the data set that had bursting strengths below 200 psi and reanalyzed the data. The results are shown in the SAS printout, Figure SIA7.3. The 95% confidence interval for the difference, ($\mu_{\text{NEW}} - \mu_{\text{OLD}}$), is now (5.57, 6.60). Thus, when the early mortality data is removed, the mean bursting strength of the new mold design exceeds the mean bursting strength of the old design anywhere from 5.57 to 6.60 psi.

Because of this analysis, the owners of the facility were motivated to solve the early mortality problem for the new mold.

**FIGURE SIA7.3**
SAS 95% Confidence Interval for Difference Between Mean Bursting Strengths of the Two Mold Designs – Early Mortality Data Removed

```
Sample Statistics

     Group           N       Mean    Std. Dev.   Std. Error
     ------------------------------------------------------
     NEW           555   223.1063       4.4783       0.1901
     OLD           758   217.0172        4.865       0.1767

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are    t statistic      Df        Pr > t
     ------------------------------------------------------
     Equal                    23.163     1311       <.0001
     Not Equal                23.461   1244.8       <.0001

95% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                   5.57           6.60
```

# Quick Review

## Key Terms

(*Note: Items marked with an asterisk (*) are from the optional section in this chapter.*)

- *Bayes credible (probability) intervals 354
- *Bayesian estimation 352
- Biased estimator 291
- *Bootstrap estimation 350
- Confidence coefficient 301
- Confidence interval 301
- Interval estimator 290
- Jackknife estimators 299
- Least-squares method 488
- Likelihood function 296
- Lower confidence limit 302
- Matched pairs 322
- Maximum likelihood method 297
- Mean squared error 291
- Method of moments 294
- Minimum variance unbiased estimator 291
- Paired observations 322
- *Percentile bootstrap confidence intervals 352
- Pivotal method 301
- Pivotal statistic 301
- Point estimator 289
- Pooled estimate of variance 316
- Population moment 294
- *Posterior probability distribution 354
- *Prior probability distribution 354
- *Resampling 350
- Robust estimators 300
- Sample moment 294
- *Squared error loss 353
- Unbiased estimator 291
- Upper confidence limit 302

## Key Formulas
## Summary of Estimation Procedures: One-Sample Case

| Parameter $\theta$ | Estimator $\hat{\theta}$ | $E(\hat{\theta})$ | $\sigma_{\hat{\theta}}$ | Approximation to $\sigma_{\hat{\theta}}$ | $(1-\alpha)100\%$ Confidence Interval | Sample Sizes | Additional Assumptions | |
|---|---|---|---|---|---|---|---|---|
| Mean $\mu$ | $\bar{y}$ | $\mu$ | $\frac{\sigma}{\sqrt{n}}$ | $\frac{s}{\sqrt{n}}$ | $\bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ | $n \geq 30$ | None | 308 |
| | | | | | $\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ where $t_{\alpha/2}$ is based on $(n-1)$ df | $n < 30$ | Normal population | 309 |
| Binomial proportion $p$ | $\hat{p} = \frac{y}{n}$ | $p$ | $\sqrt{\frac{pq}{n}}$ | $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ | $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ | $n\hat{p} \geq 4, n\hat{q} \geq 4$ | None | 329 |
| Variance $\sigma^2$ | $s^2$ | $\sigma^2$ | Not needed | Not needed | $\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$ where $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are based on $(n-1)$ df | All $n$ | Normal population | 337 |

## Summary of Estimation Procedures: Two-Sample Case

| Parameter $\theta$ | Estimator $\hat{\theta}$ | $E(\hat{\theta})$ | $\sigma_{\hat{\theta}}$ | Approximation to $\sigma_{\hat{\theta}}$ | $(1-\alpha)100\%$ Confidence Interval | Sample Sizes | Additional Assumptions | |
|---|---|---|---|---|---|---|---|---|
| $(\mu_1 - \mu_2)$ Difference between population means: Independent samples | $(\bar{y}_1 - \bar{y}_2)$ Difference between sample means | $(\mu_1 - \mu_2)$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ | $(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ | $n_1 \geq 30, n_2 \geq 30$ | None | 316 |
| | | | $\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ | $\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ | $(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ df | Either $n_1 < 30$ or $n_2 < 30$, or both | Both populations normal with equal variances $(\sigma_1^2 = \sigma_2^2)$ | 318 |
| $\mu_d = (\mu_1 - \mu_2)$ Difference between population means: Matched pairs | $\bar{d} = \Sigma d_i/n$ Mean of sample differences | $\mu_d$ | $\frac{\sigma_d}{\sqrt{n_d}}$ | $\frac{s_d}{\sqrt{n_d}}$ where $s_d$ is the standard deviation of the sample of differences | $\bar{d} \pm z_{\alpha/2}\left(\frac{s_d}{\sqrt{n_d}}\right)$ | $n_d > 30$ | None | 323 |
| | | | | | $\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n_d}}\right)$ where $t_{\alpha/2}$ is based on $(n_d - 1)$ df | $n_d < 30$ | Population of differences $d_i$ is normal | |

### Summary of Estimation Procedures: Two-Sample Case (*continued*)

| Parameter $\theta$ | Estimator $\hat{\theta}$ | $E(\hat{\theta})$ | $\sigma_{\hat{\theta}}$ | Approximation to $\sigma_{\hat{\theta}}$ | $(1-\alpha)100\%$ Confidence Interval | Sample Size | Additional Assumption | |
|---|---|---|---|---|---|---|---|---|
| $(p_1 - p_2)$ Difference between two binomial parameters | $(\hat{p}_1 - \hat{p}_2)$ Difference between the sample proportions $\hat{p}_1 = y_1/n_1$ and $\hat{p}_2 = y_2/n_2$ | $(p_1 - p_2)$ | $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$ | $\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$ | $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$ | $n_1\hat{p}_1 \geq 4, n_2\hat{p}_2 \geq 4$ $n_1\hat{q}_1 \geq 4, n_2\hat{q}_2 \geq 4$ | Independent samples | 332 |
| $\sigma_1^2/\sigma_2^2$ Ratio of population variances | $s_1^2/s_2^2$ Ratio of sample variances | $\sigma_1^2/\sigma_2^2$ | Not needed | Not needed | $\left(\frac{s_1^2}{s_2^2}\right)\frac{1}{F_{\alpha/2(\nu_1,\nu_2)}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \left(\frac{s_1^2}{s_2^2}\right) F_{\alpha/2(\nu_2,\nu_1)}$ where $F_{\alpha/2(\nu_1,\nu_2)}$ is based on $\nu_1 = (n_1 - 1)$ numerator and $\nu_2 = (n_1 - 1)$ denominator d.f. and $F_{\alpha/2(\nu_2,\nu_1)}$ is based on $\nu_2 = (n_2 - 1)$ numerator and $\nu_1 = (n_1 - 1)$ denominator d.f. | All $n_1$ and $n_2$ | Independent samples from two normal populations | 342 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $\theta$ | theta | General population parameter |
| $\hat{\theta}$ | theta-hat | Point estimate of a population parameter |
| $b(\theta)$ | bias of $\theta$ | Bias of an estimator, $\hat{\theta}$ |
| $\mu$ | mu | Population mean |
| $p$ | | Population proportion |
| $H$ | | Half-width of confidence interval |
| $\alpha$ | alpha | $(1 - \alpha)$ represents the confidence coefficient |
| $z_{\alpha/2}$ | $z$ of alpha over 2 | $z$ value used in a $100(1 - \alpha)\%$ large-sample confidence interval |
| $t_{\alpha/2}$ | $t$ of alpha over 2 | $t$ value used in a $100(1 - \alpha)\%$ small-sample confidence interval |
| $\bar{y}$ | $y$-bar | Sample mean; point estimate of $\mu$ |
| $\hat{p}$ | $p$-hat | Sample proportion; point estimate of $p$ |
| $\sigma^2$ | sigma squared | Population variance |
| $s^2$ | | Sample variance; point estimate of $\sigma^2$ |
| $\nu$ | nu | Degrees of freedom for $t$ and $\chi^2$ statistic |
| $\mu_1 - \mu_2$ | mu-1 minus mu-2 | Difference between population means |
| $\bar{y}_1 - \bar{y}_2$ | $y$-bar-1 minus $y$-bar-2 | Difference between sample means; point estimate of $(\mu_1 - \mu_2)$ |
| $s_p^2$ | $s$-$p$ squared | Pooled sample variance |
| $\mu_d$ | mu $d$ | Difference between population means, paired data |
| $\bar{d}$ | $d$-bar | Mean of sample differences; point estimate of $\mu_d$ |
| $s_d$ | $s$-$d$ | Standard deviation of sample differences |
| $p_1 - p_2$ | $p$-1 minus $p$-2 | Difference between population proportions |
| $\hat{p}_1 - \hat{p}_2$ | $p$-1 hat minus $p$-2 hat | Difference between sample proportions; point estimate of $p_1 - p_2$ |
| $F_\alpha$ | $F$-alpha | Critical value of $F$ associated with tail area $\alpha$ |
| $\nu_1$ | nu-1 | Numerator degrees of freedom for $F$ statistic |
| $\nu_2$ | nu-2 | Denominator degrees of freedom for $F$ statistic |
| $\dfrac{\sigma_1^2}{\sigma_2^2}$ | sigma-1 squared over sigma-2 squared | Ratio of two population variances |
| $m_k$ | $m$-sub-$k$ | $k$th sample moment |
| $E(y^k)$ | | $k$th population moment |
| $L$ | | Likelihood function |
| $\hat{\theta}_i$ | theta-hat $i$ | Point estimate obtained by omitting the $i$th observation |
| LCL | | Lower confidence limit |
| UCL | | Upper confidence limit |
| $\chi^2_{\alpha/2}$ | $\chi^2$ of alpha over 2 | $\chi^2$ value used in $100(1 - \alpha)\%$ confidence interval |

## Chapter Summary Notes

- A point estimator $\hat{\theta}$ of a population parameter $\theta$ is **unbiased** if $E(\hat{\theta}) = \theta$; otherwise, the estimator is **biased**.
- The **minimum variance unbiased estimator (MVUE)** of a population parameter $\theta$ has the smallest variance among all unbiased estimators.
- Methods of estimation: **pivotal method** (either the **method of moments** or the **maximum likelihood method**), **jackknife method, robust estimation methods, bootstrapping**, and **Bayes' estimation**.
- **Confidence interval**—an interval that encloses an unknown population parameter with a certain level of confidence
- **Confidence coefficient**—the probability that a randomly selected confidence interval encloses the value of the population parameter
- Interpretation of the phrase "*$(1 - \alpha)100\%$ confident*": In repeated sampling, $(1 - \alpha)100\%$ of all similarly constructed intervals will enclose the true parameter value.
- Key words for identifying $\boldsymbol{\mu}$ as the parameter of interest: *mean, average*.
- Key words/phrases for identifying $\boldsymbol{\mu_1 - \mu_2}$ as the parameter of interest: *difference between means or averages, compare two means using independent samples*.
- Key words/phrases for identifying $\boldsymbol{\mu_d}$ as the parameter of interest: *mean or average of paired differences, compare two means using matched pairs*.
- Key words for identifying $\boldsymbol{p}$ as the parameter of interest: *proportion, percentage, rate*.
- Key words/phrases for identifying $\boldsymbol{p_1 - p_2}$ as the parameter of interest: *difference between proportions or percentages, compare two proportions using independent samples*.
- Key words for identifying $\boldsymbol{\sigma^2}$ as the parameter of interest: *variance, spread, variation*.
- Key words/phrases for identifying $\sigma_1^2/\sigma_2^2$ as the parameter of interest: *difference between variances, compare variation in two populations using independent samples*.

## Supplementary Exercises

7.106 *Effectiveness of a passive sampler.* Chemical engineers at the University of Murcia (Spain) conducted a series of experiments to determine the most effective membrane to use in a passive sampler (*Environmental Science & Technology*, Vol. 27, 1993). The effectiveness of a passive sampler was measured by the sampling rate, recorded in cubic centimeters per minute. In one experiment, six passive samplers were positioned with their faces parallel to the air flow and with an air velocity of 90 centimeters per second. After 6 hours, the sampling rate of each was determined. Based on the results, a 95% confidence interval for the mean sampling rate was calculated to be (49.66, 51.48).

a. What is the confidence coefficient for the interval?
b. Give a theoretical interpretation of the confidence coefficient, part **a**.
c. Give a practical interpretation of the confidence interval.
d. What assumptions, if any, are required for the interval to yield valid inferences?

7.107 *Water pollution testing.* The EPA wants to test a randomly selected sample of $n$ water specimens and estimate $\mu$, the mean daily rate of pollution produced by a mining operation. If the EPA wants a 95% confidence interval estimate with a sampling error of 1 milligram per liter (mg/L), how many water specimens are required in the sample? Assume that prior knowledge indicates that pollution readings in water samples taken during a day are approximately normally distributed with a standard deviation equal to 5 (mg/L).

7.108 *Lead and copper in drinking water.* Periodically, the Hillsborough County (Florida) Water Department tests the drinking water of homeowners for contaminants such as lead and copper. The lead and copper levels in water specimens collected for a sample of 10 residents of the Crystal Lake Manors subdivision are shown next.

**LEADCOPP**

| Lead ($\mu$g/L) | Copper (mg/L) |
|---|---|
| 1.32 | .508 |
| 0 | .279 |
| 13.1 | .320 |
| .919 | .904 |
| .657 | .221 |
| 3.0 | .283 |
| 1.32 | .475 |
| 4.09 | .130 |
| 4.45 | .220 |
| 0 | .743 |

*Source:* Hillsborough County Water Department Environmental Laboratory, Tampa, Florida.

a. Construct a 99% confidence interval for the mean lead level in water specimens from Crystal Lake Manors.
b. Construct a 99% confidence interval for the mean copper level in water specimens from Crystal Lake Manors.
c. Interpret the intervals, parts **a** and **b,** in the words of the problem.
d. Discuss the meaning of the phrase, "99% confident."

7.109 *Accuracy and precision of an instrument.* When new instruments are developed to perform chemical analyses of products (food, medicine, etc.), they are usually evaluated with respect to two criteria: accuracy and precision. *Accuracy* refers to the ability of the instrument to identify correctly the nature and amounts of a product's components. *Precision* refers to the consistency with which the instrument will identify the components of the same material. Thus, a large variability in the identification of a single sample of a product indicates a lack of precision. Suppose a pharmaceutical firm is considering two brands of an instrument designed to identify the components of certain drugs. As part of a comparison of precision, 10 test-tube samples of a well-mixed batch of a drug are selected and then 5 are analyzed by instrument A and 5 by instrument B. The data shown in the table are the percentages of the primary component of the drug given by the instruments.

**DRUGPCT**

| | | | | | |
|---|---|---|---|---|---|
| *Instrument A* | 43 | 48 | 37 | 52 | 45 |
| *Instrument B* | 46 | 49 | 43 | 41 | 48 |

a. Construct a 90% confidence interval to compare the precision of the two instruments.
b. Based on the interval estimate of part **a**, what would you infer about the precision of the two instruments?
c. What assumptions must be satisfied to ensure the validity of any inferences derived from the estimate?

7.110 *Attributes of forest access roads.* In Ireland, the majority of commercial forests are located in remote areas on predominantly peat soils. These roads exhibit rapid deterioration when traversed by logging vehicles or other heavy machinery. A study of the attributes of forest access roads in Ireland was published in the *International Journal of Forest Engineering* (July 1999). One measure of the strength of pavement is transient surface deflection—the higher the surface deflection, the weaker the pavement. The type of pavement (mineral subgrade or peat subgrade) was determined for a sample of 72 forest access roads, then each was analyzed for surface deflection (measured in millimeters). The results are summarized in the next table.

a. Compare the two pavement types with a 95% confidence interval. Which pavement subgrade, mineral or peat, is stronger?

| | Pavement Subgrade | |
|---|---|---|
| | Mineral | Peat |
| *Number of roads* | 32 | 40 |
| *Mean surface deflection (mm)* | 1.53 | 3.80 |
| *Standard deviation* | 3.39 | 14.3 |

*Source:* Martin, A. M., et al. "Estimation of the serviceability of forest access roads." *International Journal of Forest Engineering*, Vol. 10, No. 2, July 1999 (adapted from Table 3).

b. Compare the surface deflection variances of the two pavement types with a 95% confidence interval. Which pavement subgrade, mineral or peat, has the largest surface deflection variance?

7.111 *Roofing injuries.* According to a study conducted by the California Division of Labor Research and Statistics, roofing is one of the most hazardous occupations. Of 2,514 worker injuries that caused absences for a full workday or shift after the injury, 23% were attributable to falls from high elevations on level surfaces, 21% to falling hand tools or other materials, 19% to overexertion, and 20% to burns or scalds. Assume that the 2,514 injuries can be regarded as a random sample from the population of all roofing injuries in California.

a. Construct a 95% confidence interval for the proportion of all injuries that are due to falls.
b. Construct a 95% confidence interval for the proportion of all injuries that are due to burns or scalds.

7.112 *Air bags pose danger for children.* By law, all new cars must be equipped with both driver-side and passenger-side safety air bags. There is concern, however, over whether air bags pose a danger for children sitting on the passenger side. In a National Highway Traffic Safety Administration (NHTSA) study of 55 people killed by the explosive force of air bags, 35 were children seated on the front-passenger side. (*Wall Street Journal*, January 22, 1997.) This study led some car owners with children to disconnect the passenger-side air bag. Consider all fatal automobile accidents in which it is determined that air bags were the cause of death. Let $p$ represent the true proportion of these accidents involving children seated on the front-passenger side.

a. Use the data from the NHTSA study to estimate $p$.
b. Construct a 99% confidence interval for $p$.
c. Interpret the interval, part **b**, in the words of the problem.
d. NHTSA investigators determined that 24 of 35 children killed by the air bags were not wearing seat belts or were improperly restrained. How does this information impact your assessment of the risk of an air bag fatality?

7.113 *Jitter in water power systems. Jitter* is a term used to describe the variation in conduction time of a modular pulsed-water power system. Low throughput jitter is critical to successful waterline technology. An investigation of throughput jitter in the plasma opening switch of a prototype system (*Journal of Applied Physics*, Sept. 1993) yielded the following descriptive statistics on conduction time for $n = 18$ trials: $\bar{y} = 334.8$ nanoseconds, $s = 6.3$ nanoseconds. (*Note*: Conduction time is defined as the length of time required for the downstream current to equal 10% of the upstream current.)

a. Construct a 95% confidence interval for the true standard deviation of conduction times of the prototype system.
b. A system is considered to have low throughput jitter if the true conduction time standard deviation is less than 7 nanoseconds. Does the prototype system satisfy this requirement? Explain.

7.114 *Solar irradiation study.* The *Journal of Environmental Engineering* (Feb. 1986) reported on a heat transfer model designed to predict winter heat loss in wastewater treatment clarifiers. The analysis involved a comparison of clear-sky solar irradiation for horizontal surfaces at different sites in the Midwest. The day-long solar irradiation levels (in BTU/sq. ft.) at two midwestern locations of different latitudes (St. Joseph, Missouri, and Iowa Great Lakes) were recorded on each of seven clear-sky winter days. The data are given in the table. Find a 95% confidence interval for the mean difference between the day-long clear-sky solar irradiation levels at the two sites. Interpret the results.

**SOLARAD**

| Date | St. Joseph, Mo. | Iowa Great Lakes |
|---|---|---|
| December 21 | 782 | 593 |
| January 6 | 965 | 672 |
| January 21 | 948 | 750 |
| February 6 | 1,181 | 988 |
| February 21 | 1,414 | 1,226 |
| March 7 | 1,633 | 1,462 |
| March 21 | 1,852 | 1,698 |

*Source:* Wall, D. J., and Peterson, G. "Model for winter heat loss in uncovered clarifiers." *Journal of Environmental Engineering*, Vol. 112. No. 1, Feb. 1986, p. 128.

7.115 *Sampling iron ore consignments.* A large steel corporation conducted an experiment to compare the average iron contents of two consignments of lumpy iron ore. In accordance with industrial standards, $n$ increments of iron ore were randomly selected from each consignment and measured for iron content. From previous experiments, it is known that iron contents vary over a range of roughly 3%. How large should $n$ be if the steel company wants to estimate the difference in mean iron contents of the two consignments correct to within .05% with 95% confidence? (*Hint*: To obtain an approximate value for $\sigma_1$ and $\sigma_2$, set $\sigma_1 = \sigma_2 = \sigma$ and set Range $= 4\sigma$. Then $3 \approx 4\sigma$ and $\sigma \approx \frac{3}{4}$.)

7.116 *Diazinon residue in orchards.* Pesticides applied to an extensively grown crop can result in inadvertent areawide air contamination. *Environmental Science & Technology* (Oct. 1993) reported on air deposition residues of the insecticide diazinon used on dormant orchards in the San Joaquin Valley, California. Ambient air samples were collected and analyzed at an orchard site for each of 11 days during the most intensive period of spraying. The levels of diazinon residue (in mg/m$^3$) during the day and at night are recorded in the table. The researchers want to know whether the mean diazinon residue levels differ from day to night.

**DIAZINON**

| | Diazinon Residue | |
|---|---|---|
| Date | Day | Night |
| Jan. 11 | 5.4 | 24.3 |
| 12 | 2.7 | 16.5 |
| 13 | 34.2 | 47.2 |
| 14 | 19.9 | 12.4 |
| 15 | 2.4 | 24.0 |
| 16 | 7.0 | 21.6 |
| 17 | 6.1 | 104.3 |
| 18 | 7.7 | 96.9 |
| 19 | 18.4 | 105.3 |
| 20 | 27.1 | 78.7 |
| 21 | 16.9 | 44.6 |

*Source:* Selber, J. N., et al. "Air and fog deposition residues for organophosphate insecticides used on dormant orchards in the San Joaquin Valley, California." *Environmental Science & Technology*, Vol. 27, No. 10, Oct. 1993, p. 2240 (Table IV).

a. Analyze the data using a 90% confidence interval.
b. What assumptions are necessary for the validity of the interval estimation procedure of part **a**?
c. Use the interval, part **a**, to answer the researchers' question.

7.117 *Extracting toxic compounds.* A technique, called matrix solid-phase dispersion (MSPD), has been developed for chemically extracting trace organic (toxic) compounds from fish specimens (*chromatographia*, Mar. 1995). Uncontaminated fish fillets were injected with a toxic substance and the MSPD method was used to extract the toxicant. To estimate the precision of the method, seven measurements were obtained on a single fish fillet. Summary statistics on percent of the toxin recovered are given as follows: $\bar{y} = 99\%$, $s = 9\%$. Find an estimate of the recovery percentage variance when using the MSPD method. Use a 95% confidence interval.

**NZBIRDS**

7.118 *Extinct New Zealand birds.* Refer to the *Evolutionary Ecology Research* (July 2003) study of the New Zealand bird population prior to European contact, Exercise 1.12 (p. 6). Two quantitative variables measured for each of the

**MINITAB Output for Exercise 7.118**

**Descriptive Statistics: Body Mass, Egg Length**

| Variable | N | Mean | StDev | Minimum | Maximum |
|---|---|---|---|---|---|
| Body Mass | 116 | 9113 | 31457 | 7.00 | 200000 |
| Egg Length | 115 | 61.06 | 45.46 | 16.00 | 236.00 |

116 bird species were body mass (grams) and egg length (millimeters). Descriptive statistics for these variables are shown on the MINITAB printout above.

a. Use a random number generator to select a random sample of 35 species from the **NZBIRDS** file.
b. Calculate the mean and standard deviation for the 35 sampled values of body mass. Then, use this information to construct a 95% confidence interval for the mean body mass of all 116 bird species.
c. Give a practical interpretation of the interval, part **b**.
d. Check to see if the true mean, $\mu$ (shown on the MINITAB printout), is included in the confidence interval, part **b**. Explain why the interval is very likely to contain $\mu$.
e. Repeat parts b–d for the 35 sampled values of egg length.
f. Ecologists also want to compare the proportions of flightless birds for two New Zealand bird populations—those that are extinct and those that are not extinct. Use the sample information in the table below to form a 95% confidence interval for the difference between the proportion of flightless birds for extinct and nonextinct species.
g. The ecologists are investigating the theory that the proportion of flightless birds will be greater for extinct species than for nonextinct species. Does the confidence interval, part f, support this theory? Explain.

| Bird Population | Number of Species Sampled | Number of Flightless Species |
|---|---|---|
| Extinct | 38 | 21 |
| Nonextinct | 78 | 7 |

7.119 *Quality assurance.* It costs more to produce defective items—since they must be scrapped or reworked—than it does to produce nondefective items. This simple fact suggests that manufacturers should ensure the quality of their products by perfecting their production processes instead of depending on inspection of finished products (Deming, 1986). In order to better understand a particular metal stamping process, a manufacturer wishes to estimate the mean length of items produced by the process during the past 24 hours.

a. How many parts should be sampled in order to estimate the population mean to within .1 millimeter (mm) with 90% confidence? Previous studies of this machine have indicated that the standard deviation of lengths produced by the stamping operation is about 2 mm.
b. Time permits the use of a sample size no larger than 100. If a 90% confidence interval for $\mu$ is constructed using $n = 100$, will it be wider or narrower than would have been obtained using the sample size determined in part **a**? Explain.
c. If management requires that $\mu$ be estimated to within .1 mm and that a sample size of no more than 100 be used, what is (approximately) the maximum confidence level that could be attained for a confidence interval that meets management's specifications?

7.120 *Strength of epoxy-repaired joints.* The methodology for conducting a stress analysis of newly designed timber structures is well known. However, few data are available on the actual or allowable stress for repairing damaged structures. Consequently, design engineers often propose a repair scheme (e.g., gluing) without any knowledge of its structural effectiveness. To partially fill this void, a stress analysis was conducted on epoxy-repaired truss joints (*Journal of Structural Engineering*, Feb. 1986). Tests were conducted on epoxy-bonded truss joints made of various species of wood to determine actual glue-line shear stress recorded in pounds per square inch (psi). Summary information for independent random samples of southern pine and ponderosa pine truss joints is given in the accompanying table.

| | Southern Pine | Ponderosa Pine |
|---|---|---|
| *Sample Size* | 100 | 47 |
| *Mean Shear Stress, psi* | 1,312 | 1,352 |
| *Standard Deviation* | 422 | 271 |

*Source:* Avent, R. R. "Design criteria for epoxy repair of timber structures." *Journal of Structural Engineering*, Vol. 112, No. 2, Feb. 1986, p. 232.

a. Estimate the difference between the mean shear strengths of epoxy-repaired truss joints for the two species of wood with a 90% confidence interval.
b. Construct a 90% confidence interval for the ratio of the shear stress variances of epoxy-repaired truss joints for the two species of wood. Based on this interval, is there evidence to indicate that the two shear stress variances differ? Explain.

7.121 *Cell growth experiment.* Geneticists at Duke University Medical Center have identified the E2F1 transcription factor as an important component of cell proliferation control (*Nature*, Sept. 23, 1993). The researchers induced DNA synthesis in two batches of serum-starved cells. Each cell in one batch was micro-injected with the E2F1 gene, whereas the cells in the second batch (the controls) were not exposed to E2F1. After 30 hours, the number of cells in each batch that exhibited altered growth was determined. The results of the experiment are summarized in the table on p. 366.

| | Control | E2F1 Treated Cells |
|---|---|---|
| *Total Number of Cells* | 158 | 92 |
| *Number of Growth-Altered Cells* | 15 | 41 |

*Source:* Johnson, D. G., et al. "Expression of transcription factor E2F1 induces quiescent to enter S phase." *Nature*, Vol. 365, No. 6444, Sept. 23, 1993, p. 351 (Table 1).

a. Compare the percentages of cells exhibiting altered growth in the two batches with a 90% confidence interval.
b. Use the interval, part **a**, to make an inference about the ability of the E2F1 transcription factor to induce cell growth.

7.122 *Iodine concentration study.* An experiment was conducted to investigate the precision of measurements of a saturated solution of iodine after an extended period of continuous stirring. The data shown in the table represent $n = 10$ iodine concentration measurements on the same solution. The population variance $\sigma^2$ measures the variability—i.e., the precision—of a measurement. Find a 95% confidence interval for $\sigma^2$. Interpret the result.

**IODINE**

| Run | Concentration |
|---|---|
| 1 | 5.507 |
| 2 | 5.506 |
| 3 | 5.500 |
| 4 | 5.497 |
| 5 | 5.506 |
| 6 | 5.527 |
| 7 | 5.504 |
| 8 | 5.490 |
| 9 | 5.500 |
| 10 | 5.497 |

7.123 *Bacteria in bottled water.* Is the bottled water you drink safe? According to *U.S. News & World Report*, the Natural Resources Defense Council warns that the bottled water you are drinking may contain more bacteria and other potentially carcinogenic chemicals than allowed by state and federal regulations. Of the more than 1,000 bottles studied, nearly one-third exceeded government levels. Suppose that the Natural Resources Defense Council wants an updated estimate of the population proportion of bottled water that violates at least one government standard. Determine the sample size (number of bottles) needed to estimate this proportion to within $\pm 0.01$ with 99% confidence.

7.124 *Heat transfer study.* The theoretical relationship between heat flux and temperature gradient for homogeneous materials is well known and described by a Fourier equation. However, this relationship does not hold for nonhomogeneous materials such as porous-capillary bodies, cellular systems, suspensions, and pastes. An experiment was conducted to estimate the mean thermal relaxation time (defined as the mean time needed for accumulating the thermal energy required for propagative transfer of heat) for several nonhomogeneous materials (*Journal of Heat Transfer*, Aug. 1990). A 95% confidence interval for the mean thermal relaxation time of sand was found to be $20.2 \pm 6.4$ seconds.

a. Give a practical interpretation of the 95% confidence interval.
b. Give a theoretical interpretation of the 95% confidence interval.

7.125 *Muscle fiber study.* Marine biochemists at the University of Tokyo studied the properties of crustacean skeletal muscles (*The Journal of Experimental Zoology*, Sept. 1993). It is well known that certain muscles contract faster than others. The main purpose of the experiment was to compare the biochemical properties of fast and slow muscles of crayfish. Using crayfish obtained from a local supplier, 12 fast-muscle fiber bundles were extracted and each fiber bundle tested for uptake of the protein $Ca^{2+}$. Twelve slow-muscle fiber bundles were extracted from a second sample of crayfish and $Ca^{2+}$ uptake measured. The results of the experiment are summarized here. (All $Ca^{2+}$ measurements are in moles per milligram.) Analyze the data using a 95% confidence interval. Make an inference about the difference between the protein uptake means of fast and slow muscles.

| Fast Muscle | Slow Muscle |
|---|---|
| $n_1 = 12$ | $n_2 = 12$ |
| $\bar{y}_1 = .57$ | $\bar{y}_2 = .37$ |
| $s_1 = .104$ | $s_2 = .035$ |

*Source:* Ushio, H., and Watabe, S. "Ultra-structural and biochemical analysis of the sarcoplasmic reticulum from crayfish fast and slow striated muscles." *The Journal of Experimental Zoology*, Vol. 267, Sept. 1993, p. 16 (Table 1).

## Theoretical Exercises

7.126 Let $\bar{y}_1$ be the mean of a random sample of $n_1$ observations from a Poisson distribution with mean $\lambda_1$, and let $\bar{y}_2$ be the mean of a random sample of $n_2$ observations from a Poisson distribution with mean $\lambda_2$. Assume the samples are independent.

a. Show that $(\bar{y}_1 - \bar{y}_2)$ is an unbiased estimator of $(\lambda_1 - \lambda_2)$.
b. Find $V(\bar{y}_1 - \bar{y}_2)$. How could you estimate this variance?
c. Construct a large-sample $(1 - \alpha)100\%$ confidence interval for $(\lambda_1 - \lambda_2)$. [*Hint*: Consider

$$\frac{(\bar{y}_1 - \bar{y}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\dfrac{\bar{y}_1}{n_1} + \dfrac{\bar{y}_2}{n_2}}}$$

as a pivotal statistic.]

7.127 Let $y_1, y_2, \ldots, y_n$ denote a random sample from a uniform distribution with probability density

$$f(y) = \begin{cases} 1 & \text{if } \theta \le y \le \theta + 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Show that $\bar{y}$ is a biased estimator of $\theta$, and compute the bias.

b. Find $V(\bar{y})$.

c. What function of $\bar{y}$ is an unbiased estimator of $\theta$?

7.128 Suppose $y$ is a random sample of size $n = 1$ from a normal distribution with mean 0 and unknown variance $\sigma^2$.

a. Show that $y^2/\sigma^2$ has a chi-square distribution with 1 degree of freedom. (*Hint*: The result follows directly from Theorem 6.11.)

b. Derive a 95% confidence interval for $\sigma^2$ using $y^2/\sigma^2$ as a pivotal statistic.

7.129 Suppose $y$ is a random sample of size $n = 1$ from a gamma distribution with parameters $\alpha = 1$ and arbitrary $\beta$.

a. Show that $2y/\beta$ has a gamma distribution with parameters $\alpha = 1$ and $\beta = 2$. (*Hint:* Use the distribution function approach of Section 6.7.)

b. Use the result of part **a** to show that $2y/\beta$ has a chi-square distribution with 2 degrees of freedom. (*Hint*: The result follows directly from Section 5.7.)

c. Derive a 95% confidence interval for $\beta$ using $2y/\beta$ as a pivotal statistic.

7.130 Suppose $y$ is a single observation from a normal distribution with mean $\mu$ and variance 1. Use $y$ to find a 95% confidence interval for $\mu$. [*Hint:* Start with the pivotal statistic $(y - \mu)$ and show

$$P(-z_{.025} \le y - \mu \le z_{.025}) = .95$$

Then follow the method of Example 7.6.]

7.131 A confidence interval for $\theta$ is said to be *unbiased* if the expected value of the interval midpoint is equal to $\theta$.

a. Show that the small-sample confidence interval for $\mu$,

$$\bar{y} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{y} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

is unbiased.

b. Show that the confidence interval for $\sigma^2$,

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

is biased.

7.132 Suppose $y$ is a single observation from a uniform distribution defined on the interval from 0 to $\theta$. Find a 95% confidence limit LCL for $\theta$ such that $P(\text{LCL} < \theta < \infty) = .95$. [*Hint*: Start with the pivotal statistic $y/\theta$ and show (using the method of Chapter 6) that $y/\theta$ is uniformly distributed on the interval from 0 to 1. Then observe that

$$P\left(0 < \frac{y}{\theta} < .95\right) = \int_0^{.95} (1)\,dy = .95$$

and proceed to obtain LCL.]

CHAPTER 8

# Tests of Hypotheses

**OBJECTIVE**

*To introduce the basic concepts of a statistical test of a hypothesis; to present statistical tests for several common population parameters and to illustrate their use in practical sampling situations*

**CONTENTS**



*STATISTICS IN ACTION*
Comparing Methods for Dissolving Drug Tablets—Dissolution Method Equivalence Testing

## *STATISTICS IN ACTION*

### Comparing Methods for Dissolving Drug Tablets—Dissolution Method Equivalence Testing

In the pharmaceutical industry, quality engineers are responsible for maintaining the quality of drug products produced in the manufacturing process. The key to quality is an assessment of product characteristics through repeated measurements of the variable of interest. When the variable is the concentration of a particular constituent in a mixture, the process is called an assay. For this "Statistics in Action", we focus on a chemical assay to determine how fast a solid-dosage pharmaceutical product (e.g., an aspirin tablet or capsule) dissolves. Since variation in the dissolution of the drug can have harmful side effects on the patient, quality inspectors require a test that accurately measures dissolution.

In "Dissolution Method Equivalence" (Chapter 4, *Statistical Case Studies: A Collaboration between Academe and Industry,* ASA-SIAM Series on Statistics and Applied Probability, 1998), statisticians Russell Reeve and Francis Giesbrecht explored the dissolution characteristics of a new immediate-release drug product manufactured by a well-known pharmaceutical company. An immediate-release product is designed to dissolve and enter the bloodstream as fast as possible. To test for dissolution of the solid-dosage drug, the company uses an apparatus with six vessels or tubes, each tube containing a dissolving solution. Drug tablets or capsules are dropped in the tubes. Then, at predetermined times, a small amount of the solution is withdrawn from each tube and analyzed using high performance liquid chromatography (HPLC). The HPLC device quantifies how much of the drug is in the solution; this value is expressed as percent of label strength (%LS).

Initially, the process described above is typically performed in a laboratory at the company's research and development center. Once dissolution of the drug has been deemed satisfactory, the process is transferred to the manufacturing facility. However, federal regulations require that quality engineers at the manufacturing site produce results equivalent to those at the R&D center. In fact, the company must provide documentation that verifies that any two sites using the dissolution test produce equivalent assay results.

Dissolution test data for an analgesic in tablet form conducted at two manufacturing sites (New Jersey and Puerto Rico) are listed in Table SIA8.1. (These data are saved in the **DISSOLVE** file.) Note that %LS values were obtained at four different points in time – after 20 minutes, after 40 minutes, after 60 minutes, and after 120 minutes – for each of the six vessels. Based on the sample data, do the two sites produce equivalent assay results?

In the Statistics in Action Revisited section later in this chapter, we demonstrate how the methods outlined in this chapter can be used to make the comparison required by the quality control engineers at the pharmaceutical company.

**DISSOLVE**

**TABLE SIA8.1 Dissolution Test Data (Percent Label Strength)**

| Site | Time (min) | Vessel 1 | Vessel 2 | Vessel 3 | Vessel 4 | Vessel 5 | Vessel 6 |
|---|---|---|---|---|---|---|---|
| *New Jersey* | 20 | 5 | 10 | 2 | 7 | 6 | 0 |
| | 40 | 72 | 79 | 81 | 70 | 72 | 73 |
| | 60 | 96 | 99 | 93 | 95 | 96 | 99 |
| | 120 | 99 | 99 | 96 | 100 | 98 | 100 |
| *Puerto Rico* | 20 | 10 | 12 | 7 | 3 | 5 | 14 |
| | 40 | 65 | 66 | 71 | 70 | 74 | 69 |
| | 60 | 95 | 99 | 98 | 94 | 90 | 92 |
| | 120 | 100 | 102 | 98 | 99 | 97 | 100 |

*Source:* Reeve, R., and Giesbrecht, F. "Dissolution Method Equivalence." *Statistical Case Studies: A Collaboration between Academe and Industry*, (editors: R. Peck, L. Haugh, and A. Goodman), ASA-SIAM Series on Statistics and Applied Probability, 1998 (Chapter 4, Table 4).

## 8.1 The Relationship Between Statistical Tests of Hypotheses and Confidence Intervals

As stated in Chapter 7, there are two general methods available for making inferences about population parameters. We can estimate their values using confidence intervals (the subject of Chapter 7) or we can make decisions about them. Making decisions about specific values of the population parameters—**testing hypotheses** about these values—is the topic of this chapter.

Confidence intervals and hypothesis tests are related and can be used to make decisions about parameters. For example, suppose an investigator for the Environmental Protection Agency (EPA) wants to determine whether the mean level $\mu$ of a certain type of pollutant released into the atmosphere by a chemical company meets the EPA guidelines. If 3 parts per million is the upper limit allowed by the EPA, the investigator would want to use sample data (daily pollution measurements) to decide whether the company is violating the law, i.e., to decide whether $\mu > 3$. If, say, a 99% confidence interval for $\mu$ contained only numbers greater than 3, then the EPA would be confident that the mean exceeds the established limit.

As a second example, consider a manufacturer that purchases electric light fuses in lots of 10,000, and suppose that the supplier of the fuses guarantees that no more than 1% of the fuses in any given lot are defective. Since the manufacturer cannot test each of the 10,000 fuses in a lot, he must decide whether to accept or reject a lot based on an examination of a sample of fuses selected from the lot. If the number $Y$ of defective fuses in a sample of, say, $n = 100$, is large, he will reject the lot and send it back to the supplier. Thus, he wants to decide whether the proportion $p$ of defectives in the lot exceeds .01, based on information contained in a sample. If a confidence interval for $p$ falls below .01, then the manufacturer will accept the lot and be confident that the proportion of defectives is less than 1%; otherwise, he will reject it.

The examples in the preceding paragraphs illustrate how a confidence interval can be used to make a decision about a parameter. Note that both applications are one-directional; the EPA wants to determine whether $\mu > 3$ and the manufacturer wants to know if $p > .01$. (In contrast, if the manufacturer is interested in determining whether $p > .01$ or $p < .01$, the inference would be two-directional.)

Recall, from Chapter 7, that to find the value of $z$ (or $t$) used in a $(1 - \alpha)100\%$ confidence interval, the value of $\alpha$ is divided in half and $\alpha/2$ is placed in both the upper and lower tails of the $Z$ (or $T$) distribution. Consequently, confidence intervals are designed to be two-directional. Use of a two-directional technique in a situation where a one-directional method is desired will lead the researcher (e.g., the EPA or the manufacturer) to understate the level of confidence associated with the method. As we will explain in this chapter, hypothesis tests are appropriate for either one- or two-directional decisions about a population parameter.

## 8.2 Elements and Properties of a Statistical Test

We now return to the EPA example to introduce the concepts involved in a test of a hypothesis. We will use a method analogous to proof by contradiction. The theory the EPA wants to support, called the **alternative** (or **research**) **hypothesis**, is that $\mu > 3$, where $\mu$ is the true mean level of pollution in parts per million. The alternative hypothesis is denoted by the symbol $H_a$. The theory contradictory to the alternative hypothesis, that $\mu$ is at most equal to 3, say, $\mu = 3$, is called the **null hypothesis** and is denoted by the symbol $H_0$. Thus, the EPA hopes to show support for the alternative hypothesis, $\mu > 3$, by obtaining sample evidence indicating that the null hypothesis, $\mu = 3$, is false. That is, the EPA wants to test

$$H_0: \quad \mu = 3$$

$$H_a: \quad \mu > 3$$

The decision whether to reject the null hypothesis is based on a statistic, called a **test statistic**, computed from sample data. For example, suppose the EPA plans to base its decision on a sample of $n = 30$ daily pollution readings. If the sample mean $\bar{y}$ of the 30 pollution measurements is much larger than 3, the EPA would tend to reject the null hypothesis and conclude that $\mu > 3$. However, if $\bar{y}$ is smaller than 3, say, $\bar{y} = 2.8$ parts per million, there is insufficient evidence to refute the null hypothesis. Thus, the sample mean $\bar{y}$ serves as a test statistic.

The values that the test statistic $\bar{y}$ can assume will be divided into two sets. Those larger than some specified value, say, $\bar{y} \geq 3.1$, will imply rejection of the null hypothesis and acceptance of the alternative hypothesis. This set of values of the test statistic is known as the **rejection region** for the test. A test of the null hypothesis, $H_0$: $\mu = 3$, against the alternative hypothesis, $H_a$: $\mu > 3$, employing the sample mean $\bar{y}$ as a test statistic and $\bar{y} \geq 3.1$ as a rejection region, represents one particular test that possesses specific properties. If we change the rejection region to $\bar{y} \geq 3.2$, we obtain a different test with different properties.

The preceding discussion indicates that a statistical test consists of the five elements summarized in the box.

### Elements of a Statistical Test

1. **Null hypothesis, $H_0$**, about one or more population parameters
2. **Alternative hypothesis, $H_a$**, that we will accept if we decide to reject the null hypothesis
3. **Test statistic**, computed from sample data
4. **Rejection region**, indicating the values of the test statistic that will imply rejection of the null hypothesis
5. **Conclusion,** the decision made on whether to accept or reject the null hypothesis

Since a statistical test can result in one of only two outcomes—rejecting or not rejecting the null hypothesis—the test conclusion is subject to only two types of error. In the preceding example, the EPA wants to test $H_0$: $\mu = 3$ against $H_a$: $\mu > 3$. If the EPA investigator concludes that $H_a$ is true (i.e., if he rejects $H_0$), then the EPA will charge the company with violating its pollution standards. The two errors that the EPA can make are shown in Table 8.1.

The EPA might reject the null hypothesis if, in fact, it is true. That is, the EPA might charge the company with violating its standards, when, in fact, the company is innocent (Type I error). Or the EPA might decide to accept the null hypothesis if, in fact, it is false. That is, the EPA may conclude that the company is not in violation of

**TABLE 8.1 Conclusions and Consequences for the EPA's Test of Hypothesis**

| | True State of Nature | |
|---|---|---|
| EPA Decision | Company Not in Violation ($H_0$ true) | Company in Violation ($H_a$ true) |
| Company in Violation (Reject $H_0$) | Type I error | Correct decision |
| Company Not in Violation (Accept $H_0$) | Correct decision | Type II error |

the pollution standards when, in fact, the company is in violation (Type II error). The probabilities of making these two types of errors measure the risks of making incorrect decisions when we perform a test of hypothesis and, consequently, provide measures of the goodness of this inferential decision-making procedure.

**Definition 8.1**

Rejecting the null hypothesis if it is true is a **Type I error**. The probability of making a Type I error is denoted by the symbol $\alpha$.

**Definition 8.2**

Accepting the null hypothesis if it is false is a **Type II error**. The probability of making a Type II error is denoted by the symbol $\beta$.

Which of the two errors, Type I or Type II, is more serious? From the EPA's perspective, the Type I error is the more serious error. If the EPA falsely accuses the company of violating the pollution limits, a costly lawsuit will likely occur. On the other hand, the residents who live near the chemical company would probably view the Type II error as more serious; if this error occurs, the EPA is failing to charge the company when it is, in fact, polluting the surrounding air. In either case, it is important to compute the probabilities, $\alpha$ and $\beta$, to assess the reliability of inferences derived from the hypothesis test. The next four examples illustrate how to compute these probabilities.

## Example 8.1

**Elements of a Statistical Test: Proportion of Software Purchasers**

A manufacturer of notebook computers believes that it can sell a particular software package to more than 20% of the buyers of its computers. Ten prospective purchasers of the notebook computer were randomly selected and questioned about their interest in the software package. Of these, four indicated that they planned to buy the package. Does this sample provide sufficient evidence to indicate that more than 20% of the computer purchasers will buy the software package?

**Solution**

Let $p$ be the true proportion of all prospective notebook computer buyers who will purchase the software package. Since we want to show that $p > .2$, we choose $H_a$: $p > .2$ for the alternative hypothesis and $H_0$: $p = .2$ for the null hypothesis. We will use the binomial random variable $Y$, the number of prospective purchasers in the sample who plan to buy the software, as the test statistic and will reject $H_0$: $p = .2$ if $Y$ is large. A graph of $p(y)$ for $n = 10$ and $p = .2$ is shown in Figure 8.1.

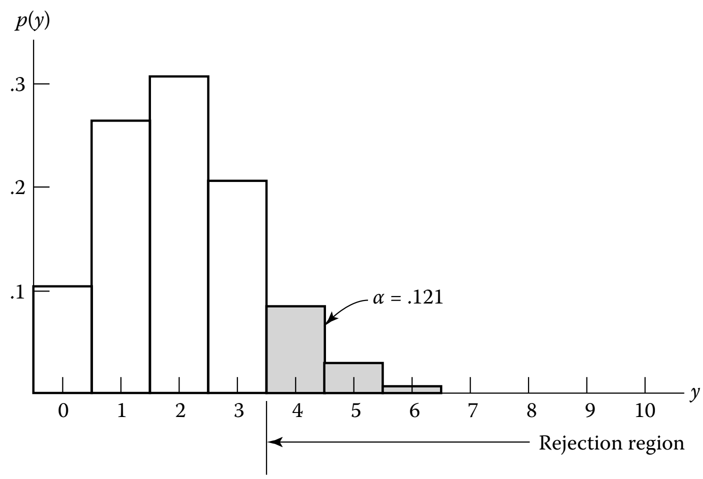



**FIGURE 8.1**
Graph of $p(y)$ for $n = 10$ and $p = .2$, i.e., if the null hypothesis is true

Large values of $Y$ will support the alternative hypothesis, $H_a$: $p > .2$, but what values of $Y$ should we include in the rejection region? Suppose that we select values of $Y \geq 4$ as the rejection region. Then the elements of the test are

$H_0$: $p = .2$
$H_a$: $p > .2$
*Test statistic*: $Y = y$
*Rejection region*: $y \geq 4$

To conduct the test, we note that the observed value of $Y$, $y = 4$, falls in the rejection region. Thus, for this test procedure, we reject the null hypothesis, $H_0$: $p = .2$, and conclude that the manufacturer is correct, i.e., $p > .2$.

**Example 8.2**

Computing the Type I Error Rate

What is the probability that the statistical test procedure of Example 8.1 would lead us to an incorrect decision if, in fact, the null hypothesis is true?

Solution

We will calculate the probability $\alpha$ that the test procedure would lead us to make a Type I error, i.e., to reject $H_0$ if, in fact, $H_0$ is true. This is the probability that $y$ falls in the rejection region if in fact $p = .2$:

$$\alpha = P(Y \geq 4 \mid p = .2) = 1 - \sum_{y=0}^{3} p(y)$$

The partial sum $\sum_{y=0}^{3} p(y)$ for a binomial random variable with $n = 10$ and $p = .2$ is given in Table 2 of Appendix B as .879. Therefore,

$$\alpha = 1 - \sum_{y=0}^{3} p(y) = 1 - .879 = .121$$

The probability that the test procedure would lead us to conclude that $p > .2$, if in fact it is not, is .121. This probability corresponds to the area of the shaded region in Figure 8.1.

In Example 8.1, we computed the probability $\alpha$ of committing a Type I error. *The probability $\beta$ of making a Type II error, i.e., failing to detect a value of p greater than .2, depends on the value of p.* For example, if $p = .20001$, it will be very difficult to detect this small deviation from the null hypothesized value of $p = .2$. In contrast, if $p = 1.0$, then *every* prospective purchaser of the minicomputer will want to buy the software package, and in such a case it will be very evident from the sample information that $p > .2$. We will illustrate the procedure for calculating $\beta$ in Example 8.3.

**Example 8.3**

Computing the Type II Error Rate

Refer to Example 8.2 and suppose that $p$ is actually equal to .60. What is the probability $\beta$ that the test procedure will fail to reject $H_0$: $p = .2$ if, in fact, $p = .6$?

Solution

The binomial probability distribution $p(y)$ for $n = 10$ and $p = .6$ is shown in Figure 8.2. The probability that we will fail to reject $H_0$ is equal to the probability that $Y = 0, 1, 2,$ or $3$, i.e., the probability that $Y$ does not fall in the rejection region. This probability, $\beta$, corresponds to the shaded area under the probability histogram in the figure. Therefore,

$$\beta = P(Y \leq 3 \mid p = .6) = \sum_{y=0}^{3} p(y) \qquad \text{for } n = 10 \text{ and } p = .6$$

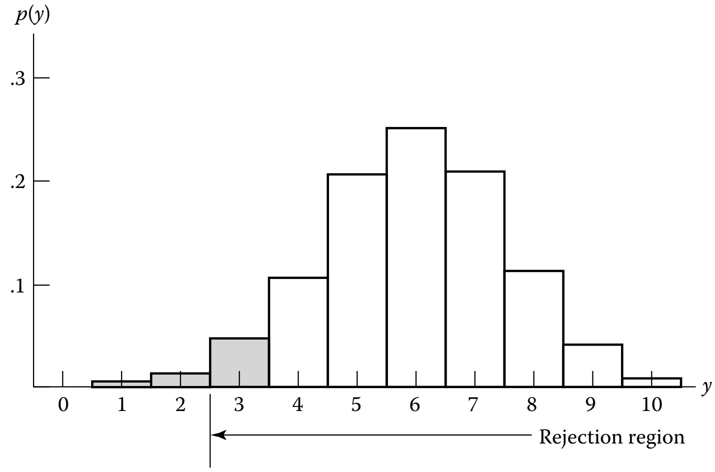



**FIGURE 8.2**
Graph of $p(y)$ for $n = 10$ and $p = .6$, i.e., if the alternative hypothesis is true

This partial sum, given in Table 2 of Appendix B for a binomial random variable with $n = 10$ and $p = .6$, is .055. Therefore, the probability that we will fail to reject $H_0$: $p = .2$ if $p$ is as large as .6 is $\beta = .055$.

Another important property of a statistical test is its ability to detect departures from the null hypothesis when they exist. This is measured by the probability of rejecting $H_0$ when, in fact, $H_0$ is false. Note that this probability is simply $(1 - \beta)$:

$$\begin{aligned} P(\text{Reject } H_0 \text{ when } H_0 \text{ is false}) &= 1 - P(\text{Accept } H_0 \text{ when } H_0 \text{ is false}) \\ &= 1 - P(\text{Type II error}) \\ &= 1 - \beta \end{aligned}$$

The probability $(1 - \beta)$ is called the **power of the test**. The higher the power, the greater the probability of detecting departures from $H_0$ when they exist.

**Definition 8.3**

The **power** of a statistical test, $(1 - \beta)$, is the probability of rejecting the null hypothesis $H_0$ when, in fact, $H_0$ is false.

**Example 8.4**

Computing the Power of a Test

Refer to the test of hypothesis in Example 8.1. Find the power of the test if in fact $p = .3$.

Solution

From Definition 8.3, the power of the test is the probability $(1 - \beta)$. The probability of making a Type II error, i.e., failing to reject $H_0$: $p = .2$, if in fact $p = .3$, will be larger than the value of $\beta$ calculated in Example 8.3 because $p = .3$ is much closer to the hypothesized value of $p = .2$. Thus,

$$\beta = P(Y \leq 3 \mid p = .3) = \sum_{y=0}^{3} p(y) \qquad \text{for } n = 10 \text{ and } p = .3$$

The value of this partial sum, given in Table 2 of Appendix B for a binomial random variable with $n = 10$ and $p = .3$, is .650. Therefore, the probability that we will fail to reject $H_0$: $p = .2$ if in fact $p = .3$ is $\beta = .650$ and the power of the test is $(1 - \beta) = (1 - .650) = .350$. You can see that the closer the actual value of $p$ is to the hypothesized null value, the more unlikely it is that we will reject $H_0$: $p = .2$.

The preceding examples indicate how we can calculate $\alpha$ and $\beta$ for a simple statistical test and thereby measure the risks of making Type I and Type II errors. These probabilities describe the properties of this inferential decision-making procedure and enable us to compare one test with another. For two tests, each with a rejection region selected so that $\alpha$ is equal to some specified value, say, .10, we would select the test that, for a specified alternative, has the smaller risk of making a Type II error, i.e., one that has the smaller value of $\beta$. This is equivalent to choosing the test with the higher power.

We will present a number of statistical tests in the following sections. In each case, the probability $\alpha$ of making a Type I error is known, i.e., $\alpha$ is selected by the experimenter and the rejection region is determined accordingly. In contrast, the value of $\beta$ for a specific alternative is often difficult to calculate. This explains why we attempt to show that $H_a$ is true by showing that the data do not support $H_0$. We hope that the sample evidence will support the alternative (or research) hypothesis. If it does, we will be concerned only about making a Type I error, i.e., rejecting $H_0$ if it is true. The probability $\alpha$ of committing such an error will be known.

## Applied Exercises

8.1 *Miscellaneous.* Define $\alpha$ and $\beta$ for a statistical test of hypothesis.

8.2 *Miscellaneous.* Explain why each of the following statements is incorrect:

a. The probability that the null hypothesis is correct is equal to $\alpha$.
b. If the null hypothesis is rejected, then the test proves that the alternative hypothesis is correct.
c. In all statistical tests of hypothesis, $\alpha + \beta = 1$.

8.3 *Screening new drugs.* Pharmaceutical companies are continually searching for new drugs. Testing the thousands of compounds for the few that might be effective is known in the pharmaceutical industry as *drug screening*. Dunnett (1978) views the drug-screening procedure in its preliminary stage in terms of a statistical decision problem: "In drug screening, two actions are possible: (1) to 'reject' the drug, meaning to conclude that the tested drug has little or no effect, in which case it will be set aside and a new drug selected for screening; and (2) to 'accept' the drug provisionally, in which case it will be subjected to further, more refined experimentation."* Since it is the goal of the researcher to find a drug that affects a cure, the null and alternative hypotheses in a statistical test would take the following form:

$H_0$: Drug is ineffective in treating a particular disease
$H_a$: Drug is effective in treating a particular disease

Dunnett comments on the possible errors associated with the drug-screening procedure: "To abandon a drug when in fact it is a useful one (a *false negative*) is clearly undesirable, yet there is always some risk in that. On the other hand, to go ahead with further, more expensive testing of a drug that is in fact useless (a *false positive*) wastes time and money that could have been spent on testing other compounds."

a. A false negative corresponds to which type of error, Type I or Type II?
b. A false positive corresponds to which type of error, Type I or Type II?
c. Which of the two errors is more serious? Explain.

8.4 *Pascal array variables.* Pascal is a high-level programming language used frequently in microprocessors. An experiment was conducted to investigate the proportion of Pascal variables that are *array* variables (in contrast to *scalar* variables, which are less efficient in terms of execution time). Twenty variables are randomly selected from a set of Pascal programs and $Y$, the number of array variables, is recorded. Suppose we want to test the hypothesis that Pascal is a more efficient language than Algol, in which 20% of the variables are array variables. That is, we will test $H_0$: $p = .20$ against $H_a$: $p > .20$, where $p$ is the probability of observing an array variable on each trial. (Assume that the 20 trials are independent.)

a. Find $\alpha$ for the rejection region $Y \geq 8$.
b. Find $\alpha$ for the rejection region $Y \geq 5$.
c. Find $\beta$ for the rejection region $Y \geq 8$ if $p = .5$. (*Note*: Past experience has shown that approximately half the variables in most Pascal programs are array variables.)
d. Find $\beta$ for the rejection region $Y \geq 5$. if $p = .5$.
e. Which of the rejection regions, $Y \geq 8$ or $Y \geq 5$, is more desirable if you want to minimize the probability of a Type I error? Type II error?
f. Find the rejection region of the form $Y \geq \alpha$ so that $\alpha$ is approximately equal to .01.
g. For the rejection region determined in part **f**, find the power of the test, if in fact $p = .4$.
h. For the rejection region determined in part **f**, find the power of the test, if in fact $p = .7$.

8.5 *Defective power meters.* A manufacturer of power meters, which are used to regulate energy thresholds of a data-communications system, claims that when its production process is operating correctly, only 10% of the power

*From Tanur, J. M., et al., eds. *Statistics: A Guide to the Unknown*. San Francisco: Holden-Day, 1978.

meters will be defective. A vendor has just received a shipment of 25 power meters from the manufacturer. Suppose the vendor wants to test $H_0$: $p = .10$ against $H_a$: $p > .10$, where $p$ is the true proportion of power meters that are defective. Use $Y \geq 6$ as the rejection region.

a. Determine the value of $\alpha$ for this test procedure.
b. Find $\beta$ if in fact $p = .2$. What is the power of the test for this value of $p$?
c. Find $\beta$ if in fact $p = .4$. What is the power of the test for this value of $p$?

8.6 *Authorizing computer users.* At high-technology industries, computer security is achieved by using a *password*—a collection of symbols (usually letters and numbers) that must be supplied by the user before the computer permits access to the account. The problem is that persistent hackers can create programs that enter millions of combinations of symbols into a target system until the correct password is found. The newest systems solve this problem by requiring authorized users to identify themselves by unique body characteristics. For example, a system developed by Palmguard, Inc. tests the hypothesis

$H_0$: The proposed user is authorized
$H_a$: The proposed user is unauthorized

by checking characteristics of the proposed user's palm against those stored in the authorized users' data bank.

a. Define a Type I error and Type II error for this test. Which is the more serious error? Why?
b. Palmguard reports that the Type I error rate for its system is less than 1%, whereas the Type II error rate is .00025%. Interpret these error rates.
c. Another successful security system, the EyeDentifyer, "spots authorized computer users by reading the one-of-a-kind patterns formed by the network of minute blood vessels across the retina at the back of the eye." The EyeDentifyer reports Type I and II error rates of .01% (1 in 10,000) and .005% (5 in 100,000), respectively. Interpret these rates.

### Theoretical Exercise

8.7 Show that for a fixed sample size $n$, $\alpha$ increases as $\beta$ decreases, and vice versa.

## 8.3 Finding Statistical Tests: Classical Methods

To find a statistical test about one or more population parameters, we must (1) find a suitable test statistic and (2) specify a rejection region. Classical statisticians use a method proposed by R. A. Fisher for finding a reasonable test statistic for testing a hypothesis. For example, suppose we want to test a hypothesis about the sole parameter $\theta$ of a probability function $p(y)$ or density function $f(y)$, and let $L$ represent the likelihood function of the sample. Then to test the null hypothesis, $H_0$: $\theta = \theta_0$, Fisher's **likelihood ratio test statistic** is

$$\lambda = \frac{\text{Likelihood assuming } \theta = \theta_0}{\text{Likelihood assuming } \theta = \hat{\theta}} = \frac{L(\theta_0)}{L(\hat{\theta})}$$

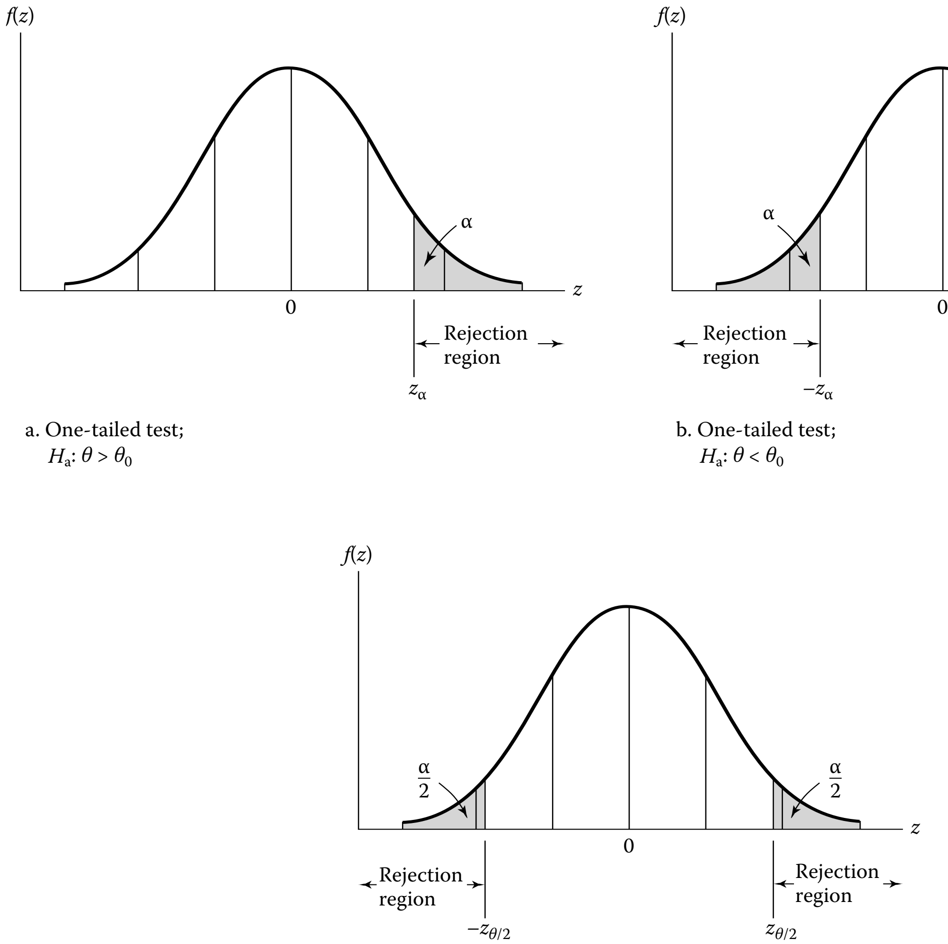
where $\hat{\theta}$ is the maximum likelihood estimator of $\theta$. Fisher reasoned that if $\theta$ differs from $\theta_0$, then the value of the likelihood $L$ when $\theta = \hat{\theta}$ will be larger than when $\theta = \theta_0$. Thus, the rejection region for the test contains values of $\lambda$ that are small—say, smaller than some value $\lambda_R$.

If you are interested in learning more about Fisher's likelihood ratio test, consult the references at the end of this chapter. Fortunately, most of the statistics that we would choose intuitively for test statistics are functions of the corresponding likelihood ratio statistic $\lambda$. These are the pivotal statistics used to construct confidence intervals in Chapter 7.

Recall that most of the pivotal statistics in Chapter 7 have approximately normal sampling distributions for large samples. This fact allows us to easily derive a large-sample statistical test of hypothesis. To illustrate, suppose that we want to test a hypothesis, $H_0$: $\theta = \theta_0$, about a parameter $\theta$ and that the estimator $\hat{\theta}$ possesses a normal sampling distribution with mean $\theta$ and standard deviation $\sigma_{\hat{\theta}}$. We will further assume that $\sigma_{\hat{\theta}}$ is known or that we can obtain a good approximation for it when the sample size(s) is (are) large. It can be shown (proof omitted) that the likelihood ratio test statistic $\lambda$ reduces to the standard normal variable $Z$:

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

The location of the rejection region for this test can be deduced by examining the formula for the test statistic $Z$. The farther $\hat{\theta}$ departs from $\theta_0$, i.e., the larger the absolute value of the deviation $|\hat{\theta} - \theta_0|$, the greater will be the weight of evidence to indicate that $\theta$ is not equal to $\theta_0$. If we want to detect values of $\theta$ larger than $\theta_0$, i.e., $H_a$: $\theta > \theta_0$, we locate the rejection region in the upper tail of the sampling distribution of the standard normal $z$ test statistic (see Figure 8.3a). If we want to detect only values of $\theta$ less than $\theta_0$, i.e., $H_a$: $\theta < \theta_0$, we locate the rejection region in the lower tail of the $z$ distribution (see Figure 8.3b). These two tests are called **one-tailed statistical tests** because the entire rejection region is located in only one tail of the $Z$ distribution. However, if we want to detect *either* a value of $\theta$ larger than $\theta_0$ or a value smaller than $\theta_0$, i.e., $H_a$: $\theta \neq \theta_0$, we locate the rejection region in both the upper and the lower tails of the $z$ distribution (see Figure 8.3c). This is called a **two-tailed statistical test**.




a. One-tailed test; $H_a$: $\theta > \theta_0$

b. One-tailed test; $H_a$: $\theta < \theta_0$

c. Two-tailed test; $H_\theta$: $\theta \neq \theta_0$

**FIGURE 8.3**
Rejection regions for one- and two-tailed tests

The large-sample statistical test that we have described is summarized in the following box. Many of the population parameters and test statistics discussed in the remaining sections of this chapter satisfy the assumptions of this test. We will illustrate the use of the test with a practical example on the population mean $\mu$.

### A Large-Sample Test Based on the Standard Normal $z$ Test Statistic

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\theta = \theta_0$ | $H_0$: $\theta = \theta_0$ |
| $H_a$: $\theta > \theta_0$ (or $H_a$: $\theta < \theta_0$) | $H_a$: $\theta \neq \theta_0$ |
| *Test statistic:* $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$ | *Test statistic:* $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$ |
| *Rejection region:* $Z > z_\alpha$ (or $Z < -z_\alpha$) where $P(Z > z_\alpha) = \alpha$ | *Rejection region:* $\|Z\| > z_{\alpha/2}$ where $P(Z > z_{\alpha/2}) = \alpha/2$ |

**Example 8.5**

**Testing $\mu$: Mean Number of Heavy Freight Trucks Traveling per Hour**

The Department of Highway Improvements, responsible for repairing a 25-mile stretch of interstate highway, wants to design a surface that will be structurally efficient. One important consideration is the volume of heavy freight traffic on the interstate. State weigh stations report that the average number of heavy-duty trailers traveling on a 25-mile segment of the interstate is 72 per hour. However, the section of highway to be repaired is located in an urban area and the department engineers believe that the volume of heavy freight traffic for this particular section is greater than the average reported for the entire interstate. To validate this theory, the department monitors the highway for 50 1-hour periods randomly selected throughout the month. Suppose the sample mean and standard deviation of the heavy freight traffic for the 50 sampled hours are

$$\bar{y} = 74.1 \qquad s = 13.3$$

Do the data support the department's theory? Use $\alpha = .10$.

**Solution**

For this example, the parameter of interest is $\mu$, the average number of heavy-duty trailers traveling on the 25-mile stretch of interstate highway. Recall that the sample mean $\bar{y}$ is used to estimate $\mu$ and that for large $n$, $\bar{y}$ has an approximately normal sampling distribution. Thus, we can apply the large-sample test outlined in the box.

The elements of the test are

$H_0$: $\mu = 72$

$H_a$: $\mu > 72$

*Test statistic*: $$Z = \frac{\bar{y} - 72}{\sigma_{\bar{y}}} = \frac{\bar{y} - 72}{\sigma/\sqrt{n}} \approx \frac{\bar{y} - 72}{s/\sqrt{n}}$$

*Rejection region*: $Z > 1.28$

(since $z_{.10} = 1.28$, from Table 5 of Appendix B)

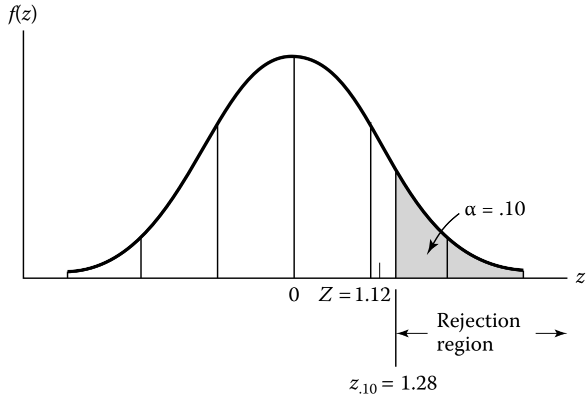
We now substitute the sample statistics into the test statistic to obtain

$$Z \approx \frac{74.1 - 72}{13.3/\sqrt{50}} = 1.12$$

Thus, although the average number of heavy freight trucks per hour in the sample exceeds the state's average by more than 2, the $Z$ value of 1.12 does not fall in the rejection region (see Figure 8.4). Therefore, this sample does not provide sufficient evidence at $\alpha = .10$ to support the Department of Highway Improvements theory.

What is the risk of making an incorrect decision in Example 8.5? If we reject the null hypothesis, then we know that the probability of making a Type I error (rejecting $H_0$ if it is true) is $\alpha = .10$. However, we failed to reject the null hypotheses in

**FIGURE 8.4**
Location of the test statistic for Example 8.5




Example 8.5 and, consequently, we must be concerned about the possibility of making a Type II error (accepting $H_0$ if, in fact, it is false). We will evaluate the risk of making a Type II error in Example 8.6.

## Example 8.6

Calculating $\beta$ for the Traveling Trucks Test

Refer to the one-tailed test for $\mu$, Example 8.5. If the mean number $\mu$ of heavy freight trucks traveling a particular 25-mile stretch of interstate highway is in fact 78 per hour, what is the probability that the test procedure of Example 8.5 would fail to detect it? That is, what is the probability $\beta$ that we would fail to reject $H_0$: $\mu = 72$ in this one-tailed test if $\mu$ is actually equal to 78?

**Solution**

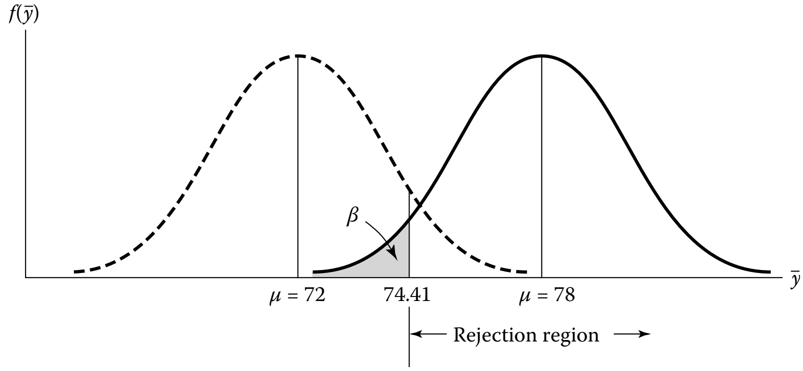
To calculate $\beta$ for the large-sample $Z$ test, we need to specify the rejection region in terms of the point estimator $\hat{\theta}$, where, for this example, $\hat{\theta} = \bar{y}$. From Figure 8.4, you can see that the rejection region consists of values of $Z \geq 1.28$. To determine the value of $\bar{y}$ corresponding to $z = 1.28$, we substitute into the equation

$$Z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \approx \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad \text{or} \quad 1.28 = \frac{\bar{y} - 72}{13.3/\sqrt{50}}$$

Solving for $\bar{y}$, we obtain $\bar{y} = 74.41$. Therefore, the rejection region for the test is $Z \geq 1.28$ or, equivalently, $\bar{y} \geq 74.41$.

The dotted curve in Figure 8.5 is the sampling distribution for $\bar{y}$ if $H_0$: $\mu = 72$ is true. This curve was used to locate the rejection region for $\bar{y}$ (and, equivalently, $z$), i.e., values of $\bar{y}$ contradictory to $H_0$: $\mu = 72$. The solid curve is the sampling distribution for $\bar{y}$ if $\mu = 78$. Since we want to find $\beta$ if $H_0$ is in fact false and $\mu = 78$, we want to find the probability that $\bar{y}$ does not fall in the rejection region if $\mu = 78$. This

**FIGURE 8.5**
The probability $\beta$ of making a Type II error if $\mu = 78$ in Example 8.6



probability corresponds to the shaded area under the solid curve for values of $\bar{y} < 74.41$. To find this area under the normal curve, we need to find the area A corresponding to

$$Z = \frac{\bar{y} - 78}{\sigma/\sqrt{n}} \approx \frac{74.41 - 78}{13.3/\sqrt{50}} = -1.91$$

The value of A, given in Table 5 of Appendix B, is .4719. Then from Figure 8.5, it can be seen that

$$\beta = .5 - A = .5 - .4719 = .0281$$

Therefore, the probability of failing to reject $H_0$: $\mu = 72$ if $\mu$ is, in fact, as large as $\mu = 78$ is only .0281.

Example 8.6 illustrates that it is not too difficult to calculate $\beta$ for various alternatives for the large-sample $Z$ test (see box). However, it may be extremely difficult to calculate $\beta$ for other tests. Although sophisticated techniques are available for evaluating the risk of making a Type II error when the exact value of $\beta$ is unavailable or is difficult to calculate, they are beyond the scope of this text. Consult the references at the end of this chapter if you are interested in learning about these methods.

### Calculating $\beta$ for a Large-Sample $Z$ Test

Consider a large-sample test of $H_0$: $\theta = \theta_0$ at significance level $\alpha$. The value of $\beta$ for a specific value of the alternative $\theta = \theta_a$ is calculated as follows:

**Upper-tailed test:** $\beta = P\left(Z < \dfrac{\hat{\theta}_0 - \theta_a}{\sigma_{\hat{\theta}}}\right)$

where $\hat{\theta}_0 = \theta_0 + z_\alpha \sigma_{\hat{\theta}}$ is the value of the estimator corresponding to the border of the rejection region

**Lower-tailed test:** $\beta = P\left(Z > \dfrac{\hat{\theta}_0 - \theta_a}{\sigma_{\hat{\theta}}}\right)$

where $\hat{\theta}_0 = \theta_0 - z_\alpha \sigma_{\hat{\theta}}$ is the value of the estimator corresponding to the border of the rejection region

**Two-tailed test:** $\beta = P\left(\dfrac{\hat{\theta}_{0,\mathrm{L}} - \theta_a}{\sigma_{\hat{\theta}}} < Z < \dfrac{\hat{\theta}_{0,\mathrm{U}} - \theta_a}{\sigma_{\hat{\theta}}}\right)$

where $\hat{\theta}_{0,\mathrm{U}} = \theta_0 + z_\alpha \sigma_{\hat{\theta}}$ and $\hat{\theta}_{0,\mathrm{L}} = \theta_0 - z_\alpha \sigma_{\hat{\theta}}$ are the values of the estimator corresponding to the borders of the rejection region

## Theoretical Exercises

8.8 Suppose $y_1, y_2, \ldots, y_n$ is a random sample from a normal distribution with unknown mean $\mu$ and variance $\sigma^2 = 1$, i.e.,

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2}$$

Show that the likelihood $L$ of the sample is

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\Sigma_{i=1}^{n} (y_i-\mu)^2/2}$$

8.9 Refer to Exercise 8.8. Suppose we want to test $H_0$: $\mu = 0$ against the alternative $H_a$: $\mu > 0$. Since the estimator of $\mu$ is $\hat{\mu} = \bar{y}$, the likelihood ratio test statistic is

$$\lambda = \frac{L(\mu_0)}{L(\hat{\mu})} = \frac{L(0)}{L(\bar{y})}$$

Show that

$$\lambda = e^{-n(\bar{y})^2/2}$$

[*Hint:* Use the fact that $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$.]

8.10 Refer to Exercise 8.8 and 8.9. Show that the rejection region $\lambda \le \lambda_\alpha$ is equivalent to the rejection region $\bar{y} \ge \bar{y}_\alpha$, where $P(\lambda \le \lambda_\alpha) = \alpha$ and $P(\bar{y} \ge \bar{y}_\alpha) = \alpha$. (*Hint:* Use the fact that $e^{-a^2} \to 0$ as $|a| \to \infty$ .)

## 8.4 Choosing the Null and Alternative Hypotheses

Now that you have conducted a large-sample statistical test of hypothesis and have seen how to calculate the value of $\beta$—the probability of failing to reject $H_0$: $\theta = \theta_0$ if $\theta$ is in fact equal to some alternative value, $\theta = \theta_a$—the logic for choosing the null and alternative hypotheses may make more sense to you. The theory that we want to support (or detect if true) is usually chosen as the alternative hypothesis because, if the data support $H_a$ (i.e., if we reject $H_0$), we immediately know the value of $\alpha$, the probability of incorrectly rejecting $H_0$ if it is true. For example, in Example 8.5, the Department of Highway Improvements theorized that the mean number of heavy-duty vehicles traveling a certain segment of interstate exceeds 72 per hour. Consequently, the department set up the alternative hypothesis as $H_a$: $\mu > 72$. In contrast, if we choose the null hypothesis as the theory that we want to support, and if the data support this theory, i.e., the test leads to nonrejection of $H_0$, then we would have to investigate the values of $\beta$ for some specific alternatives. Clearly, we want to avoid this tedious and sometimes extremely difficult task, if possible.

Another issue that arises in a practical situation is whether to conduct a one- or a two-tailed test. The decision depends on what you want to detect. For example, suppose you operate a chemical plant that produces a variable amount $Y$ of product per day and that if $E(Y) = \mu$ is less than 100 tons per day, you will eventually be bankrupt. If $\mu$ exceeds 100 tons per day, you are financially safe. To determine whether your process is leading to financial disaster, you will want to detect whether $\mu < 100$ tons, and you will conduct a one-tailed test of $H_0$: $\mu = 100$ versus $H_a$: $\mu < 100$. If you were to conduct a two-tailed test for this situation, you would reduce your chance of detecting values of $\mu$ less than 100 tons, i.e., you would increase the values of $\beta$ for alternative values of $\mu < 100$ tons.

As a different example, suppose you have designed a new drug so that its mean potency is some specific level, say, 10%. As the mean potency tends to exceed 10%, you lose money. If it is less than 10% by some specified amount, the drug becomes ineffective as a pharmaceutical (and you lose money). To conduct a test of the mean potency $\mu$ for this situation, you would want to detect values of $\mu$ either larger than or smaller than $\mu = 10$. Consequently, you would select $H_a$: $\mu \neq 10$ and conduct a two-tailed statistical test (or alternatively, construct a confidence interval).

These examples demonstrate that a statistical test is an attempt to detect departures from $H_0$; the key to the test is to define the specific *alternatives* that you want to detect. We must stress, however, that $H_0$ and $H_a$ *should be constructed prior to obtaining and observing the sample data*. If you use information in the sample data to aid in selecting $H_0$ and $H_a$, the prior information gained from the sample biases the test results—specifically, the true probability of a Type I error will be larger than the preselected value of $\alpha$.

**Example 8.7**

**Choosing $H_0$ and $H_a$ for Testing the Mean Diameter of Bearings**

A metal lathe is checked periodically by quality control inspectors to determine whether it is producing machine bearings with a mean diameter of .5 inch. If the mean diameter of the bearings is larger or smaller than .5 inch, then the process is out of control and needs to be adjusted. Formulate the null and alternative hypotheses that could be used to test whether the bearing production process is out of control.

**Solution** The hypotheses must be stated in terms of a population parameter. Thus, we define

$$\mu = \text{true mean diameter (in inches) of all bearings produced by the lathe}$$

If either $\mu > .5$ or $\mu < .5$, then the metal lathe's production process is out of control. Since we wish to be able to detect either possibility, the null and alternative hypotheses would be

$H_0$: $\mu = .5$ (i.e., the process is in control)
$H_a$: $\mu \neq .5$ (i.e., the process is out of control)

In Sections 8.5–8.12, we will present applications of the hypothesis-testing logic developed in this chapter. The cases to be considered are those for which we developed estimation procedures in Chapter 7. Since the theory and reasoning involved are based on the developments of Chapter 7 and Sections 8.1–8.4, we will present only a summary of the hypothesis-testing procedure for one-tailed and two-tailed tests in each situation.

## Applied Exercises

In Exercises 8.11–8.16, formulate the appropriate null and alternative hypotheses.

8.11 *Strength of natural fiber composites.* An article in *ACS Sustainable Chemistry & Engineering* (Vol. 1, 2013) investigated the use of natural fiber composites produced from switchgrass. Researchers want to know if the mean tensile strength of this fiber composite exceeds 20 megapascals.

8.12 *Egg-hatching rate of frogs.* A herpetologist wants to determine whether the egg-hatching rate for a certain species of frog exceeds .5 when the eggs are exposed to ultraviolet radiation.

8.13 *Testing fishing line.* A manufacturer of fishing line wants to show that the mean breaking strength of a competitor's 22-pound line is really less than 22 pounds.

8.14 *Loaded casino dice.* A craps player who has experienced a long run of bad luck at the craps table wants to test whether the casino dice are "loaded," i.e., whether the proportion of "sevens" occurring in many tosses of the two dice is different from $\frac{1}{6}$ (if the dice are fair, the probability of tossing a "seven" is $\frac{1}{6}$).

8.15 *Software vendor ratings.* Each year, *Computerworld* magazine reports the Datapro ratings of all computer software vendors. Vendors are rated on a scale from 1 to 4 ($1 =$ poor, $4 =$ excellent) in such areas as reliability, efficiency, ease of installation, and ease of use by a random sample of software users. A software vendor wants to determine whether its product has a higher mean Datapro rating than a rival vendor's product.

8.16 *Radium in soil.* The Environmental Protection Agency wishes to test whether the mean amount of radium-226 in soil in a Florida county exceeds the maxium allowable amount, 4 pCi/L.

8.17 *Real-time scheduling.* Industrial engineers want to compare two methods of real-time scheduling in a manufacturing operation. Specifically, they want to determine whether the mean number of items produced differs for the two methods.

## 8.5 The Observed Significance Level for a Test

According to the statistical test procedures described in the preceding sections, the rejection region and the corresponding value of $\alpha$ are selected prior to conducting the test and the conclusion is stated in terms of rejecting or not rejecting the null hypothesis. A second method of presenting the result of a statistical test is one that reports the extent to which the test statistic disagrees with the null hypothesis and leaves the reader the task of deciding whether to reject the null hypothesis. This measure of disagreement is called the **observed significance level** (or ***p*-value**) for the test.*

**Definition 8.4**

The **observed significance level**, or ***p*-value**, for a specific statistical test is the probability (assuming $H_0$ is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the one computed from the sample data.

When publishing the results of a statistical test of hypothesis in journals, case studies, reports, etc., many researchers make use of $p$-values. Instead of selecting $\alpha$ a priori and then conducting a test as outlined in this chapter, the researcher may compute and report the value of the appropriate test statistic and its associated $p$-value. It is left to the reader of the report to judge the significance of the result, i.e., the reader must determine whether to reject the null hypothesis in favor of the alternative hypothesis, based on the reported $p$-value. Usually, *the null hypothesis will be rejected only if the observed significance level is* ***less*** *than the fixed significance level* $\alpha$ *chosen by the reader*. There are two inherent advantages of reporting test results in this manner: (1) Readers are permitted to select the maximum value of $\alpha$ that they would be willing to tolerate if they actually carried out a standard test of hypothesis in the manner outlined in this chapter, and (2) it is an easy way to present the results of test calculations performed by a computer. Most statistical software packages perform the calculations for a test, give the observed value of the test statistic, and leave it to the reader to formulate a conclusion. Others give the observed significance level for the test, a procedure that makes it easy for the user to decide whether to reject the null hypothesis.

### Interpreting *p*-Values

1. Choose the maximum value of $\alpha$ you are willing to tolerate.
2. Find the observed significance level ($p$-value) of the test.
3. Regret the null hypothesis if $\alpha > p$-value.

**Example 8.8**

Finding a One-Tailed *p*-value

Find the observed significance level for the statistical test of Example 8.5 and interpret the result.

Solution

In Example 8.5, we tested a hypothesis about the mean $\mu$ of the number of heavy freight trucks per hour using a particular 25-mile stretch of interstate highway. Since we wanted to detect values of $\mu$ larger than $\mu_0 = 72$, we conducted a one-tailed test, rejecting $H_0$ for large values of $\bar{y}$, or equivalently, large values of $Z$. The observed value of $Z$, computed from the sample of $n = 50$ randomly selected 1-hour periods, was $Z = 1.12$. Since any value of $Z$ larger than $Z = 1.12$ would be even more contradictory to $H_0$, the observed significance level for the test is

$$p\text{-value} = P(Z \geq 1.12)$$

This value corresponds to the shaded area in the upper tail of the $z$ distribution shown in Figure 8.6. The area $A$ corresponding to $z = 1.12$, given in Table 5 of Appendix B, is .3686. Therefore, the observed significance level is

$$p\text{-value} = P(Z \geq 1.12) = .5 - A = .5 - .3686 = .1314$$

*The term *p-value* or *probability value* was coined by users of statistical methods. The $p$ in the expression *p-value* should not be confused with the binomial parameter $p$.

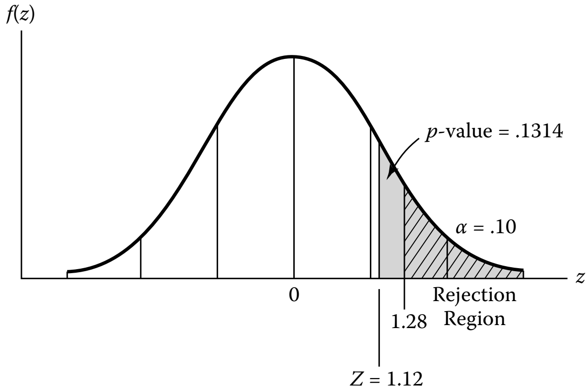



**FIGURE 8.6**
Finding the $p$-value for an upper-tailed test when $z = 1.12$

This result indicates that the probability of observing a $z$ value at least as contradictory to $H_0$ as the one observed in this (if $H_0$ is in fact true) is .1314. Therefore, we will reject $H_0$ only for preselected values of $\alpha$ greater than .1314. Recall that the Department of Highway Improvements selected a Type I error probability of $\alpha = .10$. Since $p$-value $= .1314$ exceeds $\alpha = .10$, the department has insufficient evidence to reject $H_0$. Note that this conclusion agrees with that of Example 8.5, as shown in Figure 8.6.

## Example 8.9

**Finding a Two-Tailed $p$-value**

Suppose that the test of Example 8.5 had been a two-tailed test, i.e., suppose that the alternative of interest had been $H_a$: $\mu \neq 72$. Find the observed significance level for the test and interpret the result. Assume that $\alpha = .10$, as in Example 8.5.

**Solution** If the test were two-tailed, either very large or very small values of $Z$ would be contradictory to the null hypothesis $H_0$: $\mu = 72$. Consequently, values of $Z \geq 1.12$ or $Z \leq -1.12$ would be more contradictory to $H_0$ than the observed value of $Z = 1.12$. Therefore, the observed significance level for the test (shaded in Figure 8.7) is

$$\begin{aligned} p\text{-value} &= P(Z \geq 1.12) + P(Z \leq -1.12) \\ &= 2(.1314) = .2628 \end{aligned}$$

Since we want to conduct the two-tailed test at $\alpha = .10$, the rejection region is $|Z| > 1.96$, as shown in Figure 8.7. Note that the $p$-value exceeds $\alpha$; we again have insufficient evidence to reject $H_0$.

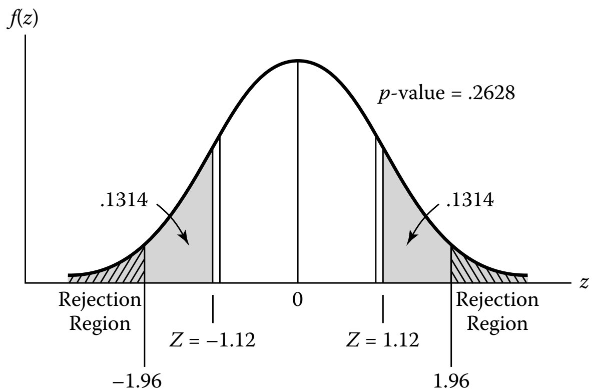



**FIGURE 8.7**
Finding the $p$-value for a two-tailed test when $z = 1.12$

Observed significance levels are more easily obtained using statistical software. The exact $p$-value for the one-tailed test of Example 8.8 is shown (shaded) on the MINITAB printout, Figure 8.8. Typically, a researcher will utilize statistical software, rather than probability tables or formulas, to find $p$-values.

**FIGURE 8.8**
MINITAB Output for One-Tailed Test of a Population Mean

```
One-Sample Z

Test of mu = 72 vs > 72
The assumed standard deviation = 13.3


                     95% Lower
 N   Mean  SE Mean       Bound     Z      P
50  74.10     1.88       71.01  1.12  0.132
```

*Note:* Some statistical software packages (e.g., SPSS) will conduct only two-tailed tests of hypothesis. For these packages, you obtain the $p$-value for a one-tailed test as shown in the next box.

### Converting a Two-Tailed $p$-Value from a Printout to a One-Tailed $p$-Value

$$p = \frac{\text{Reported } p\text{-value}}{2} \qquad \text{if } \begin{cases} H_a \text{ is of form } > \text{ and } z \text{ is positive} \\ H_a \text{ is of form } < \text{ and } z \text{ is negative} \end{cases}$$

$$p = 1 - \left(\frac{\text{Reported } p\text{-value}}{2}\right) \qquad \text{if } \begin{cases} H_a \text{ is of form } > \text{ and } z \text{ is negative} \\ H_a \text{ is of form } < \text{ and } z \text{ is positive} \end{cases}$$

## Applied Exercises

8.18 *One-tailed p-value.* For a large-sample test of $H_0$: $\theta = \theta_0$ versus $H_a$: $\theta > \theta_0$, compute the $p$-value associated with each of the following test statistic values:

a. $z = 1.96$
b. $z = 1.645$
c. $z = 2.67$
d. $z = 1.25$

8.19 *Two-tailed p-value.* For a large-sample test of $H_0$: $\theta = \theta_0$ versus $H_a$: $\theta \neq \theta_0$, compute the $p$-value associated with each of the following test statistic values:

a. $z = -1.01$
b. $z = -2.37$
c. $z = 4.66$
d. $z = 1.45$

8.20 *Comparing "α" to p-value.* For each $\alpha$ and observed significance level ($p$-value) pair, indicate whether the null hypothesis would be rejected.

a. $\alpha = .05$, $p$-value $= .10$
b. $\alpha = .10$, $p$-value $= .05$
c. $\alpha = .01$, $p$-value $= .001$
d. $\alpha = .025$, $p$-value $= .05$
e. $\alpha = .10$, $p$-value $= .45$

8.21 *Converting a two-tailed p-value.* In a test of $H_0$: $\mu = 75$ performed using the computer, SPSS reports a two-tailed $p$-value of .1032. Make the appropriate conclusion for each of the following situations:

a. $H_a$: $\mu < 75$, $z - -1.63$, $\alpha = .05$
b. $H_a$: $\mu < 75$, $z = 1.63$, $\alpha = .10$
c. $H_a$: $\mu < 75$, $z = -1.63$, $\alpha = .10$
d. $H_a$: $\mu < 75$, $z = -1.63$, $\alpha = .01$

8.22 *p-value interpretation.* An analyst tested the null hypothesis $\mu \geq 20$ against the alternative hypothesis that $\mu < 20$. The analyst reported a $p$-value of .06. What is the smallest value of $\alpha$ for which the null hypothesis would be rejected?

## 8.6 Testing a Population Mean

In Example 8.5, we developed a large-sample test for a population mean based on the standard normal $z$ statistic. The elements of this test are summarized in the box.

**Large-Sample ($n \geq 30$) Test of Hypothesis About a Population Mean $\mu$**

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\mu = \mu_0$ | $H_0$: $\mu = \mu_0$ |
| $H_a$: $\mu > \mu_0$ (or $H_a$: $\mu < \mu_0$) | $H_a$: $\mu \neq \mu_0$ |
| *Test statistic:* $Z = \dfrac{\bar{y} - \mu_0}{\sigma_{\bar{y}}} \approx \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ | *Test statistic:* $Z = \dfrac{\bar{y} - \mu_0}{\sigma_{\bar{y}}} \approx \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ |
| *Rejection region:* $Z > z_\alpha$ (or $Z < -z_\alpha$) | *Rejection region:* $\lvert Z \rvert > z_{\alpha/2}$ |
| $p$-value $= P(Z > z_c)$ [or, $P(Z < z_c)$] | $p$-value $= 2P(Z > \lvert z_c \rvert)$ |

where $P(Z > z_\alpha) = \alpha$, $P(Z > z_{\alpha/2}) = \alpha/2$, $\mu_0$ is our symbol for the particular numerical value specified for $\mu$ in the null hypothesis, and $z_c$ is the computed value of the test statistic.

*Assumptions:* None (since the central limit theorem guarantees that $\bar{y}$ is approximately normal regardless of the distribution of the sampled population)

### Example 8.10

**Large-Sample Test of $\mu$: Mean Length-to-Width Ratio of Bones**

Humerus bones from the same species of animal tend to have approximately the same length-to-width ratios. When fossils of humerus bones are discovered, archeologists can often determine the species of animal by examining the length-to-width ratios of the bones. It is known that species A has a mean ratio of 8.5. Suppose 41 fossils of humerus bones were unearthed at an archeological site in East Africa, where species A is believed to have inhabited. (Assume that the unearthed bones are all from the same unknown species.) The length-to-width ratios of the bones were measured and are listed in Table 8.2.

We wish to test the hypothesis that $\mu$, the population mean ratio of all bones of this particular species, is equal to 8.5 against the alternative that it is different from 8.5, i.e., we wish to test whether the unearthed bones are from species A.

a. Suppose we want a very small chance of rejecting $H_0$, if, in fact, $\mu$ is equal to 8.5. That is, it is important that we avoid making a Type I error. Select an appropriate value of the significance level, $\alpha$.

b. Test whether $\mu$, the population mean length-to-width ratio, is different from 8.5, using the significance level selected in part **a.**

BONES

**TABLE 8.2 Length-to-Width Ratios of a Sample of Humerus Bones**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 10.73 | 8.89 | 9.07 | 9.20 | 10.33 | 9.98 | 9.84 | 9.59 |
| 8.48 | 8.71 | 9.57 | 9.29 | 9.94 | 8.07 | 8.37 | 6.85 |
| 8.52 | 8.87 | 6.23 | 9.41 | 6.66 | 9.35 | 8.86 | 9.93 |
| 8.91 | 11.77 | 10.48 | 10.39 | 9.39 | 9.17 | 9.89 | 8.17 |
| 8.93 | 8.80 | 10.02 | 8.38 | 11.67 | 8.30 | 9.17 | 12.00 |
| 9.38 | | | | | | | |

Solution

a. The hypothesis-testing procedure that we have developed gives us the advantage of being able to choose any significance level that we desire. Since the significance level, $\alpha$, is also the probability of a Type I error, we will choose $\alpha$ to be very small. In general, researchers who consider a Type I error to have very serious practical consequences should perform the test at a very low $\alpha$ value—say, $\alpha = .01$. Other researchers may be willing to tolerate an $\alpha$ value as high as .10 if a Type I error is not deemed a serious error to make in practice. For this example, we will test at $\alpha = .01$.

b. We formulate the following hypotheses:

$$H_0\text{: } \mu = 8.5$$
$$H_a\text{: } \mu \neq 8.5$$

Note that this is a two-tailed test, since we want to detect departures from $\mu = 8.5$ in either direction. The sample size is large ($n = 41$); thus, we may proceed with the large-sample test about $\mu$.

At significance level $\alpha = .01$, we will reject the null hypothesis for this two-tailed test if

$$|Z| > z_{\alpha/2} = z_{.005}$$

i.e., if $Z < -2.58$ or if $Z > 2.58$. This rejection region is shown in Figure 8.9.

After entering the data of Table 8.2 into a computer, we obtained the summary statistics shown in the SAS printout, Figure 8.7. The values $\bar{y} = 9.257$ and $s = 1.203$ (shaded in the printout) are used to compute the test statistic

$$Z \approx \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{9.257 - 8.5}{1.203/\sqrt{41}} = 4.03$$

This test statistic value is also shaded on Figure 8.10, as well as the $p$-value of the test, $p$-value $= .002$.

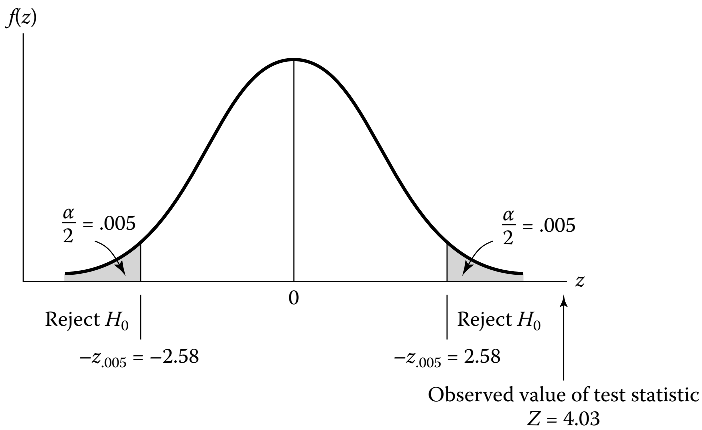



**FIGURE 8.9**
Rejection region for Example 8.10

Note that the test statistic lies within the rejection region (see Figure 8.9), and, $\alpha = .01$ exceeds the $p$-value. Consequently, we reject $H_0$ and conclude that the mean length-to-width ratio of all humerus bones of this particular species is significantly different from 8.5. If the null hypothesis is in fact true (i.e., if $\mu = 8.5$), then the probability that we have incorrectly rejected it is equal to $\alpha = .01$.

```
Sample Statistics for LWRATIO

     N          Mean        Std. Dev.       Std. Error
   ------------------------------------------------------
    41          9.26          1.20             0.19

Hypothesis Test

   Null hypothesis:   Mean of LWRATIO =  8.5
   Alternative:       Mean of LWRATIO ^= 8.5

          t Statistic       Df      Prob > t
          ----------------------------------
             4.030          40       0.0002
```

**FIGURE 8.10**
SAS printout for Example 8.10

The *practical* implications of the result obtained in Example 8.10 remain to be studied further. Perhaps the animal discovered at the archeological site is of some species other than A. Alternatively, the unearthed humerus bones may have larger than normal length-to-width ratios because of unusual feeding habits of species A. **It is not always the case that a statistically significant result implies a practically significant result.** The researcher must retain objectivity and judge the practical significance using, among other criteria, knowledge of the subject matter and the phenomenon under investigation.

A small-sample statistical test for making inferences about a population mean is (like its associated confidence interval of Section 7.4) based on the assumption that the sample data are independent observations on a normally distributed random variable. The test statistic is based on the $T$ distribution given in Section 7.4.

The elements of the statistical test are listed in the accompanying box. As we suggested in Chapter 7, the small-sample test will possess the properties specified in the box even if the sampled population is moderately nonnormal. However, for data that departs greatly from normality (i.e., highly skewed data), we must resort to one of the nonparametric techniques discussed in Chapter 15.

### Small-Sample Test of Hypothesis About a Population Mean $\mu$

| One-Tailed Test | Two-Tailed Test |
|---|---|
| $H_0$: $\mu = \mu_0$ | $H_0$: $\mu = \mu_0$ |
| $H_a$: $\mu > \mu_0$ (or $H_a$: $\mu < \mu_0$) | $H_a$: $\mu \neq \mu_0$ |

$$\textit{Test statistic: } T = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

| | |
|---|---|
| *Rejection region:* $T > t_\alpha$ (or $T < -t_\alpha$) | *Rejection region:* $\lvert T\rvert > t_{\alpha/2}$ |
| $p$-value $= P(T \geq t_c)$ [or, $P(T \leq t_c)$] | $p$-value $= 2\mathrm{P}(T \geq \lvert t_c\rvert)$ |

where the distribution of $t$ is based on $(n - 1)$ degrees of freedom; $P(T > t_\alpha) = \alpha$; $P(T > t_{\alpha/2}) = \alpha/2$, and $t_c$ is the computed value of the test-statistic.

*Assumption:* The relative frequency distribution of the population from which the sample was selected is approximately normal.

*Warning:* If the data depart greatly from normality, this small-sample test may lead to erroneous inferences. In this case, use the nonparametric sign test that is discussed in Section 15.2.

## Example 8.11

Small-Sample Test of $\mu$: Mean Benzene Content

Scientists have labeled benzene, a chemical solvent commonly used to synthesize plastics, as a possible cancer-causing agent. Studies have shown that people who work with benzene more than 5 years have 20 times the incidence of leukemia than the general population. As a result, the federal government has lowered the maximum allowable level of benzene in the workplace from 10 parts per million (ppm) to 1 ppm. Suppose a steel manufacturing plant, which exposes its workers to benzene daily, is under investigation by the Occupational Safety and Health Administration (OSHA). Twenty air samples, collected over a period of 1 month and examined for benzene content, yielded the data in Table 8.3. Is the steel manufacturing plant in violation of the new government standards? Test the hypothesis that the mean level of benzene at the steel manufacturing plant is greater than 1 ppm, using $\alpha = .05$.

**BENZENE**

**TABLE 8.3 Benzene Content for 20 Air Samples**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.21 | 1.44 | 2.54 | 2.97 | 0.00 | 3.91 | 2.24 | 2.41 | 4.50 | 0.15 |
| 0.30 | 0.36 | 4.50 | 5.03 | 0.00 | 2.89 | 4.71 | 0.85 | 2.60 | 1.26 |

Solution

The OSHA wants to establish the research hypothesis that the mean level of benzene, $\mu$, at the steel manufacturing plant exceeds 1 ppm. The elements of this small-sample one-tailed test are

$H_0$: $\mu = 1$

$H_a$: $\mu > 1$

*Test statistic:* $T = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$

*Assumption*: The relative frequency distribution of the population of benzene levels for all air samples at the steel manufacturing plant is approximately normal.

*Rejection region*: For $\alpha = .05$ and df $= (n - 1) = 19$, reject $H_0$ if $T > t_{.05} = 1.729$ (see Figure 8.11)

Summary statistics for the sample data are shown on the MINITAB printout, Figure 8.12. The values of $\bar{y}$ and $s$ (highlighted) are $\bar{y} = 2.143$ and $s = 1.736$.

We now calculate the test statistic:

$$T = \frac{\bar{y} - 1}{s/\sqrt{n}} = \frac{2.143 - 1}{1.736/\sqrt{20}} = 2.95$$

The value of $T$ is also shown (highlighted) on Figure 8.12 as well as the $p$-value of the test, .004.

Note that the test statistic value falls into the rejection region (see Figure 8.11), and $\alpha = .05$ exceeds the $p$-value of the test. Therefore the OSHA concludes that $\mu > 1$ part per million and the plant is in violation of the new government standards.

**FIGURE 8.11**
Rejection region for Example 8.11

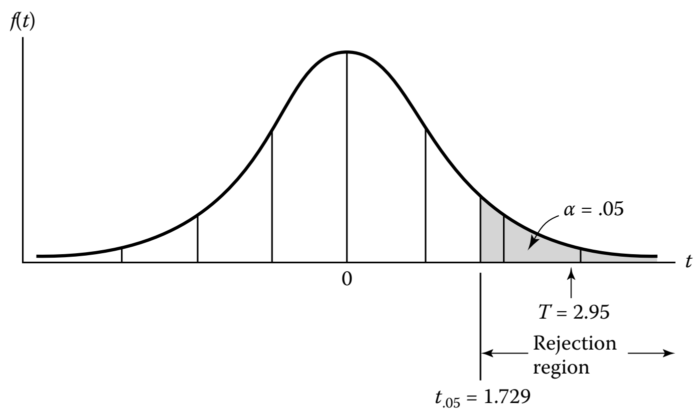


**FIGURE 8.12**
MINITAB printout for Example 8.11

**One-Sample T: BENZENE**

Test of mu = 1 vs > 1

| Variable | N | Mean | StDev | SE Mean | 95% Lower Bound | T | P |
|---|---|---|---|---|---|---|---|
| BENZENE | 20 | 2.14350 | 1.73602 | 0.38819 | 1.47228 | 2.95 | 0.004 |

The reliability associated with this inference is $\alpha = .05$. This implies that if the testing procedure was applied repeatedly to random samples of data collected at the plant, the OSHA would falsely reject $H_0$ for only 5% of the tests. Consequently, the OSHA is highly confident (95% confident) that the plant is violating the new standards.

## Applied Exercises

**FUP**

8.23 *Stability of compounds in new drugs.* Refer to the *ACS Medicinal Chemistry Letters* (Vol. 1, 2010) study of the metabolic stability of drugs, Exercise 2.16 (p. 36). Recall that two important values computed from the testing phase are the fraction of compound unbound to plasma (*fup*) and the fraction of compound unbound to microsomes (*fumic*). A key formula for assessing stability assumes that the fup/fumic ratio is 1. Pharmacologists at Pfizer Global Research and Development tested 416 drugs and reported the fup/fumic ratio for each. These data are saved in the **FUP** file and summary statistics are provided in the MINITAB printout shown below. Suppose the pharmacologists want to determine if the true mean ratio, $\mu$, differs from 1.

a. Specify the null and alternative hypothesis for this test.
b. Descriptive statistics for the sample ratios are provided in the accompanying MINITAB printout. Note that the sample mean, $\bar{y} = .327$ is less than 1. Consequently, a pharmacologist wants to reject the null hypothesis. What are the problems with using such a decision rule?
c. Locate values of the test statistic and corresponding $p$-value on the printout.
d. Select a value of $\alpha$, the probability of a Type I error. Interpret this value in the words of the problem.
e. Give the appropriate conclusion, based on the results of parts c and d.
f. What conditions must be satisfied for the test results to be valid?

8.24 *Surface roughness of pipe.* Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated interior pipe used in oil fields, Exercise 7.26 (p. 311). The data (in micrometers) for 20 sampled pipe sections are reproduced in the table on p. 391.

a. Give the null and alternative hypotheses for testing whether the mean surface roughness of coated interior pipe, $\mu$, differs from 2 micrometers.
b. The results of the test, part a, are shown in the MINITAB printout at the bottom of the page. Locate the test statistic and $p$-value on the printout.

MINITAB Output for Exercise 8.23

**One-Sample T: RATIO**

Test of mu = 1 vs not = 1

| Variable | N | Mean | StDev | SE Mean | 95% CI | T | P |
|---|---|---|---|---|---|---|---|
| RATIO | 416 | 0.3269 | 0.2915 | 0.0143 | (0.2988, 0.3550) | -47.09 | 0.000 |

MINITAB Output for Exercise 8.24

**One-Sample T: ROUGH**

Test of mu = 2 vs not = 2

| Variable | N | Mean | StDev | SE Mean | 95% CI | T | P |
|---|---|---|---|---|---|---|---|
| ROUGH | 20 | 1.88100 | 0.52391 | 0.11715 | (1.63580, 2.12620) | -1.02 | 0.322 |

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

c. Give the rejection region for the hypothesis test, using $\alpha = .05$.
d. State the appropriate conclusion for the hypothesis test.
e. In Exercise 7.26 you found a 95% confidence interval for $\mu$. Explain why the confidence interval and test statistic lead to the same conclusion about $\mu$.

**DISTILL**

8.25 *Water distillation with solar energy.* In countries with a water shortage, converting salt water to potable water is a critical problem. The standard method of water distillation is with a single slope solar still. Several enhanced solar energy water distillation systems were investigated in *Applied Solar Energy* (Vol. 46, 2010). One new system employs a sun tracking meter and a step-wise basin. The new system was tested over three randomly selected days at a location in Amman, Jordan. The daily amounts of distilled water collected by the new system over the three days were 5.07, 5.45, and 5.21 liters per square meter ($l/m^2$). Suppose it is known that the mean daily amount of distilled water collected by the standard method at the same location in Jordan is $\mu = 1.4\ l/m^2$.

a. Set up the null and alternative hypotheses for determining whether the mean daily amount of distilled water collected by the new system is greater than 1.4.
b. For this test, give a practical interpretation of the value $\alpha = .10$.
c. Find the mean and standard deviation of the distilled water amounts for the sample of three days. (The data are saved in the **DISTILL** file.)
d. Use the information from part c to calculate the test statistic.
e. Find the observed significance level ($p$-value) of the test.
f. State, practically, the appropriate conclusion.
g. Find the value of $\beta$ for $\mu_a = 5\ l/m^2$. Interpret this value.
h. Find the power of the test for $\mu_a = 5\ l/m^2$. Interpret this value.

**YIELD**

8.26 *Yield strength of steel connecting bars.* To protect against earthquake damage, steel beams are typically fitted and connected with plastic hinges. However, these plastic hinges are prone to deformations and are difficult to inspect and repair. An alternative method of connecting steel beams—one that uses high strength steel bars with clamps—was investigated in *Engineering Structures* (July 2013). Mathematical models for predicting the performance of these steel connecting bars assume the bars have a mean yield strength of 300 megapascals (MPa). To verify this assumption, the researchers conducted material property tests on the steel connecting bars. In a sample of three tests, the yield strengths were 354, 370, and 359 MPa. (These data are saved in the **YIELD** file.) Do the data indicate that the true mean yield strength of the steel bars exceeds 300 MPa? Test using $\alpha = .01$.

8.27 *Cheek teeth of extinct primates.* Refer to the *American Journal of Physical Anthropology* (Vol. 142, 2010) study of the characteristics of cheek teeth (e.g., molars) in an extinct primate species, Exercise 2.14 (p. 35). Recall that the researchers recorded the dentary depth of molars (in millimeters) for a sample of 18 cheek teeth extracted from skulls. These depth measurements are reproduced in the accompanying table. Anthropologists know that the mean dentary depth of molars in an extinct primate species—called Species "A"—is 15 millimeters. Is there evidence to indicate that the sample of 18 cheek teeth come from some other extinct primate species (i.e., some species other than Species "A")? Use the accompanying SPSS printout to answer the question.

**CHEEKTEETH**

Data on Dentary Depth (mm) of Molars

| | |
|---|---|
| 18.12 | 16.55 |
| 19.48 | 15.70 |
| 19.36 | 17.83 |
| 15.94 | 13.25 |
| 15.83 | 16.12 |
| 19.70 | 18.13 |
| 15.76 | 14.02 |
| 17.00 | 14.04 |
| 13.96 | 16.20 |

*Source:* Boyer, D.M., Evans, A.R., and Jernvall, J. "Evidence of Dietary Differentiation Among Late Paleocene-Early Eocene Plesiadapids (Mammalia, Primates)", *American Journal of Physical Anthropology,* Vol. 142, 2010. (Table A3.)

## SPSS Output for Exercise 8.27

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| M2Depth | 18 | 16.4994 | 1.97042 | .46443 |

**One-Sample Test**

| | Test Value = 15 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| M2Depth | 3.229 | 17 | .005 | 1.49944 | .5196 | 2.4793 |

8.28 *Dissolved organic compound in lakes.* The level of dissolved oxygen in the surface water of a lake is vital to maintaining the lake's ecosystem. Environmentalists from the University of Wisconsin monitored the dissolved oxygen levels over time for a sample of 25 lakes in the state (*Aquatic Biology,* May 2010). To ensure a representative sample, the environmentalists focused on several lake characteristics, including dissolved organic compound (DOC). The DOC data (measured in grams per cubic-meters) for the 25 lakes are listed in the accompanying table. The population of Wisconsin lakes has a mean DOC value of 15 grams/m$^3$.

a. Use a hypothesis test (at $\alpha = .10$) to make an inference about whether the sample is representative of all Wisconsin lakes for the characteristic, dissolved organic compound.

b. What is the likelihood that the test, part a, will detect a mean that differs from 15 grams/m$^3$ if, in fact, $\mu_a =$ 14 grams/m$^3$?

**WISCLAKES**

| LAKE | DOC | LAKE | DOC |
|---|---|---|---|
| Allequash | 9.6 | Muskellunge | 18.4 |
| Big Muskellunge | 4.5 | Northgate Bog | 2.7 |
| Brown | 13.2 | Paul | 4.2 |
| Crampton | 4.1 | Peter | 30.2 |
| Cranberry Bog | 22.6 | Plum | 10.3 |
| Crystal | 2.7 | Reddington Bog | 17.6 |
| EastLong | 14.7 | Sparkling | 2.4 |
| Helmet | 3.5 | Tenderfoot | 17.3 |
| Hiawatha | 13.6 | Trout Bog | 38.8 |
| Hummingbird | 19.8 | Trout Lake | 3.0 |
| Kickapoo | 14.3 | Ward | 5.8 |
| Little Arbor Vitae | 56.9 | West Long | 7.6 |
| Mary | 25.1 | | |

*Source:* Langman, O.C., et al. "Control of dissolved oxygen in northern temperate lakes over scales ranging from minutes to days", *Aquatic Biology,* Vol. 9, May 2010 (Table 1).

8.29 *Cooling method for gas turbines.* During periods of high electricity demand, especially during the hot summer months, the power output from a gas turbine engine can drop dramatically. One way to counter this drop in power is by cooling the inlet air to the gas turbine. An increasingly popular cooling method uses high-pressure inlet fogging. The performance of a sample of 67 gas turbines augmented with high-pressure inlet fogging was investigated in the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005). One measure of performance is heat rate (kilojoules per kilowatt per hour). Heat rates for the 67 gas turbines are listed in the table on the bottom of page. Suppose that a standard gas turbine has, on average, a heat rate of 10,000 kJ/kWh. Conduct a test to determine if the mean heat rate of gas turbines augmented with high-pressure inlet fogging exceeds 10,000 kJ/kWh. Use $\alpha = .05$.

8.30 *Alkalinity of river water.* In Exercise 5.36 (p. 205) you learned that the mean alkalinity level of water specimens collected from the Han River in Seoul, Korea, is 50 milligrams per liter. (*Environmental Science & Engineering*, September 1, 2000.) Consider a random sample of 100 water specimens collected from a tributary of the Han River. Suppose the mean and standard deviation of the alkalinity levels for the sample are $\bar{y} = 67.8$ mg/L and $s = 14.4$ mg/L. Is there sufficient evidence (at $\alpha = .01$) to indicate that the population mean alkalinity level of water in the tributary exceeds 50 mg/L?

8.31 *Walking straight into circles.* When people get lost in unfamiliar terrain, do they really walk in circles, as is commonly believed? To answer this question, researchers conducted a field experiment and reported the results in *Current Biology* (September 29, 2009). Fifteen volunteers were blindfolded and asked to walk as straight as possible in a certain direction in a large field. Walking trajectories were monitored every second for 50 minutes using GPS and the average directional bias (degrees per second) recorded for each walker. The data are shown in the table on p. 393. A strong tendency to veer consistently in the same direction will cause walking in circles. A mean directional bias of 0 indicates that walking trajectories were random. Consequently, the researchers tested whether the

**GASTURBINE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14622 | 13196 | 11948 | 11289 | 11964 | 10526 | 10387 | 10592 | 10460 | 10086 |
| 14628 | 13396 | 11726 | 11252 | 12449 | 11030 | 10787 | 10603 | 10144 | 11674 |
| 11510 | 10946 | 10508 | 10604 | 10270 | 10529 | 10360 | 14796 | 12913 | 12270 |
| 11842 | 10656 | 11360 | 11136 | 10814 | 13523 | 11289 | 11183 | 10951 | 9722 |
| 10481 | 9812 | 9669 | 9643 | 9115 | 9115 | 11588 | 10888 | 9738 | 9295 |
| 9421 | 9105 | 10233 | 10186 | 9918 | 9209 | 9532 | 9933 | 9152 | 9295 |
| 16243 | 14628 | 12766 | 8714 | 9469 | 11948 | 12414 | | | |

**CIRCLES**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| −4.50 | −1.00 | −0.50 | −0.15 | 0.00 | 0.01 | 0.02 | 0.05 | 0.15 |
| 0.20 | 0.50 | 0.50 | 1.00 | 2.00 | 3.00 | | | |

Source: Souman, J.L., Frissen, I., Sreenivasa, M.N., & Ernst, M.O. "Walking straight into circles", Current Biology, Vol. 19, No. 18, Sep. 29, 2009 (Figure 2).

true mean bias differed significantly from 0. A SAS printout of the analysis is shown below.

a. Interpret the results of the hypothesis test for the researchers. Use $\alpha = .10$.
b. Although most volunteers showed little overall bias, the researchers produced maps of the walking paths showing that each occasionally made several small circles during the walk. Ultimately, the researchers supported the "walking into circles" theory. Explain why the data in the table is insufficient for testing whether an individual walks into circles.

**SAS Output for Exercise 8.31**

The TTEST Procedure

Variable: BIAS

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 15 | 0.0853 | 1.6031 | 0.4139 | -4.5000 | 3.0000 |

| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| 0.0853 | -0.8024 | 0.9731 | 1.6031 | 1.1737 | 2.5282 |

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 14 | 0.21 | 0.8396 |

8.32 *Deep hole drilling.* "Deep hole" drilling is a family of drilling processes used when the ratio of hole depth to hole diameter exceeds 10. Successful deep hole drilling depends on the satisfactory discharge of the drill chip. An experiment was conducted to investigate the performance of deep hole drilling when chip congestion exists (*Journal of Engineering for Industry*, May 1993). The length (in millimeters) of 50 drill chips resulted in the following summary statistics: $\bar{y} = 81.2$ mm, $s = 50.2$ mm. Conduct a test to determine whether the true mean drill chip length, $\mu$, differs from 75 mm. Use a significance level of $\alpha = .01$.

## Theoretical Exercise

8.33 Refer to Exercises 8.8–8.10 (p. 380, 381). Show that the rejection region for the likelihood ratio test is given by $Z > z_\alpha$, where $P(Z > z_\alpha) = \alpha$.

(*Hint:* Under the assumption that $H_0$: $\mu = 0$ is true, show that $\sqrt{n}(\bar{y})$ is a standard normal random variable.)

## 8.7 Testing the Difference Between Two Population Means: Independent Samples

Consider independent random samples from two populations with means $\mu_1$ and $\mu_2$, respectively. When the sample sizes are large (i.e., $n_1 \geq 30$ and $n_2 \geq 30$), a test of hypothesis for the difference between the population means $(\mu_1 - \mu_2)$ is based on the pivotal $z$ statistic given in Section 7.5. A summary of the large-sample test is provided in the box.

**Large-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Independent Samples**

*One-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) > D_0$

[or $H_a$: $(\mu_1 - \mu_2) < D_0$]

*Two-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) \neq D_0$

$$\text{Test statistic: } Z = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sigma_{(\bar{y}_1 - \bar{y}_2)}} \approx \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

| *Rejection region:* | *Rejection region:* |
|---|---|
| $Z > z_\alpha \quad (\text{or } Z < -z_\alpha)$ | $\lvert Z \rvert > z_{\alpha/2}$ |
| $p\text{-value} = P(Z > z_c)\ [\text{or, } P(Z < z_c)]$ | $p\text{-value} = 2P(Z > \lvert z_c \rvert)$ |

where $P(Z > z_\alpha) = \alpha$, $P(Z > z_{\alpha/2}) = \alpha/2$, $\mu_0$ is our symbol for the particular numerical value specified for $\mu$ in the null hypothesis, and $z_c$ is the computed value of the test statistic.

(*Note:* $D_0$ is our symbol for the particular numerical value specified for $(\mu_1 - \mu_2)$ in the null hypothesis. In many practical applications, we wish to hypothesize that there is no difference between the population means; in such cases, $D_0 = 0$.)

*Assumptions:* 1. The sample sizes $n_1$ and $n_2$ are sufficiently large—say, $n_1 \geq 30$ and $n_2 \geq 30$.
2. The two samples are selected randomly and independently from the target populations.

## Example 8.12

Testing $\mu_1 - \mu_2$: Comparing Two Leavening Processes

To reduce costs, a bakery has implemented a new leavening process for preparing commercial bread loaves. Loaves of bread were randomly sampled and analyzed for calorie content both before and after implementation of the new process. A summary of the results of the two samples is shown in the Table 8.4. Do these samples provide sufficient evidence to conclude that the mean number of calories per loaf has decreased since the new leavening process was implemented? Test using $\alpha = .05$.

**TABLE 8.4 Summary of Calories per Loaf of Bread, Example 8.12**

| *New Process* | *Old Process* |
|---|---|
| $n_1 = 50$ | $n_2 = 30$ |
| $\bar{y}_1 = 1{,}255$ calories | $\bar{y}_2 = 1{,}330$ calories |
| $s_1 = 215$ calories | $s_2 = 238$ calories |

Solution

We can best answer this question by performing a test of a hypothesis. Defining $\mu_1$ as the mean calorie content per loaf manufactured by the new process and $\mu_2$ as the mean calorie content per loaf manufactured by the old process, we will attempt to support the research (alternative) hypothesis that $\mu_2 > \mu_1$ [i.e., that $(\mu_1 - \mu_2) < 0$]. Thus, we will test the null hypothesis that $(\mu_1 - \mu_2) = 0$, rejecting this hypothesis if $(\bar{y}_1 - \bar{y}_2)$ equals a large negative value. The elements of the test are as follows:

$$H_0\text{:} \quad (\mu_1 - \mu_2) = 0 \quad (\text{i.e., } D_0 = 0)$$

$$H_a\text{:} \quad (\mu_1 - \mu_2) < 0 \quad (\text{i.e., } \mu_1 < \mu_2)$$

*Test statistic:* $$Z = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sigma_{(\bar{y}_1 - \bar{y}_2)}} = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sigma_{(\bar{y}_1 - \bar{y}_2)}}$$

(since both $n_1$ and $n_2$ are greater than or equal to 30)

*Rejection region:* $Z < -z_\alpha = -1.645$ (see Figure 8.13)

*Assumptions:* The two samples of bread loaves are independently selected.

We now calculate the test statistic:

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sigma_{(\bar{y}_1 - \bar{y}_2)}} = \frac{(1{,}255 - 1{,}330)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

$$\approx \frac{-75}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{-75}{\sqrt{\dfrac{(215)^2}{50} + \dfrac{(238)^2}{30}}} = \frac{-75}{53.03} = -1.41$$

**FIGURE 8.13**
Rejection region for Example 8.12

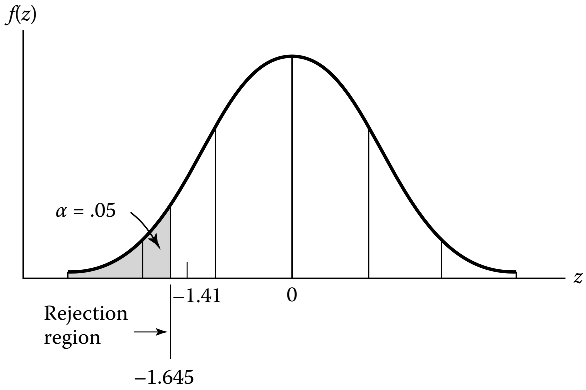



**FIGURE 8.14**
MINITAB Test to Compare Means, Example 8.12

```
Two-Sample T-Test and CI

                          SE
Sample   N  Mean  StDev  Mean
1       50  1255    215    30
2       30  1330    238    43


Difference = mu (1) - mu (2)
Estimate for difference:  -75.0
95% upper bound for difference:  13.7
T-Test of difference = 0 (vs <): T-Value = -1.41  P-Value = 0.081  DF = 56
```

This value is shaded on the MINITAB printout of the analysis, Figure 8.14. Note that the $p$-value of the test (also shaded) is .081.

As you can see in Figure 8.13, the calculated $z$ value does not fall in the rejection region. Also, $\alpha = .05$ is less than $p$-value $= .081$. Consequently, we fail to reject $H_0$; the samples do not provide sufficient evidence (at $\alpha = .05$) to conclude that the new process yields a loaf with fewer mean calories.

When the sample sizes $n_1$ and $n_2$ are inadequate to permit use of the large-sample procedure of Example 8.12, modifications may be made to perform a small-sample test of hypothesis about the difference between two population means. The test procedure is based on assumptions that are, again, more restrictive than in the large-sample case. The elements of the hypothesis test and the assumptions required are listed in the box. *Reminder:* When the assumption of normal population is grossly violated, the small-sample test outlined here will be invalid. In this case, we must resort to a nonparametric method (Chapter 15).

### Small-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Independent Samples

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $(\mu_1 - \mu_2) = D_0$ | $H_0$: $(\mu_1 - \mu_2) = D_0$ |
| $H_a$: $(\mu_1 - \mu_2) > D_0$ | $H_a$: $(\mu_1 - \mu_2) \neq D_0$ |
| [or $H_a$: $(\mu_1 - \mu_2) < D_0$] | |

*Test statistic*: $$T = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

| | |
|---|---|
| *Rejection region:* $T > t_\alpha$ [or $T < -t_\alpha$] | *Rejection region:* $\lvert T\rvert > t_{\alpha/2}$ |
| $p$-value $= P(T \geq t_c)$ [or, $P(T \leq t_c)$] | $p$-value $= 2P(T > \lvert t_c\rvert)$ |

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

the distribution of $T$ is based on $n_1 + n_2 - 2$ df, and $t_c$ is the computed value of the test statistic.

*Assumptions:*
1. The populations from which the samples are selected both have approximately normal relative frequency distributions.
2. The variances of the two populations are equal, i.e., $\sigma_1^2 = \sigma_2^2$
3. The random samples are selected in an independent manner from the two populations.

*Warning:* When the assumption of normal populations is violated, the test may lead to erroneous inferences. In this case, use the nonparametric Wilcoxon test described in Section 15.3.

## Example 8.13

Testing $\mu_1 - \mu_2$: Comparing Gas and Electric Energy

An industrial plant wants to determine which of two types of fuel—gas or electric—will produce more useful energy at the lower cost. One measure of economical energy production, called the *plant investment per delivered quad*, is calculated by taking the amount of money (in dollars) invested in the particular utility by the plant and dividing by the delivered amount of energy (in quadrillion British thermal units). The smaller this ratio, the less an industrial plant pays for its delivered energy. Independent random samples of 11 plants using electrical utilities and 16 plants using gas utilities were taken, and the plant investment/quad was calculated for each. The data are listed in Table 8.5. Do these data provide sufficient evidence at $\alpha = .05$ to indicate a difference in the average investment/quad between all plants using gas and all those using electric utilities?

**INVQUAD**

**TABLE 8.5 Data on Plant Investment/Quad, Example 8.13**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Electric: | 204.15 | 0.57 | 62.76 | 89.72 | 0.35 | 85.46 | | |
| | 0.78 | 0.65 | 44.38 | 9.28 | 78.60 | | | |
| Gas: | 0.78 | 16.66 | 74.94 | 0.01 | 0.54 | 23.59 | 88.79 | 0.64 |
| | 0.82 | 91.84 | 7.20 | 66.64 | 0.74 | 64.67 | 165.60 | 0.36 |

Solution

Let $\mu_1$ represent the mean investment/quad for all plants with electric utilities and let $\mu_2$ represent the mean investment/quad for all plants with gas utilities. Then, we want to conduct the test:

$$H_0: \quad (\mu_1 - \mu_2) = 0 \qquad (\text{i.e., } \mu_1 = \mu_2)$$
$$H_a: \quad (\mu_1 - \mu_2) \neq 0 \qquad (\text{i.e., } \mu_1 > \mu_2 \text{ or } \mu_1 < \mu_2)$$

Summary statistics for the two samples were produced using SPSS. The resulting SPSS printout is shown in Figure 8.15. Note that $\bar{y}_1 = 52.43$, $\bar{y}_2 = 37.74$, $s_1 = 62.43$, and $s_2 = 49.05$.

To obtain the test statistics, we first calculate

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(11 - 1)(62.43)^2 + (16 - 1)(49.05)^2}{11 + 16 - 2}$$

$$= \frac{75{,}051.31}{25} = 3002.05$$

Then, if we can assume that the distributions of the investment/quad data for the two plant types are both approximately normal with equal variances, the test statistic is

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{52.43 - 37.74}{\sqrt{3002.05\left(\frac{1}{11} + \frac{1}{16}\right)}} = \frac{14.69}{21.46} = .68$$

Note that this test statistic and the corresponding $p$-value for the test are both shaded on the SPSS printout in Figure 8.15. Since the two-tailed $p$-value (for the equal variances case), $p$-value $= .500$ exceeds $\alpha = .05$, there is insufficient evidence to reject $H_0$.

That is, we cannot conclude (at $\alpha = .05$) that the mean investment/quad levels for those plants with electric and gas utilities are different.

**Group Statistics**

| | UTILITY | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| INV_QUAD | Electric | 11 | 52.4273 | 62.42553 | 18.82200 |
| | Gas | 16 | 37.7388 | 49.04545 | 12.26136 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| INV_QUAD | Equal variances assumed | .264 | .612 | .684 | 25 | .500 | 14.68852 | 21.46024 | -29.50968 | 58.88672 |
| | Equal variances not assumed | | | .654 | 18.114 | .521 | 14.68852 | 22.46350 | -32.48433 | 61.86138 |

**FIGURE 8.15**
SPSS printout for Example 8.13

Recall from Section 7.5 that valid small-sample inferences about $(\mu_1 - \mu_2)$ can still be made when the assumption of equal variances is violated. We conclude this section by giving the modifications required to obtain approximate small-sample tests about $(\mu_1 - \mu_2)$ when $\sigma_1^2 \neq \sigma_2^2$ for the two cases described in Section 7.5: $n_1 = n_2$ and $n_1 \neq n_2$.

## Modifications to Small-Sample Tests About $(\mu_1 - \mu_2)$ When $\sigma_1^2 \neq \sigma_2^2$: Independent Samples

$n_1 = n_2 = n$

*Test statistic:*

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{1}{n}(s_1^2 + s_2^2)}}$$

*Degrees of freedom:* $\nu = n_1 + n_2 - 2 = 2(n-1)$

$n_1 \neq n_2$

*Test statistic:*

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

*Degrees of freedom:* $$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}\right]}$$

[*Note:* The value of $\nu$ will generally not be an integer. Round down to the nearest integer to use the $T$ table (Table 7 of Appendix B).]

## Applied Exercises

8.34 *Drug content assessment.* Refer to Exercise 7.39 (p. 319) and the *Analytical Chemistry* (Dec. 15, 2009) study in which scientists used high-performance liquid chromatography to determine the amount of drug in a tablet. Recall that 25 tablets were produced at each of two different, independent sites, and drug concentration (measured as a percentage) was determined for each tablet. These data are reproduced in the accompanying table. In Exercise 7.39 you used a 95% confidence interval to determine whether there is any difference between the mean drug concentration in tablets produced at the two sites. Now analyze the data using a statistical test of hypothesis at $\alpha = .05$. (See the accompanying MINITAB printout.) Do the inferences drawn from the test of hypothesis and confidence interval agree?

**DRUGCON**

*Site 1*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 91.28 | 92.83 | 89.35 | 91.90 | 82.85 | 94.83 | 89.83 | 89.00 | 84.62 |
| 86.96 | 88.32 | 91.17 | 83.86 | 89.74 | 92.24 | 92.59 | 84.21 | 89.36 |
| 90.96 | 92.85 | 89.39 | 89.82 | 89.91 | 92.16 | 88.67 | | |

*Site 2*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 89.35 | 86.51 | 89.04 | 91.82 | 93.02 | 88.32 | 88.76 | 89.26 | 90.36 |
| 87.16 | 91.74 | 86.12 | 92.10 | 83.33 | 87.61 | 88.20 | 92.78 | 86.35 |
| 93.84 | 91.20 | 93.44 | 86.77 | 83.77 | 93.19 | 81.79 | | |

*Source:* Borman, P.J., Marion, J.C., Damjanov,I., & Jackson, P. "Design and analysis of method equivalence studies", *Analytical Chemistry,* Vol. 81, No. 24, December 15, 2009 (Table 3).

### MINITAB Output for Exercise 8.34

**Two-Sample T-Test and CI: Content, Site**

```
Two-sample T for Content

Site   N   Mean  StDev  SE Mean
1     25  89.55   3.07     0.61
2     25  89.03   3.34     0.67


Difference = mu (1) - mu (2)
Estimate for difference:  0.515
95% CI for difference:  (-1.308, 2.338)
T-Test of difference = 0 (vs not =): T-Value = 0.57  P-Value = 0.573  DF = 48
Both use Pooled StDev = 3.2057
```

8.35 *Time required to complete a task.* When asked, "How much time will you require to complete this task", cognitive theory posits that people (e.g., an electrical engineer) will typically underestimate the time required. Would the opposite theory hold if the question was phrased in terms of how much work could be completed in a given amount of time? This was the question of interest to researchers writing in *Applied Cognitive Psychology* (Vol. 25, 2011). For one study conducted by the researchers, each in a sample of forty University of Oslo students was asked how many minutes it would take to read a 32-page technical report. In a second study, forty-two students were asked how many pages of a lengthy technical report they could read in 48 minutes. (The students in either study did not actually read the report.) Numerical descriptive statistics (based on summary information published in the article) for both studies are provided in the accompanying table.

| | Estimated Time (minutes) | Estimated Number of Pages |
|---|---|---|
| Sample size, $n$ | 40 | 42 |
| Sample mean, $\bar{x}$ | 60 | 28 |
| Sample standard deviation, s | 41 | 14 |

a. The researchers determined that the actual mean time it takes to read the report is $\mu = 48$ minutes. Is there evidence to support the theory that the students, on average, will overestimate the time it takes to read the report? Test using $\alpha = .10$.
b. The researchers also determined that the actual mean number of pages of the report that are read within the allotted time is $\mu = 32$ pages. Is there evidence to support the theory that the students, on average, will underestimate the number of report pages that can be read? Test using $\alpha = 10$.
c. The researchers noted that the distribution of both estimated time and estimated number of pages is highly skewed (i.e., not normally distributed). Does this fact impact the inferences derived in parts a and b? Explain.

8.36 *Do video game players have superior visual attention skills?* Researchers at Griffin University (Australia) conducted a study to determine whether video game players have superior visual attention skills than non-video game players. (*Journal of Articles in Support of the Null Hypothesis,* Vol. 6, 2009.) Two groups of male psychology students—32 video game players (VGP group) and 28 non-players (NVGP group)—were subjected to a series of visual attention tasks that included the attentional blink test. A test for the difference between two means yielded $t = -.93$ and $p$-value $= .358$. Consequently, the researchers' reported that "no statistically significant differences in the mean test performances of the two groups were found". Summary statistics for the comparison are provided in the next table. Do you agree with the researchers' conclusion?

| | *VGP* | *NVGP* |
|---|---|---|
| Sample size: | 32 | 28 |
| Mean score: | 84.81 | 82.64 |
| Standard deviation: | 9.56 | 8.43 |

*Source:* Murphy, K. and Spencer, A. "Playing video games does not make for better visual attention skills", *Journal of Articles in Support of the Null Hypothesis,* Vol. 6, No. 1, 2009.

8.37 *Index of Biotic Integrity.* Refer to the *Journal of Agricultural, Biological, and Environmental Sciences* (June 2005) analysis of the Index of Biotic Integrity (IBI), Exercise 7.42 (p. 320). Recall that the IBI measures the biological condition or health of an aquatic region. Summary data on the IBI for sites located in two Ohio river basins, Muskingum and Hocking, are reproduced in the next table. Conduct a test of hypothesis (at $\alpha = .10$) to compare the mean IBI values of the two river basins. Explain why the result will agree with the inference derived from the 90% confidence interval, Exercise 7.42.

| River Basin | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| Muskingum | 53 | .035 | 1.046 |
| Hocking | 51 | .340 | .960 |

*Source:* Boone, E. L., Keying, Y., and Smith, E. P. "Evaluating the relationship between ecological and habitat conditions using hierarchical models." *Journal of Agricultural, Biological, and Environmental Sciences*, Vol. 10, No. 2, June 2005 (Table 1).

8.38 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercise 2.23 (p. 38). Fifty solutions of deionized water were prepared both with and without calcium/gypsum, and the level of flotation of silica in the solution was measured using a variable called *zeta potential* (measured in millivolts, mV). The data (simulated, based on information provided in the journal article) are reproduced in the table on the next page. Conduct a test of hypothesis to compare the mean zeta potential values of the two types of solutions. Can you conclude that the addition of calcium/gypsum to the solution impacts silica flotation level?

**GASTURBINE**

8.39 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of gas turbines augmented with high-pressure inlet fogging, Exercise 8.29 (p. 392). The researchers classified gas turbines into three categories: traditional, advanced, and aeroderivative. Summary statistics on heat rate (kilojoules per kilowatt per hour) for each of the three types of gas turbines in the sample are shown in the MINITAB printout on the next page.

### Data for Exercise 8.38

**SILICA**

Without calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −47.1 | −53.0 | −50.8 | −54.4 | −57.4 | −49.2 | −51.5 | −50.2 | −46.4 | −49.7 |
| −53.8 | −53.8 | −53.5 | −52.2 | −49.9 | −51.8 | −53.7 | −54.8 | −54.5 | −53.3 |
| −50.6 | −52.9 | −51.2 | −54.5 | −49.7 | −50.2 | −53.2 | −52.9 | −52.8 | −52.1 |
| −50.2 | −50.8 | −56.1 | −51.0 | −55.6 | −50.3 | −57.6 | −50.1 | −54.2 | −50.7 |
| −55.7 | −55.0 | −47.4 | −47.5 | −52.8 | −50.6 | −55.6 | −53.2 | −52.3 | −45.7 |

With calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −9.2 | −11.6 | −10.6 | −8.0 | −10.9 | −10.0 | −11.0 | −10.7 | −13.1 | −11.5 |
| −11.3 | −9.9 | −11.8 | −12.6 | −8.9 | −13.1 | −10.7 | −12.1 | −11.2 | −10.9 |
| −9.1 | −12.1 | −6.8 | −11.5 | −10.4 | −11.5 | −12.1 | −11.3 | −10.7 | −12.4 |
| −11.5 | −11.0 | −7.1 | −12.4 | −11.4 | −9.9 | −8.6 | −13.6 | −10.1 | −11.3 |
| −13.0 | −11.9 | −8.6 | −11.3 | −13.0 | −12.2 | −11.3 | −10.5 | −8.8 | −13.4 |

### MINITAB Output for Exercise 8.39

**Descriptive Statistics: HEATRATE**

| Variable | ENGINE | N | Mean | StDev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| HEATRATE | Advanced | 21 | 9764 | 639 | 9105 | 11588 |
| | Aeroderiv | 7 | 12312 | 2652 | 8714 | 16243 |
| | Traditional | 39 | 11544 | 1279 | 10086 | 14796 |

a. Is there sufficient evidence of a difference between the mean heat rates of traditional augmented gas turbines and aeroderivative augmented gas turbines? Test using $\alpha = .05$.

b. Is there sufficient evidence of a difference between the mean heat rates of advanced augmented gas turbines and aeroderivative augmented gas turbines? Test using $\alpha = .05$.

**VOLTAGE**

8.40 *Process voltage readings.* Refer to the Harris Corporation/University of Florida comparison of the mean process voltage readings at two locations, Exercise 7.46 (p. 321). The data for 30 production runs at both the old and new locations are saved in the **VOLTAGE** file. The SAS printout of the analysis is reproduced below. Find and interpret the $p$-value for the test to compare the mean process voltage readings. What do you conclude? Does your answer agree with Exercise 7.46?

8.41 *Shopping vehicle and judgment.* Refer to the *Journal of Marketing Research* (Dec., 2011) study of shopping cart design, Exercise 2.43 (p. 50). Design engineers want to know whether you may be more likely to purchase a vice product (e.g., a candy bar) when your arm is flexed (as

### SAS Output for Exercise 8.40

**Sample Statistics**

| Group | N | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| NEW | 30 | 9.422333 | 0.4789 | 0.0874 |
| OLD | 30 | 9.803667 | 0.5409 | 0.0988 |

**Hypothesis Test**

Null hypothesis: Mean 1 - Mean 2 = 0
Alternative: Mean 1 - Mean 2 ^= 0

| If Variances Are | t statistic | Df | Pr > t |
|---|---|---|---|
| Equal | -2.891 | 58 | 0.0054 |
| Not Equal | -2.891 | 57.16 | 0.0054 |

**90% Confidence Interval for the Difference between Two Means**

| Lower Limit | Upper Limit |
|---|---|
| -0.60 | -0.16 |

when carrying a shopping basket) than when your arm is extended (as when pushing a shopping cart). To test this theory, the researchers recruited 22 consumers and had each push their hand against a table while they were asked a series of shopping questions. Half of the consumers were told to put their arm in a flex position (similar to a shopping basket) and the other half were told to put their arm in an extended position (similar to a shopping cart). Participants were offered several choices between a vice and a virtue (e.g., a movie ticket vs. a shopping coupon, pay later with a larger amount vs. pay now) and a choice score (on a scale of 0 to 100) was determined for each. (Higher scores indicate a greater preference for vice options.) The average choice score for consumers with a flexed arm was 59, while the average for consumers with an extended arm was 43.

a. Suppose the standard deviations of the choice scores for the flexed arm and extended arm conditions are 4 and 2, respectively. In Exercise 2.43a you were asked whether this information supports the researchers' theory. Now answer the question by conducting a hypothesis test. Use $\alpha = .05$.

b. Suppose the standard deviations of the choice scores for the flexed arm and extended arm conditions are 10 and 15, respectively. In Exercise 2.43b you were asked whether this information supports the researchers' theory. Now answer the question by conducting a hypothesis test. Use $\alpha = .05$.

8.42 *Computer-mediated communication study.* Computer-mediated communication (CMC) is a form of interaction that heavily involves technology (e.g., instant messaging, email). A study was conducted to compare relational intimacy in people interacting via CMC to people meeting face-to-face (FTF). (*Journal of Computer-Mediated Communication*, Apr. 2004.) Participants were 48 undergraduate students, of which half were randomly assigned to the CMC group and half assigned to the FTF group. Each group was given a task that required communication with their group members. Those in the CMC group communicated using the "chat" mode of instant-messaging software; those in the FTF group met in a conference room. The variable of interest, relational intimacy score, was measured (on a 7-point scale) for each participant after each of three different meeting sessions. Summary statistics for the first meeting session are given here. The researchers hypothesized that, after the first meeting, the mean relational intimacy score for participants in the CMC group would be lower than the mean relational intimacy score for participants in the FTF group. Test the researchers' hypothesis using $\alpha = .10$.

| | CMC | FTF |
|---|---|---|
| *Number of Participants* | 24 | 24 |
| *Sample Mean* | 3.54 | 3.53 |
| *Standard Deviation* | .49 | .38 |

8.43 *Wastewater treatment study.* In *Ecological Engineering* (Feb. 2004), the potential of floating aquatic plants to treat dairy manure wastewater was investigated. For one part of the study, 16 treated wastewater samples were randomly divided into two groups—a control algal was cultured in half the samples and water hyacinth was cultured in the other half. The rate of increase in the amount of total phosphorus was measured in each water sample; a summary of the results is given in the accompanying table. Conduct a test to determine if there is a difference in mean rates of increase of total phosphorus for the two aquatic plants. Use $\alpha = .05$.

| | Control Algal | Water Hyacinth |
|---|---|---|
| *Number of Water Samples* | 8 | 8 |
| *Sample Mean* | .036 | .026 |
| *Standard Deviation* | .008 | .006 |

*Source:* Sooknah, R., and Wilkie, A. "Nutrient removal by floating aquatic macrophytes cultured in anaerobically digested flushed dairy manure wastewater." *Ecological Engineering*, Vol. 22, No. 1, Feb. 2004 (Table 5).

8.44 *Insecticides used in orchards.* *Environmental Science & Technology* (Oct. 1993) reported on a study of insecticides used on dormant orchards in the San Joaquin Valley, California. Ambient air samples were collected and analyzed daily at an orchard site during the most intensive period of spraying. The thion and oxon levels (in ng/m$^3$) in the air samples are recorded in the table, as well as the oxon/thion ratios. Compare the mean oxon/thion ratios of foggy and clear/cloudy conditions at the orchard using a test of hypothesis. Use $\alpha = .05$.

**ORCHARD**

| Date | Condition | Thion | Oxon | Oxon/ Thion Ratio |
|---|---|---|---|---|
| Jan. 15 | Fog | 38.2 | 10.3 | .270 |
| 17 | Fog | 28.6 | 6.9 | .241 |
| 18 | Fog | 30.2 | 6.2 | .205 |
| 19 | Fog | 23.7 | 12.4 | .523 |
| 20 | Fog | 62.3 | (Air sample lost) | — |
| 20 | Clear | 74.1 | 45.8 | .618 |
| 21 | Fog | 88.2 | 9.9 | .112 |
| 21 | Clear | 46.4 | 27.4 | .591 |
| 22 | Fog | 135.9 | 44.8 | .330 |
| 23 | Fog | 102.9 | 27.8 | .270 |
| 23 | Cloudy | 28.9 | 6.5 | .225 |
| 25 | Fog | 46.9 | 11.2 | .239 |
| 25 | Clear | 44.3 | 16.6 | .375 |

*Source:* Selber, J. N., et al. "Air and fog deposition residues of four organophosphate insecticides used on dormant orchards in the San Joaquin Valley, California." *Environmental Science & Technology*, Vol. 27, No. 10, Oct. 1993, p. 2240 (Table V).

## 8.8 Testing the Difference Between Two Population Means: Matched Pairs

It may be possible to acquire more information on the difference between two population means by using data collected in matched pairs instead of independent samples. Consider, for example, an experiment to investigate the effectiveness of cloud seeding in the artificial production of rainfall. Two farming areas with similar past meteorological records were selected for the experiment. One is seeded regularly; the other is left unseeded. The monthly precipitation at the farms will be recorded for 6 randomly selected months. The resulting data, matched on months, can be used to test a hypothesis about the difference between the mean monthly precipitation in the seeded and unseeded areas. The appropriate procedures are summarized in the boxes.

### Large-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Matched Pairs

*One-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) > D_0$ [or $H_a$: $(\mu_1 - \mu_2) < D_0$]

*Two-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) \neq D_0$

$$\textit{Test statistic:}\quad Z = \frac{\bar{d} - D_0}{\sigma_d/\sqrt{n}} \approx \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

where $\bar{d}$ and $s_d$ represent the mean and standard deviation of the sample of differences.

*Rejection region:* $Z > z_\alpha$ [or $Z < -z_\alpha$] — *Rejection region:* $|Z| > z_{\alpha/2}$

$p$-value $= P(Z > z_c)$ [or, $P(Z < z_c)$] — $p$-value $= 2P(Z > |z_c|)$

where $P(Z > z_\alpha) = \alpha$, $P(Z > z_{\alpha/2}) = \alpha/2$ and $z_c$ is the computed value of the test statistic.

[*Note:* $D_0$ is our symbol for the particular numerical value specified for $(\mu_1 - \mu_2)$ in $H_0$. In many applications, we want to hypothesize that there is no difference between the population means; in such cases, $D_0 = 0$.]

### Small-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Matched Pairs

*One-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) > D_0$ [or $H_a$: $(\mu_1 - \mu_2) < D_0$]

*Two-Tailed Test*

$H_0$: $(\mu_1 - \mu_2) = D_0$

$H_a$: $(\mu_1 - \mu_2) \neq D_0$

$$\textit{Test statistic:}\quad T = \frac{\bar{d} - D_0}{\sigma_d/\sqrt{n}} \approx \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

where $\bar{d}$ and $s_d$ represent the mean and standard deviation of the sample of differences.

*Rejection region:* $T > t_\alpha$ [or $T > -t_\alpha$] — *Rejection region:* $|T| > t_{\alpha/2}$

$p$-value $= P(T \geq t_c)$ [or, $P(T \leq t_c)$] — $p$-value $= 2P(T \geq |t_c|)$

where the $T$-distribution is based on $(n - 1)$ degrees of freedom, $P(T > t_\alpha) = \alpha$, $P(T > t_{\alpha/2}) = \alpha/2$, $\mu_0$ is our symbol for the particular numerical value specified for $\mu$ in the null hypothesis, and $t_c$ is the computed value of the test statistic.

[*Note:* $D_0$ is our symbol for the particular numerical value specified for $(\mu_1 - \mu_2)$ in the null hypothesis. In many practical applications, we want to hypothesize that there is no difference between the population means; in such cases, $D_0 = 0$.]

*Assumptions:* 1. The relative frequency distribution of the population of differences is approximately normal.
2. The paired differences are randomly selected from the population of differences.

*Warning:* When the assumption of normality is grossly violated, the $t$ test may lead to erroneous inferences. In this case, use the nonparametric Wilcoxon test described in Section 15.4.

## Example 8.14

Testing $\mu_d$: Cloud Seeding

Consider the cloud seeding experiment to compare monthly precipitation at the two farm areas. Do the data given in Table 8.6 provide sufficient evidence to indicate that the mean monthly precipitation at the seeded farm area exceeds the corresponding mean for the unseeded farm area? Test using $\alpha = .05$.

CLOUDSEED

**TABLE 8.6 Monthly Precipitation Data (in Inches) for Example 8.14**

| Farm Area | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *Seeded* | 1.75 | 2.12 | 1.53 | 1.10 | 1.70 | 2.42 |
| *Unseeded* | 1.62 | 1.83 | 1.40 | .75 | 1.71 | 2.33 |
| $d$ | .13 | .29 | .13 | .35 | −.01 | .09 |

Solution

Let $\mu_1$ and $\mu_2$ represent the mean monthly precipitation values for the seeded and unseeded farm areas, respectively. Since we want to be able to detect $\mu_1 > \mu_2$, we will conduct the one-tailed test:

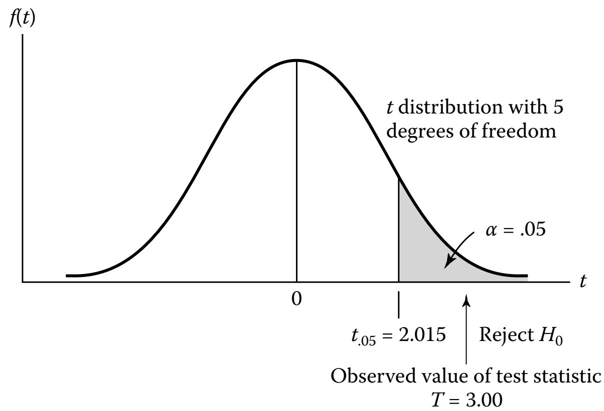
$$H_0\colon\ (\mu_1 - \mu_2) = 0$$
$$H_a\colon\ (\mu_1 - \mu_2) > 0$$

Assuming the differences in monthly precipitation values for the two areas are from an approximately normal distribution, the test statistic will have a $t$ distribution based on $(n - 1) = (6 - 1) = 5$ degrees of freedom. We will reject the null hypothesis if

$$T > t_{.05} = 2.015 \quad \text{(see Figure 8.16)}$$

To conduct the test by hand, we must first calculate the difference $d$ in monthly precipitation at the two farm areas for each month. These differences (where the observations for the unseeded farm area is subtracted from the observation for the seeded area within each pair) are shown in the last row of Table 8.6. Next, we would calculate the mean $\bar{d}$ and standard deviation $s_d$ for this sample of $n = 6$ differences to obtain the test statistic.

**FIGURE 8.16**
Rejection region for Example 8.14



```
Paired T for SEEDED - UNSEEDED

              N      Mean     StDev   SE Mean
SEEDED        6   1.77000   0.45974   0.18769
UNSEEDED      6   1.60667   0.52164   0.21296
Difference    6  0.163333  0.133066  0.054324


95% lower bound for mean difference: 0.053868
T-Test of mean difference = 0 (vs > 0): T-Value = 3.01  P-Value = 0.015
```

**FIGURE 8.17**
MINITAB printout for Example 8.14

Rather than perform these calculations, we will rely on the output from a computer. The MINITAB printout for the analysis is shown in Figure 8.17. The test statistic, shaded in Figure 8.17, is $T = 3.01$.

Since this value of the test statistic exceeds the critical value $t_{.05} = 2.015$, there is sufficient evidence (at $\alpha = .05$) to indicate that the mean monthly precipitation at the seeded farm area exceeds the mean for the unseeded farm area.

The same conclusion can be reached by examining the $p$-value of the test. The one-tailed $p$-value, shaded on the MINITAB printout, is .015. Since this value is less than the chosen $\alpha$ level (.05), we reject $H_0$. In fact, we will reject $H_0$ for any $\alpha$ larger than $p$-value $= .015$.

In the experiment of Example 8.14, why did we collect the data in matched pairs rather than use independent random samples of months, with some assigned to only the seeded area and others to only the unseeded area? The answer is that we expected some months to have more rain than others. To cancel out this variation from month to month, the experiment was designed so that precipitation at both farm areas would be recorded during the same months. Then both farm areas would be subjected to the same weather pattern in a given month. By comparing precipitation *within* each month, we were able to obtain more information on the difference in mean monthly precipitation than we could have obtained by independent random sampling.

## Applied Exercises

8.45 *Estimating well scale deposits.* Scale deposits can cause a serious reduction in the flow performance of a well. A study published in the *Journal of Petroleum and Gas Engineering* (April 2013) compared two methods of estimating the damage from scale deposits (called *skin factor*). One method of estimating the well skin factor uses a series of Excel spreadsheets, while the second method employs EPS computer software. Skin factor data was obtained from applying both methods to 10 randomly selected oil wells—5 vertical wells and 5 horizontal wells. The results are supplied in the accompanying table.

a. Compare the mean skin factor values for the two estimation methods using all 10 sampled wells. Test at $\alpha = .05$. What do you conclude?
b. Repeat part a, but analyze the data for the 5 horizontal wells only.
c. Repeat part a, but analyze the data for the 5 vertical wells only.

**SKIN**

| Well (Type) | Excel Spreadsheet | EPS Software |
|---|---|---|
| 1 (Horizontal) | 44.48 | 37.77 |
| 2 (Horizontal) | 18.34 | 13.31 |
| 3 (Horizontal) | 19.21 | 7.02 |
| 4 (Horizontal) | 11.70 | 4.77 |
| 5 (Horizontal) | 9.25 | 1.96 |
| 6 (Vertical) | 317.40 | 281.74 |
| 7 (Vertical) | 181.44 | 192.16 |
| 8 (Vertical) | 154.65 | 140.84 |
| 9 (Vertical) | 77.43 | 56.86 |
| 10 (Vertical) | 49.37 | 45.01 |

*Source:* Rahuma, K.M., et al. " Comparison between spreadsheet and specialized programs in calculating the effect of scale deposition on the well flow performance", *Journal of Petroleum and Gas Engineering,* Vol. 4, No. 4, April 2013 (Table 2).

8.46 *Computer-mediated communication study.* Refer to the *Journal of Computer-Mediated Communication* (Apr. 2004) study to compare relational intimacy in people interacting via computer-mediated communication (CMC) to people meeting face-to-face (FTF), Exercise 8.42 (p. 401). Recall that a relational intimacy score was measured (on a 7-point scale) for each participant after each of three different meeting sessions. The researchers also hypothesized that the mean relational intimacy score for participants in the CMC group will significantly increase between the first and third meetings, but the difference between the first and third meetings will not significantly change for participants in the FTF group.

a. For the CMC group comparison, give the null and alternative hypotheses of interest.
b. The researchers made the comparison, part **a**, using a paired $t$ test. Explain why the data should be analyzed as matched pairs.
c. For the CMC group comparison, the reported test statistic was $t = 3.04$ with $p$-value $= .003$. Interpret these results. Is the researchers' hypothesis supported?
d. For the FTF group comparison, give the null and alternative hypotheses of interest.
e. For the FTF group comparison, the reported test statistic was $t = .39$ with $p$-value $= .70$. Interpret these results. Is the researchers' hypothesis supported?

8.47 *Twinned drill holes.* Refer to the *Exploration and Mining Geology* (Vol. 18, 2009) study of drilling twin holes, Exercise 7.49 (p. 326). Recall that geologists use data collected at both holes to estimate the total amount of heavy minerals (THM) present at the drilling site. Data (THM percentages) for a sample of 15 twinned holes drilled at a diamond mine in Africa are repeated in the accompanying table. The geologists want to know if there is any evidence of a difference in the true THM means of all original holes and their twin holes drilled at the mine.

a. Conduct the appropriate test of hypothesis for the geologists. Use $\alpha = .10$.
b. In Exercise 7.49d, you formed a 90% confidence interval for the true mean difference ("1st hole" minus "2nd hole") in THM measurements and used this interval to answer the question of interest to the geologists. Do the inferences derived from the hypothesis test and confidence interval agree? Is this a surprising result? Explain.

**TWINHOLE**

| Location | 1st Hole | 2nd Hole |
|---|---|---|
| 1 | 5.5 | 5.7 |
| 2 | 11.0 | 11.2 |
| 3 | 5.9 | 6.0 |
| 4 | 8.2 | 5.6 |
| 5 | 10.0 | 9.3 |
| 6 | 7.9 | 7.0 |
| 7 | 10.1 | 8.4 |
| 8 | 7.4 | 9.0 |
| 9 | 7.0 | 6.0 |
| 10 | 9.2 | 8.1 |
| 11 | 8.3 | 10.0 |
| 12 | 8.6 | 8.1 |
| 13 | 10.5 | 10.4 |
| 14 | 5.5 | 7.0 |
| 15 | 10.0 | 11.2 |

8.48 *Settlement of shallow foundations.* Refer to the *Environmental & Engineering Geoscience* (Nov. 2012) study of methods for predicting settlement of shallow foundations on cohesive soil, Exercise 7.50 (p. 326). Actual settlement values for a sample of 13 structures built on a shallow foundation were determined, and these values compared to settlement predictions made using a formula that accounts for dimension, rigidity, and embedment depth of the foundation. The data (in millimeters) are reproduced in the table below. Use the SAS printout on the next page to test the hypothesis of no difference between the mean actual and mean predicted settlement values. Test using $\alpha = .05$.

**SHALLOW**

| Structure | Actual | Predicted |
|---|---|---|
| 1 | 11 | 11 |
| 2 | 11 | 11 |
| 3 | 10 | 12 |
| 4 | 8 | 6 |
| 5 | 11 | 9 |
| 6 | 9 | 10 |
| 7 | 9 | 9 |
| 8 | 39 | 51 |
| 9 | 23 | 24 |
| 10 | 269 | 252 |
| 11 | 4 | 3 |
| 12 | 82 | 68 |
| 13 | 250 | 264 |

*Source:* Ozur, M. "Comparing Methods for Predicting Immediate Settlement of Shallow Foundations on Cohesive Soils Based on Hypothetical and Real Cases", *Environmental & Engineering Geoscience,* Vol. 18, No. 4, November 2012 (from Table 4).

### SAS Output for Exercise 8.48

The TTEST Procedure

Difference: ACTUAL - PREDICTED

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 13 | 0.4615 | 8.3528 | 2.3166 | -14.0000 | 17.0000 |

| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| 0.4615 | -4.5860 | 5.5091 | 8.3528 | 5.9897 | 13.7883 |

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 12 | 0.20 | 0.8454 |

8.49 *Light to dark transition of genes.* *Synechocystis*, a type of cyanobacterium that can grow and survive in a wide range of conditions, is used by scientists to model DNA behavior. In the *Journal of Bacteriology* (July 2002), scientists isolated genes of the bacterium responsible for photosynthesis and respiration and investigated the sensitivity of the genes to light. Each gene sample was grown to midexponential phase in a growth incubator in "full light." The lights were extinguished and growth measured after 24 hours in the dark ("full dark"). The lights were then turned back on for 90 minutes ("transient light") followed immediately by an additional 90 minutes in the dark ("transient dark"). Standardized growth measurements in each light/dark condition were obtained for 103 genes. The complete data set is saved in the **GENEDARK** file. Data for the first 10 genes are shown in the accompanying table.

**GENEDARK**
(*First 10 Observations Shown*)

| Gene ID | Full-Dark | Tr-Light | Tr-Dark |
|---|---|---|---|
| SLR2067 | −0.00562 | 1.40989 | −1.28569 |
| SLR1986 | −0.68372 | 1.83097 | −0.68723 |
| SSR3383 | −0.25468 | −0.79794 | −0.39719 |
| SLL0928 | −0.18712 | −1.20901 | −1.18618 |
| SLR0335 | −0.20620 | 1.71404 | −0.73029 |
| SLR1459 | −0.53477 | 2.14156 | −0.33174 |
| SLL1326 | −0.06291 | 1.03623 | 0.30392 |
| SLR1329 | −0.85178 | −0.21490 | 0.44545 |
| SLL1327 | 0.63588 | 1.42608 | −0.13664 |
| SLL1325 | −0.69866 | 1.93104 | −0.24820 |

*Source:* Gill, R. T., et al. "Genome-wide dynamic transcriptional profiling of the light to dark transition in *Synechocystis Sp.* PCC6803." *Journal of Bacteriology*, Vol. 184, No. 13, July 2002.

a. Treat the data for the first 10 genes as a random sample collected from the population of 103 genes and test the hypothesis of no difference between the mean standardized growth of genes in the full-dark condition and genes in the transient-light condition. Use $\alpha = .01$.
b. Use a statistical software package to compute the mean difference in standardized growth of the 103 genes in the full-dark condition and the transient-light condition. Did the test, part **a**, detect this difference?
c. Repeat parts **a** and **b** for a comparison of the mean standardized growth of genes in the full-dark condition and genes in the transient-dark condition.
d. Repeat parts **a** and **b** for a comparison of the mean standardized growth of genes in the transient-light condition and genes in the transient-dark condition.

8.50 *Testing electronic circuits.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) comparison of two methods of testing electronic circuits, Exercise 7.52 (p. 327). Each of 11 circuits was tested using the standard compression/depression method and the new Huffman-based coding method, and the compression ratio recorded. The data are reproduced in the accompanying table. In theory, the Huffman coding method will yield a smaller mean compression ratio.

a. Test the theory using $\alpha = .05$.
b. Does your conclusion, part **a**, agree with the inference derived from the 95% confidence interval found in Exercise 7.52?

**CIRCUITS**

| Circuit | Standard Method | Huffman Coding Method |
|---|---|---|
| 1 | .80 | .78 |
| 2 | .80 | .80 |
| 3 | .83 | .86 |
| 4 | .53 | .53 |
| 5 | .50 | .51 |
| 6 | .96 | .68 |
| 7 | .99 | .82 |
| 8 | .98 | .72 |
| 9 | .81 | .45 |
| 10 | .95 | .79 |
| 11 | .99 | .77 |

*Source:* Ichihara, H., Shintani, M., and Inoue, T. "Huffman-based test response coding." *IEICE Transactions on Information & Systems*, Vol. E88-D, No. 1, Jan. 2005 (Table 3).

8.51 *Concrete pavement response to temperature.* Civil engineers at West Virginia University have developed a 3D model to predict the response of jointed concrete pavement to temperature variations. (*The International Journal of Pavement Engineering*, Sept. 2004.) To validate the model, model predictions were compared to field measurements on key concrete stress variables taken at a newly constructed highway. One variable measured was slab top transverse strain (i.e., change in length per unit length per unit time) at a distance of 1 meter from the longitudinal joint. The 5-hour changes (8:20 P.M. to 1:20 A.M.) in slab top transverse strain for 6 days are listed in the next table. Is there a significant difference between the mean daily transverse strain changes from field measurements and the 3D model? Test using $\alpha = .05$.

**SLABSTRAIN**

| | | Change in Transverse Strain | |
|---|---|---|---|
| Day | Change in Temperature (°C) | *Field Measurement* | *3D Model* |
| *Oct. 24* | −6.3 | −58 | −52 |
| *Dec. 3* | 13.2 | 69 | 59 |
| *Dec. 15* | 3.3 | 35 | 32 |
| *Feb. 2* | −14.8 | −32 | −24 |
| *Mar. 25* | 1.7 | −40 | −39 |
| *May 24* | −.2 | −83 | −71 |

*Source:* Shoukry, S., William, G., and Riad, M. "Validation of 3DFE model of jointed concrete pavement response to temperature variations." *The International Journal of Pavement Engineering*, Vol. 5, No. 3, Sept. 2004 (Table IV).

8.52 *Solar energy generation along highways.* The potential of using solar panels constructed above national highways to generate energy was explored in the *International Journal of Energy and Environmental Engineering* (Dec. 2013). Two-layer solar panels (with 1 meter separating the panels) were constructed above sections of both east-west and north-south highways in India. The amount of energy (kilo-Watt hours) supplied to the country's grid by the solar panels above the two types of highways was determined each month. The data for several randomly selected months are provided in the table. The researchers concluded that the "two-layer solar panel energy generation is more viable for the north-south oriented highways as compared to east-west oriented roadways". Do you agree?

**SOLAR**

| Month | East-West | North-South |
|---|---|---|
| February | 8658 | 8921 |
| April | 7930 | 8317 |
| July | 5120 | 5274 |
| September | 6862 | 7148 |
| October | 8608 | 8936 |

Source: Sharma, P. and Harinarayana, T. "Solar energy generation potential along national highways", *International Journal of Energy and Environmental Engineering,* Vol. 49, No. 1, December 2013 (Table 3).

8.53 *Modeling transport of gases.* In *AIChE Journal* (Jan. 2005), chemical engineers published a new method for modeling multicomponent transport of gases. Twelve gas mixtures consisting of neon, argon, and helium were prepared at different ratios and at different temperatures. The viscosity of each mixture ($10^{-5}$Pa · s) was measured experimentally and was calculated with the new model. The results are shown in the table below. The chemical engineers concluded that there is "an excellent agreement between our new calculation and experiments." Do you agree? Your answer should include a discussion of practical versus statistical significance.

**VISCOSITY**

| | Viscosity Measurements | | | Viscosity Measurements | |
|---|---|---|---|---|---|
| Mixture | Experimental | New Method | Mixture | Experimental | New Method |
| 1 | 2.740 | 2.736 | 7 | 2.886 | 2.910 |
| 2 | 2.569 | 2.575 | 8 | 2.957 | 2.965 |
| 3 | 2.411 | 2.432 | 9 | 3.790 | 3.792 |
| 4 | 2.504 | 2.512 | 10 | 3.574 | 3.582 |
| 5 | 3.237 | 3.233 | 11 | 3.415 | 3.439 |
| 6 | 3.044 | 3.050 | 12 | 3.470 | 3.476 |

*Source:* Kerkhof, P., and Geboers, M. "Toward a unified theory of isotropic molecular transport phenomena." *AIChE Journal*, Vol. 51, No. 1, January 2005 (Table 2).

## 8.9 Testing a Population Proportion

In Section 8.2, we gave several examples of a statistical test of hypothesis for a population proportion $p$ (e.g., the proportion of PC note book purchasers who buy a particular software package.). When the sample size is large, the sample proportion of successes $\hat{p}$ is approximately normal and the general formulas for conducting a large-sample $z$ test (given in Section 8.2) can be applied.

The procedure for testing a hypothesis about a population proportion $p$ based on a large sample from the target population is described in the box. (Recall that $p$ represents the probability of success in a binomial experiment.) For the procedure to be valid, the sample size must be sufficiently large to guarantee approximate normality of the sampling distribution of the sample proportion, $\hat{p}$. As with confidence intervals, a general rule of thumb for determining whether $n$ is "sufficiently large" is that both $n\hat{p}$ and $n\hat{q}$ are greater than or equal to 4.

### Large-Sample Test of Hypothesis About a Population Proportion

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $p = p_0$ | $H_0$: $p = p_0$ |
| $H_a$: $p > p_0$ [or $H_a$: $p < p_0$] | $H_a$: $p \neq p_0$ |

*Test statistic*: $Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}}$

where $q_0 = 1 - p_0$

| | |
|---|---|
| *Rejection region:* $Z > z_\alpha$ [or $Z < -z_\alpha$] | *Rejection region:* $\lvert Z\rvert > z_{\alpha/2}$ |
| $p$-value $= P(Z > z_c)$ [or, $P(Z < z_c)$] | $p$-value $= 2P(Z > \lvert z_c\rvert)$ |

where $P(Z > z_\alpha) = \alpha$, $P(Z > z_{\alpha/2}) = \alpha/2$, $p_0$ is our symbol for the particular numerical value specified for $p$ in the null hypothesis, and $z_c$ is the computed value of the test statistic.

*Assumption*: The sample size $n$ is sufficiently large so that the approximation is valid. As a rule of thumb, the condition of "sufficiently large" will be satisfied when $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$.

**Example 8.15**

Testing a Proportion: Steel Highway Bridges

Controversy surrounds the use of weathering steel in the construction of highway bridges. Critics have recently cited serious corrosive problems with weathering steel and are currently urging states to prohibit its use in bridge construction. On the other hand, the steel corporations claim that these charges are exaggerated and report that 95% of all weathering steel bridges in operation show "good" performance, with no major corrosive damage. To test this claim, a team of engineers and steel industry experts evaluated 60 randomly selected weathering steel bridges and found 54 of them showing "good" performance. Is there evidence, at $\alpha = .05$, that the true proportion of weathering steel highway bridges that show "good" performance is less than .95, the figure quoted by the steel corporations?

Solution

The parameter of interest is a population proportion, $p$. We want to test

$H_0$: $p = .95$

$H_a$: $p < .95$

where $p$ is the true proportion of all weathering steel highway bridges that show "good" performance.

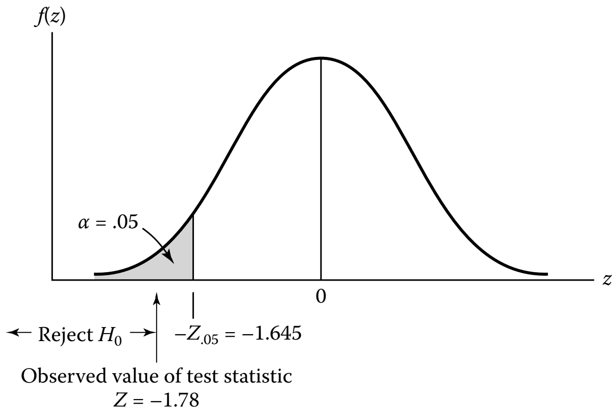
At significance level $\alpha = .05$, the null hypothesis will be rejected if

$$Z < -z_{.05}$$

that is, $H_0$ will be rejected if

$$Z < -1.645 \qquad \text{(see Figure 8.18)}$$

**FIGURE 8.18**
Rejection region for Example 8.15




The sample proportion of bridges that show "good" performance is

$$\hat{p} = \frac{54}{60} = .90$$

Thus, the test statistic has the value

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}} = \frac{.90 - .95}{\sqrt{(.95)(.05)/60}} = -1.78$$

This value of the test statistic is shown (shaded) on a MINITAB printout of the analysis, Figure 8.19. The *p*-value of the test (also shaded on the printout) is .038. Of course, we know we can conduct the test using either the rejection region or the *p*-value approach. Since $\alpha = .05$ exceeds *p*-value $= .038$, the null hypothesis can be rejected. There is sufficient evidence to support the hypothesis that the proportion of weathering steel highway bridges that show "good" performance is less than .95. [Note that both $n\hat{p} = 60(.90) = 54$ and $n\hat{q} = 60(.10) = 6$ exceed 4. Thus, the sample size is clearly large enough to guarantee the validity of the hypothesis test.]

**FIGURE 8.19**
MINITAB Test of a Population Proportion, Example 8.15

**Test and CI for One Proportion**

Test of p = 0.95 vs p < 0.95

| Sample | X | N | Sample p | 95% Upper Bound | Z-Value | P-Value |
|---|---|---|---|---|---|---|
| 1 | 54 | 60 | 0.900000 | 0.963705 | -1.78 | 0.038 |

Using the normal approximation.

Although small-sample procedures are available for testing hypotheses about a population proportion, the details are omitted from our discussion. It is our experience that they are of limited utility, since most surveys of binomial populations (for example, opinion polls) performed in the real world use samples that are large enough to employ the techniques of this section.

## Applied Exercises

8.54 *Annual survey of computer crimes.* The Computer Security Institute (CSI) conducts an annual survey of computer crime at United States businesses. CSI sends survey questionnaires to computer security personnel at all U.S. corporations and government agencies. A total of 351 organizations responded to the 2010 CSI survey. Of these, 144 admitted unauthorized use of computer systems at their firms during the year. (*CSI Computer Crime and Security Survey, 2010/2011*.) Let $p$ represent the true proportion of U.S. organizations that experience unauthorized use of computer systems at their firms.

a. Calculate a point estimate for $p$.
b. Set up the null and alternative hypothesis to test whether the value of $p$ differs from .35.
c. Calculate the test statistic for the test, part **b.**
d. Find the rejection region for the test if $\alpha = .05$.
e. Use the results of parts **c** and **d** to make the appropriate conclusion.
f. Find the $p$-value of the test and confirm that the conclusion based on the $p$-value agrees with the conclusion in part **e.**

8.55 *Toxic chemical incidents.* Refer to the *Process Safety Progress* (Sept. 2004) study of an emergency response system for incidents involving toxic chemicals in Taiwan, Exercise 3.5 (p. 86). In a sample of 250 toxic chemical incidents logged since the system was implemented, 15 occurred in a school laboratory. Suppose you want to conduct a test of hypothesis to determine if the true percentage of toxic chemical incidents in Taiwan that occur in a school laboratory is less than 10%.

a. Set up the null and alternative hypothesis for the test.
b. Give the rejection region for $\alpha = .01$.
c. Compute the value of the test statistic.
d. Give the appropriate conclusion for the test.

8.56 *Underwater acoustic communication.* Refer to the *IEEE Journal of Oceanic Engineering* (April 2013) study of the characteristics of subcarriers—telecommunication signals carried on top of one another—for underwater acoustic communications, Exercise 4.43 (p. 158). Recall that a subcarrier can be classified as either a data subcarrier (used for data transmissions), a pilot subcarrier (used for channel estimation and synchronization), or a null subcarrier (used for direct current and guard banks transmitting no signal). In a sample of 1,024 subcarrier signals transmitted off the coast of Martha's Vineyard, 672 were determined to be data subcarriers, 256 pilot subcarriers, and 96 null subcarriers. Suppose a communications engineer who works near Martha's Vineyard believes that fewer than 70% of all subcarrier signals transmitted in the area are data subcarriers. Is there evidence to support this belief? Test using $\alpha = .05$.

8.57 *Wiki usage in engineering education.* A *wiki* is a web information depository with content that can be updated and edited through a web browser. Engineering faculty at a university in Portugal investigated the degree to which wiki tools are accepted in an academic environment (*Computer Applications in Engineering Education,* Vol. 20, 2012). An online survey was made available to both professors and students that were involved in engineering courses that make use of a wiki-based tool. A total of 136 students responded to the survey. One of the survey questions asked, "Have you ever edited content in a wiki-based tool?" Of the 136 respondents, 72 answered "yes". Do the survey results support the claim that more than half of engineering students edit content in wiki-based tools? Test using $\alpha = .10$.

8.58 *Killing insects with low oxygen.* A group of Australian entomological toxicologists investigated the impact of exposure to low oxygen on the mortality of insects. (*Journal of Agricultural, Biological, and Environmental Statistics*, Sept. 2000.) Thousands of adult rice weevils were placed in a chamber filled with wheat grain and the chamber was exposed to nitrogen gas for 4 days. Insects were assessed as dead or alive 24 hours after exposure. The results: 31,386 dead weevils and 35 weevils found alive. Previous studies have shown a 99% mortality rate in adult rice weevils exposed to carbon dioxide for 4 days. Is the mortality rate for adult rice weevils exposed to nitrogen higher than 99%? Test using $\alpha = .10$.

8.59 *Friction in paper-feeding process.* Researchers at the University of Rochester studied the friction that occurs in the paper-feeding process of a photocopier (*Journal of Engineering for Industry*, May 1993). The experiment involved monitoring the displacement of individual sheets of paper in a stack fed through the copier. If no sheet except the top one moved more than 25% of the total stroke distance, the feed was considered successful. In a stack of 100 sheets of paper, the feeding process was successful 94 times. The success rate of the feeder is designed to be .90. Test to determine whether the true success rate of the feeder exceeds .90. Use $\alpha = .10$.

8.60 *Dehorning of dairy calves.* For safety reasons, calf dehorning has become a routine practice at dairy farms. A 2009 report by Europe's Standing Committee on the Food Chain and Animal Health (SANKO) stated that 80% of European dairy farms carry out calf dehorning. A later study, published in the *Journal of Dairy Science* (Vol. 94, 2011), found that in a sample of 639 Italian dairy farms, 515 dehorn calves. Does the *Journal of Dairy Science* study support or refute the figure reported by SANKO? Explain.

8.61 *Identifying organisms using a computer. National Science Education Standards* recommend that all life science students be exposed to methods of identifying unknown biological specimens. Due to certain limitations of traditional identification methods, biology professors at Slippery Rock University (SRU) developed a computer-aided system for identifying common conifers (deciduous trees) called Confir ID. (*The American Biology Teacher,* May 2010.) A sample of 171 life science students were exposed to both a traditional method of identifying conifers and Confir ID and then asked which method they preferred. The results: 138 students indicated their preference for Confir ID. In order to change the life sciences curriculum at SRU to include Confir ID, the biology department requires that more than 70% of the students prefer the new, computerized method. Should Confir ID be added to the curriculum at SRU? Explain your reasoning.

8.62 *Study of lunar soil. Meteoritics* (March 1995) reported the results of a study of lunar soil evolution. Data were obtained from the *Apollo 16* mission to the moon, during which a 62-cm core was extracted from the soil near the landing site. Monomineralic grains of lunar soil were separated out and examined for coating with dust and glass fragments. Each grain was then classified as coated or uncoated. Of interest is the "coat index," that is, the proportion of grains that are coated. According to soil evolution theory, the coat index will exceed .5 at the top of the core, equal .5 in the middle of the core, and fall below .5 at the bottom of the core. Use the summary data in the accompanying table to test each part of the three-part theory. Use $\alpha = .05$ for each test.

| | Location (depth) | | |
|---|---|---|---|
| | Top (4.25 cm) | Middle (28.1 cm) | Bottom (54.5 cm) |
| *Number of Grains Sampled* | 84 | 73 | 81 |
| *Number Coated* | 64 | 35 | 29 |

*Source:* Basu, A., and McKay, D.S. "Lunar soil evolution processes and Apollo 16 core 60013/60014." *Meteoritics*, Vol. 30, No. 2, Mar. 1995, p. 166 (Table 2).

## 8.10 Testing the Difference Between Two Population Proportions

Consider a transportion engineer who wants to compare the proportion of cars traveling with two or more people prior to adding a car-pool only lane on a major highway to the proportion a month after the car-pool lane was added. Let $p_1$ and $p_2$ represent the proportions prior to and after adding the car-pool lane, respectively. The method for performing a large-sample test of hypothesis about $(p_1 - p_2)$, the difference between two binomial proportions, is outlined in the box (p. 412).

When testing the null hypothesis that $(p_1 - p_2)$ equals some specified difference—say, $D_0$—we make a distinction between the case $D_0 = 0$ and the case $D_0 \neq 0$. For the special case $D_0 = 0$, i.e., when we are testing $H_0$: $(p_1 - p_2) = 0$ or, equivalently, $H_0$: $p_1 = p_2$, the best estimate of $p_1 = p_2 = p$ is found by dividing the total number of successes in the combined samples by the total number of observations in the two samples. That is, if $y_1$ is the number of successes in sample 1 and $y_2$ is the number of successes in sample 2, then

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

In this case, the best estimate of the standard deviation of the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is found by substituting $\hat{p}$ for both $p_1$ and $p_2$:

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

For all cases in which $D_0 \neq 0$ [for example, when testing $H_0$: $(p_1 - p_2) = .2$], we use $\hat{p}_1$ and $\hat{p}_2$ in the formula for $\sigma_{(\hat{p}_1 - \hat{p}_2)}$. However, in most practical situations, we will want to test for a difference between proportions—that is, we will want to test $H_0$: $(p_1 - p_2) = 0$.

### Large-Sample Test of Hypothesis About $(p_1 - p_2)$: Independent Samples

*One-Tailed Test*

$H_0$: $(p_1 - p_2) = D_0$

$H_a$: $(p_1 - p_2) > D_0$ [or $H_a$: $(p_1 - p_2) < D_0$]

*Two-Tailed Test*

$H_0$: $(p_1 - p_2) = D_0$

$H_a$: $(p_1 - p_2) \neq D_0$

$$\textit{Test statistic:} \quad Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

*Rejection region:* $Z > z_\alpha$ [or $Z < -z_\alpha$]

*Rejection region:* $|Z| > z_{\alpha/2}$,

$p$-value $= P(Z > z_c)$ [or, $P(Z < z_c)$]

$p$-value $= 2 \cdot P(Z > |z_c|)$

where $P(Z > z_\alpha) = \alpha$, $P(Z > z_{\alpha/2}) = \alpha/2$ and $z_c$ is the computed value of the test statistic.

When $D_0 \neq 0$,

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$.

When $D_0 = 0$,

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where the total number of successes in the combined sample is $(y_1 + y_2)$ and

$$\hat{p}_1 = \hat{p}_2 = \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

*Assumption*: The sample sizes, $n_1$ and $n_2$, are sufficiently large. This will be satisfied if $n_1\hat{p}_1 \geq 4$, $n_1\hat{q}_1 \geq 4$, and $n_2\hat{p}_2 \geq 4$, $n_2\hat{q}_2 \geq 4$.

The sample sizes $n_1$ and $n_2$ must be sufficiently large to ensure that the sampling distributions of $\hat{p}_1$ and $\hat{p}_2$, and hence of the difference $(\hat{p}_1 - \hat{p}_2)$, are approximately normal. The rule of thumb used to determine if the sample sizes are "sufficiently large" is the same as that given in Section 7.8, namely, that the quantities $n_1\hat{p}_1$, $n_2\hat{p}_2$, $n_1\hat{q}_1$, and $n_2\hat{q}_2$ are all greater than or equal to 4. (*Note:* If the sample sizes are not sufficiently large, $p_1$ and $p_2$ can be compared using a technique to be discussed in Chapter 9.)

## Example 8.16

Testing $p_1 - p_2$: Carpooling Study

Recently there have been intensive campaigns encouraging people to save energy by carpooling to work. Some cities have created an incentive for carpooling by designating certain highway traffic lanes as "car-pool only" (i.e., only cars with two or more passengers can use these lanes). To evaluate the effectiveness of this plan, toll booth personnel in one city monitored 2,000 randomly selected cars prior to establishing car-pool-only lanes and 1,500 cars after the car-pool-only lanes were established. The results of the study are shown in Table 8.7, where $y_1$ and $y_2$ represent the numbers of cars with two or more passengers (i.e., car-pool riders) in the "before" and "after" samples, respectively. Do the data indicate that the fraction of cars with car-pool riders has increased over this period? Use $\alpha = .05$.

**TABLE 8.7 Results of Carpooling Study, Example 8.17**

| | Before Car-Pool Lanes Established | After Car-Pool Lanes Established |
|---|---|---|
| ***Sample Size*** | $n_1 = 2{,}000$ | $n_2 = 1{,}500$ |
| ***Car-Pool Riders*** | $y_1 = 652$ | $y_2 = 576$ |

Solution

If we define $p_1$ and $p_2$ as the true proportions of cars with car-pool riders before and after establishing car-pool lanes, respectively, the elements of our test are

$H_0$: $(p_1 - p_2) = 0$
$H_a$: $(p_1 - p_2) < 0$

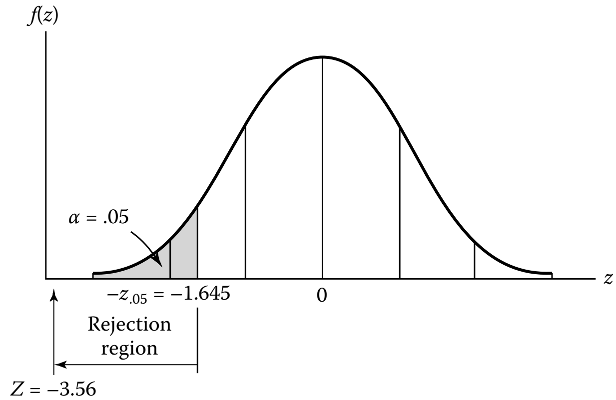
(The test is one-tailed since we are interested only in determining whether the proportion of cars with car-pool riders has increased, i.e., whether $p_2 > p_1$.)

*Test statistic*: $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$

*Rejection region*: $\alpha = .05$

$$Z < -z_\alpha = -z_{.05} = -1.645 \quad \text{(see Figure 8.20)}$$

**FIGURE 8.20**
Rejection region for Example 8.16




We now calculate the sample proportions of cars with car-pool riders:

$$\hat{p}_1 = \frac{652}{2{,}000} = .326 \qquad \hat{p}_2 = \frac{576}{1{,}500} = .384$$

The test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}} \approx \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

where

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{652 + 576}{2{,}000 + 1{,}500} = .351$$

Thus,

$$Z = \frac{.326 - .384}{\sqrt{(.351)(.649)\left(\dfrac{1}{2{,}000} + \dfrac{1}{1{,}500}\right)}} = \frac{-.058}{.0163} = -3.56$$

The test statistic value is also shown (shaded) on the MINITAB printout of the analysis, Figures 8.21. The $p$-value of the test (also highlighted on the printout) is approximately 0. Note that $Z = -3.56$ falls in the rejection region and $\alpha = .05$ exceeds the

**FIGURE 8.21**
MINITAB Test of Difference between Population Proportions, Example 8.16

```
Test and CI for Two Proportions

Sample     X     N  Sample p
1        652  2000  0.326000
2        576  1500  0.384000


Difference = p (1) - p (2)
Estimate for difference:  -0.058
95% upper bound for difference:  -0.0310948
Test for difference = 0 (vs < 0):  Z = -3.55  P-Value = 0.000
```

$p$-value. Thus, there is sufficient evidence at $\alpha = .05$ to conclude that the proportion of all cars with car-pool riders has increased after establishing car-pool lanes. We could place a confidence interval on $(p_1 - p_2)$ if we were interested in estimating the extent of the increase.

## Applied Exercises

8.63 *Producer willingness to supply biomass.* Refer to the *Biomass and Energy* (Vol. 36, 2012) study of the willingness of producers to supply biomass products such as surplus hay, Exercise 7.67 (p. 334). Recall that independent samples of Missouri producers and Illinois producers were surveyed and the number of producers willing to supply windrowing (mowing and piling) of hay was determined for each sample. Of the 558 Missouri producers surveyed, 187 were willing to offer windrowing services; of the 940 Illinois producers surveyed, 380 were willing to offer windrowing services. In Exercise 7.67, you obtained a 99% confidence interval for the difference between the proportions of producers who are willing to offer windrowing services in Missouri and Illinois from a MINITAB printout (reproduced below). Now, use the information on the printout to conduct a statistical test to determine if the proportion of producers who are willing to offer windrowing services to the biomass market differ for the two areas. For what value of $\alpha$ will the inferences derived from the test and confidence interval agree? Carry out the test of hypothesis and make the appropriate inference.

**MTBE**

8.64 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in New Hampshire wells, Exercise 7.66 (p. 334). Recall that 223 wells were classified according to well class (public or private) and detectable level of MTBE (below limit or detect). The SPSS printout below gives the number of wells in the sample with a detectable level of MTBE for both the 120 public wells and the 103 private wells.

a. Conduct a two-tailed test of hypothesis to compare the true proportion of public wells with a detectable level of MTBE to the true proportion of private wells with a detectable level of MTBE. Use $\alpha = .05$.
b. In Exercise 7.66, you compared the two proportions with a 95% confidence interval. Explain why the inference derived from the two-tailed test, part **a**, will agree with the inference derived from the confidence interval.

**MINITAB Output for Exercise 8.63**

```
Test and CI for Two Proportions

Sample     X    N  Sample p
1        187  558  0.335125
2        380  940  0.404255


Difference = p (1) - p (2)
Estimate for difference:  -0.0691299
99% CI for difference:  (-0.135079, -0.00318070)
Test for difference = 0 (vs not = 0):  Z = -2.67  P-Value = 0.008

Fisher's exact test: P-Value = 0.008
```

**SPSS Output for Exercise 8.64**

**DETECT * CLASS Crosstabulation**

Count

| | | CLASS | | |
|---|---|---|---|---|
| | | Private | Public | Total |
| DETECT | Below Limit | 81 | 72 | 153 |
| | Detect | 22 | 48 | 70 |
| Total | | 103 | 120 | 223 |

8.65 *Study of armyworm pheromones.* Refer to the *Journal of Chemical Ecology* (March 2013) study to determine the effectiveness of pheromones produced by two different strains of fall armyworms, Exercise 7.68 (p. 334). Recall that both corn-strain and rice-strain male armyworms were released into a field containing a synthetic pheromone made from a corn-strain blend. A count of the number of males trapped by the pheromone was then determined. The experiment was conducted once in a corn field, then again in a grass field. The results are repeated in the accompanying table. In Exercise 7.78 you compared the proportions of corn-strain and rice-strain males trapped by the pheromone.

a. Now, the researchers want to compare the proportion of corn-strain males trapped in the corn field to the proportion of corn-strain males trapped in the grass field. Carry out this comparison using a hypothesis test (at $\alpha = .10$). What inference can you draw from the data?
b. Repeat part a for the proportions of rice-strain males trapped by the pheromone.

| | Corn Field | Grass Field |
|---|---|---|
| Number of corn-strain males released | 112 | 215 |
| Number trapped | 86 | 164 |
| Number of rice-strain males released | 150 | 669 |
| Number trapped | 92 | 375 |

8.66 *Fluoride toxicity in Pakistan drinking water.* The results of an evaluation of the drinking water quality in Pakistan was reported in *Drinking Water Engineering and Science* (Vol. 6, 2013). Due to high levels of fluoride in the drinking water, Pakistanis are susceptible to fluoride toxicity (fluorosis)—which occurs when the fluoride level exceeds 1.5 milligrams per liter of water (mg/l). Water specimens were collected from various surface or groundwater sources (e.g., hand pumps, wells, springs, dams, etc.) of major cities of the country. The table gives the results for two cities—Lahore and Faisalabad. Is there evidence to indicate that the fraction of water specimens that exceed 1.5 mg/l of fluoride differs for the two cities? Test using $\alpha = .10$.

| | Lahore | Faisalabad |
|---|---|---|
| Number of water specimens sampled | 79 | 30 |
| Number exceeding 1.5 mg/l of fluoride | 21 | 4 |

8.67 *Traffic sign maintenance.* The Federal Highway Administration (FHWA) recently issued new guidelines for maintaining and replacing traffic signs. Civil engineers at North Carolina State University conducted a study of the effectiveness of various sign maintenance practices developed to adhere to the new guidelines and published the results in the *Journal of Transportation Engineering* (June 2013). One portion of the study focused on the proportion of traffic signs that fail the minimum FHWA retroreflectivity requirements. Of 1,000 signs maintained by the North Carolina Department of Transportation (NCDOT), 512 were deemed failures. Of 1,000 signs maintained by county-owned roads in North Carolina, 328 were deemed failures. Conduct a test of hypothesis to determine whether the true proportions of traffic signs that fail the minimum FHWA retroreflectivity requirements differ depending on whether the signs are maintained by the NCDOT or by the county. Test using $\alpha = .05$.

8.68 *Inactive oil and gas structures.* Refer to the *Oil & Gas Journal* (Jan. 3, 2005) study of 3,400 oil and gas structures in the Gulf of Mexico, Exercise 3.19 (p. 93). The accompanying table breaks down these structures by type (caisson, well protector, or fixed platform) and status (active or inactive). Assume the 3,400 structures are a representative sample of all oil and gas structures worldwide.

| | Structure Type | | | |
|---|---|---|---|---|
| | Caisson | Well Protector | Fixed Platform | Totals |
| *Active* | 503 | 225 | 1,447 | 2,175 |
| *Inactive* | 598 | 177 | 450 | 1,225 |
| *Totals* | 1,101 | 402 | 1,897 | 3,400 |

*Source:* Kaiser, M., and Mesyanzhinov, D. "Study tabulates idle Gulf of Mexico structures." *Oil & Gas Journal*, Vol. 103, No. 1, Jan. 3, 2005 (Table 2).

a. Conduct a test (at $\alpha = .10$) to determine if the proportion of caisson structures that are inactive exceeds the proportion of well protector structures that are inactive.
b. Conduct a test (at $\alpha = .10$) to determine if the proportion of caisson structures that are inactive exceeds the proportion of fixed platform structures that are inactive.
c. Conduct a test (at $\alpha = .10$) to determine if the proportion of well protector structures that are inactive differs from the proportion of fixed platform structures that are inactive.

8.69 *Killing insects with low oxygen.* Refer to the *Journal of Agricultural, Biological, and Environmental Statistics* (Sept. 2000) study of the mortality of rice weevils exposed to low oxygen. Exercise 8.58 (p. 410). Recall that 31,386 of 31,421 rice weevils were found dead after exposure to nitrogen gas for 4 days. In a second experiment, 23,516 of 23,676 rice weevils were found dead after exposure to nitrogen gas for 3.5 days. Conduct a test of hypothesis to compare the mortality rates of adult rice weevils exposed to nitrogen at the two exposure times. Is there a significant difference (at $\alpha = .10$) in the mortality rates?

8.70 *Vulnerability of relying party websites.* When you sign on to your Facebook account, you are granted access to more than 1 million *relying party* (RP) websites. This single sign-on (SSO) scheme is enabled by OAuth 2.0, an open and standardized web resource authorization protocol. Although the protocol claims to be secure, there is anecdotal evidence of critical vulnerabilities that allow an attacker to gain unauthorized access to the user's profile and allow the attacker to impersonate the victim on the RP website. Computer and systems engineers at the University of British Columbia investigated the vulnerability of relying party websites and presented their results at the *Proceedings of the 5th AMC Workshop on Computers & Communication Security* (Oct. 2012). RP websites were categorized as server-flow or client-flow websites. Of the 40 server-flow sites studied, 20 were found to be vulnerable to impersonation attacks. Of the 54 client-flow sites examined, 41 were found to be vulnerable to impersonation attacks. Do these results indicate that a client-flow website is more likely to be vulnerable to an impersonation attack than a server-flow website? Test using $\alpha = .01$.

8.71 *Engineering vs. technology degrees.* In addition to the traditional bachelor of engineering (BE) degree, many universities worldwide offer a bachelor of technology (BTech) degree for students who wish to work as an engineering technician. There is a perception that BTech students are not as "academically strong" as BE students. This issue was addressed in the *International Journal of Continuing Engineering Education and Lifelong Learning* (Vol. 13, 2003). The researchers compared BE and BTech students at an Australian university on a variety of academic-related outcomes. The following table gives the percentages of BE and BTech students who withdrew from two traditionally rigorous courses, engineering mathematics and engineering graphics/CAD.

| Engineering Mathematics | BE Students | BTech Students |
|---|---|---|
| *Number Enrolled* | 537 | 117 |
| *Percentage Withdrawn* | 27.8% | 19.7% |

| Engineering Graphics/CAD | BE Students | BTech Students |
|---|---|---|
| *Number Enrolled* | 727 | 374 |
| *Percentage Withdrawn* | 39.5% | 52.1% |

*Source:* Palmer, S., and Bray, S. "Comparative academic performance of engineering and technology students at Deakin University, Australia." *International Journal of Continuing Engineering Education and Lifelong Learning*, Vol. 13, No. 1–2, 2003 (Tables 5 and 8).

a. Is there sufficient evidence of a difference between the percentage of BE students and percentage of BTech students who withdraw from engineering mathematics? Test using $\alpha = .05$.
b. Is there sufficient evidence of a difference between the percentage of BE students and percentage of BTech students who withdraw from engineering graphics/CAD? Test using $\alpha = .05$.

## 8.11 Testing a Population Variance

In this section we consider a hypothesis test for a population variance, $\sigma^2$ (e.g., the variation in daily amount of rainfall). Recall from Section 7.9 that the pivotal statistic for estimating a population variance $\sigma^2$ does not possess a standard normal ($Z$) distribution. Therefore, we cannot apply the procedure outlined in Section 8.3 when testing hypotheses about $\sigma^2$.

When the sample is selected from a normal population, however, the pivotal statistic possesses a chi-square ($\chi^2$) distribution and the test can be conducted as outlined in the box. Note that the assumption of normality is required regardless of whether the sample size $n$ is large or small.

### Test of Hypothesis About a Population Variance $\sigma^2$

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\sigma^2 = \sigma_0^2$ | $H_0$: $\sigma^2 = \sigma_0^2$ |
| $H_a$: $\sigma^2 > \sigma_0^2$ | $H_a$: $\sigma^2 \neq \sigma_0^2$ |
| [or $H_a$: $\sigma^2 < \sigma_0^2$ | |

$$\textit{Test statistic}: \quad \chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

*Rejection region:*

$$\chi^2 > \chi^2_\alpha \quad (\text{or } \chi^2 < \chi^2_{1-\alpha})$$

$$p\text{-value} = P(\chi^2 > \chi^2_c)$$
$$[\text{or}, P(\chi^2 < \chi^2_c)]$$

*Rejection region:*

$$\chi^2 < \chi^2_{1-\alpha/2} \text{ or } \chi^2 > \chi^2_{\alpha/2}$$

$$p\text{-value} = 2\min\{P(\chi^2 > \chi^2_c),$$
$$P(\chi^2 < \chi^2_c)\}$$

where $\chi^2_\alpha$ and $\chi^2_{1-\alpha}$ are values of $\chi^2$ that locate an area of $\alpha$ to the right and $\alpha$ to the left, respectively, of a chi-square distribution based on $(n-1)$ degrees of freedom, and $\chi^2_c$ is the calculated value of the test statistic.

(*Note:* $\sigma_0^2$ is our symbol for the particular numerical value specified for $\sigma^2$ in the null hypothesis.)

*Assumption:* The population from which the random sample is selected has an approximately normal distribution.

## Example 8.17

**Testing $\sigma^2$: Variation in Fill Measurements**

Refer to Example 7.15 (p. 337) concerning the variability of the amount of fill at a cannery. Suppose regulatory agencies specify that the standard deviation of the amount of fill should be less than .1 ounce. The quality control supervisor sampled $n = 10$ cans and measured the amount of fill in each. The data are reproduced in Table 8.8. Does this information provide sufficient evidence to indicate that the standard deviation $\sigma$ of the fill measurements is less than .1 ounce?

FILLWTS

**TABLE 8.8 Fill Weights of Cans**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7.96 | 7.90 | 7.98 | 8.01 | 7.97 | 7.96 | 8.03 | 8.02 | 8.04 | 8.02 |

**Solution** Since the null and alternative hypotheses must be stated in terms of $\sigma^2$ (rather than $\sigma$), we will want to test the null hypothesis that $\sigma^2 = .01$ against the alternative that $\sigma^2 < .01$. Therefore, the elements of the test are

$$H_0\text{:}\quad \sigma^2 = .01 \quad (\text{i.e., } \sigma = .1)$$
$$H_a\text{:}\quad \sigma^2 < .01 \quad (\text{i.e., } \sigma < .1)$$

*Assumption:* The populaton of fill amounts is approximately normal.

*Test statistic:* $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

*Rejection region:* The smaller the value of $s^2$ we observe, the stronger the evidence in favor of $H_a$. Thus, we reject $H_0$ for "small values" of the test statistic. With $\alpha = .05$ and 9 df, the $\chi^2$ value for rejection is found in Table 8 of Appendix B and pictured in Figure 8.22. We will reject

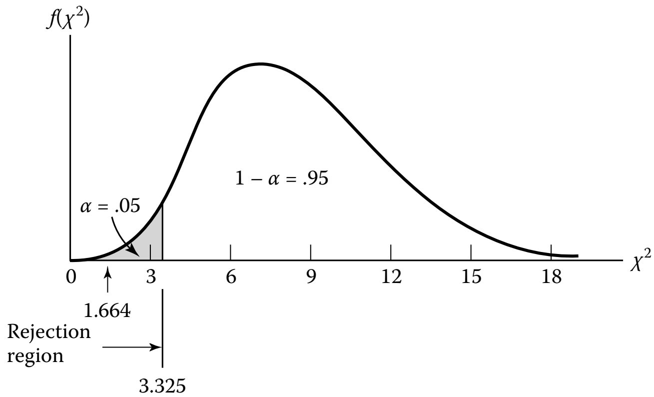



**FIGURE 8.22** Rejection region for Example 8.17

$H_0$ if $\chi^2 < 3.32511$. (Remember that the area given in Table 9 of Appendix B is the area to the *right* of the numerical value in the table. Thus, to determine the lower-tail value that has $\alpha = .05$ to its *left*, we use the $\chi^2_{.95}$ column in Table 8.)

To compute the test statistic, we need to find the sample standard deviation, $s$. Numerical descriptive statistics for the sample data are provided in the MINITAB printout shown in Figure 8.23. The value of $s$ (shaded on the printout) is $s = .043$. Substituting $s = .043$, $n = 10$, and $\sigma_0^2 = .01$ into the formula for the test statistic, we obtain

$$\chi^2 = \frac{(10 - 1)(.043)^2}{.01} = 1.67$$

Note that this test statistic and the corresponding $p$-value of the test (.004) are both given (shaded) at the bottom of the MINITAB printout, Figure 8.23.

*Conclusion*: Since the test statistic, $\chi^2 = 1.67$, is less than 3.32511 (or, since $\alpha = .05 > p\text{-value} = .004$), the supervisor can conclude (at $\alpha = .05$) that the variance of the population of all amounts of fill is less than .01 ($\sigma < .1$). If this procedure is repeatedly used, it will incorrectly reject $H_0$ only 5% of the time. Thus, the quality control supervisor is confident in the decision that the cannery is operating within the desired limits of variability.

**FIGURE 8.23**
MINITAB Test of a Population Variance, Example 8.17

**Test and CI for One Variance: WEIGHT**

```
Method

Null hypothesis          Sigma-squared = 0.01
Alternative hypothesis   Sigma-squared < 0.01

The chi-square method is only for the normal distribution.
The Bonett method is for any continuous distribution.


Statistics

Variable   N   StDev  Variance
WEIGHT    10  0.0431   0.00185


95% One-Sided Confidence Intervals

                       Upper Bound   Upper Bound
Variable  Method         for StDev  for Variance
WEIGHT    Chi-Square        0.0708       0.00502
          Bonett            0.0784       0.00614


Tests

                             Test
Variable  Method        Statistic  DF  P-Value
WEIGHT    Chi-Square         1.67   9    0.004
          Bonett                –   –    0.015
```

## Applied Exercises

8.72 *Characteristics of a rock fall.* Refer to the *Environmental Geology* (Vol. 58, 2009) simulation study of how far a block from a collapsing rock wall will bounce down a soil slope, Exercise 2.29 (p. 43). Rebound lengths (in meters) were estimated for 13 rock bounces. The data are repeated in the table. Descriptive statistics for the rebound lengths are shown on the accompanying SAS printout. Consider a test of hypothesis for the variation in rebound lengths for the theoretical population of rock bounces from the collapsing rock wall. In particular, a geologist wants to determine if the variance differs from 10 m$^2$.

**ROCKFALL**

| | | | | | | |
|---|---|---|---|---|---|---|
| 10.94 | 13.71 | 11.38 | 7.26 | 17.83 | 11.92 | |
| 11.87 | 5.44 | 13.35 | 4.90 | 5.85 | 5.10 | 6.77 |

*Source:* Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy", *Environmental Geology,* Vol. 58, 2009. (Table 2.)

a. Define the parameter of interest.
b. Specify the null and alternative hypothesis.
c. Compute the value of the test statistic.
d. Determine the rejection region for the test using $\alpha = .10$.
e. Make the appropriate conclusion.
f. What condition must be satisfied in order for the inference, part e, to be valid?

**PONDICE**

8.73 *Albedo of ice meltponds.* Refer to the National Snow and Ice Data Center (NSIDC) collection of data on the albedo of ice meltponds, Exercise 7.80 (p. 340). The visible albedo values for a sample of 504 ice meltponds located in the Canadian Arctic are saved in the **PONDICE** file.

a. Conduct a test (at $\alpha = .10$) to determine if the true variance of the visible albedo values of all Canadian Arctic ice ponds differs from .0225. (*Note:* For 503 df, $\chi^2_{.95} = 451.991$ and $\chi^2_{.05} = 556.283$.)
b. Discuss the practical significance of the test in part **a.** (*Hint:* Use the 90% confidence interval you found in Exercise 7.80.)

8.74 *Oil content of fried sweet potato chips.* Refer to the *Journal of Food Engineering* (Sep., 2013) study of the characteristics of sweet potato chips fried at different temperatures, Exercise 7.75 (p. 338). Recall that a sample of 6 sweet potato slices were fried at 130º using a vacuum fryer and the internal oil content (gigagrams) was measured for each slice. The results were: $\bar{y} = .178$ g/g and $s = .011$ g/g.

a. Conduct a test of hypothesis to determine if the standard deviation, $\sigma$, of the population of internal oil contents for sweet potato slices fried at 130º differs from .1. Use $\alpha = .05$
b. In Exercise 7.75 you formed a 95% confidence interval for the true standard deviation of the internal oil content distribution for the sweet potato chips. Use this interval to make an inference about whether $\sigma = .1$. Does the result agree with the test, part a?

8.75 *Strand bond performance of pre-stressed concrete.* An experiment was carried out to investigate the strength of pre-stressed, bonded concrete after anchorage failure has occurred and the results published in *Engineering Structures* (June 2013). The maximum strand force, measured in kiloNewtons (kN), achieved after anchorage failure for 8 pre-stressed concrete strands is given in the accompanying table. Conduct a test of hypothesis to determine if the true standard deviation of the population of maximum strand forces is less than 5 kN. Test using $\alpha = .10$

**FORCE**

| | | | | | |
|---|---|---|---|---|---|
| 158.2 | 161.5 | 166.5 | 158.4 | 159.9 | 161.9 |
| 162.8 | 161.2 | 160.1 | 175.6 | 168.8 | 163.7 |

8.76 *Deep-hole drilling.* Refer to the *Journal for Engineering for Industry* (May 1993) study of deep hole drilling under drill chip congestion, Exercise 8.32 (p. 393). Test to determine whether the true standard deviation of drill chip lengths differs from 75 mm. Recall that for $n = 50$ drill chips, $s = 50.2$.

8.77 *Electrical signal theory.* Recording electrical activity of the brain is important in clinical problems as well as in neurophysiological research. To improve the signal-to-noise ratio (SNR) in the electrical activity, it is necessary to repeatedly stimulate subjects and average the responses—a procedure that assumes that single responses are homogeneous. A study was conducted to test the homogeneous signal theory (*IEEE Engineering in Medicine and Biology*

### SAS Output for Exercise 8.72

The MEANS Procedure

Analysis Variable : LENGTH

| N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|
| 13 | 9.7169231 | 4.0947291 | 4.9000000 | 17.8300000 |

*Magazine*, Mar. 1990). The null hypothesis is that the variance of the SNR readings of subjects equals the "expected" level under the homogeneous signal theory. For this study, the "expected" level was assumed to be .54. If the SNR variance exceeds this level, the researchers will conclude that the signals are nonhomogeneous.

a. Set up the null and alternative hypotheses for the researchers.
b. SNRs recorded for a sample of 41 normal children ranged from .03 to 3.0. Use this information to obtain an estimate of the sample standard deviation. (*Hint:* Assume that the distribution of SNRs is normal and that most of the SNRs in the population will fall within $\mu \pm 2\sigma$, i.e., from $\mu - 2\sigma$ to $\mu + 2\sigma$. Note that the range of the interval equals $4\sigma$.)
c. Use the estimate of $s$ in part **b** to conduct the test of part **a.** Test using $\alpha = .10$.

8.78 *Measuring PCBs.* Polychlorinated biphenyls (PCBs), used in the manufacture of large electrical transformers and capacitors, are extremely hazardous contaminants when released into the environment. The Environmental Protection Agency (EPA) is experimenting with a new device for measuring PCB concentration in fish. To check the precision of the new instrument, seven PCB readings were taken on the same fish sample. The data are recorded here (in parts per million):

**PCBFISH**

| | | | | | | |
|---|---|---|---|---|---|---|
| 6.2 | 5.8 | 5.7 | 6.3 | 5.9 | 5.8 | 6.0 |

Suppose the EPA requires an instrument that yields PCB readings with a variance of less than .1. Does the new instrument meet the EPA's specifications? Test at $\alpha = .05$.

8.79 *Rubber cement canning.* A company produces a fast-drying rubber cement in 32-ounce aluminum cans. A quality control inspector is interested in testing whether the variance of the amount of rubber cement dispensed into the cans is more than .3. If so, the dispensing machine is in need of adjustment. Since inspection of the canning process requires that the dispensing machines be shut down, and shutdowns for any lengthy period of time cost the company thousands of dollars in lost revenue, the inspector is able to obtain a random sample of only 10 cans for testing. After measuring the weights of their contents, the inspector computes the following summary statistics:

$$\bar{y} = 31.55 \text{ ounces} \qquad s = .48 \text{ ounce}$$

a. Does the sample evidence indicate that the dispensing machines are in need of adjustment? Test at significance level $\alpha = .05$.
b. What assumption is necessary for the hypothesis test of part **a** to be valid?

## 8.12 Testing the Ratio of Two Population Variances

As in the one-sample case, the pivotal statistic for comparing two population variances, $\sigma_1^2$ and $\sigma_2^2$, has a nonnormal sampling distribution. Recall from Section 7.10 that the ratio of the sample variances $s_1^2/s_2^2$ possesses, under certain conditions, an $F$ distribution.

The elements of the hypothesis test for the ratio of two population variances, $\sigma_1^2/\sigma_2^2$, are given in the box.

### Test of Hypothesis for the Ratio of Two Population Variances $\sigma_1^2/\sigma_2^2$: Independent Samples

*One-Tailed Test*

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a: \frac{\sigma_1^2}{\sigma_2^2} > 1$$

$$\left[\text{or, } H_a: \frac{\sigma_1^2}{\sigma_2^2} < 1\right]$$

*Two-Tailed Test*

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

| | |
|---|---|
| *Test statistic:* $F = \frac{s_1^2}{s_2^2}\left[\text{or, } F = \frac{s_2^2}{s_1^2}\right]$ | *Test statistic:* $F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}} = \begin{cases} \frac{s_1^2}{s_2^2} & \text{when } s_1^2 > s_2^2 \\ \frac{s_2^2}{s_1^2} & \text{when } s_2^2 > s_1^2 \end{cases}$ |
| *Rejection region:* $F > F_\alpha$; $p\text{-value} = P(F > F_c)$ | *Rejection region:* $F > F_{\alpha/2}$; $p\text{-value} = 2 \cdot P(F > F_c)$ |

where $F_\alpha$ and $F_{\alpha/2}$ are values that locate area $\alpha$ and $\alpha/2$, respectively, in the upper tail of the $F$ distribution with $\nu_1$ = numerator degrees of freedom (i.e., the df for the sample variance in the numerator) and $\nu_2$ = denominator degrees of freedom (i.e., the df for the sample variance in the denominator) and $F_c$ is the computed value of the test statistic.

*Assumptions:*
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The random samples are selected in an independent manner from the two populations.

## Example 8.18

A Test to Compare Variances: Hospital Sterilization

Heavy doses of ethylene oxide (ETO) in rabbits have been shown to alter significantly the DNA structure of cells. Although it is a known mutagen and suspected carcinogen, ETO is used quite frequently in sterilizing hospital supplies. A study was conducted to investigate the effect of ETO on hospital personnel involved with the sterilization process. Thirty-one subjects were randomly selected and assigned to one of two tasks. Thirteen subjects were assigned the task of opening and unloading a sterilizer gun filled with ETO (task 1). The remaining 18 subjects were assigned the task of opening a sterilization package containing ETO (task 2). After the tasks were performed, researchers measured the amount of ETO (in milligrams) present in the bloodstream of each subject. A summary of the results appears in Table 8.9. Do the data provide sufficient evidence to indicate a difference in the variability of the ETO levels in subjects assigned to the two tasks? Test using $\alpha = .10$.

**TABLE 8.9 Summary Data for Example 8.18**

| | Task 1 | Task 2 |
|---|---|---|
| **Sample Size** | 13 | 18 |
| **Mean** | 5.60 | 5.90 |
| **Standard Deviation** | 3.10 | 1.93 |

Solution

Let

$\sigma_1^2$ = Population variance of ETO levels in subjects assigned task 1
$\sigma_2^2$ = Population variance of ETO levels in subjects assigned task 2

For this test to yield valid results, we must assume that both samples of ETO levels come from normal populations and that the samples are independent.

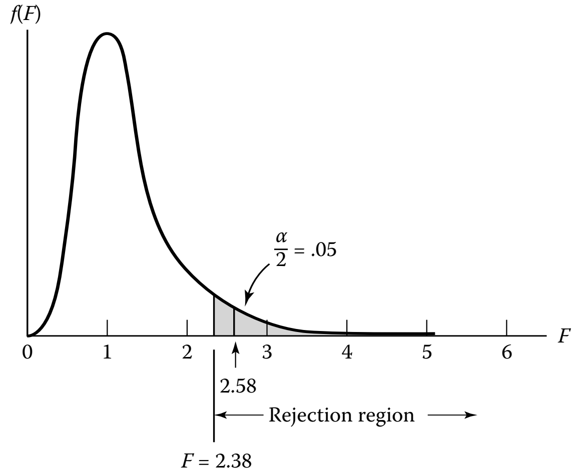
The hypotheses of interest are, then,

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad (\sigma_1^2 = \sigma_2^2)$$

$$H_a\colon \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \qquad (\sigma_1^2 \neq \sigma_2^2)$$

The nature of the $F$ tables given in Appendix B affects the form of the test statistic. To form the rejection region for a two-tailed $F$ test, we want to make certain that the upper tail is used, because only the upper-tail values of $F$ are shown in Tables 9–12 of Appendix B. To accomplish this, **we will always place the larger sample variance in the numerator of the $F$ test statistic**. This has the effect of doubling the tabulated

**FIGURE 8.24**
Rejection region for Example 8.18




value for $\alpha$, since we double the probability that the $F$ ratio will fall in the upper tail by always placing the larger sample variance in the numerator. That is, we make the test two-tailed by putting the larger variance in the numerator rather than establishing rejection regions in both tails.

Thus, for our example, we have a numerator $s_1^2$ with df $= n_1 - 1 = 12$ and a denominator $s_2^2$ with df $= n_2 - 1 = 17$. Therefore, the test statistic will be

$$F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}} = \frac{s_1^2}{s_2^2}$$

and we will reject $H_0$: $\sigma_1^2 = \sigma_2^2$ for $\alpha = .10$ when the calculated value of $F$ exceeds the tabulated value:

$$F_{\alpha/2} = F_{.05} = 2.38$$

We can now calculate the value of the test statistic and complete the analysis:

$$F = \frac{s_1^2}{s_2^2} = \frac{(3.10)^2}{(1.93)^2} = \frac{9.61}{3.72} = 2.58$$

When we compare this to the rejection region shown in Figure 8.24, we see that $F = 2.58$ falls in the rejection region. Therefore, the data provide sufficient evidence to indicate that the population variances differ. It appears that hospital personnel involved with opening the sterilization package (task 2) have less variable ETO levels than those involved with opening and unloading the sterilizer gun (task 1).
[*Note:* You can also use the $p$-value of the test to make the appropriate conclusion. The $p$-value for this two-tailed $F$ test is shown (shaded) on the MINITAB printout, Figure 8.25. Since $p$-value $= .073$ is less than $\alpha = .10$, there is sufficient evidence to reject $H_0$.]

**FIGURE 8.25**
MINITAB printout for Example 8.18

**Test for Equal Variances**

```
90% Bonferroni confidence intervals for standard deviations

Sample   N     Lower     StDev     Upper
     1  13   2.22297   3.10000   5.11728
     2  18   1.44730   1.92873   2.89144


F-Test (normal distribution)
Test statistic = 2.58, p-value = 0.073
```

What would you have concluded in Example 8.18 if the value of $F$ calculated from the samples had not fallen in the rejection region? Would you conclude that the null hypothesis of equal variances is true? No, because then you risk the possibility of a Type II error (failing to reject $H_0$ if $H_a$ is true) without knowing the value of $\beta$, the probability of failing to reject $H_0$: $\sigma_1^2 = \sigma_2^2$ if in fact it is false. Since we will not consider the calculation of $\beta$ for specific alternatives, when the $F$ statistic does not fall in the rejection region, we simply conclude that insufficient sample evidence exists to refute the null hypothesis that $\sigma_1^2 = \sigma_2^2$.

Example 8.18 illustrates the technique for calculating the test statistic and rejection region for a two-tailed test to avoid the problem of locating an $F$ value in the lower tail of the $F$ distribution. In a one-tailed test this is much easier to accomplish since we can control how we specify the ratio of the population variances in $H_0$ and $H_a$. That is, we can always make a one-tailed test an *upper-tailed* test. For example, if we want to test whether $\sigma_1^2$ is greater than $\sigma_2^2$, then we write the alternative hypothesis as

$$H_a: \ \frac{\sigma_1^2}{\sigma_2^2} > 1 \qquad (\text{i.e., } \sigma_1^2 > \sigma_2^2)$$

and the appropriate test statistic is $F = s_1^2/s_2^2$. Conversely, if we want to test whether $\sigma_1^2$ is less than $\sigma_2^2$ (i.e, whether $\sigma_2^2$ is greater than $\sigma_1^2$), we write

$$H_a: \ \frac{\sigma_2^2}{\sigma_1^2} > 1 \qquad (\text{i.e., } \sigma_2^2 > \sigma_1^2)$$

and the corresponding test statistic is $F = s_2^2/s_1^2$.

## Applied Exercises

**DRUGCON**

8.80 *Drug content assessment.* Refer to Exercise 7.84 (p. 344) and the *Analytical Chemistry* (Dec. 15, 2009) study in which scientists used high-performance liquid chromatography to determine the amount of drug in a tablet. Recall that 25 tablets were produced at each of two different, independent sites. In Exercise 7.84 you used a 95% confidence interval to determine if the two sites produce drug concentrations with different variances. Now make the inference with a test of hypothesis at $\alpha = .05$. Use the information provided in the MINITAB printout on p. 424.

8.81 *Attributes of forest access roads.* Refer to the *International Journal of Forest Engineering* (July 1999) study of the attributes of forest access roads in Ireland, Exercise 7.110 (p. 363). Recall that the transient surface deflection (millimeters) was measured for independent random samples of 32 mineral subgrade access roads and 40 peat subgrade access roads. The results are reproduced in the accompanying table.

a. Compare the surface deflection variances of the two pavement types with a two-tailed test of hypothesis using $\alpha = .05$.

b. In Exercise 7.110, you used a 95% confidence interval to compare the surface deflection variances. Demonstrate that the inferences derived from the test and confidence interval are identical. Will this always be the case? Explain.

| | Pavement Subgrade | |
|---|---|---|
| | Mineral | Peat |
| *Number of Roads* | 32 | 40 |
| *Mean Surface Deflection (mm)* | 1.53 | 3.80 |
| *Standard Deviation* | 3.39 | 14.3 |

*Source:* Martin, A. M., et al. "Estimation of the serviceability of forest access roads." *International Journal of Forest Engineering*, Vol. 10, No. 2, July 1999 (adapted from Table 3).

8.82 *Hippo grazing patterns in Kenya.* Refer to the *Landscape & Ecology Engineering* (Jan., 2013) study of hippopotamus grazing patterns in Kenya, Exercise 7.85 (p. 344). Recall that plots of land were sampled in two areas—a national reserve and a pastoral ranch—and the number of hippo trails from a water source was determined for each plot. Sample statistics are reproduced in the table on p. 424. In Exercise 7.85 you found a 90% confidence interval for $\sigma_1^2/\sigma_2^2$, the ratio of the variances associated with the two areas, and used it to determine if the variability in number of hippo trails from a water source in the national reserve differs from the variability in number of hippo trails from a water source in the pastoral ranch. Explain why a test of hypothesis at $\alpha = .10$ will result in the same inference, then carry out the test to verify your results.

**MINITAB Output for Exercise 8.80**

```
Test and CI for Two Variances: Content vs Site

Method

Null hypothesis         Variance(1) / Variance(2) = 1
Alternative hypothesis  Variance(1) / Variance(2) not = 1
Significance level      Alpha = 0.05


Statistics

Site   N  StDev  Variance
1     25  3.067     9.406
2     25  3.339    11.147

Ratio of standard deviations = 0.919
Ratio of variances = 0.844


95% Confidence Intervals

                                      CI for
Distribution    CI for StDev         Variance
of Data            Ratio               Ratio
Normal          (0.610, 1.384)  (0.372, 1.915)
Continuous      (0.497, 1.315)  (0.247, 1.729)


Tests

                                                 Test
Method                          DF1  DF2  Statistic  P-Value
F Test (normal)                  24   24       0.84    0.681
Levene's Test (any continuous)    1   48       0.64    0.427
```

**Table for Exercise 8.82**

| | National Reserve | Pastoral Ranch |
|---|---|---|
| Sample size: | 406 | 230 |
| Mean number of trails: | .31 | .13 |
| Standard deviation: | .40 | .30 |

*Source:* Kanga, E.M., et al. "Hippopotamus and livestock grazing: influences on riparian vegetation and facilitation of other herbivores in the Mara Region of Kenya", *Landscape & Ecology Engineering,* Vol. 9, No. 1, January 2013.

8.83 *Analyzing human inspection errors.* Tests of product quality using human inspectors can lead to serious inspection error problems (*Journal of Quality Technology*). To evaluate the performance of inspectors in a new company, a quality manager had a sample of 12 novice inspectors evaluate 200 finished products. The same 200 items were evaluated by 12 experienced inspectors. The quality of each item—whether defective or nondefective—was known to the manager. The next table lists the number of inspection errors (classifying a defective item as nondefective or vice versa) made by each inspector.

a. Prior to conducting this experiment, the manager believed the variance in inspection errors was lower for experienced inspectors than for novice inspectors. Do the sample data support her belief? Test using $\alpha = .05$.

b. What is the appropriate $p$-value of the test you conducted in part a?

**ERRORS**

| Novice Inspectors | | | | Experienced Inspectors | | | |
|---|---|---|---|---|---|---|---|
| 30 | 35 | 26 | 40 | 31 | 15 | 25 | 19 |
| 36 | 20 | 45 | 31 | 28 | 17 | 19 | 18 |
| 33 | 29 | 21 | 48 | 24 | 10 | 20 | 21 |

**GASTURBINE**

8.84 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of gas turbines augmented with high-pressure inlet fogging, Exercise 8.39 (p. 399). Heat rate data (kilojoules per kilowatt per hour) for each of three types of gas turbines (advanced, aeroderivative, traditional) are saved in the **GASTURBINE** file. In order to compare the mean heat rates of two types of gas turbines, you assumed that the heat rate variances were equal.

a. Conduct a test (at $\alpha = .05$) for equality of heat rate variances for traditional and aeroderivative augmented gas turbines. Use the result to make a statement about the validity of the inference derived in Exercise 8.33 **a.**

b. Conduct a test (at $\alpha = .05$) for equality of heat rate variances for advanced and aeroderivative augmented gas turbines. Use the result to make a statement about the validity of the inference derived in Exercise 8.39 **b**.

**ORCHARD**

8.85 *Insecticides used in orchards.* Refer to Exercise 8.44 (p. 401). Recall that an *Environmental Science & Technology* study was conducted to compare the mean oxon/thion ratios at a California orchard under two weather conditions—foggy and clear/cloudy. The data are saved in the **ORCHARD** file. Test the assumption of equal variances required for the comparison of means to be valid. Use $\alpha = .05$.

8.86 *Oil content of fried sweet potato chips.* Refer to the *Journal of Food Engineering* (Sep. 2013) study of the characteristics of fried sweet potato chips, Exercise 8.74 (p. 419). Recall that a sample of 6 sweet potato slices fried at 130º using a vacuum fryer yielded the following statistics on internal oil content (measured in gigagrams): $\bar{y}_1 = .178$ g/g and $s_1 = .011$ g/g. A second sample of 6 sweet potato slices was obtained, only these were subjected to a two-stage frying process (again, at 130º) in an attempt to improve texture and appearance. Summary statistics on internal oil content for this second sample follows: $\bar{y}_2 = .140$ g/g and $s_2 = .002$ g/g. The researchers want to compare the mean internal oil contents of sweet potato chips fried with the two methods using a $t$-test. Do you recommend the researchers carry out this analysis? Explain. (Recall your answer to Exercise 7.86.)

8.87 *Cracking torsion of T-beams.* An experiment was conducted to study the effect of reinforced flanges on the torsional capacity of reinforced concrete T-beams (*Journal of the American Concrete Institute*, Jan.–Feb. 1986). Several different types of T-beams were used in the experiment, each type having a different flange width. The beams were tested under combined torsion and bending until failure (cracking). One variable of interest is the cracking torsion moment at the top of the flange of the T-beam. Cracking torsion moments for eight beams with 70-cm slab widths and eight beams with 100-cm slab widths follow:

**TBEAMS**

| | |
|---|---|
| *70-cm Slab Width:* | 6.00, 7.20, 10.20, 13.20, 11.40, 13.60, 9.20, 11.20 |
| *100-cm Slab Width:* | 6.80, 9.20, 8.80, 13.20, 11.20, 14.90, 10.20, 11.80 |

a. Is there evidence of a difference in the variation in the cracking torsion moments of the two types of T-beams? Use $\alpha = .10$.

b. What assumptions are required for the test to be valid?

8.88 *Shopping vehicle and judgment.* Refer to the *Journal of Marketing Research* (Dec., 2011) study of shopping cart design, Exercise 8.41 (p. 400). Recall that design engineers want to know whether the mean choice of vice-over-virtue score is higher when a consumer's arm is flexed (as when carrying a shopping basket) than when the consumer's arm is extended (as when pushing a shopping cart). The average choice score for the $n_1 = 11$ consumers with a flexed arm was $\bar{y}_1 = 59$, while the average for the $n_2 = 11$ consumers with an extended arm was $\bar{y}_2 = 43$. In which scenario is the assumption required for a t-test to compare means more likely to be violated, $s_1 = 4$ and $s_2 = 2$, or, $s_1 = 10$ and $s_2 = 15$? Explain.

## Theoretical Exercises

8.89 Suppose we want to test $H_0$: $\sigma_1^2 = \sigma_2^2$ versus $H_a$: $\sigma_1^2 \neq \sigma_2^2$. Show that the rejection region given by

$$\frac{s_1^2}{s_2^2} > F_{\alpha/2} \quad \text{or} \quad \frac{s_1^2}{s_2^2} < F_{(1-\alpha/2)}$$

where $F$ depends on $\nu_1 = (n_1 - 1)$ df and $\nu_2 = (n_2 - 1)$ df, is equivalent to the rejection region given by

$$\frac{s_1^2}{s_2^2} > F_{\alpha/2}$$

where $F$ depends on $\nu_1$ numerator df and $\nu_2$ denominator df, or

$$\frac{s_2^2}{s_1^2} > F^*_{\alpha/2}$$

where $F^*$ depends on $\nu_2$ numerator df and $\nu_1$ denominator df. [*Hint:* Use the fact (proof omitted) that

$$F_{(1-\alpha/2)} = \frac{1}{F^*_{\alpha/2}}$$

where $F$ depends on $\nu_1$ numerator df and $\nu_2$ denominator df and $F^*$ depends on $\nu_2$ numerator df and $\nu_1$ denominator df.]

8.90 Use the results of Exercise 8.89 to show that

$$P\left(\frac{\text{Larger sample variance}}{\text{Smaller sample variance}} > F_{\alpha/2}\right) = \alpha$$

where $F$ depends on numerator df $= [(\text{Sample size for numerator sample variance}) - 1]$ and denominator df $= [(\text{Sample size for denominator sample variance}) - 1]$. [*Hint:* First write

$$P\left(\frac{\text{Larger sample variance}}{\text{Smaller sample variance}} > F_{\alpha/2}\right)$$

$$= P\left(\frac{s_1^2}{s_2^2} > F_{\alpha/2} \quad \text{or} \quad \frac{s_2^2}{s_1^2} > F_{\alpha/2}\right)$$

Then use the fact that $P(F > F_{\alpha/2}) = \alpha/2$.]

## 8.13 Alternative Testing Procedures: Bootstrapping and Bayesian Methods (*Optional*)

In optional Section 7.14, we introduced two alternative methods for finding confidence intervals: the *bootstrapping* method and a *Bayesian* method. These procedures can also be used to conduct a statistical test of hypothesis. In certain sampling situations, the conclusions drawn from one or both of these methods may be more valid than those produced using the classical tests of Sections 8.4–8.12, especially when the data do not adhere to the underlying assumptions.

### Bootstrap Hypothesis Tests

Recall that the bootstrap is a Monte Carlo method that involves *resampling*—that is, taking repeated samples of size $n$ (with replacement) from the original sample data set. The bootstrap testing procedure uses resampling to find an approximation for the observed significance level ($p$-value) of the test. The steps required to obtain the bootstrap $p$-value estimate for a test on a population mean are listed in the box.

**Bootstrap $p$-Value for Testing a Population Mean, $H_0$: $\mu = \mu_0$**

Let $y_1, y_2, y_3, \ldots, y_n$ represent a random sample of size $n$ from a population with mean $E(Y) = \mu$.

*Step 1* Calculate the value of the test statistic for the sample: $t_c = (\bar{y} - \mu_0)/(s/\sqrt{n})$ where $\bar{y}$ is the sample mean and $s$ is the sample standard deviation.

*Step 2* Select $j$, where $j$ is the number of times you will resample. (Usually, $j$ is a very large number, say, $j = 1{,}000$ or $j = 3{,}000$.)

*Step 3* Transform each of the sample $y$ values as follows: $x_i = y_i - \bar{y} + \mu_0$. That is, take each sample $y$ value, subtract the sample mean, then add $\mu_0$. (This step will generate sample values with a mean equal to the hypothesized mean in $H_0$.)

*Step 4* Randomly sample, with replacement, $n$ values of $X$ from the transformed sample data set $x_1, x_2, x_3, \ldots, x_n$.

*Step 5* Repeat step 4 a total of $j$ times.

*Step 6* For each bootstrap sample, compute the test statistic: $t_j = (\bar{x}_j - \mu_0)/(s_j/\sqrt{n})$, where $\bar{x}_j$ and $s_j$ are the mean and standard deviation, respectively, of bootstrap sample $j$.

*Step 7* Find the bootstrap estimated $p$-value—called the **achieved significance level (ASL)**—as follows:

Upper-tailed test ($H_a$: $\mu > \mu_0$): ASL = (Number of times $t_j > t_c$)/$j$

Lower-tailed test ($H_a$: $\mu < \mu_0$): ASL = (Number of times $t_j < t_c$)/$j$

Two-tailed test ($H_a$: $\mu \neq \mu_0$):

$$\text{ASL} = \frac{(\text{Number of times } t_j > |t_c|) + (\text{Number of times } t_j < -|t_c|)}{j}$$

The bootstrap ASL in step 7 is based on the definition of a $p$-value given in Section 8.6 (Definition 8.4): The $p$-value is the probability of observing a value of the test statistic that is more contradictory to $H_0$ than the value calculated in the sample. In the case of an upper-tailed test, more contradictory to $H_0$ implies a test statistic value that is greater than the calculated value in the sample. We illustrate the bootstrap procedure in the next example.

**Example 8.19**

Bootstrap Test for $\mu$: Benzene Contamination

**BENZENE**

Refer to Example 8.11 and the investigation of benzene contamination at a steel manufacturing plant. The benzene level (parts per million) was determined for each in a random sample of 20 air samples. (The data are saved in the **BENZENE** file.) Recall that the OSHA wants to test $H_0$: $\mu = 1$ against $H_a$: $\mu > 1$. Find the bootstrap ASL for this upper-tailed test. Make the appropriate conclusion using $\alpha = .05$.

Solution

To find the bootstrap ASL, we follow the steps outlined above.

*Step 1* From Example 8.11, the calculated value of the test statistic is $t_c = 2.95$.

*Step 2* We chose $j = 1{,}000$ for resampling.

*Step 3* Now $\bar{y} = 2.14$ (see Example 8.11) and $\mu_0 = 1$. Thus, we transform each of the 20 sampled benzene levels as follows: $x_i = y_i - \bar{y} + \mu_0 = y_i - 2.14 + 1$. The original sample data and the transformed values are shown in the MINITAB worksheet, Figure 8.26

*Steps 4–5* SAS was programmed to generate 1,000 random samples of size $n = 20$ (selecting observations with replacement) from the transformed sample data in Figure 8.26. The data for the first three resamples are shown in Table 8.10.

**FIGURE 8.26**
MINITAB worksheet with transformed benzene levels

| ↓ | C1 | C2 | C3 |
|---|---|---|---|
| | SAMPLE | BENZENE | TRANSFORM |
| 1 | 1 | 0.21 | -0.93 |
| 2 | 2 | 1.44 | 0.30 |
| 3 | 3 | 2.54 | 1.40 |
| 4 | 4 | 2.97 | 1.83 |
| 5 | 5 | 0.00 | -1.14 |
| 6 | 6 | 3.91 | 2.77 |
| 7 | 7 | 2.24 | 1.10 |
| 8 | 8 | 2.41 | 1.27 |
| 9 | 9 | 4.50 | 3.36 |
| 10 | 10 | 0.15 | -0.99 |
| 11 | 11 | 0.30 | -0.84 |
| 12 | 12 | 0.36 | -0.78 |
| 13 | 13 | 4.50 | 3.36 |
| 14 | 14 | 5.03 | 3.89 |
| 15 | 15 | 0.00 | -1.14 |
| 16 | 16 | 2.89 | 1.75 |
| 17 | 17 | 4.71 | 3.57 |
| 18 | 18 | 0.85 | -0.29 |
| 19 | 19 | 2.60 | 1.46 |
| 20 | 20 | 1.26 | 0.12 |

**TABLE 8.10 Bootstrap Resampling from Transformed Data in Figure 8.26 (First 3 Samples)**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Sample* 1: | -1.14 | 3.89 | 1.75 | 0.12 | 0.12 | 3.36 | 3.57 | 1.46 | 3.57 | 1.4 |
| | 0.3 | 1.1 | -1.14 | 1.83 | 3.89 | -0.29 | 0.12 | 1.4 | -1.14 | 3.36 |
| *Sample* 2: | 0.12 | 1.75 | -0.93 | 1.4 | 3.36 | 0.3 | -0.29 | -0.99 | -0.29 | -0.93 |
| | 0.3 | -0.93 | -1.14 | -0.84 | 3.36 | -0.29 | 1.4 | 1.27 | 1.1 | 1.46 |
| *Sample* 3: | 3.57 | 0.3 | -0.93 | 0.12 | 1.1 | 1.75 | 3.57 | 3.36 | 1.27 | 1.27 |
| | -0.78 | -1.14 | 1.83 | -0.78 | 1.46 | 1.27 | 2.77 | -0.29 | -0.29 | 3.57 |

*Step 6* Next, we used SAS to obtain the mean and standard deviation for each of the 1,000 samples. Then, we programmed SAS to compute the test statistic from these values as follows: $t_j = (\bar{x}_j - 1)/(s_j/\sqrt{20}),\ j = 1, 2, 3, \ldots, 1{,}000.)$

*Step 7* Each of the $t$ values in step 6 was compared to the calculated test statistic, $t_c = 2.95$. Only three $t$ values (those associated with samples 126, 962, and 966) exceeded 2.95. Consequently, the bootstrap ASL value is ASL $= 3/1{,}000 = .003$.

The bootstrap-achieved significance level provides an estimate of the true $p$-value of the test. (*Note:* The $p$-value obtained in Example 8.11 was .004.) Since $\alpha = .05$ exceeds the ASL value, we have sufficient evidence to reject the null hypothesis and to conclude that $\mu > 1$.

The general procedure for obtaining a bootstrap $p$-value for a test on any population parameter $\theta$ is beyond the scope of this text. Consult the references if you wish to learn about these methods. The procedure for testing a difference between two means, $(\mu_1 - \mu_2)$, however, is very similar to the procedure for a single mean, $\mu$. We list the steps in the box.

### Bootstrap $p$-Value for Testing Equality of Population Means, $H_0$: $(\mu_1 - \mu_2) = 0$

Let $\bar{y}_1$ and $s_1$ represent the mean and standard deviation of a random sample of size $n_1$ from a population with mean $\mu_1$. Let $\bar{y}_2$ and $s_2$ represent the mean and standard deviation of a random sample of size $n_2$ from a population with mean $\mu_2$.

*Step 1* Calculate the value of the test statistic for the sample,

$$t_c = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

*Step 2* Select $j$, where $j$ is the number of times you will resample.

*Step 3* Find the mean $\bar{\bar{y}}$ of the combined samples, then transform each of the sample values as follows:

$$\text{Sample 1: } x_i = y_i - \bar{y}_1 + \bar{\bar{y}} \qquad \text{Sample 2: } x_i = y_i - \bar{y}_2 + \bar{\bar{y}}$$

(That is, take each sample value, subtract its sample mean, then add $\bar{\bar{y}}$.)

*Step 4* Randomly sample, with replacement, $n_1$ transformed values from the first sample. Randomly sample, with replacement, $n_2$ transformed values from the second sample.

*Step 5* Repeat step 4 a total of $j$ times.

*Step 6* For each bootstrap sample, compute the test statistic:

$$t_j = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

where $\bar{x}_1$ and $s_1$ are the mean and standard deviation, respectively, of bootstrap sample $j$ for sample 1, and $\bar{x}_2$ and $s_2$ are the mean and standard deviation, respectively, of bootstrap sample $j$ for sample 2.

*Step 7* Find the bootstrap estimated $p$-value—called the **achieved significance level (ASL)**—as follows:

Upper-tailed test ($H_a$: $\mu_1 - \mu_2 > 0$): ASL = (Number of times $t_j > t_c$)/$j$
Lower-tailed test ($H_a$: $\mu_1 - \mu_2 < 0$): ASL = (Number of times $t_j < t_c$)/$j$
Two-tailed test ($H_a$: $\mu_1 - \mu_2 \neq 0$):

$$\text{ASL} = \frac{(\text{Number of times } t_j > |t_c|) + (\text{Number of times } t_j < -|t_c|)}{j}$$

## Bayesian Testing Procedures

Let $y_1, y_2, y_3, \ldots, y_n$ represent a random sample of size $n$ selected from a population with unknown population parameter $\theta$. The Bayesian approach to testing a hypothesis about $\theta$ considers $\theta$ as a random variable with a known *prior distribution*, $h(\theta)$. As with interval estimation, we need to find the *posterior distribution*, $g(\theta | y_1, y_2, y_3, \ldots, y_n)$. As shown in optional Section 7.14, the posterior distribution is

$$g(\theta | y_1, y_2, y_3, \ldots, y_n) = \frac{f(y_1, y_2, y_3, \ldots, y_n | \theta) \cdot h(\theta)}{f(y_1, y_2, y_3, \ldots, y_n)}$$

where $f(y_1, y_2, y_3, \ldots, y_n) = \int f(y_1, y_2, y_3, \ldots, y_n | \theta) \cdot h(\theta)\, d\theta$

Suppose you want to test $H_0$: $\theta \leq \theta_0$ versus $H_a$: $\theta > \theta_0$. The simplest Bayesian test uses the posterior distribution $g(\theta | y_1, y_2, y_3, \ldots, y_n)$ to find the following conditional probabilities:

$$P(\theta \leq \theta_0 | y_1, y_2, y_3, \ldots, y_n) \quad \text{and} \quad P(\theta > \theta_0 | y_1, y_2, y_3, \ldots, y_n)$$

In other words, the posterior distribution is used to find the likelihoods of $H_0$ and $H_a$ occurring. A simple rule is to accept the hypothesis that is associated with the largest conditional probability. That is,

*Accept* $H_0$ if $P(\theta \leq \theta_0 | y_1, y_2, y_3, \ldots, y_n) \geq P(\theta > \theta_0 | y_1, y_2, y_3, \ldots, y_n)$

*Reject* $H_0$ (i.e., *Accept* $H_a$ if $P(\theta \leq \theta_0 | y_1, y_2, y_3, \ldots, y_n) < P(\theta > \theta_0 | y_1, y_2, y_3, \ldots, y_n)$

We illustrate the Bayesian testing method in the next example.

**Example 8.20**

Bayesian Test of $\mu$

Consider a random sample of size 20 selected from a Bernoulli probability distribution with unknown probability of success $p$. The data (measured as zeros and ones) are shown in Table 8.11. Assume that the prior distribution for $p$ is a beta probability distribution with parameters $\alpha = 1$ and $\beta = 2$. Use the sum of the Bernoulli values to conduct a Bayesian test of $H_0$: $p \leq .5$ versus $H_a$: $p > .5$.

**TABLE 8.11 Sample of 20 Values from a Bernoulli Distribution**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |

Solution We know from Example 7.20 (p. 307), that $X$, the sum of the Bernoulli random variables, has a binomial distribution with $n = 20$ and probability of success $p$. We also know that $p$ has a prior beta distribution with $\alpha = 1$ and $\beta = 2$. In Example 7.20, we showed that the posterior distribution of $p$, $g(p|x)$, has a beta distribution with parameters $\alpha = (X + 1)$ and $\beta = (n - X + 2)$. Summing the sample Bernoulli values in Table 8.11, we obtain $X = 15$. Therefore, the posterior distribution of $p$ is a beta distribution with $\alpha = (X + 1) = 16$ and $\beta = (n - X + 2) = 7$.

Since the null and alternative hypotheses are $H_0$: $p \leq .5$ and $H_a$: $p > .5$, we need to find the conditional probabilities, $P(p \leq .5|X = 15)$ and $P(p > .5|X = 15)$. Most statistical software packages have routines for computing probabilities for a wide variety of probability distributions. We use MINITAB to find $P(p \leq .5)$ for a beta distribution with $\alpha = 16$ and $\beta = 7$. The result (highlighted) is shown in Figure 8.27. You can see that $P(p \leq .5|X = 15) = .026$. Hence, $P(p > .5|X = 15) = 1 - .026 = .974$. Since the conditional probability associated with $H_a$: $p > .5$ is larger, we reject $H_0$ in favor of $H_a$ and conclude that the probability of success, $p$, exceeds .5.

**Cumulative Distribution Function**

```
Beta with first shape parameter = 16 and second = 7

  x   P( X <= x )
0.5     0.0262394
```

**FIGURE 8.27**
MINITAB calculation of $P(p \leq .5)$ using beta $(\alpha = 16, \beta = 7)$ probability function

Another approach to Bayesian testing is to use the posterior distribution to find a $(1 - \alpha)$*100% credible interval* for the parameter being tested. (See optional Section 7.14.) For example, a 90% credible interval for the probability of success $p$ in Example 8.20 is $P(\mathrm{L} < p < \mathrm{U}) = .90$, where L and U are the 5th and 95th percentiles of a beta ($\alpha = 16$, $\beta = 7$) distribution. Using the inverse beta function of MINITAB, we find that the 90% credible interval for $p$ is (.53, .84). Note the interval does not contain the null hypothesized value of .5. All values of $p$ in the credible interval exceed .5, supporting the alternative hypothesis.

## Applied Exercises

8.91 *Bearing strength of concrete FRP strips.* Refer to the *Composites Fabrication Magazine* (Sept. 2004) study of the strength of fiber-reinforced polymer (FRP) composite materials, Exercise 7.100 (p. 354). Recall that 10 specimens of pultruded FRP strips were mechanically fastened to highway bridges and tested for bearing strength. The strength measurements (recorded in mega pascal units, MPa) are reproduced in the table. Use the bootstrap procedure to test (at $\alpha = .10$) whether the true mean strength of mechanically fastened FRP strips exceeds 230 MPa.

**FRP**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 240.9 | 248.8 | 215.7 | 233.6 | 231.4 | 230.9 | 225.3 | 247.3 | 235.5 | 238.0 |

*Source:* Data are simulated from summary information provided in *Composites Fabrication Magazine*, Sept. 2004, p. 32 (Table 1).

8.92 *Surface roughness of pipe.* Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated interior pipe used in oil fields, Exercise 8.24 (p. 390). The data (in micrometers) for

20 sampled pipe sections are reproduced in the table. Use the bootstrap procedure to test (at $\alpha = .05$) whether the mean surface roughness of coated interior pipe, $\mu$, differs from 2 micrometers. Compare the bootstrap ASL to the $p$-value obtained from the test in Exercise 8.24.

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

8.93 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of three types of gas turbines augmented with high-pressure inlet fogging, Exercise 8.39 (p. 399). Heat rate data (kilojoules per kilowatt per hour) for advanced and aeroderivative gas turbines are shown in the table. Use the bootstrap procedure to test (at $\alpha = .05$) for a difference between the mean heat rates of advanced augmented gas turbines and aeroderivative augmented gas turbines. Compare the bootstrap ASL to the $p$-value obtained from the test in Exercise 8.39**b**.

**GASTURBINE**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Advanced:* | 9722 | 10481 | 9812 | 9669 | 9643 | 9115 | 9115 | 11588 |
| | 10888 | 9738 | 9295 | 9421 | 9105 | 10233 | 10186 | 9918 |
| | 9209 | 9532 | 9933 | 9152 | 9295 | | | |
| *Aeroderivative:* | 16243 | 14628 | 12766 | 8714 | 9469 | 11948 | 12414 | |

8.94 *Plant investment per delivered quad.* Refer to Example 8.13 (p. 382) and the comparison of electric and gas utility plants. The data on plant investment per delivered quad for 11 plants using electrical utilities and 16 plants using gas utilities are reproduced in the next table. Use the bootstrap procedure (at $\alpha = .05$) to test for a difference in the average investment/quad between all plants using gas and all those using electric utilities.

**INVQUAD**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***Electric:*** | 204.15 | 0.57 | 62.76 | 89.72 | 0.35 | 85.46 | | |
| | 0.78 | 0.65 | 44.38 | 9.28 | 78.60 | | | |
| ***Gas:*** | 0.78 | 16.66 | 74.94 | 0.01 | 0.54 | 23.59 | 88.79 | 0.64 |
| | 0.83 | 91.84 | 7.20 | 66.64 | 0.74 | 64.67 | 165.60 | 0.36 |

8.95 *Study of lunar soil.* Refer to the *Meteoritics* (Mar. 1995) study of lunar soil evolution, Exercise 8.62 (p. 411). Recall that one theory is that the proportion $p$ of lunar soil grains that are coated with dust and/or glass fragments will be less than .5 at the bottom of the lunar core soil sample. Assuming that the prior distribution for $p$ is a beta probability distribution with parameters $\alpha = 1$ and $\beta = 2$, conduct a Bayesian test of the hypothesis of interest. (*Note:* From Exercise 8.62, 29 of 81 grains sampled from the bottom of the lunar core were coated.)

## Theoretical Exercises

8.96 Let $y_1, y_2, y_3, \dots, y_n$ represent a random sample of size $n$ selected from a Poisson probability distribution with unknown mean $\lambda$. Let $X$ represent the sum of the Poisson values, $X = \sum y_i$. Then $X$ has a Poisson distribution with mean $n\lambda$. Assume that the prior distribution for $\lambda$ is an exponential probability distribution with parameter $\beta$. Find a Bayesian decision rule for testing $H_0$: $\lambda = \lambda_0$ versus $H_a$: $\lambda > \lambda_0$. [*Hint:* Use the posterior distribution, $g(\lambda|x)$, found in Exercise 7.104 (p. 355).]

8.97 Let $y_1, y_2, y_3, \dots, y_n$ represent a random sample of size $n$ selected from a normal probability distribution with unknown mean $\mu$ and variance $\sigma^2 = 1$. Then the sample mean, $\bar{y}$, has a normal distribution with mean $\mu$ and variance $\sigma^2 = 1/n$. Assume that the prior distribution for $\mu$ is a normal distribution with a mean of 5 and a variance of 1. Find a Bayesian decision rule for testing $H_0$: $\mu = \mu_0$ versus $H_a$: $\mu < \mu_0$. [*Hint:* Use the posterior distribution, $g(\mu|\bar{y})$, found in Exercise 7.105 (p. 355).]

## *STATISTICS IN ACTION REVISITED*
### Comparing Methods for Dissolving Drug Tablets—Dissolution Method Equivalence Testing

We now return to the drug assay problem outlined in the *Statistics in Action* application discussed at the beginning of this chapter (p. 369). Recall that a pharmaceutical company first measures the dissolution of a new drug in a Research and Development (R&D) laboratory by quantifying how much of the drug is contained in a dissolving solution; this value is expressed as percent of label strength (%LS). The process is then repeated at a manufacturing facility. Federal regulations require that quality engineers at the manufacturing site produce results equivalent to those at any other site (including the R&D lab).

Dissolution test data for an analgesic in tablet form conducted at two manufacturing sites (New Jersey and Puerto Rico) were listed in Table SIA8.1 (p. 433) and are saved in the **DISSOLVE** file. Recall that %LS values were obtained at four different points in time – after 20 minutes, after 40 minutes, after 60 minutes, and after 120 minutes – for each of the six test vessels. Based on the sample data, do the two sites produce equivalent assay results?

An initially appealing approach to answering this question is to conduct a test of hypothesis on the difference between the mean %LS measurements at the two sites. Let $\mu_1$ represent the population mean %LS for tests conducted at the New Jersey site and let $\mu_2$ represent the population mean %LS for tests conducted at the Puerto Rico site. If the test results at the two sites are equivalent, then $\mu_1 = \mu_2$. The null and alternative hypotheses can be stated:

$H_0$: $(\mu_1 - \mu_2) = 0$ (i.e., dissolution equivalence)

$H_a$: $(\mu_1 - \mu_2) \neq 0$ (i.e., non equivalence)

To simplify the analysis, the statisticians suggested conducting this test at each of the four time periods separately. The above test was conducted using SAS at each time point, with the results shown in Figure SIA8.1 on the next two pages. The *p*-values for the two-tailed tests (highlighted on the printout) for 20, 40, 60, and 120 minutes of dissolving time are .1528, .0395, .3499, and .4956, respectively. If we select a Type I error rate of $\alpha = .05$, then we fail to reject $H_0$ (*p*-value $> .05$) for three of the four time points; only when dissolving time is set at 40 minutes is there sufficient evidence to conclude that the mean %LS values for the two sites differ. In other words, one might reasonably conclude from the hypothesis tests that the two sites produce equivalent results at dissolving times of 20, 60, and 120 minutes, but *do not* produce equivalent results at a dissolving time of 40 minutes.

There are several caveats to this hypothesis testing approach, as the statisticians warned in their chapter, “Dissolution Method Equivalence”. First, the idea of equivalence in the above test is established by “accepting $H_0$”. Recall that a measure of reliability for the conclusion “accept $H_0$” is $\beta$ = P(Type II error) = P(Accept $H_0$ | $H_0$ is false). For this application, $\beta$ is the probability of saying $\mu_1 = \mu_2$, when, in fact, the means differ. Since the sampling distribution of $\mu_1 - \mu_2$ is unknown when the alternative condition, $\mu_1 \neq \mu_2$, is true, the exact value of $\beta$ is unknown. Second, the notion of “practical significance” is ignored in the hypothesis test. That is, although the population means may be statistically different at $\alpha = .05$, the true difference may be small and not considered a meaningful difference in practice. Finally, the test above may have the unfortunate effect of penalizing a testing site with a small (smaller than average) %LS variance. You can see this by examining the formula for the test statistic in Section 8.7. When the difference in sample means is divided by a small standard error (which will likely occur if one site has a small variance), the resulting *t*-value will be large (and likely to be significant).

To overcome these problems, pharmaceutical companies have developed alternative approaches to the equivalence problem. One method, suggested by the statisticians in their chapter, requires that you first find a 90% confidence interval for $\mu_1 - \mu_2$. If the confidence interval for the difference between mean %LS values lies within equivalence limits established by the company, then accept the assays of the two sites as being equivalent. The company in this application uses the equivalence limits in Table SIA8.2. Note that the limits depend on the magnitude of the mean %LS.

Using the equivalence limits of Table SIA8.2, we will accept the assays of the two sites as being equivalent if the 90% confidence interval for $\mu_1 - \mu_2$ : (a) lies between $-15$ and 15 when the mean %LS is less than 90, or (b) lies between $-7$ and 7 when the mean %LS is greater than or equal to 90. Note that this approach is equivalent to testing the following hypotheses (for those assays with mean $<$90%) :

$H_0$: $(\mu_1 - \mu_2) < -15$ or $(\mu_1 - \mu_2) > 15$ (i.e., nonequivalence)

$H_a$: $-15 < (\mu_1 - \mu_2) < 15$ (i.e., dissolution equivalence)

For this reason, this methodology is referred to as the **two one-sided t-test (TOST)**.

**TABLE SIA8.2 Determining Dissolution Equivalence**

| If Mean %LS is | Dissolution Equivalence Occurs If Mean Difference Is Between: |
|---|---|
| <90% | −15% and 15% |
| ≥90% | −7% and 7% |

```
----------------------- TIME = 20 -----------------------

Sample Statistics

     Group           N      Mean     Std. Dev.   Std. Error
     ----------------------------------------------------
     New_Jersey      6         5       3.5777       1.4606
     Puerto_Rico     6       8.5       4.2308       1.7272

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are    t statistic     Df       Pr > t
     ----------------------------------------------------
     Equal                  -1.547         10      0.1528
     Not Equal              -1.547       9.73      0.1537

90% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                  -7.60           0.60


----------------------- TIME = 40 -----------------------

Sample Statistics

     Group           N      Mean     Std. Dev.   Std. Error
     ----------------------------------------------------
     New_Jersey      6      74.5       4.4159       1.8028
     Puerto_Rico     6  69.16667       3.3116        1.352

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are    t statistic     Df       Pr > t
     ----------------------------------------------------
     Equal                   2.367         10      0.0395
     Not Equal               2.367       9.27      0.0414

90% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                   1.25           9.42
```

**FIGURE SIA8.1**
SAS Dissolution Equivalence Hypothesis Tests

```
------------------------ TIME = 60 ------------------------


Sample Statistics

     Group           N       Mean     Std. Dev.   Std. Error
     -------------------------------------------------------
     New_Jersey      6   96.33333       2.3381       0.9545
     Puerto_Rico     6   94.66667       3.4448       1.4063

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are     t statistic      Df        Pr > t
     ----------------------------------------------------
     Equal                      0.981        10       0.3499
     Not Equal                  0.981      8.80       0.3530

90% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                  -1.41           4.75



----------------------- TIME = 120 -----------------------


Sample Statistics

     Group           N       Mean     Std. Dev.   Std. Error
     -------------------------------------------------------
     New_Jersey      6   98.66667       1.5055       0.6146
     Puerto_Rico     6   99.33333       1.7512       0.7149

Hypothesis Test

     Null hypothesis:    Mean 1 - Mean 2 =  0
     Alternative:        Mean 1 - Mean 2 ^= 0

     If Variances Are     t statistic      Df        Pr > t
     ----------------------------------------------------
     Equal                     -0.707        10       0.4956
     Not Equal                 -0.707      9.78       0.4960

90% Confidence Interval for the Difference between Two Means

               Lower Limit    Upper Limit
               -----------    -----------
                  -2.38           1.04
```

**FIGURE SIA8.1**
SAS Dissolution Equivalence Hypothesis Tests (Continued)

To apply TOST to the data of Table SIA8.1, we find the 90% confidence intervals for $\mu_1 - \mu_2$. These confidence intervals, as well as the mean %LS values, are also shaded in the SAS printout, Figure SIA8.1. The confidence intervals for each of the four time points are all within their respective equivalence limits (i.e., between −15 and 15 for time points 20 and 40 minutes, and between −7 and 7 for time points 60 and 120 minutes). Consequently, the data support dissolution assay equivalence between the two sites for all four dissolution times.

TOST is now considered the standard method for bioequivalence testing of pharmaceutical products and is becoming widely accepted in process engineering, chemistry and environmental science. An excellent tutorial on TOST is given in "Beyond the t-Test: Statistical Equivalence Testing", *Analytical Chemistry* (June 1, 2005). There, the authors provide insight into TOST sample size determination and on how to choose the all-important equivalence limits.

# Quick Review

## Key Terms

*Note: Starred (*) terms are from the optional section in this chapter.*



## Key Formulas

### Summary of Hypothesis Tests: One-Sample Case

| Parameter ($\theta$) | Null Hypothesis ($H_0$) | Point Estimator ($\hat{\theta}$) | Test Statistic | Sample Size | Additional Assumptions |
|---|---|---|---|---|---|
| $\mu$ | $\mu = \mu_0$ | $\bar{y}$ | $Z = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \approx \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ | $n \geq 30$ | None 386 |
| | | | $T = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ where $T$ is based on $\nu = (n - 1)$ df | $n < 30$ | Normal population 388 |
| $p$ | $p = p_0$ | $\hat{p} = \dfrac{y}{n}$ | $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$ | $n$ large enough so that $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$ | None 408 |
| $\sigma^2$ | $\sigma^2 = \sigma_0^2$ | $s^2$ | $\chi^2 = \dfrac{(n - 1)s^2}{\sigma_0^2}$ where $\chi^2$ is based on $\nu = (n - 1)$ d.f. | All $n$ | Normal population 416 |

## Summary of Hypothesis Tests: Two-Sample Case

| Parameter ($\theta$) | Null Hypothesis ($H_0$) | Point Estimator ($\hat{\theta}$) | Test Statistic | Sample Size | Additional Assumptions | |
|---|---|---|---|---|---|---|
| $(\mu_1 - \mu_2)$<br>Independent samples | $(\mu_1 - \mu_2) = D_0$<br>(If we want to detect a difference between $\mu_1$ and $\mu_2$, then $D_0 = 0$.) | $(\bar{y} - \bar{y}_2)$ | $Z = \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \approx \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ | $n_1 \geq 30$, $n_2 \geq 30$ | None | 395 |
| | | | $T = \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{s_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$<br>where $T$ is based on $\nu = n_1 + n_2 - 2$ df and<br>$s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ | Either $n_1 < 30$ or $n_2 < 30$ or both | Both populations normal with equal variances ($\sigma_1^2 = \sigma_2^2$) (For situations in which $\sigma_1^2 \neq \sigma_2^2$, see the modifications listed in the box on p. 000.) | 395 |
| $\mu_d =$ $(\mu_1 - \mu_2)$<br>Matched pairs | $\mu_d = D_0$ (If we want to detect a difference between $\mu_1$ and $\mu_2$, then $D_0 = 0$.) | $\bar{d} = \sum_{i=1}^{n} d_i/n$<br>Mean of sample differences | $T = \dfrac{\bar{d} - D_0}{s_d/\sqrt{n_d}}$<br>where $T$ is based on $\nu = (n_d - 1)$ df | All $n_d$ (If $n_d \geq 30$, then the standard normal ($z$) test may be used.) | Population of differences $d_i$ is normal | 402 |
| $(p_1 - p_2)$ | $(p_1 - p_2) = D_0$<br>(If we want to detect a difference between $p_1$ and $p_2$, then $D_0 = 0$.) | $(\hat{p}_1 - \hat{p}_2)$ | For $D_0 = 0$:<br>$Z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$<br>where $\hat{p} = \dfrac{y_1 + y_2}{n_1 + n_2}$<br>For $D_0 \neq 0$:<br>$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}}$ | $n_1$ and $n_2$ large enough so that $n_1\hat{p}_1 \geq 4$, $n_1\hat{q}_1 \geq 4$ and $n_2\hat{p}_2 \geq 4$, $n_2\hat{q}_2 \geq 4$ | Independent samples | 412 |
| $\dfrac{\sigma_1^2}{\sigma_2^2}$ | $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$<br>(i.e., $\sigma_1^2 = \sigma_2^2$) | $\dfrac{s_1^2}{s_2^2}$ | For $H_a$: $\sigma_1^2 > \sigma_2^2$: $F = \dfrac{s_1^2}{s_2^2}$<br>For $H_a$: $\sigma_2^2 > \sigma_1^2$: $F = \dfrac{s_2^2}{s_1^2}$<br>For $H_a$: $\sigma_1^2 \neq \sigma_2^2$:<br>$F = \dfrac{\text{Larger } s^2}{\text{Smaller } s^2}$<br>where $F$ is based on $\nu_1 =$ numerator df and $\nu_2 =$ denominator df | All $n_1$ and $n_2$ | Independent random samples from normal populations | 420 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $H_0$ | h−oh | Null hypothesis |
| $H_a$ | h−a | Alternative hypothesis |
| $\alpha$ | alpha | Probability of Type I error |
| $\beta$ | beta | Probability of Type II error |
| $\theta_0$ | theta naught | Hypothesized value of population parameter in $H_0$ |
| $\mu_0$ | mu naught | Hypothesized value of population mean in $H_0$ |
| $D_0$ | d naught | Hypothesized value of population difference in $H_0$ |
| $\sigma_0^2$ | sigma-squared naught | Hypothesized value of population variance in $H_0$ |

## Chapter Summary Notes

- Elements of a **test of hypothesis: null hypothesis, alternative hypothesis, test statistic, significance level ($\boldsymbol{\alpha}$), rejection region, *p*-value, and conclusion**.
- Two types of errors in a hypothesis test: **Type I error** (reject $H_0$ when $H_0$ is true), **Type II error** (accept $H_0$ when $H_0$ is false).
- Probabilities of errors: $\boldsymbol{\alpha} = \boldsymbol{P}(\textbf{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ true})$, $\boldsymbol{\beta} = \boldsymbol{P}(\textbf{Type II error}) = \text{P}(\text{Accept } H_0 | H_0 \text{ false})$.
- Three forms of the alternative hypothesis: **lower-tailed test (<), upper-tailed test (>), two-tailed test (≠)**.
- **Observed significance level (*p*-value)** is the smallest value of $\alpha$ that can be used to reject the null hypothesis.
- **Decision rule for rejecting $\boldsymbol{H_0}$**: (1) test statistic falls into rejection region, or (2) $p$-value $< \alpha$.
- **Power of the test** $= 1 - \beta = P(\text{Reject } H_0 | H_0 \text{ false})$.
- Key words for identifying $\boldsymbol{\mu}$ as the parameter of interest: *mean, average*.
- Key words/phrases for identifying $\boldsymbol{\mu_1 - \mu_2}$ as the parameter of interest: *difference between means or averages, compare two means using independent samples*.
- Key words/phrases for identifying $\boldsymbol{\mu_d}$ as the parameter of interest: *mean or average of paired differences, compare two means using matched pairs*.
- Key words for identifying $\boldsymbol{p}$ as the parameter of interest: *proportion, percentage, rate*.
- Key words/phrases for identifying $\boldsymbol{p_1 - p_2}$ as the parameter of interest: *difference between proportions or percentages, compare two proportions using independent samples*.
- ***Key words for identifying*** $\boldsymbol{\sigma^2}$ as the parameter of interest: *variance, spread, variation*.
- Key words/phrases for identifying $\boldsymbol{\sigma_1^2/\sigma_2^2}$ as the parameter of interest: *difference between variances, compare variation in two populations using independent samples*.

## Supplementary Exercises

8.98 *Mongolian desert ants.* The *Journal of Biogeography* (Dec. 2003) published a study of ants in Mongolia (Central Asia). Botanists placed seed baits at five sites in the Dry Steppe region and six sites in the Gobi Desert and observed the number of ant species attracted to each site. These data are listed in the table on p. 438. Is there evidence to conclude that a difference exists between the average number of ant species found at sites in the two regions of Mongolia? Draw the appropriate conclusion using $\alpha = .05$.

## Data for Exercise 8.98

**GOBIANTS**

| Site | Region | Number of Ant Species |
|---|---|---|
| 1 | Dry Steppe | 3 |
| 2 | Dry Steppe | 3 |
| 3 | Dry Steppe | 52 |
| 4 | Dry Steppe | 7 |
| 5 | Dry Steppe | 5 |
| 6 | Gobi Desert | 49 |
| 7 | Gobi Desert | 5 |
| 8 | Gobi Desert | 4 |
| 9 | Gobi Desert | 4 |
| 10 | Gobi Desert | 5 |
| 11 | Gobi Desert | 4 |

*Source:* Pfeiffer, M., et al. "Community organization and species richness of ants in Mongolia along an ecological gradient from steppe to Gobi desert." *Journal of Biogeography*, Vol. 30, No. 12, Dec. 2003.

8.99 *Mongolian desert ants (continued).* Refer to the *Journal of Biogeography* (Dec. 2003) study of ants in Mongolia (Central Asia), Exercise 8.98, where you compared the mean number of ants at two desert sites. Since the sample sizes were small, the variances of the populations at the two sites must be equal in order for the inference to be valid.

a. Set up $H_0$ and $H_a$ for determining whether the variances are the same.
b. Use the data in the **GOBIANTS** file to find the test statistic for the test.
c. Give the rejection region for the test if $\alpha = .05$.
d. Find the approximate $p$-value of the test.
e. Make the appropriate conclusion in the words of the problem.
f. What conditions are required for the test results to be valid?

8.100 *Coverage of fluid mechanics.* The *Journal of Professional Issues in Engineering Education and Practice* (Apr. 2005) reported on the results of a 2005 survey of courses offered at undergraduate engineering programs. Of the 90 engineering programs that participated in the 2005 survey, 68 covered fluid mechanics. In a survey taken 20 years earlier (*Engineering Education*, Apr. 1986), 66 of the 100 undergraduate engineering programs covered fluid mechanics. Conduct a test to determine whether the fraction of undergraduate engineering programs covering fluid mechanics increased from 1986 to 2005. Use $\alpha = .01$.

8.101 *General engineering program.* The *European Journal of Engineering Education* (Vol. 38, 2013) published a study of the feasibility of adding a general engineering program to a university's specialized engineering programs (e.g., civil, mechanical, electrical engineering). A preliminary study found that about half (50%) of engineering students responded favorably to a general engineering program. Let $Y$ represent the number of students in a sample of 10 who favor a general engineering program and let $p$ represent the true proportion of all students who favor a general engineering program. Suppose you want to test $H_0$: $p = .5$ against $H_a$: $p \neq .5$. One possible procedure is to reject $H_0$ if $Y \leq 1$ or $Y \geq 8$.

a. Find $\alpha$ for this test.
b. Find $\beta$ if $p = .4$. What is the power of the test?
c. Find $\beta$ if $p = .8$. What is the power of the test?

8.102 *Accuracy of wet samplers.* Wet samplers are standard devices used to measure the chemical composition of precipitation. The accuracy of the wet deposition readings, however, may depend on the number of samplers stationed in the field. Experimenters in The Netherlands collected wet deposition measurements using anywhere from one to eight identical wet samplers (*Atmospheric Environment*, Vol. 24A, 1990). For each sampler (or sampler combination), data were collected every 24 hours for an entire year; thus, 365 readings were collected per sampler (or sampler combination). When one wet sampler was used, the standard deviation of the hydrogen readings (measured as percentage relative to the average reading from all eight samplers) was 6.3%. When three wet samplers were used, the standard deviation of the hydrogen readings (measured as percentage relative to the average reading from all eight samplers) was 2.6%. Conduct a test to compare the variation in hydrogen readings for the two sampling schemes (i.e., one wet sampler versus three wet samplers). Test using $\alpha = .05$.

8.103 *Perceptions of automation problems.* According to a popular model of managerial behavior, the current state of automation in a manufacturing firm influences managers' perceptions of problems of automation. To investigate this proposition, researchers at Concordia University (Montreal) surveyed managers at firms with a high level of automation and at firms with a low level of automation (*IEEE Transactions on Engineering Management*, Aug. 1990). Each manager was asked to give his or her perception of the problems of automation at the firm. Responses were measured on a 5-point scale (1: No problem, . . . , 5: Major problem). Summary statistics for the two groups of managers, provided in the table, were used to test the hypothesis of no difference in the mean perceptions of automation problems between managers of highly automated and less automated manufacturing firms.

| | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| *Low Level* | 17 | 3.274 | .762 |
| *High Level* | 8 | 3.280 | .721 |

*Source:* Farhoomand, A. F., Kira D., and Williams, J. "Managers' perceptions towards automation in manufacturing." *IEEE Transactions on Engineering Management*, Vol. 37, No. 3, Aug, 1990, p. 230.

a. Conduct the test for the researchers, assuming that the perception variances for the two groups of managers are equal. Use $\alpha = .01$.
b. Conduct the test for the researchers, if it is known that the perception variances differ for managers at low-level and high-level firms.

8.104 *Real-time scheduling with robots.* Researchers at Purdue University compared human real-time scheduling in a processing environment to an automated approach that utilizes computerized robots and sensing devices (*IEEE Transactions*, Mar. 1993). The experiment consisted of eight simulated scheduling problems. Each task was performed by a human scheduler and by the automated system. Performance was measured by the *throughput rate*, defined as the number of good jobs produced weighted by product quality. The resulting throughput rates are shown in the accompanying table. Analyze the data using a test of hypothesis.

**THRUPUT**

| Task | Human Scheduler | Automated Method |
|---|---|---|
| 1 | 185.4 | 180.4 |
| 2 | 146.3 | 248.5 |
| 3 | 174.4 | 185.5 |
| 4 | 184.9 | 216.4 |
| 5 | 240.0 | 269.3 |
| 6 | 253.8 | 249.6 |
| 7 | 238.8 | 282.0 |
| 8 | 263.5 | 315.9 |

*Source:* Yih, Y., Liang, T., and Moskowitz, H. "Robot scheduling in a circuit board production line: A hybrid OR/ANN approach." *IEEE Transactions*, Vol. 25, No. 2, March 1993, p. 31 (Table 1).

8.105 *Radioactive water.* A problem that occurs with certain types of mining is that some by-products tend to be mildly radioactive and these products sometimes get into our fresh water supply. The EPA has issued regulations concerning a limit on the amount of radioactivity in supplies of drinking water. Particularly, the maximum level for naturally occurring radiation is 5 picocuries per liter of water. A random sample of 24 water specimens from a city's water supply produced the sample statistics $\bar{y} = 4.61$ picocuries per liter and $s = .87$ picocurie per liter.

a. Do these data provide sufficient evidence to indicate that the mean level of radiation is safe (below the maximum level set by the EPA)? Test using $\alpha = .01$.
b. Why should you want to use a small value of $\alpha$ for the test in part **a**?
c. Calculate the value of $\beta$ for the test if $\mu_a = 4.5$ picocuries per liter of water.
d. Calculate and interpret the $p$-value for the test.

8.106 *PhD's in engineering.* The National Science Foundation, in a survey of 2,237 engineering graduate students who earned their PhD degrees, found that 607 were U.S. citizens; the majority (1,630) of the PhD degrees were awarded to foreign nationals. Conduct a test to determine whether the true proportion of engineering PhD degrees awarded to foreign nationals exceeds .5. Use $\alpha = .01$.

**DDT**

8.107 *Contamination of fish.* Refer to the U.S. Army Corps of Engineers study of contaminated fish in the Tennessee River (Alabama).

a. Use a random number table (table 1 of Appendix B) to generate a random sample of $n = 40$ observations on DDT concentration in fish from the **DDT** file. Compute $\bar{y}$ and $s$ for the sample measurements.
b. The Food and Drug Administration (FDA) sets the limit for DDT content in individual fish at 5 parts per million (ppm). Does the sample of part **a** provide sufficient evidence to conclude that the average DDT content of individual fish inhabiting the Tennessee River and its creek tributaries exceeds 5 ppm? Test using a significance level of $\alpha = .01$.
c. Suppose the test of hypothesis, part **b**, was based on a random sample of only $n = 8$ fish. What are the disadvantages of conducting this small-sample test?
d. Repeat part **b** using only the information on the DDT contents of a sample of 8 fish (randomly selected from the 40 observations of part **a**). Compare the results of the large- and small-sample tests.

8.108 *Ball bearing specifications.* In the manufacture of machinery, it is essential to utilize parts that conform to specifications. In the past, diameters of the ball bearings produced by a certain manufacturer had a variance of .00156. To cut costs, the manufacturer instituted a less expensive production method. The variance of the diameters of 100 randomly sampled bearings produced by the new process was .00211. Do the data provide sufficient evidence to indicate that diameters of ball bearings produced by the new process are more variable than those produced by the old process? Test at $\alpha = .05$.

8.109 *Active versus passive solar heating.* Home solar heating systems can be categorized into two groups, *passive* solar heating systems and *active* solar heating systems. In a passive solar heating system, the house itself is a solar energy collector, whereas in an active solar heating system, elaborate mechanical equipment is used to convert the sun's rays into heat. Consider the difference between the proportions of passive solar and active solar heating systems that require less than 200 gallons of oil per year in fuel consumption. Independent random samples of 50 passive and 50 active solar-heated homes are selected and the numbers that required less than 200 gallons of oil last year are noted, with the results given in the table on the next page. Is there evidence of a difference between the proportions of passive and active solar-heated homes that required less than 200 gallons of oil in fuel consumption last year? Test at a level of significance of $\alpha = .02$.

**Table for Exercise 8.109**

| | Passive Solar | Active Solar |
|---|---|---|
| *Number of Homes* | 50 | 50 |
| *Number That Required Less Than 200 Gallons of Oil Last Year* | 37 | 46 |

8.110 *Cyanide contamination.* *Environmental Science & Technology* (Oct. 1993) reported on a study of contaminated soil in The Netherlands. A total of 72 400-gram soil specimens were sampled, dried, and analyzed for the contaminant cyanide. The cyanide concentration (milligrams per kilogram of soil) of each soil specimen was determined using an infrared microscopic method. The sample resulted in a mean cyanide level of $\bar{y} = 84$ mg/kg and a standard deviation of $s = 80$ mg/kg.

a. Test the hypothesis that the true mean cyanide level in soil in The Netherlands falls below 100 mg/kg. Use $\alpha = .10$.

b. Would you reach the same conclusion in part **a** using $\alpha = .05$? $\alpha = .01$? Explain.

8.111 *Organic carbon in sewage.* Engineers periodically analyze water samples for various types of organic material. The total organic carbon (TOC) level was measured in water samples collected at two sewage treatment sites in England. The accompanying table gives the summary information on the TOC levels (measured in mg/L) found in the rivers adjacent to the two sewage facilities. Since the river at the Foxcote sewage treatment works was subject to periodic spillovers, not far upstream of the plant's intake, it is believed that the TOC levels found at Foxcote will have greater variation than the levels at Bedford. Does the sample information support this hypothesis? Test at $\alpha = .05$.

| Bedford | Foxcote |
|---|---|
| $n_1 = 61$ | $n_2 = 52$ |
| $\bar{y}_1 = 5.35$ | $\bar{y}_2 = 4.27$ |
| $s_1 = .96$ | $s_2 = 1.27$ |

*Source:* Pinchin, M. J. "A study of the trace organics profiles of raw and potable water systems." *Journal of the Institute of Water Engineers & Scientists*, Vol. 40, No. 1, Feb. 1986, p. 87.

8.112 *Single-T swim maze.* Merck Research Labs conducted an experiment to evaluate the effect of a new drug using the Single-T swim maze. Nineteen impregnated dam rats were captured and allocated a dosage of 12.5 milligrams of the drug. One male and one female pup were randomly selected from each resulting litter to perform in the swim maze. Each rat pup is placed in water at one end of the maze and allowed to swim until it successfully escapes at the opposite end. If the rat pup fails to escape after a certain period of time, it is placed at the beginning end of the maze and given another attempt to escape. The experiment is repeated until three successful escapes are accomplished by each rat pup. The number of swims required by each pup to perform three successful escapes is reported in the table. Is there sufficient evidence (at $\alpha = .10$) of a difference between the mean number of swims required by male and female rat pups?

**RATPUPS**

| Litter | Male | Female | Litter | Male | Female |
|---|---|---|---|---|---|
| 1 | 8 | 5 | 11 | 6 | 5 |
| 2 | 8 | 4 | 12 | 6 | 3 |
| 3 | 6 | 7 | 13 | 12 | 5 |
| 4 | 6 | 3 | 14 | 3 | 8 |
| 5 | 6 | 5 | 15 | 3 | 4 |
| 6 | 6 | 3 | 16 | 8 | 12 |
| 7 | 3 | 8 | 17 | 3 | 6 |
| 8 | 5 | 10 | 18 | 6 | 4 |
| 9 | 4 | 4 | 19 | 9 | 5 |
| 10 | 4 | 4 | | | |

*Source:* Bradstreet, Thomas E. Merck Research Labs, BL 3-2, West Point, PA 19486.

8.113 *Solder joint inspections.* Current technology uses X-rays and lasers for inspection of solder-joint defects on printed circuit boards (PCBs). (*Quality Congress Transactions*, 1986.) A particular manufacturer of laser-based inspection equipment claims that its product can inspect on average at least 10 solder joints per second when the joints are spaced .1 inch apart. The equipment was tested by a potential buyer on 48 different PCBs. In each case, the equipment was operated for exactly 1 second. The number of solder joints inspected on each run follows:

**PCB**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 9 | 10 | 10 | 11 | 9 | 12 | 8 | 8 | 9 | 6 | 10 |
| 7 | 10 | 11 | 9 | 9 | 13 | 9 | 10 | 11 | 10 | 12 | 8 |
| 9 | 9 | 9 | 7 | 12 | 6 | 9 | 10 | 10 | 8 | 7 | 9 |
| 11 | 12 | 10 | 0 | 10 | 11 | 12 | 9 | 7 | 9 | 9 | 10 |

a. The potential buyer wants to know whether the sample data refute the manufacturer's claim. Specify the null and alternative hypotheses that the buyer should test.

b. In the context of this exercise, what is a Type I error? A Type II error?

c. Conduct the hypothesis test you described in part **a**, and interpret the test's results in the context of this exercise. Use $\alpha = .05$.

8.114 *Stacked menu displays.* One feature of a user-friendly computer interface is a stacked menu display. Each time a menu item is selected, a submenu is displayed partially over the parent menu, thus creating a series of "stacked" menus. The *Special Interest Group on Computer Human Interaction Bulletin* (July 1993) reported on a study to determine the effects of the presence or absence of a stacked menu structure on search time. Twenty-two subjects were randomly placed into one of two groups, and each was asked to search a menu-driven software package for a particular item. In the experimental group ($n_1 = 11$), the stacked menu format was used; in the control group ($n_2 = 11$), only the current menu was displayed.

a. The researcher's initial hypothesis is that the mean time required to find a target item does not differ for the two menu displays. Describe the statistical method appropriate for testing this hypothesis.
b. What assumptions are required for inferences derived from the analysis to be valid?
c. The mean search times for the two groups were 11.02 seconds and 11.07 seconds, respectively. Is this enough information to conduct the test? Explain.
d. The observed significance level for the test, part **a**, exceeds .10. Interpret this result.

8.115 *Performance of R&D.* Does competition between separate research and development (R&D) teams in the U.S. Department of Defense, working independently on the same project, improve performance? To answer this question, performance ratings were assigned to each of 58 multisource (competitive) and 63 sole-source R&D contracts (*IEEE Transactions on Engineering Management*, Feb. 1990). With respect to quality of reports and products, the competitive contracts had a mean performance rating of 7.62, whereas the sole-source contracts had a mean of 6.95.

a. Set up the null and alternative hypothesis for determining whether the mean quality performance rating of competitive R&D contracts exceeds the mean for sole-source contracts.
b. Find the rejection region for the test using $\alpha = .05$.
c. The $p$-value for the test was reported to be between .02 and .03. What is the appropriate conclusion?

8.116 *Strength of sewer pipe.* The building specifications in a certain city require that the sewer pipe used in residential areas have a mean breaking strength of more than 2,500 pounds per lineal foot. A manufacturer who would like to supply the city with sewer pipe has submitted a bid and provided the following additional information: An independent contractor randomly selected seven sections of the manufacturer's pipe and tested each for breaking strength. The results (pounds per lineal foot) follow:

**SEWER**

| | | | | | | |
|---|---|---|---|---|---|---|
| 2,610 | 2,750 | 2,420 | 2,510 | 2,540 | 2,490 | 2,680 |

a. Is there sufficient evidence to conclude that the manufacturer's sewer pipe meets the required specifications? Use a significance level of $\alpha = .10$.
b. Find the value of $\beta$ for $\mu_a = 2575$. What is the power of the test?
c. Find the value of $\beta$ for $\mu_a = 2800$.
d. Find the power of the test for $\mu_a = 2800$.

CHAPTER 9

# Categorical Data Analysis

**OBJECTIVE**

*To show how to analyze count data obtained by the classification of experimental observations from a multinomial experiment*

**CONTENTS**

## STATISTICS IN ACTION

### The Case of the Ghoulish Transplant Tissue – Who is Responsible for Paying Damages?

In the 1970s and 1980s, tissue engineers began working on growing replacement organs for transplantation into patients. Thirty-some years later, the worldwide tissue transplant market has grown into a big business. According to *Organ and Tissue Transplantation and Alternatives* (January 1, 2011), "the global market for transplantation products, devices, and pharmaceuticals was valued at nearly $54 billion in 2010 and is projected to grow at an 8.3% compound annual growth rate to reach $80 billion in 2015." Here in the United States, tissue implants are routinely performed to aid patients in various types of surgery, including joint replacements, spinal surgery, sports-related surgeries (tendons and ligaments), and others.

The process of obtaining a tissue transplant involves several parties. First, of course, is the donor who has agreed to have tissue removed upon death, and whose family has approved the donation. The tissue is then "harvested" by an approved tissue bank. Next, the harvested tissue is sent to a processor who sterilizes the tissue. Finally, the processor either sends it directly to the hospital/surgeon doing the implant, or to a distributor who inventories the tissue and ultimately sends it on to the hospital/surgeon. The entire process is highly regulated by the Federal Trade Commission (FTC), particularly the harvesting and processing aspects.

Given this background, we consider an actual case that began in the early 2000s when the owner of a tissue bank — Biomedical Tissue Services (BTS) — became a ringleader of a group of funeral home directors that harvested tissue illegally and without permission of donors or their families. In some cases the cadavers were cancerous, or infected with HIV or hepatitis, all of which would, of course, disqualify them as donors. BTS then sent the tissue to processors without divulging that it had obtained the tissue illegally. (Note: The owner is currently serving 18–24 years in a New York prison.) The unsuspecting processors sterilized the tissue and sent it on for use as surgical implants. When the news story broke about how the tissue had been obtained, the processors and their distributors were required to send recall notices to the hospital/surgeons who had received the tissue, using an FTC recall letter. Some of the BTS tissue was recovered; however, much of the tissue had already been implanted, and hospitals and surgeons were required to inform patients receiving implants of the potentially infectious tissue. Although few patients subsequently became infected, a number filed suit against the distributors and processors (and BTS) asking for monetary damages.

After the bulk of the lawsuits had been either tried or settled, a dispute arose between a processor and one of its distributors regarding ultimate responsibility for payment of damages to litigating patients. In particular, the processor claimed that the distributor should be held more responsible for the damages, since in its recall package the distributor had of its own volition included some salacious, inflammatory newspaper articles describing in graphic detail the "ghoulish" acts that had been committed. None of the patients who received implants that had been sterilized by this processor ever became infected, but many still filed suit.

To establish its case against the distributor, the processor collected data on the patients who had received implants of BTS tissue it had processed, and the number of those patients who subsequently filed suit: the data revealed that of a total of 7,914 patients, 708 filed suit. A consulting statistician sub-divided this information according to whether the recall notice had been sent to the patient's surgeon by the processor or one of its distributors that had sent only the notice, or by the distributor that included the newspaper articles. The breakdown was as shown in Table SIA9.1:

**TABLE SIA9.1 Data for the Tainted Tissue Case***

| Recall notice sender | Number of Patients | Number of lawsuits |
|---|---|---|
| Processor/Other Distributor | 1,751 | 51 |
| Distributor in question | 6,163 | 657 |
| Totals: | 7,914 | 708 |

*For confidentiality purposes, the parties in the case cannot be identified. Permission to use the data in this Statistics in Action has been granted by the consulting statistician.

Do these data provide evidence of a difference in the probability that a patient would file a lawsuit depending on which party sent the recall notice? If so, and if the probability is significantly higher for the distributor in question, then the processor can argue in court that the distributor who sent the inflammatory newspaper articles is more responsible for the damages.

We apply the statistical methodology presented in this chapter to solve the case of the ghoulish transplant tissue in the *Statistics in Action Revisited example* at the end of the chapter.

## 9.1 Categorical Data and Multinomial Probabilities

In Chapters 7 and 8, we discussed how to make inferences about a proportion from a single population. Recall that the population proportion $p$ is the probability of "success" in a binomial experiment—an experiment that results in one of two possible outcomes on any one trial. In this chapter, we are interested in making inferences about the unknown probabilities (or proportions) from a **multinomial experiment** with $k$ possible outcomes. That is, we want to make inferences about $p_1, p_2, \ldots, p_k$, where $p_i$ is the probability of the $i$th outcome and $p_1 + p_2 + \cdots + p_k = 1$. (See Section 4.7 for a detailed discussion of multinomial experiments.)

To illustrate, consider a motor fan blade company that manufactures impellers on one of five production lines, A, B, C, D, or E. Assume that the lines produce impellers at the same rate and volume. In a sample of $n = 103$ impellers found to be defective, 15 were manufactured on line A, 27 on line B, 31 on line C, 19 on line D, and 11 on line E (see Table 9.1). For this multinomial experiment, there are five outcomes, or categories, into which each defective can be classified, one corresponding to each of the five production lines.

**TABLE 9.1 Classification of $n = 103$ Defective Impellers According to Production Line**

IMPELLER

| *Production Line* | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| 15 | 27 | 31 | 19 | 11 |

The practical question to be answered in the study is whether the proportions of defective impellers differ among the five production lines. Do the data provide evidence to contradict the null hypothesis $H_0$: $p_1 = p_2 = \cdots = p_5$, where $p_i$ is the proportion of defectives manufactured on the $i$th production line? If the data in Table 9.1 contradict this hypothesis, the manufacturer would want to know why the rate of production of defectives is greater on some production lines than others and would take countermeasures to reduce the production of defectives.

This chapter is concerned with the analysis of categorical data—specifically, data that represent the counts for each category of a multinomial experiment. In Sections 9.2 and 9.3, we will learn how to make inferences about the category probabilities for data classified according to a single qualitative (or categorical) variable. In Sections 9.4 and 9.5, we consider inferences about the category probabilities for data classified according to two qualitative variables. The statistic used for most of these inferences is one that possesses, approximately, the familiar **chi-square distribution**. Although the proof of the adequacy of this approximation is beyond the scope of this text, some aspects of the theory can be deduced from what we have learned in earlier chapters.

## 9.2 Estimating Category Probabilities in a One-Way Table

Consider a multinomial experiment with $k$ outcomes that correspond to categories of a single qualitative variable. The data (i.e., category counts) for such an experiment would appear similar to that of Table 9.2, where $n_1, n_2, \ldots, n_k$, represent the category counts and $n = n_1 + n_2 + \cdots + n_k$. Such a table is often called a **one-way table** since only one qualitative variable is used to form the categories, or outcomes.

To estimate category probabilities in a one-way table, consider that a multinomial experiment can always be reduced to a binomial experiment by isolating one

**TABLE 9.2 One-Way Table of Category Counts**

| Category | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | ... | $k$ |
| $n_1$ | $n_2$ | $n_3$ | ... | $n_k$ |

category, say, category $i$, and then combining all others. Since we know that in a binomial experiment with number of successes, $Y$, $\hat{p} = Y/n$ is a good estimator of the binomial parameter $p$, it follows that

$$\hat{p}_i = \frac{n_i}{n}$$

is a good estimator of $p_i$, the probability associated with category $i$ in a multinomial experiment. It also follows that $\hat{p}_i$ will possess the same properties as $\hat{p}$—namely, that when $n$ is large, $\hat{p}_i$ will be approximately normally distributed (by the central limit theorem) with

$$E(\hat{p}_i) = p_i$$

and

$$V(\hat{p}_i) = \frac{p_i(1 - p_i)}{n}$$

Consequently, a large-sample confidence interval for $p_i$ may be constructed as shown in the box.

### A Large-Sample $(1 - \alpha)100\%$ Confidence Interval for $p_i$ in a One-Way Table

$$\hat{p}_i \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n}}$$

Values of $z_{\alpha/2}$ can be found in Table 5 of Appendix B.

We will estimate the difference between a pair of category probabilities, say, categories $i$ and $j$ $(i \neq j)$, using $(\hat{p}_i - \hat{p}_j)$. This linear function of $\hat{p}_i$ and $\hat{p}_j$ will be approximately normally distributed with

$$E(\hat{p}_i - \hat{p}_j) = p_i - p_j$$

and

$$V(\hat{p}_i - \hat{p}_j) = V(\hat{p}_i) + V(\hat{p}_j) - 2\,\text{Cov}(\hat{p}_i, \hat{p}_j)$$

Since the covariance of two category counts, say, $n_i$ and $n_j$ $(i \neq j)$, is given by

$$\text{Cov}(n_i, n_j) = -np_ip_j$$

(the proof is left as an exercise at the end of this section), it follows that the covariance between the corresponding estimators, $\hat{p}_i$ and $\hat{p}_j$, is

$$\begin{aligned}
\text{Cov}\,(\hat{p}_i, \hat{p}_j) &= E[(\hat{p}_i - p_i)(\hat{p}_j - p_j)] = E\left[\left(\frac{n_i}{n} - \frac{np_i}{n}\right)\left(\frac{n_j}{n} - \frac{np_j}{n}\right)\right] \\
&= E\left[\frac{1}{n^2}(n_i - np_i)(n_j - np_j)\right] = \frac{1}{n^2}E\left[(n_i - np_i)(n_j - np_j)\right] \\
&= \frac{1}{n^2}\text{Cov}(n_i, n_j) = \frac{1}{n^2}(-np_ip_j) \\
&= \frac{-p_ip_j}{n}
\end{aligned}$$

Therefore,

$$\begin{aligned}
V(\hat{p}_i - \hat{p}_j) &= V(\hat{p}_i) + V(\hat{p}_j) - 2\,\text{Cov}(\hat{p}_i, \hat{p}_j) \\
&= \frac{p_i(1 - p_i)}{n} + \frac{p_j(1 - p_j)}{n} + \frac{2p_ip_j}{n}
\end{aligned}$$

and a large-sample $(1 - \alpha)100\%$ confidence interval for $(p_i - p_j)$ is as indicated in the box.

**A Large-Sample $(1 - \alpha)100\%$ Confidence Interval for $(p_i - p_j)$ in a One-Way Table**

$$(\hat{p}_i - \hat{p}_j) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1 - \hat{p}_i) + \hat{p}_j(1 - \hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}}$$

Values of $z_{\alpha/2}$ can be found in Table 5 of Appendix B.

**Example 9.1**

Estimating a Proportion in a 1-way Table: Defective Impellers

Refer to Table 9.1 and find a 95% confidence interval for the proportion $p_1$ of all defective impellers that can be attributed to production line A. Note that $p_1$ is *not* the proportion of impellers produced by production line A that are defective. Rather, it is the proportion of all defective impellers that are produced by production line A.

Solution

From Table 9.1, we have $n_1 = 15$ and $n = 103$. Therefore, a 95% confidence interval for $p_1$ is

$$\hat{p}_1 \pm z_{.025}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n}}, \qquad \text{where } \hat{p}_1 = \frac{n_1}{n} = \frac{15}{103} = .146$$

$$= .146 \pm 1.96\sqrt{\frac{(.146)(.854)}{103}}$$

or $.146 \pm .068$. Therefore, our interval estimate for $p_1$ is from .078 to .214.

That is, we are 95% confident that the true proportion of all defective impellers that are produced on line A falls between .078 and .214.

**Example 9.2**

Estimating $(p_1 - p_2)$ in a 1-way Table: Defective Impellers

Refer to Example 9.1 and find a 95% confidence interval for $(p_1 - p_2)$, the difference between the proportions of defective impellers attributable to production lines A and B, respectively.

Solution

From Table 9.1, we have $n_2 = 27$ and $\hat{p}_2 = n_2/n = 27/103 = .262$. Then a 95% confidence interval for $(p_1 - p_2)$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2) + 2\hat{p}_1\hat{p}_2}{n}}$$

$$= (.146 - .262) \pm 1.96\sqrt{\frac{(.146)(.854) + (.262)(.738) + 2(.146)(.262)}{103}}$$

$$= -.116 \pm .121$$

Therefore, our interval estimate of $(p_1 - p_2)$, the difference in the proportions of the defective impellers attributable to production lines A and B, is $-.237$ to .005. Since this interval includes 0, there is insufficient evidence (at the 95% confidence level) to conclude that the two proportions differ.

## Applied Exercises

9.1 *Jaw dysfunction study.* A report on dental patients with temporomandibular (jaw) joint dysfunction (TMD) was published in *General Dentistry* (Jan/Feb. 2004). A random sample of 60 patients was selected for an experimental treatment of TMD. Prior to treatment, the patients filled out a survey on two nonfunctional jaw habits—bruxism (teeth grinding) and teeth clenching—that have been linked to TMD. Of the 60 patients, 3 admitted to bruxism, 11 admitted to teeth clenching, 30 admitted to both habits, and 16 claimed they had neither habit.

a. Describe the qualitative variable of interest in the study. Give the levels (categories) associated with the variable.
b. Construct a one-way table for the sample data.

c. Find and interpret a 95% confidence interval for the true proportion of dental patients who admit to both habits.
d. Find and interpret a 95% confidence interval for the difference between the true proportion of dental patients who admit to both habits and the true proportion of dental patients who claim they have neither habit.

**TYPSTYLE**

9.2 *Mobile device typing strategies.* Researchers estimate that in a typical month, about 75 billion text messages are sent in the U.S. Text messaging on mobile devices (e.g., cell phones, smart phones) often requires typing in awkward positions that may lead to health issues. A group of Temple University public health professors investigated this phenomenon and published their results in *Applied Ergonomics* (March 2012). One portion of the study focused on the typing styles of mobile device users. Typing style was categorized as (1) device held with both hands/ both thumbs typing, (2) device held with right hand/ right thumb typing, (3) device held with left hand/ left thumb typing, (4) device held with both hands/ right thumb typing, (5) device held with left hand/right index finger typing, or (6) other. In a sample of 859 college students observed typing on their mobile devices, the professors observed 396, 311, 70, 39, 18, and 25, respectively, in the six categories.

a. Construct a one-way table for the study.
b. Estimate the proportion of mobile device users who hold the device with one hand, using a 95% confidence interval. Interpret the results, practically.
c. Estimate the difference between the proportion of mobile device users who type with both thumbs and the proportion of mobile device users who type with the right thumb, using a 95% confidence interval. Interpret the results, practically.

9.3 *CAD technology.* Each month, *Mechanical Engineering* surveys its readers with an online "Question of the Month." One issue reported on the responses to the question, "Do you feel you know enough about the latest computer-aided design (CAD) technologies to do your job?" The results: 44% answered "yes," 12% answered "no, but I'm not worried about it," 35% answered "no, and it concerns me," and 9% answered "I don't need to know CAD in my job." Assume 1,000 readers responded to the online survey.

a. Find and interpret a 95% confidence interval for the proportion of readers who feel they know enough about CAD to do their job.
b. Find and interpret a 95% confidence interval for the difference between the proportions of readers who answered "no, but I'm not worried about it" and of those who answered "no, and it concerns me."

9.4 *Digital signal and image processing.* Digital signal and image processing (DSIP) has a wide variety of applications, including entertainment (video on demand), telemedicine, security/surveillance, military target recognition, wireless communications, and intelligent transportation systems. Consequently, there is a rapidly growing need for engineers trained in DSIP. The *International Journal of Electrical Engineering Education* (Apr. 2004) reported on an evaluation of the experimental DSIP undergraduate research curriculum at Western Michigan University (WMU). A sample of 50 students responded to the statement "I believe that this research experience is very valuable to my professional future." The results: 47 students agreed, 3 students were neutral, and 0 students disagreed with the statement.

a. Estimate the proportion of WMU students who agree that their DSIP research experience is valuable to their professional future. Use a 99% confidence interval.
b. Estimate the difference between the proportions of WMU students who agree and who are neutral about the statement. Use a 99% confidence interval.

**PONDICE**

9.5 *Characteristics of ice meltponds.* Refer to the National Snow and Ice Data Center (NSIDC) collection of data on 504 ice meltponds in the Canadian Arctic, Example 2.1 (p. 16). The data are saved in the **PONDICE** file. One variable of interest to environmental engineers studying the meltponds is the type of ice observed for each pond. Recall that ice type is classified as first-year ice, multiyear ice, or landfast ice. The SAS summary table for the ice types of the 504 meltponds is reproduced at the bottom of the page.

a. Use a 90% confidence interval to estimate the proportion of meltponds in the Canadian Arctic that have first-year ice.
b. Use a 90% confidence interval to estimate the difference between the proportion of meltponds in the Canadian Arctic that have first-year ice and the proportion that have multiyear ice.

9.6 *Orientation cues experiment.* Refer to the *Human Factors* (Dec. 1988) study of color brightness as a body orientation clue, Exercise 7.65 (p. 331). Ninety college students,

## SAS Output for Exercise 9.5

The FREQ Procedure

| ICETYPE | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| First-year | 88 | 17.46 | 88 | 17.46 |
| Landfast | 196 | 38.89 | 284 | 56.35 |
| Multi-year | 220 | 43.65 | 504 | 100.00 |

reclining on their backs in the dark, were disoriented when positioned on a rotating platform under a slowly rotating disk that blocked their field of vision. The subjects were asked to say "stop" when they felt as if they were right-side up. The position of the brightness pattern on the disk in relation to each student's body orientation was then recorded. Subjects selected only three disk brightness patterns as subjective vertical clues: (1) brighter side up, (2) darker side up, and (3) brighter and darker side aligned on either side of the subjects' heads. The frequency counts for the experiment are given in the accompanying table. Construct a 95% confidence interval for the difference between the proportion of subjects who select brighter side up and the proportion who select darker side up as vertical clues. Interpret the results.

**BODYCLUE**

| Disk Orientation | | |
|---|---|---|
| Brighter Side Up | Darker Side Up | Bright and Dark Side Aligned |
| 58 | 15 | 17 |

9.7 *American perspective on engineering.* The American Association of Engineering Societies (AAES) hired Harris Interactive to conduct a survey of the American public's knowledge of and interest in engineering. (AAES/Harris Poll "American Perspectives on Engineers and Engineering: Final Report." 13 Feb. 2004.) The primary objective was to determine if the American public understands what engineers do and what sources of information they use to learn about engineering. Stratified random sampling was used to obtain a representative sample of 1,000 adults. In addition to answers to the survey questions, demographic information such as gender, age, education and number of engineers known was collected for each respondent. One survey item asked for responses to the statement, " Engineers are responsible for creating things that are harmful to society". Response categories were agree strongly, agree somewhat, disagree somewhat, or disagree strongly. An overall summary of the 965 responses is shown in the accompanying table.

a. Find and interpret a 99% confidence interval for the proportion of all American adults who disagree (somewhat or strongly) with the statement.

b. Find and interpret a 99% confidence interval for the difference between the proportions of all American adults who disagree (somewhat or strongly) and agree (somewhat or strongly) with the statement.

**HARMFUL**

| Agree Strongly | Agree Somewhat | Disagree Somewhat | Disagree Strongly |
|---|---|---|---|
| 99 | 212 | 311 | 343 |

9.8 *Irrigating cropland.* Because of erratic rainfall patterns and low water-holding capacities of soils in Florida, supplemental irrigation is required for producing most crops. A research team has developed five alternative water-management strategies for irrigating cropland in central Florida. A random sample of 100 agricultural engineers was interviewed and asked which of the strategies he or she believes would yield maximum productivity. A summary of their responses is shown in the table.

**IRRIGATE**

| *Strategy* | A | B | C | D | E |
|---|---|---|---|---|---|
| *Frequency* | 17 | 27 | 22 | 15 | 19 |

a. Find a 90% confidence interval for the true proportion of agricultural engineers who recommend strategy C.

b. Find a 90% confidence interval for the difference between the true proportions of agricultural engineers who recommend strategies E and B.

c. Find a 90% confidence interval for the true difference between the percentages of agricultural engineers who recommend strategies A and D.

## Theoretical Exercise

9.9 For the multinomial probability distribution, show that

$$\text{Cov}(n_i, n_j) = -np_ip_j$$

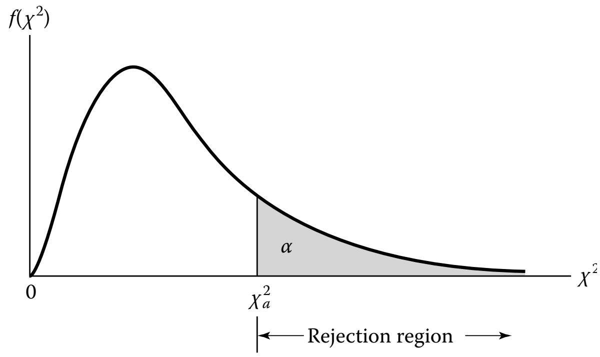
[*Hint*: First show that $E(n_in_j) = n(n-1)p_ip_j$.]

## 9.3 Testing Category Probabilities in a One-Way Table

Suppose we want to test a hypothesis about the category probabilities for the defective impeller study using the data given in Table 9.1. Specifically, we might want to test the (null) hypothesis that the proportions of defectives attributable to the five production lines are equal, i.e., $H_0$: $p_1 = p_2 = \cdots = p_5 = .2$, against the alternative hypothesis that at least two of the probabilities are unequal. Intuitively, we would choose a test statistic based on the deviations of the **observed category counts**, $n_1, n_2, \ldots, n_5$, from their expected values, or **expected category counts**

$$E(n_i) = np_i = (103)(.2) = 20.6 \qquad (i = 1, 2, \ldots, 5)$$

**FIGURE 9.1**
Rejection region for the chi-square test




Large deviations between the observed and expected category counts would provide evidence to indicate that the hypothesized category probabilities are incorrect.

The statistic used to test hypotheses about the category probabilities of a $k$-category multinomial experiment, one based on the weighted sum of squared deviations between observed and expected cell counts, is

$$\chi^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)}$$

Substituting $np_i$ for $E(n_i)$ and expanding the numerator, it can be shown (proof omitted) that

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \left( \sum_{i=1}^{k} \frac{n_i^2}{np_i} \right) - n$$

When the number $n$ of trials is large enough so that $E(n_i) \geq 5$ for $i = 1, 2, \ldots, k$, the statistic $\chi^2$ will possess (proof omitted) approximately a chi-square sampling distribution.* The value of $\chi^2$ will be larger than expected if the deviations $[n_i - E(n_i)]$ are large. Therefore, the rejection region for the test is $\chi^2 > \chi^2_\alpha$, where $\chi^2_\alpha$ is the value of $\chi^2$ that locates an area $\alpha$ in the upper tail of the chi-square distribution (see Figure 9.1).

The number of degrees of freedom for the approximating chi-square distribution will always equal $k$ less 1 degree of freedom for every linearly independent restriction placed on the category counts. For example, we always have at least one linear restriction on the category counts because their sum must equal the sample size, $n$:

$$n_1 + n_2 + \cdots + n_k = n$$

### A Test of a Hypothesis About Multinomial Probabilities: One-Way Table

$H_0$: $p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0}$, where $p_{1,0}, p_{2,0}, \ldots, p_{k,0}$, represent the hypothesized values of the multinomial probabilities

$H_a$: At least one of the multinomial probabilities differs from its null hypothesized value

*For some applications, the expected cell counts can be less than 5. More on this subject can be found in the paper by Cochran (1952) listed in the references for this chapter.

*Test statistic*: $\chi_c^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \left( \sum_{i=1}^{k} \frac{n_i^2}{np_{i,0}} \right) - n$

where $E(n_i) = np_{i,0}$, the expected number of outcomes of type $i$ assuming $H_0$ is true. The total sample size is $n$.

*Rejection region*: $\chi_c^2 > \chi_\alpha^2$, where $\chi_\alpha^2$ has $(k - 1)$ df

$p$-value: $P(\chi^2 > \chi_c^2)$

*Assumption*: For the chi-square approximation to be valid, $E(n_i) \geq 5$ for all $n_i$

Other restrictions arise if we must estimate the category probabilities. Since each estimate will involve a linear function of the category counts, the degrees of freedom for chi-square will be reduced by 1 for each category parameter that must be estimated.

A test of a hypothesis that the category probabilities assume specified values results in only a single linear restriction on the category counts—namely, $n_1 + n_2 + \cdots + n_k = n$. No category probabilities need to be estimated because their values are specified in $H_0$. The test procedure is described in the preceding box. We will illustrate this simple application of the chi-square test in Example 9.3.

## Example 9.3

**Multinomial Test: Defective Impellers**

Refer to the data provided in Table 9.1. Test the hypothesis that the proportions of all defective impellers attributable to the five production lines are equal. Test using $\alpha = .05$.

**Solution**

We want to test $H_0$: $p_1 = p_2 = \cdots = p_5 = .2$ against the alternative hypothesis, $H_a$: At least two of the category probabilities are unequal. We have already calculated

$$E(n_i) = np_i = (103)(.2) = 20.6 \qquad (i = 1, 2, \ldots, 5)$$

**TABLE 9.3 Observed and Expected Category Counts for the Data of Table 9.1**

| | | | | | |
|---|---|---|---|---|---|
| **Observed** | 15 | 27 | 31 | 19 | 11 |
| **Expected** | (20.6) | (20.6) | (20.6) | (20.6) | (20.6) |

The observed and the expected category counts (in parentheses) are shown in Table 9.3. Substituting the observed and expected values of the category counts into the formula for $\chi^2$, we obtain

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \frac{(15 - 20.6)^2}{20.6} + \frac{(27 - 20.6)^2}{20.6} + \cdots + \frac{(11 - 20.6)^2}{20.6}$$
$$= 13.36$$

The rejection region for the test is $\chi^2 > \chi_{.05}^2$, where $\chi_{.05}^2$ is based on $k - 1 = 5 - 1 = 4$ degrees of freedom. This value, found in Table 8 of Appendix B, is $\chi_{.05}^2 = 9.48773$. Since the observed value of $\chi^2$ exceeds this value, there is sufficient evidence (at $\alpha = .05$) to reject $H_0$. It appears that at least one production line is responsible for a higher proportion of defective impellers than the other lines. (*Note*: The test for a one-way table can be conducted with statistical software. The SPSS printout of the analysis is shown in Figure 9.2. Since the $p$-value, .010 (highlighted), is less than $\alpha = .05$, we reject $H_0$.)

**LINE**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| A | 15 | 20.6 | -5.6 |
| B | 27 | 20.6 | 6.4 |
| C | 31 | 20.6 | 10.4 |
| D | 19 | 20.6 | -1.6 |
| E | 11 | 20.6 | -9.6 |
| Total | 103 | | |

**Test Statistics**

| | LINE |
|---|---|
| Chi-Square[a] | 13.359 |
| df | 4 |
| Asymp. Sig. | .010 |

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 20.6.

**FIGURE 9.2**
SPSS analysis of data in Table 9.3

## Applied Exercises

**SOCROB**

9.10 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study of how engineers design social robots, Exercise 2.1 (p. 26). Recall that a social (or service) robot is designed to entertain, educate, and care for human users. In a random sample of 106 social robots obtained through a web search, the researchers found that 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels (These data are saved in the SOCROB file.) Prior to obtaining these sample results, a robot design engineer stated that 50% of all social robots produced have legs only, 30% have wheels only, 10% have both legs and wheels, and 10% have neither legs nor wheels.

a. Explain why the data collected for each sampled social robot is categorical in nature.
b. Specify the null and alternative hypothesis for testing the design engineer's claim.
c. Assuming the claim is true, determine the number of social robots in the sample you expect to fall into each design category.
d. Use the results to compute the $\chi^2$ test statistic.
e. Make the appropriate conclusion using $\alpha = .05$.

**MOBILE**

9.11 *Mobile device typing strategies.* Refer to the *Applied Ergonomics* (March 2012) study of the typing styles of mobile device users, Exercise 9.2 (p. 447). Recall that typing style was categorized as (1) device held with both hands/both thumbs typing, (2) device held with right hand/right thumb typing, (3) device held with left hand/left thumb typing, (4) device held with both hands/right thumb typing, (5) device held with left hand/right index finger typing, or (6) other. The number (in a sample of 859 college students) falling into each of the six categories was 396, 311, 70, 39, 18, and 25, respectively. (These data are saved in the **MOBILE** file.) Is this sufficient evidence to conclude that the proportions of mobile device users in the six texting style categories differ? Use $\alpha$=.10 to answer the question.

9.12 *Scanning Internet messages. Inc. Technology* reported the results of an Equifax/Harris Consumer Privacy Survey in which 328 Internet users indicated their level of agreement with the following statement: "The government needs to be able to scan Internet messages and user communications to prevent fraud and other crimes." The number of users in each response category is summarized as follows:

**SCAN**

| Agree Strongly | Agree Somewhat | Disagree Somewhat | Disagree Strongly |
|---|---|---|---|
| 59 | 108 | 82 | 79 |

a. Specify the null and alternative hypotheses you would use to determine if the opinions of Internet users are evenly divided among the four categories.
b. Conduct the test of part a using $\alpha = .05$.
c. In the context of this problem, what is a Type I error? A Type II error?
d. What assumptions must hold in order to ensure the validity of the test, part b?

**PONDICE**

9.13 *Characteristics of ice meltponds.* Refer to the study of ice meltponds in the Canadian Arctic, Exercise 9.5 (p. 404). The SAS summary table for the ice types of the 504 meltponds is reproduced at the bottom of the page. Suppose environmental engineers hypothesize that 15% of Canadian Arctic meltponds have first-year ice, 40% have landfast ice, and 45% have multiyear ice. Test the engineers' theory using $\alpha = .01$.

9.14 *Management system failures.* Refer to the *Process Safety Progress* (Dec. 2004) and U.S. Chemical Safety and Hazard Investigation Board study of industrial accidents caused by management system failures, Exercise 2.6 (p. 27). The accompanying table gives a breakdown of the root causes of a sample of 83 incidents. Are there significant differences in the percentage of incidents in the four cause categories? Test using $\alpha = .05$.

**MSFAIL**

| Management System Cause Category | Number of Incidents |
|---|---|
| Engineering & Design | 27 |
| Procedures & Practices | 24 |
| Management & Oversight | 22 |
| Training & Communication | 10 |
| *TOTAL* | 83 |

*Source*: Blair, A.S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board." *Process Safety Progress*, Vol. 23, No. 4, Dec. 2004 (Table 1).

9.15 *American perspective on engineering.* Refer to the American Association of Engineering Societies (AAES) survey of the public's knowledge of engineering, Exercise 9.7 (p. 448). Responses to the statement, "Engineers are responsible for creating things that are harmful to society" are summarized and reproduced in the table. Use these results to determine if the percentages in the four response categories are different. Test using $\alpha = .01$.

**HARMFUL**

| Agree Strongly | Agree Somewhat | Disagree Somewhat | Disagree Strongly |
|---|---|---|---|
| 99 | 212 | 311 | 343 |

9.16 *Beetles and slime molds.* A group of environmental engineers are studying mushroom-like slime molds as a potential food source for insects. (*Journal of Natural History*, May 2010.) In particular, they investigated which of six species of slime molds are most attractive to beetles inhabiting an Atlantic rain forest. A sample of 19 beetles feeding on slime mold was obtained and the species of slime mold was determined for each beetle. The number of beetles captured on each of the six species are given in the accompanying table. The researchers want to know if the relative frequency of occurrence of beetles differs for the six slime mold species.

**SLIMEMOLD**

| Slime mold species: | LE | TM | AC | AD | HC | HS |
|---|---|---|---|---|---|---|
| Number of beetles: | 3 | 2 | 7 | 3 | 1 | 3 |

a. Identify the categorical variable (and its levels) of interest in this study.
b. Set up the null and alternative hypothesis of interest to the researchers.
c. Find the test statistic and corresponding $p$-value.
d. The researchers found "no significant differences in the relative frequencies of occurrence" using $\alpha = .05$. Do you agree?
e. Comment on the validity of the inference, part d. (Determine the expected cell counts.)

**NCDOT**

9.17 *Traffic sign maintenance.* Refer to the *Journal of Transportation Engineering* (June, 2013) study of traffic sign maintenance, Exercise 8.67 (p. 415). Recall that civil engineers estimated the proportion of traffic signs maintained by the North Carolina Department of Transportation (NCDOT) that fail minimum retroreflectivity requirements. The researchers were also interested in the proportions of NCDOT signs with background colors white (regulatory signs), yellow (warning/caution), red (stop/yield/wrong way), and green (guide/information). In a random sample of 1,000 road signs maintained by the NCDOT, 373 were white, 447 were yellow, 88 were green, and 92 were red (These data are saved in the **NCDOT** file.) Suppose that NCDOT stores new signs in a warehouse for use as replacement signs; of these, 35% are white, 45% are yellow, 10% are green, and 10% are red. Does the distribution of background colors for all road signs maintained by NCDOT match the color distribution of signs in the warehouse? Test using $\alpha = .05$.

**BODYCLUE**

9.18 *Orientation clue experiment.* Refer to the *Human Factors* (Dec. 1988) study of orientation clues, Exercise 9.6 (p. 448). Conduct a test to compare the proportions of subjects that

**SAS Output for Exercise 9.13**

The FREQ Procedure

| ICETYPE | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| First-year | 88 | 17.46 | 88 | 17.46 |
| Landfast | 196 | 38.89 | 284 | 56.35 |
| Multi-year | 220 | 43.65 | 504 | 100.00 |

fall in the three disk-orientation categories. Assume you want to determine whether the three proportions differ. Use $\alpha = .05$.

9.19 *Detecting Alzheimer's disease at an early age.* Geneticists at Australian National University are studying whether the cognitive effects of Alzheimer's disease can be detected at an early age (*Neuropsychology*, Jan. 2007.) One portion of the study focused on a particular strand of DNA extracted from each in a sample of 2,097 young adults between the ages of 20 and 24. The DNA strand was classified into one of three genotypes: $E4^+/E4^+$, $E4^+/E4^-$, and $E4^-/E4^-$. The number of young adults with each genotype is shown in the accompanying table. Suppose that in adults who are not afflicted with Alzheimer's disease, the distribution of genotypes for this strand of DNA is 2% with $E4^+/E4^+$, 25% with $E4^+/E4^-$, and 73% with $E4^-/E4^-$. If differences in this distribution are detected, then this strand of DNA could lead researchers to an early test for the onset of Alzheimer's. Conduct a test (at $\alpha = .05$ ) to determine if the distribution of E4/E4 genotypes for the population of young adults differs from the norm.

**E4E4YOUNG**

| Genotype: | $E4^+/E4^+$ | $E4^+/E4^-$ | $E4^-/E4^-$ |
|---|---|---|---|
| Number of young adults: | 56 | 517 | 1524 |

## Theoretical Exercise

9.20 A general proof of the fact that $\chi^2$ possesses approximately a chi-square sampling distribution when $n$ is large is beyond the scope of this text. However, it can be justified for the binomial case $(k = 2)$. In Optional Exercise 6.118 (p. 259), we stated that if $Z$ is a standard normal random variable, then $Z^2$ is a chi-square random variable with 1 degree of freedom. Denote the two cell counts for a binomial experiment as $n_1 = Y$ and $n_2 = (n - Y)$. Then, for large $n$,

$$Z = \frac{Y - np}{\sqrt{npq}}$$

has approximately a standard normal distribution and $Z^2$ will be approximately distributed as a chi-square random variable with 1 degree of freedom. Show algebraically that for $k = 2$, $\chi^2 = Z^2$.

# 9.4 Inferences About Category Probabilities in a Two-Way (Contingency) Table

The methods presented in Section 9.2 and 9.3 are appropriate for a one-directional (or one-way) classification of the data. For example, the categories for the defective impeller data of Example 9.3 correspond to the "values" assumed by the qualitative variable, production line. Often, we may want to classify data according to two directions of classification—that is, according to two qualitative variables. The objective of such a classification usually is to determine whether the *two directions of classification are dependent*.

To illustrate, consider a questionnaire that was mailed to a sample of 150 households within 2 weeks after a nuclear mishap occurred in 1979 on Three Mile Island near Harrisburg, Pennsylvania. One question concerned residents' attitudes toward a full evacuation: "Should there have been a full evacuation of the immediate area?" Residents were classified according to the distance (in miles) of the community in which they reside from Three Mile Island and their opinion on a full evacuation. A summary of the responses for the 150 households randomly selected is shown in the **two-way table** shown in Table 9.4. This table is called a **contingency table**; it presents multinomial count data classified on two scales, or dimensions, of classification, namely, distance from Three Mile Island and responses to the full evacuation question.

**MILE3**

TABLE 9.4 **Contingency Table for Three Mile Island Survey**

| | | *Distance from Three Mile Island, miles* | | | |
|---|---|---|---|---|---|
| | | 1–6 | 7–12 | 13+ | TOTALS |
| **Full** | **Yes** | 18 | 15 | 33 | 66 |
| **Evacuation** | **No** | 20 | 19 | 45 | 84 |
| **TOTALS** | | 38 | 34 | 78 | 150 |

*Source*: Brown, S., et al. Final report on a survey of "Three Mile Island area residents." Department of Geography, Michigan State University, Aug. 1979.

**TABLE 9.5a Observed Counts for Contingency Table**

| | | Distance from Three Mile Island, miles | | | |
|---|---|---|---|---|---|
| | | 1–6 | 7–12 | 13+ | TOTALS |
| **Full** | **Yes** | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1\bullet}$ |
| **Evacuation** | **No** | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2\bullet}$ |
| **TOTALS** | | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet 3}$ | $n$ |

**TABLE 9.5b Probabilities for Contingency Table**

| | | Distance from Three Mile Island, miles | | | |
|---|---|---|---|---|---|
| | | 1–6 | 7–12 | 13+ | TOTALS |
| **Full** | **Yes** | $p_{11}$ | $p_{12}$ | $p_{13}$ | $p_{1\bullet}$ |
| **Evacuation** | **No** | $p_{21}$ | $p_{22}$ | $p_{23}$ | $p_{2\bullet}$ |
| **TOTALS** | | $p_{\bullet 1}$ | $p_{\bullet 2}$ | $p_{\bullet 3}$ | 1 |

Each cell of Table 9.4, located in a specific row and column, represents one of the $k = (2)(3) = 6$ categories of a two-directional classification of the $n = 150$ observations. The symbols representing the cell counts for the experiment in Table 9.4 are shown in Table 9.5a; the corresponding cell, row, and column probabilities are shown in Table 9.5b. Thus $n_{11}$ represents the number of residents who live within 6 miles of the accident and supported full evacuation, and $p_{11}$ represents the corresponding cell probability. The row totals (designated at $n_1$, and $n_2$,) and column totals (designated at $n_{\bullet 1}$, $n_{\bullet 2}$, and $n_{\bullet 3}$) are shown in Table 9.5a. The corresponding row and column probability totals are shown in Table 9.5b. The probability totals for the rows and columns are called **marginal probabilities**. For example, the marginal probability $p_{1\bullet}$ is the probability that a resident favored full evacuation, and the marginal probability $p_{\bullet 1}$ is the probability that a respondent lives 1–6 miles from Three Mile Island. Thus,

$$p_{1\bullet} = p_{11} + p_{12} + p_{13} = P(\text{favor full evacuation})$$

and

$$p_{\bullet 1} = p_{11} + p_{21} = P(\text{live 1–6 miles from Three Mile Island})$$

You can see that the experiment we have described is a multinomial experiment with a total of 150 trials and $(2)(3) = 6$ categories. Since the 150 residents were randomly chosen, the trials are considered independent, and the probabilities are viewed as remaining constant from trial to trial.

The objective of the study is to determine whether the two classifications, distance from Three Mile Island and opinion on full evacuation, are dependent. That is, if we know the distance from Three Mile Island, does that information provide a clue about the resident's opinion on evacuation? In a probabilistic sense, we know (Chapter 3) that independence of events $A$ and $B$ implies $P(A \cap B) = P(A)P(B)$. Similarly, in the contingency table analysis, if the two classifications are independent, the probability that an item is classified in any particular cell of the table is the product of the corresponding marginal probabilities. Thus, under the hypothesis of independence, in Table 9.5b, we must have

$$p_{11} = p_{1\bullet} p_{\bullet 1} \qquad p_{12} = p_{1\bullet} p_{\bullet 2} \qquad p_{13} = p_{1\bullet} p_{\bullet 3}$$

and so forth. Therefore, the null hypothesis that the directions of classification are independent is equivalent to the hypothesis that every cell probability is equal to the product of its respective row and column marginal probabilities. If the data disagree with the expected cell counts computed from these probabilities, there is evidence to indicate that the two directions of classification are dependent.

If we were to calculate the expected cell counts for our example, you would immediately perceive a difficulty. The marginal probabilities are unknown and must be estimated. The best estimate of the $i$th row marginal probability, call it $p_{i.}$, is

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} = \frac{\text{Row } i \text{ total}}{n}$$

Similarly, the best estimate of the $j$th marginal column probability is

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n} = \frac{\text{Column } j \text{ total}}{n}$$

Therefore, the estimated expected cell count for the cell in the $i$th row and $j$th column of the contingency table is

$$\begin{aligned}\hat{E}(n_{ij}) &= n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = n\left(\frac{n_{i\bullet}}{n}\right)\left(\frac{n_{\bullet j}}{n}\right) = \frac{n_{i\bullet}n_{\bullet j}}{n} \\ &= \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{n}\end{aligned}$$

The general form of an $r \times c$ contingency table (one containing $r$ rows and $c$ columns) is shown in Table 9.6. When $n$ is large, the test statistic

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})} = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left(n_{ij} - \dfrac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\left(\dfrac{n_{i\bullet}n_{\bullet j}}{n}\right)}$$

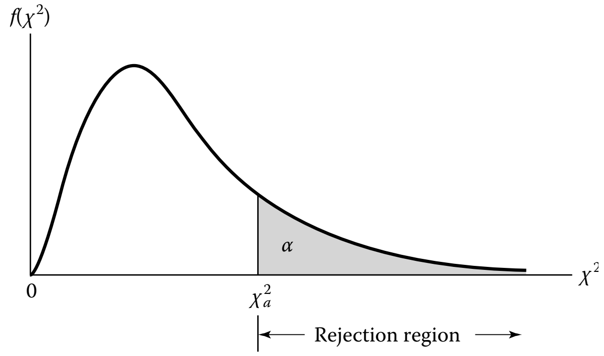
will possess approximately a chi-square distribution. The rejection region for the test will be $\chi^2 > \chi^2_\alpha$ (see Figure 9.3).

To determine the number of degrees of freedom for the approximating chi-square distribution, note that $k = rc$. From this we must subtract 1 degree of freedom because the sum of all $rc$ cell counts must equal $n$. We also subtract $(r - 1)$ because we must estimate the $(r - 1)$ row marginal probabilities. (The last row probability will then be determined because the sum of the row probabilities must equal 1.) Similarly,

**TABLE 9.6 General $r \times c$ Contingency Table**

| | | Column | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | … | c | *Row Totals* |
| | **1** | $n_{11}$ | $n_{12}$ | … | $n_{1c}$ | $n_{1\bullet}$ |
| | **2** | $n_{21}$ | $n_{22}$ | … | $n_{2c}$ | $n_{2\bullet}$ |
| **Row** | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| | ***r*** | $n_{r1}$ | $n_{r2}$ | … | $n_{rc}$ | $n_{r\bullet}$ |
| **Column Totals** | | $n_{\bullet 1}$ | $n_{\bullet 2}$ | … | $n_{\bullet c}$ | $n$ |

**FIGURE 9.3**
Rejection region for the chi-square test for dependence




we must subtract $(c - 1)$ because we must estimate $(c - 1)$ column marginal probabilities. Therefore, the degrees of freedom for chi-square will be

$$
\begin{aligned}
\text{df} &= k - \begin{pmatrix} \text{Number of linearly independent} \\ \text{restrictions on the cell counts} \end{pmatrix} \\
&= rc - (1) - (r - 1) - (c - 1) \\
&= rc - r - c + 1 \\
&= (r - 1)(c - 1)
\end{aligned}
$$

The chi-square test is summarized in the box; its use is illustrated in Example 9.4.

### General Form of a Contingency Table Analysis: A Test for Independence

$H_0$: The two classifications are independent

$H_a$: The two classifications are dependent

*Test statistic*: $$\chi_c^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

where

$$\hat{E}(n_{ij}) = \frac{n_{i\bullet} n_{\bullet j}}{n}, \qquad \begin{aligned} n_{i\bullet} &= \text{total for row } i \\ n_{\bullet j} &= \text{total for column } j \end{aligned}$$

*Rejection region*: $\chi_c^2 > \chi_\alpha^2$, where $\chi_\alpha^2$ has $(r - 1)(c - 1)$ df.

$p$-value: $P(\chi^2 > \chi_c^2)$

*Assumptions*:
1. The $n$ observed counts are a random sample from the population of interest. We may then consider this to be a multinomial experiment with $r \times c$ possible outcomes.
2. For the $\chi^2$ approximation to be valid, we require that the estimated expected counts be greater than or equal to 5 in all cells.

**Example 9.4**

Contingency Table Analysis: Nuclear Plant Evacuation

Use the data in Table 9.4 to decide whether a Harrisburg resident's opinion on full evacuation of Three Mile Island depends on how far (in miles) the resident lives from the nuclear plant.

Solution

The first step in the analysis of a contingency table is to calculate the estimated expected cell counts. For example,

$$
\begin{aligned}
E(n_{11}) &= \frac{n_{1\bullet}n_{\bullet 1}}{n} \\
&= \frac{(66)(38)}{150} = 16.72 \\
E(n_{12}) &= \frac{n_{1\bullet}\, n_{\bullet 2}}{n} \\
&= \frac{(66)(34)}{150} = 14.96 \\
&\ \ \vdots \\
E(n_{23}) &= \frac{n_{2\bullet}n_{\bullet 3}}{n} \\
&= \frac{(84)(78)}{150} = 43.68
\end{aligned}
$$

The cell counts (top number in cell) and the corresponding estimated expected values (bottom number in cell) are shown in the SAS printout of the contingency table analysis, Figure 9.4.

For this study, the $\chi^2$ test statistic is computed as follows:

$$
\begin{aligned}
\chi^2 &= \frac{[n_{11} - \hat{E}(n_{11})]^2}{\hat{E}(n_{11})} + \frac{[n_{12} - \hat{E}(n_{12})]^2}{\hat{E}(n_{12})} + \cdots + \frac{[n_{23} - \hat{E}(n_{23})]^2}{\hat{E}(n_{23})} \\
&= \sum_{j=1}^{3}\sum_{i=1}^{2} \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}
\end{aligned}
$$

Table of EVAC by DISTANCE

| EVAC / Frequency / Expected | DISTANCE 1-6 | 7-12 | OVER12 | Total |
|---|---|---|---|---|
| NO | 20<br>21.28 | 19<br>19.04 | 45<br>43.68 | 84 |
| YES | 18<br>16.72 | 15<br>14.96 | 33<br>34.32 | 66 |
| Total | 38 | 34 | 78 | 150 |

Statistics for Table of EVAC by DISTANCE

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 0.2658 | 0.8755 |
| Likelihood Ratio Chi-Square | 2 | 0.2653 | 0.8758 |
| Mantel-Haenszel Chi-Square | 1 | 0.2587 | 0.6110 |
| Phi Coefficient | | 0.0421 | |
| Contingency Coefficient | | 0.0421 | |
| Cramer's V | | 0.0421 | |

Sample Size = 150

**FIGURE 9.4**
SAS contingency table analysis for Example 9.4

Substituting the data of Figure 9.4 into this expression, we obtain

$$\chi^2 = \frac{(18 - 16.72)^2}{16.72} + \frac{(15 - 14.96)^2}{14.96} + \cdots + \frac{(45 - 43.68)^2}{43.68} = .266$$

Note that this value, $\chi^2 = .2658$, is shaded at the bottom of the SAS printout, Figure 9.4. The rejection region for the test is $\chi^2 > \chi^2_{.05} = 5.99147$, where $\chi^2_{.05}$ is based on $(r - 1)(c - 1) = (1)(2) = 2$ degrees of freedom. Since the computed value of $\chi^2$, .266, falls below this critical value, we fail to reject $H_0$; there is insufficient evidence to conclude that the two directions of data classification are dependent. It appears that opinion on full evacuation is independent of distance from Three Mile Island.

We can arrive at the same conclusion by observing that the $p$-value for the test, shaded in Figure 9.4, exceeds $\alpha = .05$.

Suppose we conclude that the two directions of classification in a contingency table are dependent. Practically speaking, this implies that the distribution of the percentages of observations falling in the categories for one of the qualitative variables depends on the level of the other variable. In the $2 \times 3$ table of Example 9.4, this means that the proportion $p_i$ of residents that favored full evacuation differed for the three distance groups. To determine the magnitude of the differences, we could construct confidence intervals for the differences, $(p_{\bullet 1} - p_{\bullet 2})$, $(p_{\bullet 1} - p_{\bullet 3})$, and $(p_{\bullet 2} - p_{\bullet 3})$, using the method of Section 7.10. In the special case of a $2 \times 2$ table, the $\chi^2$-test is equivalent to a test of the null hypothesis Ho: $p_{\bullet 1} - p_{\bullet 2} = 0$.

## Applied Exercises

9.21 *Study of orocline development.* In *Tectonics* (Oct. 2004), geologists published their research on the formation of oroclines (curved mountain belts) in the central Appalachian mountains. A comparison was made of two nappes (sheets of rock that have moved over a large horizontal distance), one in Pennsylvania and the other in Maryland. Rock samples at the mountain rim of both locations were collected and the foliation intersection axes (FIA) preserved within large mineral grains was measured for each. The accompanying table shows the number of rock samples in the different FIA measurement categories at the two locations. The geologists tested whether the distribution of FIA trends were the same for the Pennsylvania Nappe and Maryland Nappe using a chi-square test of independence.

**OROCLINE**

| | | Pennsylvania Nappe | Maryland Nappe |
|---|---|---|---|
| | *0–79°* | 20 | 6 |
| *FIA* | *80–149°* | 17 | 10 |
| | *150–179°* | 10 | 7 |

*Source*: Yeh, W., and Bell, T. "Significance of dextral reactivation of an E–W transfer fault in the formation of the Pennsylvania orocline, central Appalachians." *Tectonics*, Vol. 23, No. 5, October 2004 (Table 2).

a. Give the null and alternative hypothesis for the test.
b. The researchers reported the test statistic as $\chi^2 = 1.874$. Do you agree?
c. Find the rejection region for the test, using $\alpha = .05$.
d. Make the appropriate conclusion in the words of the problem.

9.22 *Mobile device typing strategies.* Refer to the *Applied Ergonomics* (March 2012) study of mobile device typing strategies, Exercise 9.2 (p. 447). Recall that typing style of mobile device users was categorized as (1) device held with both hands/ both thumbs typing, (2) device held with

**MOBILE**

| Typing Strategy | Number of Males | Number of Females |
|---|---|---|
| Both hands hold / both thumbs type | 161 | 235 |
| Right hand hold / right thumb type | 118 | 193 |
| Left hand hold / left thumb type | 29 | 41 |
| Both hands hold / right thumb type | 10 | 29 |
| Left hand hold / right index type | 6 | 12 |
| Other | 11 | 14 |

*Source:* Gold, J.E., et al. "Postures, typing strategies, and gender differences in mobile device usage: An observational study", *Applied Ergonomics*, Vol. 43, No. 2, March 2012 (Table 2).

right hand/ right thumb typing, (3) device held with left hand/ left thumb typing, (4) device held with both hands/ right thumb typing, (5) device held with left hand/right index finger typing, or (6) other. The researchers' main objective was to determine if there are gender differences in typing strategies. Typing strategy and gender was observed for each in a sample of 859 college students observed typing on their mobile devices. The data are summarized in the table on p. 458. Is this sufficient evidence to conclude that the proportions of mobile device users in the six texting style categories depend on whether a male or a female is texting? Use $\alpha = .10$ to answer the question.

9.23 *"Cry Wolf" effect in air traffic controlling.* Researchers at Alion Science Corporation and New Mexico State University collaborated on a study of how air traffic controllers respond to false alarms (*Human Factors*, Aug. 2009). The researchers theorize that the high rate of false alarms regarding mid-air collisions lead to the "cry wolf" effect, i.e., the tendency for air traffic controllers to ignore true alerts in the future. The investigation examined data on a random sample of 437 conflict alerts. Each alert was first classified as a "true" or "false" alert. Then, each was classified according to whether or not there was a human controller response to the alert. A summary of the responses is provided in the accompanying table. Do the data indicate that the response rate of air traffic controllers to mid-air collision alarms differs for true and false alerts? Test using $\alpha$=.05. What inference can you make concerning the "cry wolf" effect?

**ATC**

| | No Response | Response | TOTALS |
|---|---|---|---|
| True Alert | 3 | 231 | 234 |
| False Alert | 37 | 166 | 203 |
| TOTALS | 40 | 397 | 437 |

*Source:* Wickens, C.D., et al. "False alerts in air traffic control conflict alerting system: Is there a 'cry wolf' effect?", *Human Factors*, Vol. 51, No. 4, August 2009 (Table 2).

9.24 *Job satisfaction of women in construction.* The hiring of women in construction and construction-related jobs has steadily increased over the years. A study was conducted to provide employers with information designed to reduce the potential for turnover of female employees (*Journal of Professional Issues in Engineering Education & Practice*, April 2013). A survey questionnaire was emailed to members of the National Association of Women in Construction (NAWIC). A total of 477 women responded to survey questions on job challenge and satisfaction with life as an employee. The results (number of females responding in the different categories) are summarized in the accompanying table. What conclusions can you draw from the data regarding the association between an NAWIC member's satisfaction with life as an employee and their satisfaction with job challenge?

**NAWIC**

| | | Life as an Employee | |
|---|---|---|---|
| | | Satisfied | Dissatisfied |
| Job | Satisfied | 364 | 33 |
| Challege | Dissatisfied | 24 | 26 |

*Source:* Malone, E.K. & Issa, R.A. " Work-Life Balance and Organizational Commitment of Women in the U.S. Construction Industry", *Journal of Professional Issues in Engineering Education & Practice*, Vol. 139, No. 2, April 2013 (Table 11).

9.25 *Groundwater contamination in wells.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of methyl *tert*-butyl ether (MTBE) contamination in public and private New Hampshire wells, Exercise 2.12 (p. 29). Recall that data on well class (public or private), aquifer (bedrock or unconsolidated), and detectable level of MTBE (below limit or detect) were collected for a sample of 223 wells. These data are saved in the **MTBE** file. (Data for the first 10 selected wells are shown in the accompanying table.)

(Ten selected observations from 223)

**MTBE**

| Well Class | Aquifer | Detect MTBE Status |
|---|---|---|
| Private | Bedrock | Below Limit |
| Private | Bedrock | Below Limit |
| Public | Unconsolidated | Detect |
| Public | Unconsolidated | Below Limit |
| Public | Unconsolidated | Below Limit |
| Public | Unconsolidated | Below Limit |
| Public | Unconsolidated | Detect |
| Public | Unconsolidated | Below Limit |
| Public | Unconsolidated | Below Limit |
| Public | Bedrock | Detect |
| Public | Bedrock | Detect |

*Source*: Ayotte, J. D., Argue, D. M., and McGarry, F. J. "Methyl *tert*-butyl ether occurrence and related factors in public and private wells in southeast New Hampshire." *Environmental Science & Technology*, Vol. 39, No. 1, Jan. 2005.

a. Use the data in the **MTBE** file to create a contingency table for well class and detectable MTBE status.
b. Conduct a test to determine if detectable MTBE status depends on well class. Test using $\alpha = .05$.
c. Use the data in the **MTBE** file to create a contingency table for aquifer and detectable MTBE status.
d. Conduct a test to determine if detectable MTBE status depends on aquifer. Test using $\alpha = .05$.

9.26 *Flight response of geese.* Offshore oil drilling near an Alaskan estuary has led to increased air traffic—mostly large helicopters—in the area. The U.S. Fish and Wildlife

Service commissioned a study to investigate the impact these helicopters have on the flocks of Pacific brant geese that inhabit the estuary in fall before migrating. (*Statistical Case Studies: A Collaboration between Academe and Industry*, 1998.) Two large helicopters were flown repeatedly over the estuary at different altitudes and lateral distances from the flock. The flight responses of the geese (recorded as "low" or "high"), altitude (hundreds of meters), and lateral distance (hundreds of meters) for each of 464 helicopter overflights were recorded and are saved in the **PACGEESE** file. (The data for the first 10 overflights are shown in the next table.)

**PACGEESE**

(*First 10 observations shown*)

| Overflight | Altitude | Lateral Distance | Flight Response |
|---|---|---|---|
| 1 | 0.91 | 4.99 | HIGH |
| 2 | 0.91 | 8.21 | HIGH |
| 3 | 0.91 | 3.38 | HIGH |
| 4 | 9.14 | 21.08 | LOW |
| 5 | 1.52 | 6.60 | HIGH |
| 6 | 0.91 | 3.38 | HIGH |
| 7 | 3.05 | 0.16 | HIGH |
| 8 | 6.10 | 3.38 | HIGH |
| 9 | 3.05 | 6.60 | HIGH |
| 10 | 12.19 | 6.60 | HIGH |

*Source*: Erickson, W., Nick, T., and Ward, D. "Investigating flight response of Pacific brant to helicopters at Izembek Lagoon, Alaska by using logistic regression". *Statistical Case Studies: A Collaboration between Academe and Industry*, ASA-SIAM Series on Statistics and Applied Probability, 1998.)

a. The researchers categorized altitude as follows: less than 300 meters, 300–600 meters, and 600 or more meters. Summarize the data in the **PACGEESE** file by creating a contingency table for altitude category and flight response.
b. Conduct a test to determine if flight response of the geese depends on altitude of the helicopter. Test using $\alpha = .01$.
c. The researchers categorized lateral distance as follows: less than 1,000 meters, 1,000–2,000 meters, 2,000–3,000 meters, and 3,000 or more meters. Summarize the data in the **PACGEESE** file by creating a contingency table for lateral distance category and flight response.
d. Conduct a test to determine if flight response of the geese depends on lateral distance of helicopter from the flock. Test using $\alpha = .01$.
e. The current Federal Aviation Authority (FAA) minimum altitude standard for flying over the estuary is 2,000 feet (approximately 610 meters). Based on the results, parts **a–d**, what changes to the FAA regulations do you recommend in order to minimize the effects to Pacific brant geese?

**SEEDLING**

9.27 *Subarctic plant study.* The traits of seed-bearing plants indigenous to subarctic Finland was studied in *Arctic, Antarctic, and Alpine Research* (May 2004). Plants were categorized according to *type* (dwarf shrub, herb, or grass), *abundance of seedlings* (no seedlings, rare seedlings, or abundant seedlings), *regenerative group* (no vegetative reproduction, vegetative reproduction possible, vegetative reproduction ineffective, or vegetative reproduction effective), *seed weight class* (0–.1, .1–.5, .5–1.0, 1.0–5.0, and >5.0 milligrams), and *diaspore morphology* (no structure, pappus, wings, fleshy fruits, or awns/hooks). The data for a sample of 73 plants are saved in the **SEEDLING** file.

a. A contingency table for plant type and seedling abundance, produced by MINITAB, follows. (*Note*: NS = no seedlings, SA = seedlings abundant, and SR = seedlings rare.) Suppose you want to perform a chi-square test of independence to determine whether seedling abundance depends on plant type. Find the expected cell counts for the contingency table. Are the assumptions required for the test satisfied?
b. Reformulate the contingency table by combining the NS and SR categories of seedling abundance. Find the expected cell counts for this new contingency table. Are the assumptions required for the test satisfied?
c. Reformulate the contingency table of part **b** by combining the dwarf shrub and grasses categories of plant type. Find the expected cell counts for this contingency table. Are the assumptions required for the test satisfied?
d. Carry out the chi-square test for independence on the contingency table, part **c**, using $\alpha = .10$. What do you conclude?

**MINITAB Output for Exercise 9.27**

**Tabulated statistics: Abundance, Type**

```
Rows: Abundance   Columns: Type

        DwarfShrub  Grasses  Herbs  All

NS               3        1      1    5
SA               5       14     32   51
SR               5        2     10   17
All             13       17     43   73

Cell Contents:      Count
```

9.28 *American perspective on engineering.* Refer to the AAES survey of the American public's knowledge of and interest in engineering, Exercise 9.7 (p. 448). Another survey question asked, "What media sources do you use to follow stories about engineering and engineers?" Responses were categorized as either "Follow on the internet" or "Do not follow on the internet". These responses were also classified according to age of the respondent, gender, education, and familiarity with an engineer. The results are shown in

*Age (n=849 responses)*

| | 18-29 years | 30-44 years | 45-59 years | 60 years or older |
|---|---|---|---|---|
| Follow on Internet | 33 | 55 | 23 | 11 |
| Do not follow on Internet | 121 | 190 | 221 | 195 |

*Gender (n=873 responses)*

| | Male | Female |
|---|---|---|
| Follow on Internet | 77 | 44 |
| Do not follow on Internet | 310 | 442 |

*Education (n=870 responses)*

| | Less than College Grad | College Grad or Higher |
|---|---|---|
| Follow on Internet | 53 | 80 |
| Do not follow on Internet | 452 | 265 |

*Familiarity with Engineers (n=871 responses)*

| | Know 0 Engineers | Know at least 1 Engineer |
|---|---|---|
| Follow on Internet | 5 | 123 |
| Do not follow on Internet | 153 | 590 |

the accompanying tables above. In their final report, the AAES concluded that "men, more educated and younger adults, and those familiar with an engineer are more likely to mention the Internet as a way in which they followed engineering news". Do you agree?

**SWDEFECTS**

9.29 *Software defects.* The PROMISE Software Engineering Repository, hosted by the School of Information Technology and Engineering, University of Ottawa, provides

### SPSS Output for Exercise 9.29

**DEFECT * PRED_EVG Crosstabulation**

Count

| | | PRED_EVG no | PRED_EVG yes | Total |
|---|---|---|---|---|
| DEFECT | false | 441 | 8 | 449 |
| | true | 47 | 2 | 49 |
| Total | | 488 | 10 | 498 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 1.188[b] | 1 | .276 | | |
| Continuity Correction[a] | .306 | 1 | .580 | | |
| Likelihood Ratio | .948 | 1 | .330 | | |
| Fisher's Exact Test | | | | .257 | .257 |
| N of Valid Cases | 498 | | | | |

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is .98.

researchers with data sets for building predictive software models. (See *Statistics in Action*, Chapter 3.) Data on 498 modules of software code written in C language for a NASA spacecraft instrument are saved in the **SWDEFECTS** file. Recall that each module was analyzed for defects and classified as "true" if it contained defective code and "false" if not. One algorithm for predicting whether or not a module has defects is "essential complexity" (denoted EVG), where a module with at least 15 subflow graphs with D-structured primes is predicted to have a defect. When the method predicts a defect, the predicted EVG value is "yes"; otherwise, it is "no." A contingency table for the two variables, actual defective status and predicted EVG, is shown in the SPSS printout on p. 461. Interpret the results. Would you recommend the essential complexity algorithm as a predictor of defective software modules? Explain.

## 9.5 Contingency Tables with Fixed Marginal Totals

In the analysis of contingency table data, one or more of the categories may contain an insufficient number of observations. To illustrate, we will consider the study (described in Section 9.4) of the relationship between a resident's opinion of full evacuation of the area surrounding a nuclear accident and the distance the resident lives from Three Mile Island. If the random sample contains only a small number of residents that live a certain distance away, this may cause the expected cell counts for that distance to be small—perhaps less than the required 5. To guard against this possibility, experimenters often fix either the row or column totals. For our example, we would fix the column totals by randomly and independently sampling a fixed number of residents in each distance group. This would increase the likelihood that the estimated expected cell counts would be of adequate size.

For example, suppose we obtain the evacuation opinion of random samples of 100 residents in each distance group. The results might appear as shown in Table 9.7. Note the difference between this sampling procedure and the one described in Section 9.4, where we assumed that a *single* random sample of $n = 150$ residents was selected from among the population of all people residing near Three Mile Island. In this section, we have randomly and independently selected three samples, 100 residents from each distance. Therefore, the data of Table 9.7 result from three multinomial experiments, each with $k = 2$ cells (support or do not support full evacuation), corresponding to the three distances, 1–6 miles, 7–12 miles, and 13 or more miles from Three Mile Island.

A chi-square test to detect dependence between row and column classifications, when either the column or the row totals are fixed, is conducted in exactly the same way as the test of Section 9.4. It can be shown (proof omitted) that the $\chi^2$ statistic will possess a sampling distribution that is approximately a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. The test procedure is summarized in the box. An application of the test to the comparison of two or more binomial proportion is illustrated in Example 9.5.

**TABLE 9.7 Distance–Evacuation Contingency Table with Column Total Fixed**

| | | Distance from Three Mile Island, miles | | | |
|---|---|---|---|---|---|
| | | 1–6 | 7–12 | 13+ | TOTALS |
| **Full** | **Yes** | 42 | 29 | 25 | 96 |
| **Evacuation** | **No** | 58 | 71 | 75 | 204 |
| **TOTALS** | | 100 | 100 | 100 | 300 |

**General Form of Contingency Table Analysis: A Test for Independence with Row* Totals Fixed**

If row totals are fixed:

$H_0$: The row proportions in each cell do not depend on the row; that is, the distributions of observations in the column categories are the same for each row.

$H_a$: The row proportions in some (or all) of the cells depend on the row; that is, the distributions of observations in the column categories differ for at least two of the rows.

*Test statistic*: $$\chi_c^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

where

$$\hat{E}(n_{ij}) = \frac{n_{i\bullet}\, n_{j\bullet}}{n}$$

*Rejection region*: $\chi_c^2 > \chi_\alpha^2$, where $\chi_\alpha^2$ has $(r-1)(c-1)$ df

$p$-value: $P(\chi^2 > \chi_c^2)$

*Assumptions:*
1. A random sample is selected from each population for which the row totals are fixed.
2. The samples are independently selected.
3. We require the estimated expected value of each cell to be at least 5 to use the $\chi^2$ approximation.

### Example 9.5

Contingency Table with Fixed Marginals: Defective Impellers

**IMPELLER3**

To compare the proportions of defective impellers produced by three production lines, a quality control engineer randomly sampled 500 impellers from each line. The numbers of defectives for the three lines were found to be 12, 17, and 7, respectively. Do the data provide sufficient evidence to indicate differences in the proportions of defective impellers produced by the three production lines? In other words, are the two directions of classification, production line and defective status, dependent?

Solution

The data were entered as a contingency table in MINITAB, with the resulting printout shown in Figure 9.5. The objective of this experiment is to compare three binomial proportions of defectives, $p_1$, $p_2$, and $p_3$, based on three independent binomial experiments, each containing 500 observations.

The null hypothesis is that the proportions of defectives for the three production lines are identical, i.e.,

$$H_0: \quad p_1 = p_2 = p_3$$

against the alternative hypothesis

$$H_a: \quad \text{At least two of the proportions, } p_1, p_2, \text{ and } p_3, \text{ differ.}$$

Note that the null hypothesis we have specified implies that the numbers of defectives and nondefectives are independent of the production line. Therefore, we test $H_0$: $p_1 = p_2 = p_3$ using the chi-square test for a contingency table analysis.
The estimated expected cell counts are computed using the formula

$$\hat{E}(n_{ij}) = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

*Note that to obtain the procedure for conducting a $\chi^2$ analysis for fixed column totals, it is necessary only to interchange the words *column* and *row* in the box.

```
Tabulated statistics: STATUS, LINE

Using frequencies in NUMBER


Rows: STATUS    Columns: LINE

            1     2     3    All

DEFECT     12    17     7     36
           12    12    12     36

NONDEF    488   483   493   1464
          488   488   488   1464

All       500   500   500   1500
          500   500   500   1500

Cell Contents:       Count
                     Expected count


Pearson Chi-Square = 4.269, DF = 2, P-Value = 0.118
Likelihood Ratio Chi-Square = 4.399, DF = 2, P-Value = 0.111
```

**FIGURE 9.5**
MINITAB contingency table analysis for Example 9.5

Therefore,

$$\hat{E}(n_{11}) = \frac{n_{1\bullet}n_{\bullet 1}}{n} = \frac{(36)(500)}{1{,}500} = 12$$

and

$$\hat{E}(n_{12}) = \frac{n_{1\bullet}n_{\bullet 2}}{n} = \frac{(36)(500)}{1{,}500} = 12$$

These, along with the remaining estimated expected cell counts, are shown (highlighted) on the MINITAB printout, Figure 9.5.

The computed value of $\chi^2$ (also highlighted on the printout) is

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})} \\ &= \frac{(12-12)^2}{12} + \frac{(17-12)^2}{12} + \cdots + \frac{(493-488)^2}{488} \\ &= 4.269\end{aligned}$$

The rejection region for the test is $\chi^2 > \chi^2_{.05}$, where $\chi^2_{.05} = 5.99147$ is based on $(r-1)(c-1) = (1)(2) = 2$ degrees of freedom. Since the computed value of $\chi^2$ does not exceed $\chi^2_{.05}$ (and, since the $p$-value shown on the printout, .118, exceeds $\alpha = .05$), there is insufficient evidence to indicate differences in the proportions of defective impellers produced by the three production lines. Note that we do not accept $H_0$—that is, we do not conclude that $p_1 = p_2 = p_3$—because we would be concerned about the possibility of making a Type II error, failing to detect differences in the proportions of defectives if, in fact, differences exist. The test conclusion simply means that if differences exist, they were too small to detect using samples of 500 impellers from each production line.

## Applied Exercises

**HYBRID**

9.30 *Safety of hybrid cars.* According to the Highway Loss Data Institute (HLDI), "hybrid [automobiles] have a safety edge over their conventional twins when it comes to shielding their occupants from injuries in crashes". (*HLDI Bulletin*, Sept. 2011.) Consider data collected by the HLDI on Honda Accords over the past eight years. In a sample of 50,132 collision claims for conventional Accords, 5,364 involved injuries; in a sample of 1,505 collision claims for hybrid Accords, 137 involved injuries. You want to use this information to determine whether injury rate for hybrid Accords is less than the injury rate for conventional Accords.

a. Identify the two qualitative variables measured for each Honda Accord collision claim.
b. Form a contingency table for this data, giving the number of claims in each combination of the qualitative variable categories.
c. Give $H_0$ and $H_a$ for testing whether injury rate for collision claims depends on Accord model (hybrid or conventional).
d. Find the expected number of claims in each cell of the contingency table, assuming that $H_0$ is true.
e. Compute the $\chi^2$ test statistic and compare your answer to the test statistic shown on the accompanying SAS printout.
f. Find the rejection region for the test using $\alpha$=.05 and compare your answer to the critical value shown on the accompanying SAS printout.
g. Make the appropriate conclusion using both the rejection region method and the p-value (shown on the SAS printout).
h. Find a 95% confidence interval for the difference between the injury rates of conventional and hybrid Honda Accords. (See Section 8.10.) Use the interval to determine whether the injury rate for hybrid Accords is less than the injury rate for conventional Accords.

### SAS Output for Exercise 9.30

The FREQ Procedure

Table of MODEL by CLAIM

| MODEL | | CLAIM: Injury | CLAIM: No Injury | Total |
|---|---|---|---|---|
| Conventional | Frequency | 5364 | 44768 | 50132 |
| | Expected | 5340.7 | 44791 | |
| Hybrid | Frequency | 137 | 1368 | 1505 |
| | Expected | 160.33 | 1344.7 | |
| Total | Frequency | 5501 | 46136 | 51637 |

Statistics for Table of MODEL by CLAIM

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 3.9139 | 0.0479 |
| Likelihood Ratio Chi-Square | 1 | 4.0893 | 0.0432 |
| Continuity Adj. Chi-Square | 1 | 3.7480 | 0.0529 |
| Mantel-Haenszel Chi-Square | 1 | 3.9138 | 0.0479 |
| Phi Coefficient | | 0.0087 | |
| Contingency Coefficient | | 0.0087 | |
| Cramer's V | | 0.0087 | |

| Fisher's Exact Test | |
|---|---|
| Cell (1,1) Frequency (F) | 5364 |
| Left-sided Pr <= F | 0.9801 |
| Right-sided Pr >= F | 0.0246 |
| | |
| Table Probability (P) | 0.0047 |
| Two-sided Pr <= P | 0.0510 |

Sample Size = 51637

9.31 *Versatility with resistor-capacitor circuits.* Research published in the *International Journal of Electrical Engineering Education* (Oct. 2012) investigated the versatility of engineering students' knowledge of circuits with one resistor and one capacitor connected in series. Students were shown four different configurations of a resistor-capacitor circuit and then given two tasks. First, each student was asked to state the voltage at the nodes on the circuit and, second, each student was asked to graph the dynamic behavior of the circuit. Suppose that in a sample of 160 engineering students, 40 were randomly assigned to analyze Circuit 1, 40 assigned to Circuit 2, 40 assigned to Circuit 3, and 40 assigned to Circuit 4. The researchers categorized task grades as follows: Correct voltages and graph, incorrect voltages but correct graph, incorrect graph but correct voltages, incorrect voltages and incorrect graph. A summary of the results (based on information provided in the journal article) are shown in the table. Does any one circuit appear to be more difficult to analyze than any other circuit? Support your answer with a statistical test of hypothesis.

**CIRCUIT4**

| | | Circuit 1 | Circuit 2 | Circuit 3 | Circuit 4 |
|---|---|---|---|---|---|
| | Both correct | 31 | 10 | 5 | 4 |
| Answer | Incorrect voltage | 0 | 3 | 11 | 12 |
| | Incorrect graph | 5 | 17 | 16 | 14 |
| | Both incorrect | 4 | 10 | 8 | 10 |
| Total number of students | | 40 | 40 | 40 | 40 |

**SOLDER**

9.32 *Performance of solder joint inspectors.* Westinghouse Electric Company has experimented with different means of evaluating the performance of solder joint inspectors. One approach involves comparing an individual inspector's classifications with those of the group of experts that comprise Westinghouse's Work Standards Committee. In one experiment, 153 solder connections were evaluated by the committee and 111 were classified as acceptable. An inspector evaluated the same 153 connections and classified 124 as acceptable. Of the items rejected by the inspector, the committee agreed with 19. (These results are saved in the **SOLDER** file.)

a. Construct a contingency table that summarizes the classifications of the committee and the inspector.
b. Based on a visual examination of the table you constructed in part **a**, does it appear that there is a relationship between the inspector's classifications and the committee's? Explain. (A bar graph of the percentage rejected by committee and inspector will aid your examination.)
c. Conduct a chi-square test of independence for these data. Use $\alpha = .05$. Carefully interpret the results of your test in the context of the problem.

9.33 *A new dental bonding agent.* When bonding teeth, orthodontists must maintain a dry field. A new bonding adhesive (called Smartbond) has been developed to eliminate the necessity of a dry field. However, there is concern that the new bonding adhesive may not stick to the tooth as well as the current standard, a composite adhesive. (*Trends in Biomaterials & Artificial Organs*, Jan. 2003.) Tests were conducted on a sample of 10 extracted teeth bonded with the new adhesive and a sample of 10 extracted teeth bonded with the composite adhesive. The Adhesive Remnant Index (ARI), which measures the residual adhesive of a bonded tooth on a scale of 1 to 5, was determined for each of the 20 bonded teeth after 1 hour of drying. (*Note*: An ARI score of 1 implies all adhesive remains on the tooth, and a score of 5 means none of the adhesive remains on the tooth.) A breakdown of the number of bonded teeth in the five ARI categories is shown in the table.

**BONDING**

| | Adhesive Remnant Index Score | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| *Smartbond* | 2 | 8 | 0 | 0 | 0 |
| *Composite* | 1 | 5 | 3 | 1 | 0 |

*Source*: Sunny, J., and Vallathan, A. "A comparative *in vitro* study with new generation ethyl cyanoacrylate (Smartbond) and a composite bonding agent." *Trends in Biomaterials & Artificial Organs*, Vol. 16, No. 2, Jan. 2003 (Table 6).

a. Explain why the contingency table is one with fixed marginals.
b. Conduct an analysis to determine if the distribution of ARI scores differs for the two types of bonding adhesives. Use $\alpha = .05$.
c. Are the assumptions of the test satisfied? If not, how does this impact the validity of the inference derived from the test?

9.34 *Detecting Alzheimer's disease at an early age.* Refer to the *Neuropsychology* (Jan. 2007) study of whether the cognitive effects of Alzheimer's disease can be detected at an early age, Exercise 9.19 (p. 453). Recall that a particular strand of DNA was classified into one of three genotypes: $E4^+/E4^+$, $E4^+/E4^-$, and $E4^-/E4^-$. In addition to a sample of 2,097 young adults (20-24 years), two other age groups were studied: a sample of 2,182 middle age adults (40-44 years) and a sample of 2,281 elderly adults (60-64 years). The accompanying table gives a breakdown of the number of adults with the three genotypes in each age category for the total sample of 6,560 adults. The researchers concluded that "there were no significant genotype differences across the three age groups" using $\alpha = .05$. Do you agree?

**E4E4ALL**

| Age Group | $E4^+/E4^+$ Genotype | $E4^+/E4^-$ Genotype | $E4^-/E4^-$ Genotype | Sample size |
|---|---|---|---|---|
| 20-24 | 56 | 517 | 1524 | 2,097 |
| 40-44 | 45 | 566 | 1571 | 2,182 |
| 60-64 | 48 | 564 | 1669 | 2,281 |

*Source:* Jorm, A.F., et al. "APOE Genotype and Cognitive Functioning in a Large Age-Stratified Population Sample", *Neuropsychology* , Vol. 21, No. 1, January 2007 (Table 1).

9.35 *Double-blind drug study.* Seldane-D, produced by Marion Merrell Dow, Inc., is an over-the-counter drug designed to relieve sneezing, nasal congestion, and other symptoms of allergic rhinitis. General adverse effects of Seldane-D were investigated in a double-blind, controlled study of over 500 patients suffering from allergic rhinitis. A sample of 374 patients were given Seldane-D, whereas a second sample of 193 patients were given a placebo (no drug). The number of patients reporting insomnia in each of the two groups are given in the table. Test to determine whether the proportion of patients taking Seldane-D who experience insomnia differs from the corresponding proportion for patients receiving the placebo. $\alpha = .10$.

**SELDANED**

| | Seldane-D | Placebo |
|---|---|---|
| *Insomnia* | 97 | 12 |
| *No Insomnia* | 277 | 181 |
| *TOTALS* | 374 | 193 |

*Source*: Marion Merrell Dow, Incorporated. Prescription Products Division.

## 9.6 Exact Tests for Independence in a Contingency Table Analysis (*Optional*)

The procedure for testing independence in a contingency table in Sections 9.4 and 9.5 is an "approximate" test due to the fact that the $\chi^2$ test statistic has an approximate chi-square probability distribution. The larger the sample, the better the test's approximation. For this reason, the test is often called an *asymptotic* test. For small samples (e.g., samples that produce contingency tables with one or more cells that have an expected number less than 5), the *p*-value from the asymptotic chi-square test may not be a good estimate of the actual (exact) *p*-value of the test. In this case, we can employ a technique proposed by R. A. Fisher (1935).

For $2 \times 2$, or, more general $2 \times c$ contingency tables, Fisher developed a procedure for computing the exact *p*-value for the test of independence—called *Fisher's exact test*. The method, which utilizes the hypergeometric probability distribution of Chapter 4 (p. 146), is illustrated in the next example.

**Example 9.6**

Exact $\chi^2$-Test: Vaccine Trial

New, effective, AIDS vaccines are now being developed using the process of "sieving," i.e., sifting out infections with some strains of HIV. A Harvard School of Public Health statistician demonstrated how to test the efficacy of an HIV vaccine in *Chance* (Fall 2000). Table 9.8 gives the results of a preliminary HIV vaccine trial in a $2 \times 2$ contingency table. The vaccine was designed to eliminate a particular strain of the virus, called the "MN strain." The trial consisted of 7 AIDS patients vaccinated with the new drug and 31 AIDS patients who were treated with a placebo (no vaccination). The table shows the number of patients who tested positive and negative for the MN strain in the trial follow-up period.

a. Conduct a test to determine whether the vaccine is effective in treating the MN strain of HIV. Use $\alpha = .05$.

b. Are the assumptions for the test, part **a**, satisfied?

c. Consider the hypergeometric probability

$$\frac{\binom{7}{2}\binom{31}{22}}{\binom{38}{24}}$$

**HIVVAC1**

**TABLE 9.8 Contingency Table for Example 9.6**

| | MN Strain | | |
|---|---|---|---|
| Patient Group | Positive | Negative | TOTALS |
| Unvaccinated | 22 | 9 | 31 |
| Vaccinated | 2 | 5 | 7 |
| TOTALS | 24 | 14 | 38 |

*Source*: Gilbert, P. "Developing an AIDS vaccine by sieving." *Chance*, Vol. 13, No. 4, Fall 2000.

**TABLE 9.9 Alternative Contingency Tables for Example 9.6**

HIVVAC2

a.

| | MN Strain | | |
|---|---|---|---|
| Patient Group | Positive | Negative | TOTALS |
| Unvaccinated | 23 | 8 | 31 |
| Vaccinated | 1 | 6 | 7 |
| TOTALS | 24 | 14 | 38 |

HIVVAC3

b.

| | MN Strain | | |
|---|---|---|---|
| Patient Group | Positive | Negative | TOTALS |
| Unvaccinated | 24 | 7 | 31 |
| Vaccinated | 0 | 7 | 7 |
| TOTALS | 24 | 14 | 38 |

This represents the probability that 2 out of 7 vaccinated AIDS patients test positive and 22 out of 31 unvaccinated patients test positive, i.e., the probability of the table result given the null hypothesis of independence is true. Compute this probability (called the ***probability of the contingency table***).

d. Refer to part **c**. Two contingency tables (with the same marginal totals as the original table) that are more contradictory to the null hypothesis of independence than the observed table are shown in Tables 9.9a and 9.9b. Explain why these tables provide more evidence to reject $H_0$ than the original table; then, compute the probability of each table using the hypergeometric formula.

e. The $p$-value of Fisher's exact test is the probability of observing a result at least as contradictory to the null hypothesis as the observed contingency table, given the same marginal totals. Sum the probabilities of parts **c** and **d** to obtain the $p$-value of Fisher's exact test. Interpret this value in the context of the vaccine trial.

Solution

a. If the vaccine is effective in treating the MN strain of HIV, then the proportion of positive HIV patients in the vaccinated group will be smaller than the corresponding proportion for the unvaccinated group. That is, the two variables, patient group and strain test result, will be dependent. Consequently, we conducted a chi-square test for independence on the data of Table 9.8. A MINITAB printout of the analysis is displayed in Figure 9.6. The approximate $p$-value of the test (highlighted on the printout) is .036. Since this value is less than $\alpha = .05$, we reject the null hypothesis of independence and conclude that the vaccine has an effect on the proportion of patients who test positive for the MN strain of HIV.

b. The asymptotic chi-square test of part **a** is a large-sample test. Our assumption is that the sample will be large enough so that the expected cell counts are all greater than or equal to 5. These expected cell counts are highlighted on the MINITAB printout, Figure 9.6. Note that two cells have expected numbers that are less than 5. Consequently, the large-sample assumption is not satisfied. Therefore, the $p$-value produced from the test may not be a reliable estimate of the true $p$-value.

c. Using the hypergeometric distribution, the probability of the contingency table is determined as follows:

$$\frac{\binom{7}{2}\binom{31}{22}}{\binom{38}{24}} = \frac{\dfrac{7!}{2!\,5!}\,\dfrac{31!}{22!\,9!}}{\dfrac{38!}{24!\,14!}} = \frac{(21)(20{,}160{,}075)}{9{,}669{,}554{,}100} = .04378$$

**Tabulated statistics: GROUP, STRAIN**

```
Using frequencies in NUMBER


Rows: GROUP   Columns: STRAIN

           NEG     POS     All

UNVACC       9      22      31
         11.42   19.58   31.00

VACC         5       2       7
          2.58    4.42    7.00

All         14      24      38
         14.00   24.00   38.00

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 4.411, DF = 1, P-Value = 0.036
Likelihood Ratio Chi-Square = 4.289, DF = 1, P-Value = 0.038

* NOTE * 2 cells with expected counts less than 5
```

**FIGURE 9.6**
MINITAB analysis of Table 9.8

d. The two contingency tables in Table 9.9 both show fewer vaccinated patients who test positive (1 and 0 patients, respectively) than the contingency table in Tables 9.8 (2 patients). Thus, the proportion of vaccinated patients who test positive in these alternative tables ($1/24 = .042$ and $0/24 = .000$, respectively) is smaller than the corresponding proportion in the actual study $(2/24 = .083)$. Consequently, the difference between the proportions of vaccinated and unvaccinated patients who test positive will be greater in the alternative tables than in the original table. Since the null hypothesis of independence implies that the proportion of patients who test positive will be the same for both patient groups, these two contingency tables provide more evidence to reject $H_0$ than the original table.

The probability of the Table 9.9a result given the null hypothesis of independence is true, (i.e., the probability of the contingency table in Table 9.9a,) is

$$\frac{\binom{7}{1}\binom{31}{23}}{\binom{38}{24}} = \frac{\frac{7!}{1!\,6!}\,\frac{31!}{23!\,8!}}{\frac{38!}{24!\,14!}} = \frac{(7)(7{,}888{,}725)}{9{,}669{,}554{,}100} = .00571$$

Similarly, the probability of the contingency table in Tables 9.9b, is

$$\frac{\binom{7}{0}\binom{31}{24}}{\binom{38}{24}} = \frac{\frac{7!}{0!\,7!}\,\frac{31!}{24!\,7!}}{\frac{38!}{24!\,14!}} = \frac{(1)(2{,}629{,}575)}{9{,}669{,}554{,}100} = .00027$$

e. To obtain the $p$-value of Fisher's exact test, we sum contingency table probabilities for all possible contingency tables that give a result *at least* as contradictory to the null hypothesis as the observed contingency table. Since the contingency tables in Table 9.9 are the only two possible tables that give a more contradictory result, we add their hypergeometric probabilities to the hypergeometric probability for Tables 9.8 to obtain the exact $p$-value for a test of independence: $p$-value $= .04378 + .00571 + .00027 = .04976$. Since this exact $p$-value is less than $\alpha = .05$, we reject the null hypothesis of independence; there is sufficient evidence to reliably conclude that the vaccine is effective in treating the MN strain of HIV.

Fisher's exact $p$-value for this test, $p$-value $= .04976 \approx .050$, can be more easily obtained using statistical software. It is shown (highlighted) under the "Exact Sig (1-sided)" column at the bottom of the SPSS printout in Figure 9.7. (*Note*: The exact $p$-value for a two-tailed test of independence is also shown on the SPSS printout. Its value, .077, is obtained by adding the hypergeometric probability for a fourth contingency table, one with 17 unvaccinated patients testing positive and 7 vaccinated patients testing positive, to the one-tailed exact $p$-value. This table was not considered in the solution to the problem since it results in sample proportions that contradict the alternative hypothesis that the proportion of positive HIV patients in the vaccinated group is less than the proportion in the unvaccinated group.)

Fisher's exact test for a $2 \times 2$ contingency table is summarized in the box. For details on the methodology for a more general $2 \times c$ contingency table, consult the references for this chapter.

**GROUP * MNSTRAIN Crosstabulation**

| | | | MNSTRAIN | | Total |
|---|---|---|---|---|---|
| | | | NEG | POS | |
| GROUP | UNVAC | Count | 9 | 22 | 31 |
| | | Expected Count | 11.4 | 19.6 | 31.0 |
| | VACC | Count | 5 | 2 | 7 |
| | | Expected Count | 2.6 | 4.4 | 7.0 |
| Total | | Count | 14 | 24 | 38 |
| | | Expected Count | 14.0 | 24.0 | 38.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 4.411[b] | 1 | .036 | | |
| Continuity Correction[a] | 2.777 | 1 | .096 | | |
| Likelihood Ratio | 4.289 | 1 | .038 | | |
| Fisher's Exact Test | | | | .077 | .050 |
| N of Valid Cases | 38 | | | | |

a. Computed only for a 2x2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.58.

**FIGURE 9.7**
SPSS contingency table analysis for Table 9.8

### Fisher's Exact Test for Independence in a 2 × 2 Contingency Table

Suppose you observe a 2 × 2 contingency table of the form

| | *Column 1* | *Column 2* | **Row Total** |
|---|---|---|---|
| *Row 1* | $n_{11}$ | $n_{12}$ | $n_{1\bullet}$ |
| *Row 2* | $n_{21}$ | $n_{22}$ | $n_{2\bullet}$ |
| ***Column Total*** | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n$ |

*Step 1* Use the formula for the hypergeometric distribution to find the *probability of the observed contingency table*:

$$\text{Probability} = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{21}}}{\binom{n}{n_{\bullet 1}}} = \frac{\binom{n_{\bullet 1}}{n_{11}}\binom{n_{\bullet 2}}{n_{12}}}{\binom{n}{n_{1\bullet}}}$$

*Step 2* Construct all possible 2 × 2 contingency tables that have the same marginal totals as the observed table.

*Step 3* Use the hypergeometric formula to find the probability of each contingency table in step 2. The contingency tables with probabilities less than or equal to the probability of the observed table are at least as contradictory to the null hypothesis of independence as the observed table.

*Step 4* Sum the probabilities of all contingency tables that are at least as contradictory to the null hypothesis of independence as the observed table. (*Note*: Include the probability of the observed table in the sum.) This sum represents Fisher's exact $p$-value for a *two-tailed test*.

## Applied Exercises

9.36 *Drinking-water quality study.* Refer to the *Disasters* (Vol. 28, 2004) study of the effects of a tropical cyclone on the quality of drinking water on a remote Pacific island, Exercise 1.11 (p. 7). One part of the study evaluated the usefulness of a simple paper-strip, hydrogen sulphide ($H_2S$) test kit for water quality in determining the presence of fecal bacteria. (*Note*: The $H_2S$ test paper is designed to turn black when fecal bacteria is present in the water.) Each in a sample of 17 water specimens (size 500 milliliters) obtained 3 days after Cyclone Ami hit the island was tested for fecal bacteria. Both the conventional fecal coliform test and the simple $H_2S$ test were applied to each water specimen. The test results are summarized in the table.

**H2STEST**

| | | Bacteria Detected in Conventional Test | |
|---|---|---|---|
| | | Yes | No |
| | *Blackened* | 7 | 4 |
| *$H_2S$ Test Result* | *Not Blackened* | 0 | 6 |

*Source*: Mosley, L., Sharp, D. and Singh, S. "Effects of a tropical cyclone on the drinking-water quality of a remote Pacific island." *Disasters*, Vol. 28, No. 4, 2004 (from Table 3).

a. Explain why Fisher's exact test should be used to determine whether the $H_2S$ test result depends on whether or not bacteria is present in the water specimen.
b. Construct all possible contingency tables with the same marginal totals as the observed table.
c. Use the hypergeometric formula to find the probability of each of the tables, part **b**, occurring. Identify the tables that have probabilities less than or equal to the probability of the observed table. (These are the tables that provide more convincing evidence to reject the null hypothesis of independence than the observed table.)
d. Sum the hypergeometric probabilities of the tables identified in part **c**. This sum represents the $p$-value of Fisher's exact test.
e. The researchers conclude that "the $H_2S$ test showed good agreement with the conventional fecal coliform test." Do you agree? Test using $\alpha = .10$.

9.37 *A new dental bonding agent.* Refer to the *Trends in Biomaterials & Artificial Organs* (Jan. 2003) study of a new bonding adhesive for teeth, Exercise 9.33 (p. 466). Recall that the new adhesive (called Smartbond) was compared to the standard composite adhesive. The Adhesive Remnant Index (ARI) scores for 10 teeth bonded with the new adhesive and 10 teeth bonded with the composite adhesive were measured. The contingency table for the data is reproduced here.

**BONDING**

| | Adhesive Remnant Index Score | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| *Smartbond* | 2 | 8 | 0 | 0 | 0 |
| *Composite* | 1 | 5 | 3 | 1 | 0 |

*Source*: Sunny, J., and Vallathan, A. "A comparative *in vitro* study with new generation ethyl cyanoacrylate (Smartbond) and a composite bonding agent." *Trends in Biomaterials & Artificial Organs*, Vol. 16, No. 2, Jan. 2003 (Table 6).

a. Explain why Fisher's exact test for independence can (and should) be applied to this contingency table.
b. A SAS printout of the contingency table analysis is shown below. Use the information on the printout to conduct Fisher's exact test at $\alpha = .05$.

**SWDEFECTS**

9.38 *Software defects.* Refer to the study on predicting defects in software code written in C language for a NASA spacecraft instrument, Exercise 9.29 (p. 461). The SPSS contingency table for the two categorical variables, actual defective status and predicted defective status using EVG, is reproduced at the top of p. 473.

a. Show that there are 11 possible contingency tables (including the observed table) with the same marginal totals as the observed table.
b. Use the hypergeometric formula to find the probability of each of the 11 tables in part **a**.
c. Use the probabilities, part **b**, to find the $p$-value of Fisher's exact test for independence. Verify your calculations by checking the $p$-value shown on the SPSS printout.
d. Since the sample size is large, the $p$-value for the asymptotic chi-square test should be approximately equal to Fisher's exact test $p$-value. Is this true?

## SAS Output for Exercise 9.37

The FREQ Procedure

Table of ADHESIVE by ARI

| ADHESIVE / ARI<br>Frequency<br>Expected | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| COMPOSITE | 1<br>1.5 | 5<br>6.5 | 3<br>1.5 | 1<br>0.5 | 10 |
| SMARTBOND | 2<br>1.5 | 8<br>6.5 | 0<br>1.5 | 0<br>0.5 | 10 |
| Total | 3 | 13 | 3 | 1 | 20 |

Statistics for Table of ADHESIVE by ARI

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 3 | 5.0256 | 0.1699 |
| Likelihood Ratio Chi-Square | 3 | 6.5836 | 0.0864 |
| Mantel-Haenszel Chi-Square | 1 | 3.4898 | 0.0617 |
| Phi Coefficient | | 0.5013 | |
| Contingency Coefficient | | 0.4481 | |
| Cramer's V | | 0.5013 | |

WARNING: 75% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

| | |
|---|---|
| Table Probability (P) | 0.0209 |
| Pr <= P | 0.2616 |

Sample Size = 20

**SPSS Output for Exercise 9.38**

**DEFECT * PRED_EVG Crosstabulation**

Count

| | | PRED_EVG | | |
|---|---|---|---|---|
| | | no | yes | Total |
| DEFECT | false | 441 | 8 | 449 |
| | true | 47 | 2 | 49 |
| Total | | 488 | 10 | 498 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 1.188[b] | 1 | .276 | | |
| Continuity Correction[a] | .306 | 1 | .580 | | |
| Likelihood Ratio | .948 | 1 | .330 | | |
| Fisher's Exact Test | | | | .257 | .257 |
| N of Valid Cases | 498 | | | | |

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is .98.

9.39 *Job satisfaction of women in construction.* Refer to the *Journal of Professional Issues in Engineering Education & Practice* (April 2013) study of the job satisfaction of members of the National Association of Women in Construction (NAWIC), Exercise 9.24 (p. 459). The results for the survey of 477 women are reproduced in the accompanying table. Use statistical software to conduct an exact test to determine if an NAWIC member's satisfaction with life as an employee is related to their satisfaction with job challenge. Test using $\alpha = .05$.

**NAWIC**

| | | Life as an Employee | |
|---|---|---|---|
| | | Satisfied | Dissatisfied |
| Job | Satisfied | 364 | 33 |
| Challege | Dissatisfied | 24 | 26 |

*Source:* Malone, E.K. & Issa, R.A. "Work-Life Balance and Organizational Commitment of Women in the U.S. Construction Industry", *Journal of Professional Issues in Engineering Education & Practice*, Vol. 139, No. 2, April 2013 (Table 11).

## *STATISTICS IN ACTION REVISITED*

### The Case of the Ghoulish Transplant Tissue

We return to the case involving tainted transplant tissue (see p. 474). Recall that a processor of the tainted tissue filed a lawsuit against a tissue distributor, claiming that the distributor was more responsible for paying damages to litigating transplant patients. Why? Because the distributor in question had sent recall notices (as required by the FTC) to hospitals and surgeons with unsolicited newspaper articles describing in graphic detail the "ghoulish" acts that had been committed. According to the processor, by including the articles in the recall package this distributor inflamed the tissue recipients, increasing the likelihood that the patient would file a lawsuit.

To prove its case in court, the processor needed to establish a statistical link between the likelihood of a lawsuit and the sender of the recall notice. More specifically, can the processor show that the probability of a lawsuit is higher for those patients of surgeons who received the recall notice with the inflammatory articles than for those patients of surgeons who only received the recall notice?

A statistician, serving as an expert consultant for the processor, reviewed data for the 7,914 patients who received recall notices (of which 708 filed suit). These data are saved in the **GHOUL1** file. For each patient, the file contains information on the SENDER of the recall notice (Processor or Distributor) and whether or not a LAWSUIT was filed (Yes or No). Since both of these variables are qualitative, and we want to know whether the probability of a LAWSUIT depends on the SENDER of the recall notice, a contingency table analysis is appropriate.

Figure SIA9.1 shows the MINITAB contingency table analysis. The null and alternative hypotheses for the test are

$H_0$: Lawsuit and Sender are independent $\qquad$ $H_a$: Lawsuit and Sender are dependent

Both the chi-square test statistic (100.5) and *p*-value of the test (.000) are highlighted on the printout. If we conduct the test at $\alpha = .01$, there is sufficient evidence to reject $H_0$. That is, the data provide evidence to indicate that the likelihood of a tainted transplant patient filing a lawsuit is associated with the sender of the recall notice.

To determine which sender had the highest percentage of patients to file a lawsuit, examine the row percentages (highlighted) in the contingency table of Figure SIA9.1 You can see that of the 1,751 patients sent recall notices by the processor, 51 (or 2.91%) filed lawsuits. In contrast, of the 6,163 patients sent recall notices by the distributor in question, 657 (or 10.66%) filed lawsuits. Thus, the probability of a patient filing a lawsuit is almost five times higher for the distributor's patients than for the processor's patients.

Before testifying on these results in court, the statistician decided to do one additional analysis: he eliminated from the sample data any patients whose surgeon had been sent notices by both parties. Why? Since these patient's surgeons received both recall notices, the underlying reason for filing a lawsuit would be unclear. Did the patient file simply because he or she received tainted transplant tissue, or was the filing motivated by the inflammatory articles that accompanied the recall notice? After eliminating these

**Tabulated statistics: SENDER, LAWSUIT**

```
Rows: SENDER   Columns: LAWSUIT

                 No     Yes     All

Distributor    5506     657    6163
              89.34   10.66  100.00
               5612     551    6163
              1.989  20.244       *

Processor      1700      51    1751
              97.09    2.91  100.00
               1594     157    1751
              7.001  71.252       *

All            7206     708    7914
              91.05    8.95  100.00
               7206     708    7914
                  *       *       *

Cell Contents:      Count
                    % of Row
                    Expected count
                    Contribution to Chi-square


Pearson Chi-Square = 100.485, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 124.748, DF = 1, P-Value = 0.000
```

**FIGURE SIA9.1**
MINITAB Contingency Table Analysis—Likelihood of Lawsuit vs. Recall Notice Sender

**TABLE SIA9.2 Data for the Tainted Tissue Case, Dual Recall Notices Eliminated**

| Recall notice sender | Number of Patients | Number of lawsuits |
|---|---|---|
| Processor/Other Distributor | 1,522 | 31 |
| Distributor in question | 5,705 | 606 |
| Totals: | 7,227 | 637 |

patients, the data looked like that shown in Table SIA9.2. A MINITAB contingency table analysis on this reduced data set (saved in the **GHOUL2** file) is shown in Figure SIA9.2.

Like in the previous analysis, the chi-square test statistic (110.2) and *p*-value of the test (.000) — both highlighted on the printout — imply that the likelihood of a tainted transplant patient filing a lawsuit is associated with the sender of the recall notice, at $\alpha$=.01. Also, the percentage of patients filing lawsuits when sent a recall notice by the distributor (10.62%) is again five times higher than the percentage of patients filing lawsuits when sent a recall notice by the processor (2.04%).

The results of both analyses were used to successfully support the processor's claim in court. Nonetheless, we need to point out one caveat to the contingency table analyses. Be careful not to conclude that the data are proof that the inclusion of the inflammatory articles *caused* the probability of litigation to increase. Without controlling all possible variables that may related to filing a lawsuit (e.g., a patient's socioeconomic status, whether or not a patient has filed a lawsuit in the past), we can only say that the two qualitative variables, lawsuit status and recall notice sender, are statistically associated. However, the fact that the likelihood of a lawsuit is almost five times higher when the notice is sent by the distributor shifts the burden of proof to the distributor to explain why this occurred and to convince the court that it should not be held accountable for paying the majority of the damages.

**Tabulated statistics: SENDER, LAWSUIT**

```
Rows: SENDER   Columns: LAWSUIT

                   No     Yes     All

Distributor      5099     606    5705
                89.38   10.62  100.00
                 5202     503    5705
                2.045  21.160       *

Processor        1491      31    1522
                97.96    2.04  100.00
                 1388     134    1522
                7.667  79.315       *

All              6590     637    7227
                91.19    8.81  100.00
                 6590     637    7227
                    *       *       *

Cell Contents:      Count
                    % of Row
                    Expected count
                    Contribution to Chi-square


Pearson Chi-Square = 110.187, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 144.862, DF = 1, P-Value = 0.000
```

**FIGURE SIA9.2**
MINITAB Contingency Table Analysis, with Dual Recall Notices Eliminated

**FIGURE SIA9.3**
MINITAB Output with 95% Confidence Interval and Test for Difference in Proportions of Lawsuits Filed

**Test and CI for Two Proportions**

```
Sample    X     N  Sample p
1        31  1522  0.020368
2       606  5705  0.106223


Difference = p (1) - p (2)
Estimate for difference:  -0.0858547
95% CI for difference:  (-0.0965452, -0.0751641)
Test for difference = 0 (vs not = 0):  Z = -10.50  P-Value = 0.000

Fisher's exact test: P-Value = 0.000
```

*Alternative Analysis*: As mentioned in Section 9.4, a 2 $\times$ 2 contingency table analysis is equivalent to a comparison of two population proportions. In the tainted tissue case, we want to compare $p_1$, the proportion of lawsuits filed by patients who were sent recall notices by the processor, to $p_2$, the proportion of lawsuits filed by patients who were sent recall notices by the distributor who included the inflammatory articles. Both a test of the null hypothesis, $H_0$: $(p_1 - p_2) = 0$ and a 95% confidence interval for the difference, $(p_1 - p_2)$ using the reduced sample data are shown (highlighted) on the MINITAB printout, Figure SIA9.3.

The $p$-value for the test (.000) indicates that the two proportions are significantly different at $\alpha$.=05. The 95% confidence interval, (−.097, −.075), shows that the proportion of lawsuits associated with patients who were sent recall notices from the distributor ranges between .075 and .097 higher than the corresponding proportion for the processor. Both results support the processor's case, namely, that the patients who were sent recall notices with the inflammatory news articles were more likely to file a lawsuit than those who were sent only recall notices.

# Quick Review

## Key Terms

[*Note*: Items marked with an asterisk (*) are from the optional section in this chapter.]

Chi-square distribution 444
Contingency table 453
Dependence of two classifications 454
Expected category count 448
*Fisher's exact test 471
Fixed marginals 463
Marginal probabilities 454
Multinomial experiment 444
Observed cell count 449
One-way table 444
*Probability of the contingency table 468
Two-way table 453

## Key Formulas

One-way Table

*Confidence Interval for $p_i$*: $\hat{p}_i \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_i(1-\hat{p}_i)}{n}}$ 000

*Confidence Interval for $p_i - p_j$*: $(\hat{p}_i - \hat{p}_j) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}}$ 000

*Test Statistic*:

$$\chi^2 = \sum\frac{[n_i - E(n_i)]^2}{E(n_i)}$$ 000

where $n_i$ = count for cell $i$

$E(n_i) = np_{i,0}$

$p_{i,0}$ = hypothesized value of $p_i$ in $H_0$

Two-way Table

*Test Statistic*:

$$\chi^2 = \sum \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})} \qquad 000, 000$$

$$\text{*Probability of a } 2 \times 2 \text{ contingency table: } p = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{21}}}{\binom{n}{n_{\bullet 1}}} \qquad 000$$

where $n_{ij}$ = count for cell in row $i$, column $j$

$\hat{E}(n_{ij}) = n_{i\bullet}n_{\bullet j}/n$

$n_{i\bullet}$ = total for row $i$

$n_{\bullet j}$ = total for column $j$

$n$ = total sample size

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $p_{i,0}$ | *p-i*-zero | Value of multinomial probability $p_i$ hypothesized in $H_0$ |
| $\chi^2$ | Chi-square | Test statistic used in analysis of count data |
| $n_i$ | *n-i* | Number of observed outcomes in cell $i$ of one-way table |
| $E(n_i)$ | *e-n-i* | Expected number of outcomes in cell $i$ of one-way table when $H_0$ is true |
| $p_{ij}$ | *p-i-j* | Probability of an outcome in row $i$ and column $j$ of a two-way contingency table |
| $n_{ij}$ | *n-i-j* | Number of observed outcomes in row $i$ and column $j$ of a two-way contingency table |
| $\hat{E}(n_{ij})$ | Estimated *e-n-i-j* | Estimated expected number of outcomes in row $i$ and column $j$ of a two-way contingency table |
| $n_{i\bullet}$ | *n-i*-dot | Total number of outcomes in row $i$ of a contingency table |
| $n_{\bullet j}$ | *n*-dot-*j* | Total number of outcomes in column $j$ of a contingency table |

## Chapter Summary Notes

- **Multinomial data** are qualitative data that fall into more than two *categories, classes*, or *cells*.
- Properties of a **multinomial experiment**: (1) $n$ identical trials, (2) $k$ possible outcomes to each trial, (3) probabilities of the $k$ outcomes remain the same from trial to trial, (4) trials are independent, (5) the variables of interest are the *cell counts*.
- A **one-way table** is a summary table for a single qualitative variable.
- A **two-way table**, or **contingency table**, is a summary table for two qualitative variables.
- The **chi-square ($\chi^2$) statistic** is used to test probabilities associated with one-way and two-way tables.
- Conditions required for a valid **$\chi^2$-test**: (1) multinomial experiment, (2) sample size $n$ is large—satisfied when expected cell counts are all greater than or equal to 5.
- A significant $\chi^2$-test for a *two-way table* implies that the **two qualitative variables are dependent**.
- Chi-square tests for independence *cannot be used to infer that a causal relationship* exists between the two qualitative variables.
- **Fisher's exact test** can be applied to 2×2 or more general 2×$c$ contingency tables.

## Supplementary Applied Exercises

[*Note*: Exercises marked with an asterisk (*) require methods from the optional section in this chapter.]

**TURN**

9.40 *Turns at intersections.* A traffic study found that of 972 automobiles entering a busy intersection during the period from 4 P.M. to 7 P.M., 357 turned left, 321 turned right, and 294 drove straight through the intersection. (These results are saved in the **TURN** file.)

a. Construct a one-way table for the study.

b. Find a 95% confidence interval for the true proportion of automobiles that drive straight through the intersection during this period.

c. Find a 95% confidence interval for the difference between the proportions of automobiles that turn left and turn right, respectively, during this period. Interpret the interval.

d. Do the data disagree with the hypothesis that the traffic is equally divided among the three directions? Test using $\alpha = .05$.

e. Do the data provide sufficient evidence to indicate that more than one-third of all automobiles entering the intersection turn left? Test using $\alpha = .05$.

9.41 *Compressed work weeks. Compressed work weeks* are defined as "alternative work schedules in which a trade is made between the number of hours worked per day, and the number of days worked per week, in order to work the standard number of weekly hours in less than 5 days." A field study was conducted at a large, midwestern, continuous-processing (7 days/24 hours) chemical plant that had experimented with four different work schedules, two of which were compressed:

| | |
|---|---|
| Three 8-hour fixed shifts (day, evening, midnight) | Two 12-hours fixed shifts (12 A.M.–12 P.M., 12 P.M.–12 A.M.) |
| Three 8-hour rotating shifts | Two 12-hour rotating shifts |

Six hundred seventy-one hourly employees were asked to rank the four work schedules in order of preference. The accompanying table gives the number of first-place rankings for each schedule. Is there sufficient evidence to indicate that the hourly employees have a preference for one of the work schedules? Test using $\alpha = .01$.

**WORKSCHED**

| 8-hour fixed | 8-hour rotating | 12-hour fixed | 12-hour rotating |
|---|---|---|---|
| 389 | 54 | 208 | 20 |

9.42 *Fugitive dust plumes.* Fugitive dust plumes generated by farm equipment can be hazardous to human health. In the *Journal of Agricultural, Biological, and Environmental Sciences* (Mar. 2001), environmental engineers developed a model for dust particle concentrations in plumes produced by a tractor operating in a wheat field. The tractor traveled along six parallel, equilength paths in the field. A remote sensing instrument with a laser beam, placed at the edge of the field, measured the particulate matter in the dust every .5 seconds. Unfortunately, a few of the measurements were censored (i.e., higher than the signal level of the instrument). This usually occurred when the tractor was a short distance from the instrument's laser beam. The next table shows the number of censored measurements for each of the six tractor lines.

a. Calculate and compare the sample proportion of censored measurements for the six tractor lines.

b. Do the data provide sufficient evidence to indicate that the proportion of censored measurements differs for the six tractor lines? Test using $\alpha = .01$.

c. Comment on the practical versus statistical significance of the test.

**DUSTCENSOR**

| Tractor Line | Uncensored Measurements | Censored Measurements | TOTALS |
|---|---|---|---|
| 1 | 6047 | 175 | 6222 |
| 2 | 4456 | 236 | 4692 |
| 3 | 6821 | 319 | 7140 |
| 4 | 5889 | 231 | 6120 |
| 5 | 9873 | 480 | 10,353 |
| 6 | 4607 | 187 | 4794 |
| TOTALS | 37,693 | 1628 | 39,321 |

*Source*: Johns, C., Holmen, B., Niemeier, A., and Shumway, R., "Nonlinear regression for modeling censored one-dimensional concentration profiles of fugitive dust plumes." *Journal of Agricultural, Biological, and Environmental Sciences*, Vol. 6, No. 1, March 2001 (from data file provided by coauthor Brit Holmen).

9.43 *Atomic weapons exposure.* Researchers at the Oak Ridge (Tennessee) National Laboratory have developed an algorithm to estimate the numbers of expected and excess cases of thyroid cancer occurring in the lifetime of those exposed to atomic weapons tests at the Nevada Test Site in the 1950s (*Health Physics*, Jan. 1986). Of the approximately 23,000 people exposed to the weapons-testing fallout, 58 were expected to develop thyroid cancer in their remaining lifetimes. According to the algorithm, the 58 cases can be categorized by sex and level of radiation (dose) at the time of exposure as shown in the table on p. 479. Suppose that the data represent a random sample of 58 thyroid cancer patients selected from the target population. Conduct a test to determine whether the two directions of classification, sex and dose at time of exposure, are independent. Use $\alpha = .01$.

**ATOMIC**

| | | Gender | | |
|---|---|---|---|---|
| | | Male | Female | TOTALS |
| | *Less than 1* | 6 | 13 | 19 |
| *Dose, rad* | *1–10* | 8 | 18 | 26 |
| | *11 or more* | 3 | 10 | 13 |
| *TOTALS* | | 17 | 41 | 58 |

*Source*: Zeighami, E. A., and Morris, M. D. "Thyroid cancer risk in the population around the Nevada test site." *Health Physics*, Vol. 50, No. 1, Jan. 1986, p. 26 (Table 2).

9.44 *Pesticide use in orchards.* Four pesticides used in dormant California orchards are chlorpyrifos, diazinon, methidathion, and parathion. *Environmental Science & Technology* (Oct. 1993) reported the number of applications of these spray chemicals over a 6-month period in California. The data for each of three types of fruit or nut orchards are shown in the accompanying table. (Parathion has since been banned for use on deciduous fruit and nut trees.)

a. Conduct a test to determine (at $\alpha = .01$) whether pesticide used depends on type of orchard.

b. Because of the large number of pesticide applications reported, the total sample size for the test, part **a**, is extremely large $(n = 417,697)$. Consequently, a "statistically significant" result may not be "practically significant." Perform an analysis to show the magnitude of differences in the rates of methidathion application for the three orchard types.

**PESTICIDE**

| | Fruit/Nut Trees | | |
|---|---|---|---|
| Chemical | Almonds | Peaches | Nectarines |
| *Chlorpyrifos* | 41,077 | 4,419 | 11,594 |
| *Diazinon* | 102,935 | 9,651 | 5,928 |
| *Methidathion* | 21,240 | 5,198 | 1,790 |
| *Parathion* | 136,064 | 53,384 | 24,417 |

*Source*: Selber, J. N., et al. "Air and fog deposition residues of four organophosphate insecticides used on dormant orchards in the San Joaquin Valley, California." *Environmental Science & Technology*, Vol. 27, No. 10, Oct. 1993, p. 2236 (Table 1).

9.45 *Trapping grain moths.* In an experiment described in the *Journal of Agricultural, Biological, and Environmental Statistics* (Dec. 2000), bins of corn were stocked with various parasites (e.g., grain moths) in late winter. In early summer (June), three bowl-shaped traps were placed on the grain surface in order to capture the moths. All three traps were baited with a sex pheromone lure: however, one trap used an unmarked sticky adhesive, one was marked with a fluorescent red powder, and one was marked with a fluorescent blue powder. The traps were set on a Wednesday and the catch collected the following Thursday and Friday. The table shows the number of moths captured in each trap on each day. Conduct a test (at $\alpha = .10$) to determine if the percentages of moths caught by the three traps depend on day of the week.

**MOTHTRAP**

| | Adhesive—No Mark | Red Mark | Blue Mark |
|---|---|---|---|
| *Thursday* | 136 | 41 | 17 |
| *Friday* | 101 | 50 | 18 |

*Source*: Wilcyto, E. P., et al. "Self-marking recapture models for estimating closed insect populations." *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 5, No. 4, December 2000 (Table 5A).

9.46 *Species hotspots.* Refer to the *Nature* (Sept. 1993) study of animal and plant species "hot spots" in Great Britain, Exercise 3.81 (p. 129). A hot spot is defined as a 10-km square area that is species-rich, i.e., that is heavily populated by the species of interest. Similarly, a cold spot is a 10-km square area that is species-poor. The following table gives the number of butterfly hot spots and number of butterfly cold spots in a sample of 2,588 10-km square areas. In theory, 5% of the areas should be butterfly hot spots, 5% should be butterfly cold spots, with the remaining areas (90%) neutral. Test the theory using $\alpha = .01$.

**HOTSPOTS**

| | |
|---|---|
| Butterfly Hot Spots | 123 |
| Butterfly Cold Spots | 147 |
| Neutral Areas | 2,318 |
| TOTAL | 2,588 |

*Source*: Prendergast, J. R., et al. "Rare species, the coincidence of diversity hotspots and conservation strategies." *Nature* Vol. 365, No. 6444, Sept. 23, 1993, p. 335 (Table 1).

9.47 *Irrigating crop land.* Refer to the survey of agricultural engineers, Exercise 9.8 (p. 404). Do the data present sufficient evidence to indicate a preference for one or more of the five water-management strategies? Test using $\alpha = .05$.

9.48 *Salamander snout wounds. Dear enemy recognition* (DER) is the term used by naturalists and ecologists for the aggressive behavior of birds, mammals, and ants when their territorial boundaries are violated by one of their own species. DER is often followed by escalated attacks on the invading animal. A study explored the possibility that the red-backed salamander employs DER by using chemical signals to distinguish familiar from unfamiliar salamanders. In escalated contests, a salamander will attempt to bite an opponent's snout—an injury that could reduce a salamander's ability to locate prey, mates, and territorial competitors. One part of the study focused on a comparison of the proportions of males and females exhibiting wounds in the snout. One hundred forty-four salamanders were collected from a forest, killed, and inspected for scar tissue in the snout. The results are shown in the table at the top of p. 480.

**DER**

| | Male | Female | TOTALS |
|---|---|---|---|
| *Scar tissue in snout* | 5 | 12 | 17 |
| *No scar tissue in snout* | 76 | 51 | 127 |
| *TOTALS* | 81 | 63 | 144 |

*Source*: Jaeger, R. G. "Dear enemy recognition and the costs of aggression between salamanders." *The American Naturalist*, June 1981, Vol. 117, pp. 962–973. Reprinted by permission of the University of Chicago Press. © 1981 The University of Chicago.

a. Use a chi-square test to determine whether there is a difference between the proportions of males and females with scar tissue in the snout. Use $\alpha = .01$.
b. Estimate the difference between the proportions of males and females with scar tissue in the snout. Use a 99% confidence interval. Interpret the result.
c. Apply Fisher's exact test to the data. Compare the results to the test, part **a**.

9.49 *Video time compression.* Video engineers use time compression to shorten the time required for broadcasting a television commercial. But can shorter commercials be effective? To answer this question, 200 college students were randomly divided into three groups. The first group (57 students) was shown a videotape of a television program that included a 30-second commercial; the second group (74 students) was shown the same videotape but with the 24-second time-compressed version of the commercial; and the third group (69 students) was shown a 20-second time-compressed version of the commercial. Two days after viewing the tape, the three groups of students were asked to name the brand that was advertised. The numbers of students recalling the brand name for each of the three groups are given in the table below.

a. Do the data provide sufficient evidence (at $\alpha = .05$) that the two directions of classification, type of commercial and recall of brand name, are dependent? Interpret your results.
b. Construct a 95% confidence interval for the difference between the proportions recalling brand name for viewers of normal and 24-second time-compressed commercials.

9.50 *Battle simulation trials.* In order to evaluate their situational awareness, fighter aircraft pilots participate in battle simulations. At a random point in the trial, the simulator is frozen and data on situation awareness are immediately collected. The simulation is then continued until, ultimately, performance (e.g., number of kills) is measured. A study reported in *Human Factors* (Mar. 1995) investigated whether temporarily stopping the simulation results in any change in pilot performance. Trials were designed so that some simulations were stopped to collect situation awareness data while others were not stopped. Each trial was then classified according to the number of kills made by the pilot. The data for 180 trials are summarized in the contingency table below. Conduct a contingency table analysis and fully interpret the results.

**SIMKILLS**

| | Number of Kills | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | Totals |
| *Stops* | 32 | 33 | 19 | 5 | 2 | 91 |
| *No Stops* | 24 | 36 | 18 | 8 | 3 | 89 |
| *Totals* | 56 | 69 | 37 | 13 | 5 | 180 |

9.51 *Decision support system.* A decision support system (DSS) is a computerized system designed to aid in the management and analysis of large data sets. Ideally, a DSS should include four components: (1) a data extraction system, (2) a relational database organization, (3) analysis models, and (4) a user-friendly interactive dialogue between the user and the system. A state highway agency recently installed a DSS to help monitor data on road construction contract bids. As part of a self-examination, the agency selected 151 of the most recently encountered problems that could be traced directly to the DSS and classified each according to the component of origination. Can it be concluded from the data in the table that the proportions of problems are different for at least two of the four DSS components? Test using $\alpha = .05$.

**DSS**

| *Component* | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| *Number of Problems* | 31 | 28 | 45 | 47 |

**TIMECOMP**

| | | Type of Commercial | | | |
|---|---|---|---|---|---|
| | | Normal Version (30 Seconds) | Time-Compressed Version 1 (24 Seconds) | Time-Compressed Version 2 (20 Seconds) | TOTALS |
| *Recall of* | *Yes* | 15 | 32 | 10 | 57 |
| *Brand Name* | *No* | 42 | 42 | 59 | 143 |
| *TOTALS* | | 57 | 74 | 69 | 200 |

*9.52 *Characteristics of radio receivers.* An experiment was conducted to compare the fidelity and selectivity of radio receivers. Thirty receivers were tested and classified as low or high in each of the two categories. Do the data in the table provide sufficient evidence to indicate a dependence between fidelity and selectivity? Use Fisher's exact test at $\alpha = .05$.

**RADIO**

| | | Selectivity | |
|---|---|---|---|
| | | *Low* | *High* |
| *Fidelity* | *Low* | 10 | 6 |
| | *High* | 12 | 2 |

9.53 *Gastroenteritis outbreak.* A waterborne nonbacterial gastroenteritis outbreak occurred in Colorado as a result of a long-standing filter deficiency and malfunction of a sewage treatment plant. A study was conducted to determine whether the incidence of gastrointestinal disease during the epidemic was related to water consumption (*American Water Works Journal*, Jan. 1986). A telephone survey of households yielded the accompanying information on daily consumption of 8-ounce glasses of water for a sample of 40 residents who exhibited gastroenteritis symptoms during the epidemic.

**GASTRO**

| | Daily Consumption of 8-Ounce Glasses of Water | | | | |
|---|---|---|---|---|---|
| | 0 | 1–2 | 3–4 | 5 or more | Total |
| *Number of respondents with symptoms* | 6 | 11 | 13 | 10 | 40 |

*Source*: Hopkins, R. S., et al. "Gastroenteritis: Case study of a Colorado outbreak." *American Water Works Journal*, Vol. 78, No. 1, Jan. 1986, p. 42 (Table 1). Copyright © 1986, American Water Works Association. Reprinted by permission.

a. Use a 99% confidence interval to estimate the percentage of gastroenteritis cases who drink 1–2 glasses of water per day.
b. Use a 99% confidence interval to estimate the difference between the percentages of gastroenteritis cases who drink 1–2 and 0 glasses of water per day.
c. Conduct a test to determine whether the incidence of gastrointestinal disease during the epidemic is related to water consumption. Use $\alpha = .01$.

9.54 *Accident rate of workers.* Does the propensity for worker injuries depend on the length of time that a worker has been on the job? An analysis of 714 worker injuries by one manufacturer gave the results shown in the table for the distribution of injuries over the eight 1-hour time periods per shift.

**ACCIDENTS**

| *Hour of Shift* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Number of Accidents* | 93 | 71 | 79 | 72 | 98 | 89 | 102 | 110 |

a. Do the data imply that the probabilities of worker accidents are higher in some time periods than in others? Test using $\alpha = .10$.
b. Do the data provide sufficient evidence to indicate that the probability of an accident during the last 4 hours of a shift is greater than during the first 4 hours? Test using $\alpha = .10$. (*Hint*: Test $H_0$: $p_1 = .5$, where $p_1$ is the probability of an accident during the last 4 hours.)

9.55 *Manganese in the Earth's crust.* The scarce and essential metal, manganese, has been found in abundance in nodules on the deep seafloor. To investigate the relationship between the magnetic age of Earth's crust on the ocean floor and the probability of finding manganese nodules in that location, crust specimens were selected from seven magnetic age locations and the percentage of specimens containing manganese nodules was recorded for each. The data are shown in the accompanying table. Is there sufficient evidence to indicate that the probability of finding manganese nodules in the deep-sea Earth's crust is dependent on the magnetic age of the crust? Test using $\alpha = .05$.

**MANGANESE**

| Age | Number of Specimens | Percentage with Manganese Nodules |
|---|---|---|
| Miocene–recent | 389 | 5.9 |
| Oligocene | 140 | 17.9 |
| Eocene | 214 | 16.4 |
| Paleocene | 84 | 21.4 |
| Late Cretaceous | 247 | 21.1 |
| Early and Middle Cretaceous | 1,120 | 14.2 |
| Jurassic | 99 | 11.0 |

*Source:* Menard, H. W. "Time, chance, and the origin of manganese nodules." *American Scientist*, Sept.–Oct. 1976.

CHAPTER 10

# Simple Linear Regression

**OBJECTIVE**

*To present the basic concepts of regression analysis based on a simple linear relation between a response y and a single predictor variable x*

**CONTENTS**



*STATISTICS IN ACTION*

Can Dowsers Really Detect Water?

## *STATISTICS IN ACTION*

### Can Dowsers Really Detect Water?

The act of searching for and finding underground supplies of water using nothing more than a divining rod is commonly known as "dowsing." Although widely regarded among scientists as no more than a superstitious relic from medieval times, dowsing remains popular in folklore and, to this day, there are individuals who claim to have this mysterious skill.

Many dowsers in Germany claim that they respond to "earthrays" that emanate from the water source. These earthrays, say the dowsers, are a subtle form of radiation potentially hazardous to human health. As a result of these claims the German government in the mid-1980s conducted a 2-year experiment to investigate the possibility that dowsing is a genuine skill. If such a skill could be demonstrated, reasoned government officials, then dangerous levels of radiation in Germany could be detected, avoided, and disposed of.

A group of university physicists in Munich, Germany, were provided a grant of 400,000 marks (≈$250,000) to conduct the study. Approximately 500 candidate dowsers were recruited to participate in preliminary tests of their skill. To avoid fraudulent claims, the 43 individuals who seemed to be the most successful in the preliminary tests were selected for the final, carefully controlled, experiment.

The researchers set up a 10-meter-long line on the ground floor of a vacant barn, along which a small wagon could be moved. Attached to the wagon was a short length of pipe, perpendicular to the test line, that was connected by hoses to a pump with running water. The location of the pipe along the line for each trial of the experiment was assigned using a computer-generated random number. On the upper floor of the barn, directly above the experimental line, a 10-meter test line was painted. In each trial, a dowser was admitted to this upper level and required, with his or her rod, stick, or other tool of choice, to ascertain where the pipe with running water on the ground floor was located.

Each dowser participated in at least one test series, that is, a sequence of from 5 to 15 trials (typically 10), with the pipe randomly repositioned after each trial. (Some dowsers undertook only one test series, selected others underwent more than 10 test series.) Over the 2-year experimental period, the 43 dowsers participated in a total of 843 tests. The experiment was "double blind" in that neither the observer (researcher) on the top floor nor the dowser knew the pipe's location, even after a guess was made. (*Note:* Before the experiment began, a professional magician inspected the entire arrangement for potential deception or cheating by the dowsers.)

For each trial, two variables were recorded: the actual pipe location (in decimeters from the beginning of the line) and the dowser's guess (also measured in decimeters). Based on an examination of these data, the German physicists concluded in their final report that although most dowsers did not do particularly well in the experiments, "some few dowsers, in particular tests, showed an extraordinarily high rate of success, which can scarcely if at all be explained as due to chance . . . a real core of dowser-phenomena can be regarded as empirically proven . . . " (Wagner, Betz, and König, 1990).

This conclusion was critically assessed by Professor J.T. Enright of the University of California—San Diego (*Skeptical Inquirer,* Jan/Feb 1999). Enright focused on the three "best" dowsers (numbered 99, 18, and 108). Each of these dowsers performed the experiment multiple times and the best test series (sequence of trials) for each of these three dowsers was identified. These data, saved in the **DOWSING** file, are listed in Table SIA10.1.

We apply the statistical methodology presented in this chapter to the data in Table SIA10.1 in order to determine whether, in fact, dowsers can really detect water. The analysis and results are presented in the *Statistics in Action Revisited* example at the end of this chapter.

**DOWSING**

**TABLE SIA10.1 Dowsing Trial Results: Best Series for the Three Best Dowsers**

| Trial | Dowser Number | Pipe Location | Dowser's Guess |
|---|---|---|---|
| 1 | 99 | 4 | 4 |
| 2 | 99 | 5 | 87 |
| 3 | 99 | 30 | 95 |
| 4 | 99 | 35 | 74 |
| 5 | 99 | 36 | 78 |
| 6 | 99 | 58 | 65 |
| 7 | 99 | 40 | 39 |
| 8 | 99 | 70 | 75 |
| 9 | 99 | 74 | 32 |
| 10 | 99 | 98 | 100 |
| 11 | 18 | 7 | 10 |
| 12 | 18 | 38 | 40 |
| 13 | 18 | 40 | 30 |
| 14 | 18 | 49 | 47 |
| 15 | 18 | 75 | 9 |
| 16 | 18 | 82 | 95 |
| 17 | 108 | 5 | 52 |
| 18 | 108 | 18 | 16 |
| 19 | 108 | 33 | 37 |
| 20 | 108 | 45 | 40 |
| 21 | 108 | 38 | 66 |
| 22 | 108 | 50 | 58 |
| 23 | 108 | 52 | 74 |
| 24 | 108 | 63 | 65 |
| 25 | 108 | 72 | 60 |
| 26 | 108 | 95 | 49 |

*Source*: Enright, J. T. "Testing dowsing: The failure of the Munich experiments." *Skeptical Inquirer*. Jan/Feb. 1999, p.45 (Figure 6a).

## 10.1 Regression Models

One of the most important applications of statistics involves estimating the mean value of a response variable $y$ or predicting some future value of $y$ based on knowledge of a set of related independent* variables, $x_1, x_2, \dots, x_k$.

For example, the manager of a warehouse that uses automated vehicles might want to relate the congestion time $y$ (the **dependent variable**) to such variables as the number of vehicles in operation and the size of the load being moved (the **independent variables**). The objective would be to develop a **prediction equation** (or **model**) that expresses $y$ as a function of the independent variables. This would

*When the word *independent* is applied to the variables $x_1, x_2, \dots, x_k$, we mean independent in an algebraic rather than probabilistic sense.

enable the manager to predict $y$ for specific values of the independent variables and, ultimately, to use knowledge derived from a study of the prediction equation to institute policies to minimize the congestion time.

As another example, an engineer might want to relate the rate of malfunction $y$ of a mechanical assembler to such variables as its speed of operation and the assembler operator. The objective would be to develop a prediction equation relating the dependent variable $y$ to the independent variables and to use the prediction equation to predict the value of the rate of malfunction $y$ for various combinations of speed of operation and operator.

The models used to relate a dependent variable $y$ to the independent variables $x_1, x_2, \ldots, x_k$ are called **regression models** or **linear statistical models** because they express the mean value of $y$ for given values of $x_1, x_2, \ldots, x_k$ as a linear function of a set of unknown parameters. These parameters are estimated from sample data using a process to be explained in Section 10.3.

The concepts of a regression analysis are introduced in this chapter using a very simple regression model—one that relates $y$ to a single independent variable $x$. We will learn how to fit this model to a set of data using the **method of least squares** and will examine in detail the different types of inferences that can result from a regression analysis. In Chapters 11–12, we will apply this knowledge to help us understand the theoretical and practical implications of a **multiple regression analysis**—the problem of relating $y$ to two or more independent variables.

**Definition 10.1**

The variable to be predicted (or modeled), $y$, is called the **dependent** (or **response**) **variable**.

**Definition 10.2**

The variables used to predict (or model) $y$ are called **independent variables** and are denoted by the symbols $x_1$, $x_2$, $x_3$, etc.

The preceding chapters provide a foundation for a study of applied statistics. Although the derivations of the sampling distributions for many of the statistics that we will subsequently encounter are mathematically beyond the scope of this text, the knowledge that you have acquired will help you to understand how they are derived and, in many instances, will enable you to find the means and variances of these sampling distributions.

Your theoretical knowledge will be useful for another reason. The theory of statistics, like the theories of physics, engineering, economics, etc., is only a model for reality. It exactly explains reality only when the assumptions of the methodology are exactly satisfied. Since this situation rarely occurs, the application of statistics (or physics, engineering, economics, etc.) to the solution of real-world problems is an art. Thus, to apply theory to the real world, you must know the extent to which deviations from assumptions will affect the resulting statistical inferences, and you must be able to adapt the model and methodology to the conditions of a practical problem. A basic understanding of the theory underlying the methodology will help you to do this.

## 10.2 Model Assumptions

To simplify our discussion, we will postulate the following fictitious situation and data set. Suppose that the developer of a new insulation material wants to determine the amount of compression that would be produced on a 2-inch-thick specimen of the material when subjected to different amounts of pressure. Five experimental pieces of the material were tested under different pressures. The values of $x$ (in units of

**TABLE 10.1 Compression Versus Pressure for an Insulation Material**

**INSULATION**

| Specimen | Pressure $x$ | Compression $y$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 3 | 2 |
| 4 | 4 | 2 |
| 5 | 5 | 4 |

10 pounds per square inch) and the resulting amounts of compression $y$ (in units of .1 inch) are given in Table 10.1. A plot of the data, called a **scattergram** (or, **scatterplot**) is shown in Figure 10.1.

Suppose we believe that the value of $y$ tends to increase in a linear manner as $x$ increases. Then we could select a model relating $y$ to $x$ by drawing a line through the points in Figure 10.1. Such a **deterministic model**—one that does not allow for errors of prediction—might be adequate if all of the points of Figure 10.1 fell on the fitted line. However, you can see that this idealistic situation will not occur for the data of Table 10.1. No matter how you draw a line through the points of Figure 10.1, at least some of the points will deviate substantially from the fitted line.

The solution to the preceding problem is to construct a **probabilistic model** relating $y$ to $x$—one that acknowledges the random variation of the data points about a line. One type of probabilistic model, a **simple linear regression model**, makes the assumption that the mean value of $y$ for a given value of $x$ graphs as a straight line and that points deviate about this **line of means** by a random (positive or negative) amount equal to $\varepsilon$, i.e.,

$$y = \underbrace{\beta_0 + \beta_1 x}_{\substack{\text{Mean value of } y \\ \text{for a given } x}} + \underbrace{\varepsilon}_{\substack{\text{Random} \\ \text{error}}}$$

where $\beta_0$ and $\beta_1$ are unknown parameters of the deterministic (nonrandom) portion of the model. If we assume that the points deviate above and below the line of means, with some deviations positive, some negative, and with $E(\varepsilon) = 0$, then the mean value of $y$ is

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

Therefore, the mean value of $y$ for a given value of $x$, represented by the symbol $E(y)$,* graphs as a straight line with $y$-intercept equal to $\beta_0$ and slope equal to $\beta_1$. A graph of the hypothetical line of means, $E(y) = \beta_0 + \beta_1 x$, is shown in Figure 10.2.

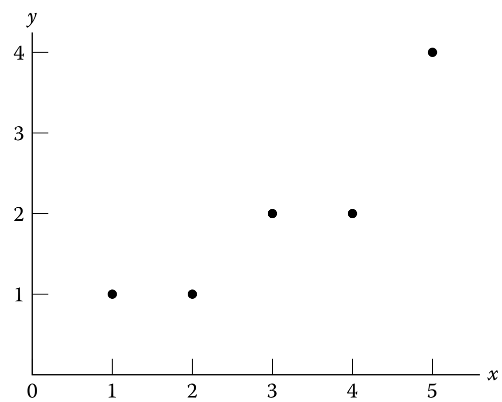
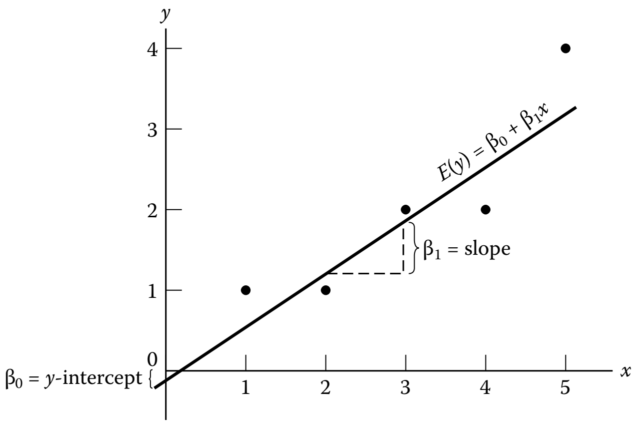



**FIGURE 10.1**
Scattergram for the data in Table 10.1

*The mean value of $y$ for a given value of $x$ should be denoted by the symbol $E(Y \mid x)$. However, this notation becomes cumbersome when the model contains more than one independent variable. Consequently, we will abbreviate the notation and represent $E(Y \mid x)$ by the symbol $E(y)$.

**FIGURE 10.2**
A graph of the data points of Table 10.1 and the hypothetical line of means, $E(y) = \beta_0 + \beta_1 x$




### A Simple Linear Regression (Probabilistic) Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

$y$ = Dependent variable

$x$ = Independent variable

$E(y) = \beta_0 + \beta_1 x$ is the deterministic component (the equation of a straight line)

$\varepsilon$ (epsilon) = **Random error** component

$\beta_0$ (beta-zero) = ***y*-intercept** of the line, i.e., point at which the line intercepts or cuts through the *y*-axis (see Figure 10.2)

$\beta_1$ (beta-one) = **Slope** of the line, i.e., amount of increase (or decrease) in the deterministic component $y$ for every 1 unit increase in $x$ (see Figure 10.2)

To fit a simple linear regression model to a set of data, we must find estimators for the unknown parameters, $\beta_0$ and $\beta_1$, of the line of means, $E(y) = \beta_0 + \beta_1 x$. Valid inferences about $\beta_0$ and $\beta_1$ will depend on the sampling distributions of the estimators, which in turn depend on the probability distribution of the random error, $\varepsilon$; consequently, we must first make specific assumptions about $\varepsilon$. These assumptions, summarized here, are basic to every statistical regression analysis.

**Assumption 1** The mean of the probability distribution of $\varepsilon$ is 0. That is, the average of the errors over an infinitely long series of experiments is 0 for each setting of the independent variable $x$. This assumption implies that the mean value of $y$, $E(y)$, for a given value of $x$ is $E(y) = \beta_0 + \beta_1 x$.

**Assumption 2** The variance of the probability distribution of $\varepsilon$ is constant for all settings of the independent variable $x$. For our straight-line model, this assumption means that the variance of $\varepsilon$ is equal to a constant, say, $\sigma^2$, for all values of $x$.

**Assumption 3** The probability distribution of $\varepsilon$ is normal.

**Assumption 4** The errors associated with any two different observations are independent. That is, the error associated with one value of $y$ has no effect on the errors associated with other $y$ values.

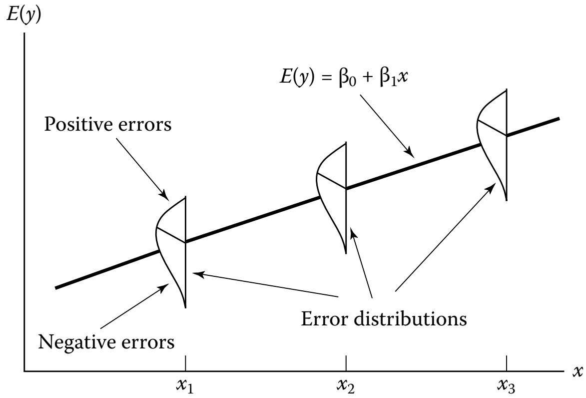



**FIGURE 10.3**
The probability distribution of $\varepsilon$

The implications of the first three assumptions can be seen in Figure 10.3, which shows distributions of errors for three particular values of $x$, namely, $x_1$, $x_2$, and $x_3$. Note that the relative frequency distributions of the errors are normal, with a mean of 0, and a constant variance $\sigma^2$ (all the distributions shown have the same amount of spread or variability). The straight line shown in Figure 10.3 is the mean value of $y$ for a given value of $x$, $E(y) = \beta_0 + \beta_1 x$.

Various techniques exist for checking the validity of these assumptions, and there are remedies to be applied when they appear to be invalid. We discuss these techniques in detail in Chapter 11. In actual practice, the assumptions need not hold exactly for least-squares estimators and test statistics (to be described subsequently) to possess the measures of reliability that we would expect from a regression analysis. The assumptions will be satisfied adequately for many practical applications.

## 10.3 Estimating $\beta_0$ and $\beta_1$: The Method of Least Squares

To choose the "best-fitting" line for a set of data, we must estimate the unknown parameters, $\beta_0$ and $\beta_1$, of the simple linear regression model. These estimators could be found using the method of maximum likelihood (Section 7.3), but the easiest method—and one that is intuitively appealing—is the **method of least squares**. When the assumptions of Section 10.2 are satisfied, then the maximum likelihood and the least-squares estimators of $\beta_0$ and $\beta_1$ are identical.*

The reasoning behind the method of least squares can be seen by examining Figure 10.4, which shows a line drawn on the scattergram of the data points of Table 10.1. The vertical line segments represent **deviations** of the points from the line. You can see by shifting a ruler around the graph that it is possible to find many lines for which the sum of deviations (or **errors**) is equal to 0, but it can be shown that there is one and only one line for which the **sum of squares of the deviations** is a minimum. The sum of squares of the deviations is called the **sum of squares for error** and is denoted by the symbol SSE. The line is called the **least-squares line**, the **regression line**, or the **least-squares prediction equation**.

To find the least-squares line for a set of data, assume that we have a sample of $n$ data points which can be identified by corresponding values of $x$ and $y$, say, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. For example, the $n = 5$ data points shown in

*The method of least squares is a valid estimation technique even when one or more of the assumptions of Section 10.2 are violated. If the assumptions are not satisfied, it is the validity of inferences derived from the estimates that is in question.

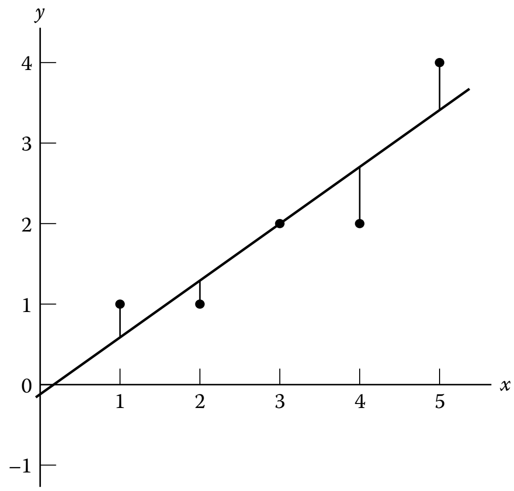



**FIGURE 10.4**
Graph showing the deviations of the points about a line

Table 10.1 are (1, 1), (2, 1), (3, 2), (4, 2), and (5, 4). The straight-line model for the response $y$ in terms of $x$ is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The line of means is $E(y) = \beta_0 + \beta_1 x$ and the fitted line, which we hope to find, is represented as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Thus, $\hat{y}$ is an estimator of the mean value of $y$, $E(y)$, and a predictor of some future value of $y$; and $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of $\beta_0$ and $\beta_1$, respectively.

For a given data point, say, the point $(x_i, y_i)$, the observed value of $y$ is $y_i$ and the predicted value of $y$ would be obtained by substituting $x_i$ into the prediction equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

And the deviation of the $i$th value of $y$ from its predicted value, called a **residual**, is

$$(y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

Then the sum of squares of the deviations of the $y$ values about their predicted values (i.e., sum of squared residuals) for all of the $n$ data points is

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ that make the SSE a minimum are called the **least-squares estimates** of the population parameters $\beta_0$ and $\beta_1$, and the prediction equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the **least-squares line**.

**Definition 10.3**

A regression **residual** $\hat{\epsilon}$ is defined as the difference between an observed $y$ value and its corresponding predicted value:

$$\hat{\epsilon} = y - \hat{y}$$

**Definition 10.4**

The **least-squares line** is one that has a smaller SSE than any other straight-line model.

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

are obtained by setting the two partial derivatives, $\partial\text{SSE}/\partial\hat{\beta}_0$ and $\partial\text{SSE}/\partial\hat{\beta}_1$, equal to 0 and solving the resulting simultaneous linear system of **least-squares equations**. To illustrate, we first compute the partial derivatives:

$$\frac{\partial\text{SSE}}{\partial\hat{\beta}_0} = \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-1)$$

$$\frac{\partial\text{SSE}}{\partial\hat{\beta}_1} = \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-x_i)$$

Setting these partial derivatives equal to 0 and simplifying, we obtain the least-squares equations:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

or

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving this pair of simultaneous linear equations for $\beta_0$ and $\beta_1$, we obtain (proof omitted) the formulas shown in the box.

### Formulas for the Least-Squares Estimates

*Slope*: $\hat{\beta}_1 = \dfrac{\text{SS}_{xy}}{\text{SS}_{xx}}$

*y-intercept*: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

where

$$\text{SS}_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$\text{SS}_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$n = \text{Sample size}$$

**Example 10.1**

Finding $\hat{\beta}_0$, $\hat{\beta}_1$, and SSE

a. Calculate the least-squares estimates of $\beta_0$ and $\beta_1$ for the data of Table 10.1. Then compute SSE.

b. Give a practical interpretation of the results.

**TABLE 10.2 Preliminary Computations for the Insulation Compression Example**

| | $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ | $y_i^2$ |
|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 4 | 2 | 1 |
| | 3 | 2 | 9 | 6 | 4 |
| | 4 | 2 | 16 | 8 | 4 |
| | 5 | 4 | 25 | 20 | 16 |
| **TOTALS** | $\sum x_i = 15$ | $\sum y_i = 10$ | $\sum x_i^2 = 55$ | $\sum x_i y_i = 37$ | $\sum y_i^2 = 26$ |

Solution

a. Preliminary computations for finding the least-squares line for the insulation compression data are contained in Table 10.2. We can now calculate*

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{5} = 37 - \frac{(15)(10)}{5}$$
$$= 37 - 30 = 7$$
$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{5} = 55 - \frac{(15)^2}{5}$$
$$= 55 - 45 = 10$$

Then, the slope of the least-squares line is

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{7}{10} = .7$$

and the $y$-intercept is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{\sum y_i}{5} - \hat{\beta}_1\frac{\left(\sum x_i\right)}{5}$$
$$= \frac{10}{5} - (.7)\frac{(15)}{5}$$
$$= 2 - (.7)(3) = 2 - 2.1 = -.1$$

The least-squares line is thus

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -.1 + .7x$$

The graph of this line is shown in Figure 10.5.

The observed and predicted values of $y$, the deviations of the $y$ values about their predicted values and the squares of these deviations for the data of Table 10.1 are shown in Table 10.3. Note that the sum of squares of the deviations, SSE, is 1.10. This is smaller than the value of SSE that would be obtained by fitting any other possible straight line to the data.

*Note*: The calculations required to obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and SSE in simple linear regression, although straightforward, can become rather tedious. Even with the use of a calculator, the process is laborious and susceptible to error, especially when the sample size is large. Most engineers and scientists rely on statistical software to run the simple

*Since summations will be used extensively from this point on, we will omit the limits on $\sum$ when the summation includes all the measurements in the sample, i.e., when the symbol is $\sum_{i=1}^{n}$, we will write $\sum$.

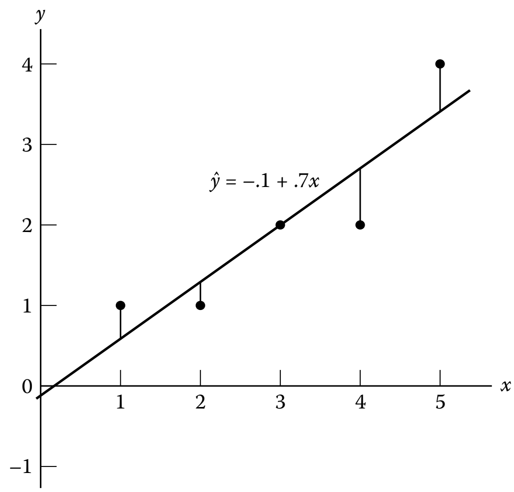



**FIGURE 10.5**
The line $\hat{y} = -.1 + .7x$ fit to the data

**TABLE 10.3 Comparing Observed and Predicted Values for the Least-Squares Model**

| $x$ | $y$ | $\hat{y} = -.1 + .7x$ | $(y - \hat{y})$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 1 | .6 | $(1 - .6) = .4$ | .16 |
| 2 | 1 | 1.3 | $(1 - 1.3) = -.3$ | .09 |
| 3 | 2 | 2.0 | $(2 - 2.0) = 0$ | .00 |
| 4 | 2 | 2.7 | $(2 - 2.7) = -.7$ | .49 |
| 5 | 4 | 3.4 | $(4 - 3.4) = .6$ | .36 |
| | | | Sum of errors $= 0$ | SSE $= 1.10$ |

linear regression. The SAS, MINITAB, SPSS, and EXCEL outputs for the analysis of the data in Table 10.1 are shown in Figure 10.6a–d. The values of $\hat{\beta}_0$, $\hat{\beta}_1$, and SSE are highlighted on the printouts. (Note that these values agree exactly with our hand-calculated values.)

b. Our interpretation of the least-squares slope, $\hat{\beta}_1 = .7$, is that compression $y$ will increase .7 unit for every 1-unit increase in pressure $x$. Since $y$ is measured in units of .1 inch and $x$ in units of 10 pounds per square inch, our interpretation is that compression increases .07 inch for every 10-pound-per-square-inch increase in pressure. We will attach a measure of reliability to this inference in Section 10.6.

   The least-squares intercept, $\hat{\beta}_0 = -.1$, is our estimate of compression $y$ when pressure is set at $x = 0$ pounds per square inch. Since level of compression can never be negative, why does such a nonsensical result occur? The reason is that we are attempting to use the least-squares model to predict $y$ for a value of $x$ ($x = 0$) that is outside the range of the sample data and impractical. (We have more to say about predicting outside the range of the sample data—called **extrapolation**—in Section 10.9.) Consequently, $\hat{\beta}_0$ will not always have a practical interpretation. Only when $x = 0$ is within the range of the $x$ values in the sample and is a practical value will $\hat{\beta}_0$ have a meaningful interpretation.

To summarize, we have defined the best-fitting straight line to be the one that satisfies the least-squares criterion; that is, the sum of the squared errors will be smaller than for any other straight-line model. This line is called the **least-squares line**, and its equation is called the **least-squares prediction equation**.

a. SAS output

```
                         The REG Procedure
                           Model: MODEL1
                    Dependent Variable: COMP_Y

             Number of Observations Read           5
             Number of Observations Used           5

                        Analysis of Variance

                                   Sum of          Mean
Source                  DF        Squares        Square    F Value    Pr > F

Model                    1        4.90000       4.90000      13.36    0.0354
Error                    3        1.10000       0.36667
Corrected Total          4        6.00000

          Root MSE              0.60553    R-Square     0.8167
          Dependent Mean        2.00000    Adj R-Sq     0.7556
          Coeff Var            30.27650

                        Parameter Estimates

                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1       -0.10000        0.63509      -0.16      0.8849
     PRESS_X       1        0.70000        0.19149       3.66      0.0354
```

b. MINITAB output

```
The regression equation is
COMP_Y = - 0.100 + 0.700 PRESS_X


Predictor     Coef  SE Coef      T      P
Constant   -0.1000   0.6351  -0.16  0.885
PRESS_X     0.7000   0.1915   3.66  0.035


S = 0.605530   R-Sq = 81.7%   R-Sq(adj) = 75.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  4.9000  4.9000  13.36  0.035
Residual Error   3  1.1000  0.3667
Total            4  6.0000
```

**FIGURE 10.6**
Statistical software printouts for the simple linear regression analysis, Example 10.1

c. SPSS output

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .904[a] | .817 | .756 | .606 |

a. Predictors: (Constant), PRESS_X

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 4.900 | 1 | 4.900 | 13.364 | .035[a] |
| | Residual | 1.100 | 3 | .367 | | |
| | Total | 6.000 | 4 | | | |

a. Predictors: (Constant), PRESS_X

b. Dependent Variable: COMP_Y

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -.100 | .635 | | -.157 | .885 |
| | PRESS_X | .700 | .191 | .904 | 3.656 | .035 |

a. Dependent Variable: COMP_Y

d. EXCEL output

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | *Regression Statistics* | | | | | | |
| 4 | Multiple R | 0.903696114 | | | | | |
| 5 | R Square | 0.816666667 | | | | | |
| 6 | Adjusted R Square | 0.755555556 | | | | | |
| 7 | Standard Error | 0.605530071 | | | | | |
| 8 | Observations | 5 | | | | | |
| 9 | | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | |
| 12 | Regression | 1 | 4.9 | 4.9 | 13.36363636 | 0.035352847 | |
| 13 | Residual | 3 | 1.1 | 0.366667 | | | |
| 14 | Total | 4 | 6 | | | | |
| 15 | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| 17 | Intercept | -0.1 | 0.635085296 | -0.15746 | 0.88488398 | -2.12112675 | 1.9211268 |
| 18 | PRESS_X | 0.7 | 0.191485422 | 3.655631 | 0.035352847 | 0.090607356 | 1.3093926 |

**FIGURE 10.6 (*Continued*)**

## Applied Exercises

10.1 *Solving for $\beta_0$ and $\beta_1$.* The equation for a straight line (deterministic) is

$$y = \beta_0 + \beta_1 x$$

If the line passes through the point (0, 1), then $x = 0$, $y = 1$ must satisfy the equation. That is,

$$1 = \beta_0 + \beta_1(0)$$

Similarly, if the line passes through the point (2, 3), then $x = 2$, $y = 3$ must satisfy the equation:

$$3 = \beta_0 + \beta_1(2)$$

Use these two equations to solve for $\beta_0$ and $\beta_1$, and find the equation of the line that passes through the points (0, 1) and (2, 3).

10.2 *Finding line equations.* In each case find the equation of the line that passes through the points. Graph each line.

a. (0, 2) and (2, 6)
b. (0, 4) and (2, 6)
c. (0, −2) and (−1, −6)
d. (0, −4) and (3, −7)

10.3 *Identifying y-intercept and slope.* Plot the following lines; give the $y$-intercept and slope of each.

a. $y = 3 + 2x$
b. $y = 1 + x$
c. $y = -2 + 3x$
d. $y = 5x$
e. $y = 4 - 2x$

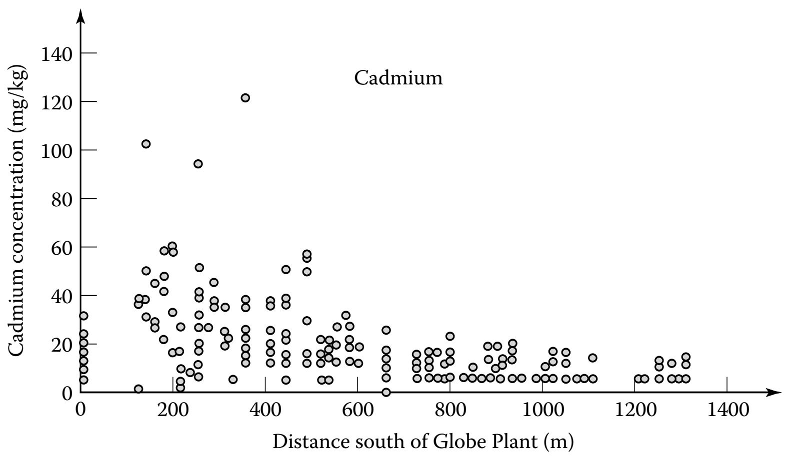
10.4 *Redshifts of Quasi-Stellar Objects.* Astronomers call a shift in the spectrum of galaxies a "redshift." A correlation between redshift level and apparent magnitude (i.e., brightness on a logarithmic scale) of a Quasi-Stellar Object (QSO) was discovered and reported in the *Journal of Astrophysics & Astronomy* (Mar./Jun. 2003). Simple linear regression was applied to data collected for a sample of over 6,000 QSOs with confirmed redshift. The analysis yielded the following results for a specific magnitude range: $\hat{y} = 18.13 + 6.21x$, where $y$ = magnitude and $x$ = redshift level.

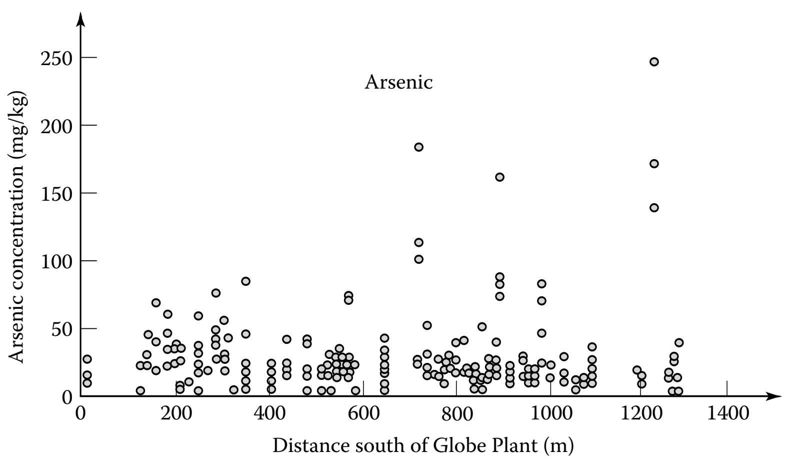
a. Graph the least-squares line. Is the slope of the line positive or negative?
b. Interpret the estimate of the $y$-intercept in the words of the problem.
c. Interpret the estimate of the slope in the words of the problem.

10.5 *Arsenic in soil.* In Denver, Colorado, environmentalists have discovered a link between high arsenic levels in soil and a crabgrass killer used in the 1950s and 1960s. (*Environmental Science & Technology*. Sept. 1, 2000.) The recent discovery was based, in part, on the scattergrams shown below. The graphs plot the level of the metals cadmium and arsenic, respectively, against the distance from a former smelter plant for samples of soil taken from Denver residential properties.

a. Normally, the metal level in soil decreases as distance from the source (e.g., a smelter plant) increases. Propose a straight-line model relating metal level $y$ to distance from the plant $x$. Based on the theory, would you expect the slope of the line to be positive or negative?

**Scattergrams for Exercise 10.5**



### SPSS Output for Exercise 10.6

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 19.393 | 1.708 | | 11.357 | .000 |
| | ABConc | -8.036 | 2.820 | -.819 | -2.849 | .046 |

a. Dependent Variable: WickLength

b. Examine the scatterplot for cadmium. Does the plot support the theory, part **a**?
c. Examine the scatterplot for arsenic. Does the plot support the theory, part **a**? (*Note*: This finding led investigators to discover the link between high arsenic levels and the use of the crabgrass killer.)

10.6 *New method for blood typing.* In *Analytical Chemistry* (May 2010), chemical engineers tested a new method of typing blood using lost cost paper. Blood drops were applied to the paper and the rate of absorption (called *blood wicking*) was measured. The table below gives the wicking lengths (millimeters) for six blood drops, each at a different antibody concentration. Let $y$ = wicking length and $x$ = antibody concentration.

a. Give the equation of the straight-line model relating $y$ to $x$.
b. An SPSS printout of the simple linear regression analysis is shown above. Give the equation of the least squares line.
c. Give practical interpretations (if possible) of the estimated $y$-intercept and slope of the line.

**BLOODTYPE**

| Droplet | Length (mm) | Concentration |
|---|---|---|
| 1 | 22.50 | 0.0 |
| 2 | 16.00 | 0.2 |
| 3 | 13.50 | 0.4 |
| 4 | 14.00 | 0.6 |
| 5 | 13.75 | 0.8 |
| 6 | 12.50 | 1.0 |

*Source:* Khan, M.S., et al. "Paper diagnostic for instant blood typing", *Analytical Chemistry,* Vol. 82, No. 10, May 2010 (adapted from Figure 4b).

10.7 *Sound waves from a basketball.* Refer to the *American Journal of Physics* (June 2010) study of sound waves in a spherical cavity, Exercise 2.29 (p. 43). The frequencies of sound waves (estimated using a mathematical formula) resulting from the first 24 resonances (echoes) after striking a basketball with a metal rod are reproduced in the next table. Recall that the researcher expects the sound wave frequency to increase as the number of resonances increase.

**BBALL**

| Resonance | Frequency |
|---|---|
| 1 | 979 |
| 2 | 1572 |
| 3 | 2113 |
| 4 | 2122 |
| 5 | 2659 |
| 6 | 2795 |
| 7 | 3181 |
| 8 | 3431 |
| 9 | 3638 |
| 10 | 3694 |
| 11 | 4038 |
| 12 | 4203 |
| 13 | 4334 |
| 14 | 4631 |
| 15 | 4711 |
| 16 | 4993 |
| 17 | 5130 |
| 18 | 5210 |
| 19 | 5214 |
| 20 | 5633 |
| 21 | 5779 |
| 22 | 5836 |
| 23 | 6259 |
| 24 | 6339 |

*Source:* Russell, D.A. "Basketballs as spherical acoustic cavities", *American Journal of Physics,* Vol. 48, No. 6, June 2010. (Table I.)

a. Hypothesize a model for frequency ($y$) as a function of number of resonances ($x$) that proposes a linearly increasing relationship.
b. According to the researcher's theory, will the slope of the line be positive or negative?
c. Estimate the beta parameters of the model and (if possible) give a practical interpretation of each.

10.8 *Estimating repair and replacement costs of water pipes.* Pipes used in a water distribution network are susceptible to breakage due to a variety of factors. When pipes break, engineers must decide whether to repair or replace the broken pipe. A team of civil engineers used regression analysis to estimate $y =$ the ratio of repair to replacement cost of commercial pipe in the *IHS Journal of Hydraulic Engineering* (September 2012). The independent variable in the regression analysis was $x =$ the diameter (in millimeters) of the pipe. Data for a sample of 13 different pipe sizes are provided in the table.

**WATERPIPE**

| DIAMETER | RATIO |
|---|---|
| 80 | 6.58 |
| 100 | 6.97 |
| 125 | 7.39 |
| 150 | 7.61 |
| 200 | 7.78 |
| 250 | 7.92 |
| 300 | 8.20 |
| 350 | 8.42 |
| 400 | 8.60 |
| 450 | 8.97 |
| 500 | 9.31 |
| 600 | 9.47 |
| 700 | 9.72 |

*Source:* Suribabu, C.R. & Neelakantan, T.R. "Sizing of water distribution pipes based on performance measure and breakage-repair replacement economics", *IHS Journal of Hydraulic Engineering,* Vol. 18, No. 3, September 2012 (Table 1).

a. Construct a scatterplot for the data.
b. Find the least-squares line relating ratio of repair to replacement cost ($y$) to pipe diameter ($x$).
c. Plot the least squares line on the graph, part a.
d. Interpret, practically, the values $\hat{\beta}_0$ and $\hat{\beta}_1$.

10.9 *Extending the life of an aluminum smelter pot.* An investigation of the properties of bricks used to line aluminum smelter pots was published in *The American Ceramic Society Bulletin* (Feb. 2005). Six different commercial bricks were evaluated. The lifelength of a smelter pot depends on the porosity of the brick lining (the less porosity, the longer the life); consequently, the researchers measured the apparent porosity of each brick specimen, as well as the mean pore diameter of each brick. The data are given in the next table.

a. Find the least-squares line relating porosity ($y$) to mean pore diameter ($x$).
b. Interpret the $y$-intercept of the line.
c. Interpret the slope of the line.
d. Predict the apparent porosity percentage for a brick with a mean pore diameter of 10 micrometers.

**SMELTPOT**

| Brick | Apparent Porosity (%) | Mean Pore Diameter (micrometers) |
|---|---|---|
| A | 18.8 | 12.0 |
| B | 18.3 | 9.7 |
| C | 16.3 | 7.3 |
| D | 6.9 | 5.3 |
| E | 17.1 | 10.9 |
| F | 20.4 | 16.8 |

*Source:* Bonadia, P., et al. "Aluminosilicate refractories for aluminum cell linings." *The American Ceramic Society Bulletin*, Vol. 84, No. 2, Feb. 2005 (Table II).

10.10 *Spreading rate of spilled liquid.* A contract engineer at DuPont Corp. studied the rate at which a spilled volatile liquid will spread across a surface (*Chemical Engineering Progress*, Jan. 2005). Assume 50 gallons of methanol spills onto a level surface outdoors. The engineer used derived empirical formulas (assuming a state of turbulent-free convection) to calculate the mass (in pounds) of the spill after a period of time ranging from 0 to 60 minutes. The calculated mass values are given in the accompanying table.

**LIQUIDSPILL**

| Time (minutes) | Mass (pounds) | Time (minutes) | Mass (pounds) |
|---|---|---|---|
| 0 | 6.64 | 22 | 1.86 |
| 1 | 6.34 | 24 | 1.60 |
| 2 | 6.04 | 26 | 1.37 |
| 4 | 5.47 | 28 | 1.17 |
| 6 | 4.94 | 30 | 0.98 |
| 8 | 4.44 | 35 | 0.60 |
| 10 | 3.98 | 40 | 0.34 |
| 12 | 3.55 | 45 | 0.17 |
| 14 | 3.15 | 50 | 0.06 |
| 16 | 2.79 | 55 | 0.02 |
| 18 | 2.45 | 60 | 0.00 |
| 20 | 2.14 | | |

*Source:* Barry, J. "Estimating rates of spreading and evaporation of volatile liquids." *Chemical Engineering Progress*, Vol. 101, No. 1, Jan. 2005.

a. Construct a scattergram for the data, with $y =$ calculated mass and $x =$ time.
b. Find the least-squares line relating mass ($y$) to time ($x$).
c. Plot the least-squares line on the graph, part **a.**
d. Interpret the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

10.11 *New method of estimating rainfall.* Accurate measurements of rainfall are critical for many hydrological and meteorological projects. Two standard methods of monitoring rainfall use rain gauges and weather radar. Both, however, can be contaminated by human and environmental interference. In the *Journal of Data Science* (Apr. 2004). researchers employed artificial neural networks (i.e., computer-based mathematical models) to estimate rainfall at a meteorological station in Montreal. Rainfall estimates were made every 5 minutes over a 70-minute period by each of the three methods. The data (in millimeters) are listed in the table.

a. Propose a straight-line model relating rain gauge amount ($y$) to weather radar rain estimate ($x$).
b. Fit the model to the data using the method of least squares.
c. Graph the least-squares line on a scattergram of the data. Is there visual evidence of a relationship between the two variables? Is the relationship positive or negative?
d. Interpret the estimates of the $y$-intercept and slope in the words of the problem.
e. Now consider a model relating rain gauge amount ($y$) to the artificial neural network rain estimate ($x$). Repeat parts **a–d** for this model.

**RAINFALL**

| Time | Radar | Rain Gauge | Neural Network |
|---|---|---|---|
| 8:00 A.M. | 3.6 | 0 | 1.8 |
| 8:05 | 2.0 | 1.2 | 1.8 |
| 8:10 | 1.1 | 1.2 | 1.4 |
| 8:15 | 1.3 | 1.3 | 1.9 |
| 8:20 | 1.8 | 1.4 | 1.7 |
| 8:25 | 2.1 | 1.4 | 1.5 |
| 8:30 | 3.2 | 2.0 | 2.1 |
| 8:35 | 2.7 | 2.1 | 1.0 |
| 8:40 | 2.5 | 2.5 | 2.6 |
| 8:45 | 3.5 | 2.9 | 2.6 |
| 8:50 | 3.9 | 4.0 | 4.0 |
| 8:55 | 3.5 | 4.9 | 3.4 |
| 9:00 A.M. | 6.5 | 6.2 | 6.2 |
| 9:05 | 7.3 | 6.6 | 7.5 |
| 9:10 | 6.4 | 7.8 | 7.2 |

*Source:* Hessami, M., et al. "Selection of an artificial neural network model for the post-calibration of weather radar rainfall estimation." *Journal of Data Science*, Vol. 2, No. 2, Apr. 2004. (Adapted from Figures 2 and 4.)

10.12 *Sweetness of orange juice.* The quality of the orange juice produced by a manufacturer (e.g., Minute Maid, Tropicana) is constantly monitored. There are numerous sensory and chemical components that combine to make the best-tasting orange juice. For example, one manufacturer has developed a quantitative index of the "sweetness" of orange juice (the higher the index, the sweeter the juice). Is there a relationship between the sweetness index and a chemical measure such as the amount of water-soluble pectin (parts per million) in the orange juice? Data collected on these two variables for 24 production runs at a juice manufacturing plant are shown in the next table. Suppose a manufacturer wants to use simple linear regression to predict the sweetness ($y$) from the amount of pectin ($x$).

a. Find the least-squares line for the data.
b. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$ in the words of the problem.
c. Predict the sweetness index if the amount of pectin in the orange juice is 300 ppm.

**OJUICE**

| Run | Sweetness Index | Pectin (ppm) |
|---|---|---|
| 1 | 5.2 | 220 |
| 2 | 5.5 | 227 |
| 3 | 6.0 | 259 |
| 4 | 5.9 | 210 |
| 5 | 5.8 | 224 |
| 6 | 6.0 | 215 |
| 7 | 5.8 | 231 |
| 8 | 5.6 | 268 |
| 9 | 5.6 | 239 |
| 10 | 5.9 | 212 |
| 11 | 5.4 | 410 |
| 12 | 5.6 | 256 |
| 13 | 5.8 | 306 |
| 14 | 5.5 | 259 |
| 15 | 5.3 | 284 |
| 16 | 5.3 | 383 |
| 17 | 5.7 | 271 |
| 18 | 5.5 | 264 |
| 19 | 5.7 | 227 |
| 20 | 5.3 | 263 |
| 21 | 5.9 | 232 |
| 22 | 5.8 | 220 |
| 23 | 5.8 | 246 |
| 24 | 5.9 | 241 |

*Note:* The data in the table are authentic. For confidentiality reasons, the manufacturer cannot be disclosed.

10.13 *Characterizing bone with fractal geometry.* In *Medical Engineering & Physics* (May 2013), researchers used fractal geometry to characterize human cortical bone. A measure of the variation in the volume of cortical bone tissue—called fractal dimension—was determined for each in a sample of 10 human ribs. The researchers used fractal dimension scores to predict the bone tissue's stiffness index, called Young's Modulus (measured in gigapascals). The experimental data are shown below. Consider the linear model, $E(y) = \beta_0 + \beta_1 x$, where $y$ = Young's Modulus and $x$ = fractal dimension score. Find an estimate of the increase (or decrease) in Young's Modulus for every 1 point increase in a bone tissue's fractal dimension score.

**CORTBONE**

| Young's Modulus (GPa) | Fractal Dimension |
|---|---|
| 18.3 | 2.48 |
| 11.6 | 2.48 |
| 32.2 | 2.39 |
| 30.9 | 2.44 |
| 12.5 | 2.50 |
| 9.1 | 2.58 |
| 11.8 | 2.59 |
| 11.0 | 2.59 |
| 19.7 | 2.51 |
| 12.0 | 2.49 |

*Source:* Sanchez-Molina, D., et al. "Fractal dimension and mechanical properties of human cortical bone", *Medical Engineering & Physics,* Vol. 35, No. 5, May 2013 (Table 1).

10.14 *Thermal performance of copper tubes.* A study was conducted to model the thermal performance of integral-fin tubes used in the refrigeration and process industries (*Journal of Heat Transfer*, Aug. 1990). Twenty-four specially manufactured integral-fin tubes having rectangular-shaped fins made of copper were used in the experiment. Vapor was released downward into each tube and the vapor-side heat transfer coefficient (based on the outside surface area of the tube) was measured. The dependent variable for the study is the heat transfer enhancement ratio $y$, defined as the ratio of the vapor-side coefficient of the fin tube to the vapor-side coefficient of a smooth tube evaluated at the same temperature. Theoretically, heat transfer will be related to the area at the top of the tube that is "unflooded" by condensation of the vapor. The data in the table are the unflooded area ratio ($x$) and heat transfer enhancement ($y$) values recorded for the 24 integral-fin tubes.

a. Find the least-squares line relating heat transfer enhancement $y$ to unflooded area ratio $x$.
b. Plot the data points and graph the least-squares line as a check on your calculations.
c. Interpret the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

**FINTUBES**

| *Unflooded Area Ratio, x* | *Heat Transfer Enhancement, y* |
|---|---|
| 1.93 | 4.4 |
| 1.95 | 5.3 |
| 1.78 | 4.5 |
| 1.64 | 4.5 |
| 1.54 | 3.7 |
| 1.32 | 2.8 |
| 2.12 | 6.1 |
| 1.88 | 4.9 |
| 1.70 | 4.9 |
| 1.58 | 4.1 |
| 2.47 | 7.0 |
| 2.37 | 6.7 |
| 2.00 | 5.2 |
| 1.77 | 4.7 |
| 1.62 | 4.2 |
| 2.77 | 6.0 |
| 2.47 | 5.8 |
| 2.24 | 5.2 |
| 1.32 | 3.5 |
| 1.26 | 3.2 |
| 1.21 | 2.9 |
| 2.26 | 5.3 |
| 2.04 | 5.1 |
| 1.88 | 4.6 |

*Source*: Marto, P. J., et. al. "An experimental study of R-113 film condensation on horizontal integral-fin tubes." *Journal of Heat Transfer*, Vol. 112, Aug. 1990, p. 763 (Table 2).

10.15 *Cracking in bottom ash waste asphalt.* A common byproduct of burning municipal solid waste is bottom ash. The use of bottom ash waste in the building of asphalt roads was investigated in the *Journal of Civil Engineering and Construction Technology* (Feb. 2013). Of particular interest was the relationship between the cracking rate of bottom ash waste asphalt and stress intensity. A mixture of bottom ash waste asphalt was prepared and 15 slabs produced. Stress intensity (number of load cycle applications per millimeter) was varied for each slab, then the cracking growth rate (millimeters per cycle) was measured. The data (simulated from information in the journal article) are listed in the table on p. 500. Crack growth rate ($y$) was modeled as a function of stress intensity ($x$) using the Paris Law power function: $y = ax^b$, where a and b are unknown constants.

a. Note that if you take the natural logarithm of both sides of the equation, you obtain the expression: $\ln(y) = \ln(a) + b\ln(x)$. This is the equation of a straight line relating log of crack growth rate to log of stress intensity. Fit this straight-line model to the data and give the least-squares prediction equation.

b. Based on the results, part a, how much would you expect the log of crack growth rate to change as the log of stress intensity increases by 1 unit?

**BOTASH**

| Slab | Stress | Crack Rate |
|---|---|---|
| 1 | 0.05 | 0.004 |
| 2 | 0.10 | 0.304 |
| 3 | 0.15 | 0.016 |
| 4 | 0.20 | 0.150 |
| 5 | 0.25 | 0.116 |
| 6 | 0.30 | 0.098 |
| 7 | 0.35 | 0.008 |
| 8 | 0.40 | 0.044 |
| 9 | 0.45 | 0.551 |
| 10 | 0.50 | 1.283 |
| 11 | 0.55 | 0.365 |
| 12 | 0.60 | 0.080 |
| 13 | 0.65 | 9.161 |
| 14 | 0.70 | 0.097 |
| 15 | 0.75 | 1.711 |

10.16 *Wind turbine blade stress.* Mechanical engineers at the University of Newcastle (Australia) investigated the use of timber in high-efficiency small wind turbine blades (*Wind Engineering*, Jan. 2004). The strengths of two types of timber—radiata pine and hoop pine—were compared. Twenty specimens (called "coupons") of each timber blade were fatigue tested by measuring the stress (in MPa) on the blade after various numbers of blade cycles. A simple linear regression analysis of the data—one conducted for each type of timber—yielded the following results (where $y$ = stress and $x$ = natural logarithm of number of cycles):

**Radiata pine:** $\hat{y} = 97.37 - 2.50x$

**Hoop pine:** $\hat{y} = 122.03 - 2.36x$

a. Interpret the estimated slope of each line.

b. Interpret the estimated $y$-intercept of each line.

c. Based on these results, which type of timber blade appears to be stronger and more fatigue-resistant? Explain.

## Theoretical Exercises

10.17 The maximum likelihood estimator of the mean $\mu$ of a normal distribution is the sample mean $\bar{y}$. Consider the model $E(y) = \mu$. Show that the least-squares estimator of $\mu$ is also $\bar{y}$. [*Hint*: Minimize SSE $= \sum(y_i - \hat{\mu})^2$ with respect to $\hat{\mu}$.]

10.18 Consider the pair of simultaneous linear equations:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum x_i = \sum y_i$$
$$\hat{\beta}_0\sum x_i + \hat{\beta}_1\sum x_i^2 = \sum x_i y_i$$

Derive the formulas for the least-squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$.

# 10.4 Properties of the Least-Squares Estimators

An examination of the formulas for the least-squares estimators reveals that they are linear functions of the observed $y$ values, $y_1, y_2, \ldots, y_n$. Since we have assumed (Section 10.2) that the random errors associated with these $y$ values, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent, normally distributed random variables with mean 0 and variance $\sigma^2$, it follows that the $y$ values will be normally distributed with mean $E(y) = \beta_0 + \beta_1 x$ and variance $\sigma^2$ and that $\hat{\beta}_0$ and $\hat{\beta}_1$ will possess sampling distributions that are normally distributed (Theorem 6.10).

The mean and the variance of the sampling distribution of $\hat{\beta}_1$ are given in Section 10.6. We will illustrate how they are acquired in Example 10.2.

**Example 10.2**

Deriving $E(\hat{\beta}_1)$ and $V(\hat{\beta}_1)$

Find the mean and variance of the sampling distribution of $\hat{\beta}_1$.

Solution

The quantity $SS_{xx}$ that appears in the formula for $\hat{\beta}_1$ involves only the $x$ values, which are assumed to be known—i.e., nonrandom. Therefore, $SS_{xx}$ can be treated as a constant when we find the expected value of $\hat{\beta}_1$. In contrast, $SS_{xy}$ is a function of the random variables, $y_1, y_2, \ldots, y_n$. Thus,

$$\begin{aligned} SS_{xy} &= \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum[(x_i - \bar{x})(y_i) - (x_i - \bar{x})\bar{y}] \\ &= \sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x}) \end{aligned}$$

But

$$\sum(x_i - \bar{x}) = \sum x_i - n\bar{x} = \sum x_i - \sum x_i = 0$$

Therefore, $SS_{xy} = \Sigma(x_i - \bar{x})y_i$. Substituting this quantity into the formula for $\hat{\beta}_1$, we obtain

$$\begin{aligned} \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} = \frac{1}{SS_{xx}}\sum(x_i - \bar{x})y_i \\ &= \frac{(x_1 - \bar{x})}{SS_{xx}}y_1 + \frac{(x_2 - \bar{x})}{SS_{xx}}y_2 + \cdots + \frac{(x_n - \bar{x})}{SS_{xx}}y_n \end{aligned}$$

This shows that $\hat{\beta}_1$ is a linear function of the normally distributed random variables, $y_1, y_2, \ldots, y_n$. The coefficients, $a_1, a_2, \ldots, a_n$, of the random variables in the linear function are

$$a_1 = \frac{(x_1 - \bar{x})}{SS_{xx}} \qquad a_2 = \frac{(x_2 - \bar{x})}{SS_{xx}} \qquad \ldots \qquad a_n = \frac{(x_n - \bar{x})}{SS_{xx}}$$

The final step in finding the mean $E(\hat{\beta}_1)$ and the variance $V(\hat{\beta}_1)$ of the sampling distribution of $\hat{\beta}_1$ is to apply Theorem 6.8, which gives the rule for finding the mean and the variance of a linear function of random variables. Thus,

$$E(\hat{\beta}_1) = E\left[\frac{(x_1 - \bar{x})}{SS_{xx}}y_1 + \frac{(x_2 - \bar{x})}{SS_{xx}}y_2 + \cdots + \frac{(x_n - \bar{x})}{SS_{xx}}y_n\right]$$

where $y_1, y_2, \ldots, y_n$ are obtained by substituting the appropriate values of $x$ into the formula for the linear model, i.e.,

$$\begin{array}{lll} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 & \text{and} & E(y_1) = \beta_0 + \beta_1 x_1 \\ y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2 & \text{and} & E(y_2) = \beta_0 + \beta_1 x_2 \\ \vdots & & \vdots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n & \text{and} & E(y_n) = \beta_0 + \beta_1 x_n \end{array}$$

Therefore,

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{(x_1 - \bar{x})}{SS_{xx}}E(y_1) + \frac{(x_2 - \bar{x})}{SS_{xx}}E(y_2) + \cdots + \frac{(x_n - \bar{x})}{SS_{xx}}E(y_n) \\ &= \frac{(x_1 - \bar{x})}{SS_{xx}}(\beta_0 + \beta_1 x_1) + \frac{(x_2 - \bar{x})}{SS_{xx}}(\beta_0 + \beta_1 x_2) + \cdots \\ &\quad + \frac{(x_n - \bar{x})}{SS_{xx}}(\beta_0 + \beta_1 x_n) \\ &= \frac{\beta_0}{SS_{xx}}\sum(x_i - \bar{x}) + \frac{\beta_1}{SS_{xx}}\sum(x_i - \bar{x})x_i \end{aligned}$$

But,

$$\text{SS}_{xx} = \sum (x_i - \bar{x})^2 = \sum [(x_i - \bar{x})x_i - \bar{x}(x_i - \bar{x})]$$
$$= \sum (x_i - \bar{x})x_i - \bar{x} \sum (x_i - \bar{x})$$

Since we have already shown that $\sum (x_i - \bar{x}) = 0$, we have $\text{SS}_{xx} = \sum (x_i - \bar{x})x_i$ and therefore,

$$E(\hat{\beta}_1) = 0 + \frac{\beta_1}{\text{SS}_{xx}}(\text{SS}_{xx}) = \beta_1$$

This shows that $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$.

Applying the formula given in Theorem 6.8 for finding the variance of a linear function of random variables, and remembering that the covariance between any pair of $y$ values will equal 0 because all pairs of $y$ values are assumed to be independent, we have

$$V(\hat{\beta}_1) = \frac{(x_1 - \bar{x})^2}{(\text{SS}_{xx})^2} V(y_1) + \frac{(x_2 - \bar{x})^2}{(\text{SS}_{xx})^2} V(y_2) + \cdots + \frac{(x_n - \bar{x})^2}{(\text{SS}_{xx})^2} V(y_n)$$

According to the assumptions made in Section 10.2 $V(y_1) = V(y_2) = \cdots = V(y_n) = \sigma^2$. Therefore,

$$V(\hat{\beta}_1) = \frac{(x_1 - \bar{x})^2}{(\text{SS}_{xx})^2}\sigma^2 + \frac{(x_2 - \bar{x})^2}{(\text{SS}_{xx})^2}\sigma^2 + \cdots + \frac{(x_n - \bar{x})^2}{(\text{SS}_{xx})^2}\sigma^2$$
$$= \frac{\sigma^2 \sum (x_i - \bar{x})^2}{(\text{SS}_{xx})^2} = \sigma^2 \frac{\text{SS}_{xx}}{(\text{SS}_{xx})^2} = \frac{\sigma^2}{\text{SS}_{xx}}$$

and

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\text{SS}_{xx}}}$$

We will use the results of Example 10.2 in Section 10.6 to test hypotheses about and to construct a confidence interval for the slope $\beta_1$ of a regression line. The practical implications of these inferences will also be explained.

## Theoretical Exercises

10.19 Show that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\text{SS}_{xx}}\right] y_i$$

[*Hint*: Note that

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\text{SS}_{xx}}$$
$$= \frac{\sum (x_i - \bar{x})y_i}{\text{SS}_{xx}} - \frac{\bar{y} \sum (x_i - \bar{x})}{\text{SS}_{xx}}$$
$$= \frac{\sum (x_i - \bar{x})y_i}{\text{SS}_{xx}}$$

since $\sum (x_i - \bar{x}) = 0$.]

10.20 We showed in Example 10.2 that $\hat{\beta}_1$, the least-squares estimator of the slope $\beta_1$, is an unbiased estimator of $\beta_1$, i.e., $E(\hat{\beta}_1) = \beta_1$. Use the result from Exercise 10.17 to show that $E(\hat{\beta}_0) = \beta_0$.

10.21 In Exercise 10.19, you showed that $\hat{\beta}_0$ could be written as a linear function of independent random variables. Use Theorem 6.8 to show that

$$V(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(\frac{\sum x_i^2}{\text{SS}_{xx}}\right)$$

## 10.5 An Estimator of $\sigma^2$

In most practical situations, the variance $\sigma^2$ of the random error $\varepsilon$ will be unknown and must be estimated from the sample data. Since $\sigma^2$ measures the variation of the $y$ values about the line $E(y) = \beta_0 + \beta_1 x$, it seems intuitively reasonable to estimate $\sigma^2$ by dividing SSE by an appropriate number. Theorem 10.1, an extension of Theorem 6.11, will be useful in obtaining an unbiased estimator.

**THEOREM 10.1**

Let $s^2 = \text{SSE}/(n-2)$. Then, when the assumptions of Section 10.2 are satisfied, the statistic

$$\chi^2 = \frac{\text{SSE}}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2}$$

possesses a chi-square distribution with $\nu = (n-2)$ degrees of freedom.

From Theorem 10.1, it follows that

$$s^2 = \frac{\chi^2\sigma^2}{n-2}$$

Then

$$E(s^2) = \frac{\sigma^2}{n-2}E(\chi^2)$$

where $E(\chi^2) = \nu = (n-2)$. Therefore,

$$E(s^2) = \frac{\sigma^2}{n-2}(n-2) = \sigma^2$$

and we conclude that $s^2$ is an unbiased estimator of $\sigma^2$.

The procedure used in Table 10.3 to calculate SSE can lead to large rounding errors. The formula for $s^2$ and an appropriate method for calculating SSE are shown in the box below. We will illustrate the calculation of $s^2$ with Example 10.3.

**Estimation of $\sigma^2$**

$$s^2 = \frac{\text{SSE}}{\text{Degrees of freedom for error}} = \frac{\text{SSE}}{n-2}$$

where

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \text{SS}_{yy} - \hat{\beta}_1\text{SS}_{xy}$$

$$\text{SS}_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

*Warning:* When performing these calculations, you may be tempted to round the calculated values of $\text{SS}_{yy}$, $\hat{\beta}_1$, and $\text{SS}_{xy}$. Be certain to carry at least six significant figures for each of these quantities to avoid substantial errors in the calculation of SSE.

**Example 10.3**

Estimating $\sigma^2$

Estimate $\sigma^2$ for the data of Table 10.1.

**Solution**

In the insulation compression example, we previously calculated SSE = 1.10 for the least-squares line $\hat{y} = -.1 + .7x$. Recalling that there were $n = 5$ data points, we have $n - 2 = 5 - 2 = 3$ df for estimating $\sigma^2$. Thus,

$$s^2 = \frac{\text{SSE}}{n - 2} = \frac{1.10}{3} = .367$$

is the estimated variance, and

$$s = \sqrt{.367} = .606$$

is the estimated standard deviation of $\varepsilon$. Both these values are highlighted on the MINITAB printout, Figure 10.7.

```
The regression equation is
COMP_Y = - 0.100 + 0.700 PRESS_X


Predictor     Coef  SE Coef      T      P
Constant   -0.1000   0.6351  -0.16  0.885
PRESS_X     0.7000   0.1915   3.66  0.035


S = 0.605530   R-Sq = 81.7%   R-Sq(adj) = 75.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  4.9000  4.9000  13.36  0.035
Residual Error   3  1.1000  0.3667
Total            4  6.0000
```

**FIGURE 10.7**
MINITAB simple linear regression of data in Table 10.1

You may be able to obtain an intuitive feeling for $s$ by recalling the interpretation given to a standard deviation in Chapter 2 and remembering that the least-squares line estimates the mean value of $y$ for a given value of $x$. Since $s$ measures the spread of the distribution of $y$ values about the least-squares line, we should not be surprised to find that most of the observations lie within $2s$, or $2(.606) = 1.21$, of the least-squares line. For this simple example (only five data points), all five data points fall within $2s$ of the least-squares line. In Section 10.9, we will use $s$ to evaluate the error of prediction when the least-squares line is used to predict a value of $y$ to be observed for a given value of $x$.

### Interpretation of $s$, the Estimated Standard Deviation of $\varepsilon$

We expect most of the observed $y$ values to lie within $2s$ of their respective least-squares predicted values, $\hat{y}$.

## Applied Exercises

**BLOODTYPE**

10.22 *New method for blood typing.* Refer to the *Analytical Chemistry* (May 2010) study in which medical researchers tested a new method of typing blood using lost cost paper, Exercise 10.6 (p. 496). The data were used to fit the straight-line model relating $y$ = wicking length to $x$ = antibody concentration.

a. Give the values of SSE, $s^2$, and $s$ shown on the SPSS printout below.
b. Give a practical interpretation of s. Recall that wicking length is measured in millimeters.

10.23 *Interpreting the standard deviation.* Calculate SSE, $s^2$, and $s$ for the least-squares lines in:

a. Exercise 10.7
b. Exercise 10.8
c. Exercise 10.9
d. Exercise 10.10
e. Exercise 10.11
f. Exercise 10.12
g. Exercise 10.13
h. Exercise 10.14
i. Exercise 10.15

Interpret the value of $s$ for each line.

10.24 *Thickness of dust on solar cells.* The performance of a solar cell can deteriorate when atmospheric dust accumulates on the solar panel surface. In the *International Journal of Energy and Environmental Engineering* (Dec. 2012), researchers at the Renewable Energy Research Laboratory, University of Lucknow (India) estimated the relationship between the dust thickness and the efficiency of a solar cell. The thickness of dust (in millimeters) collected on a solar cell was measured three times per month over a year-long period. Each time the dust thickness was measured, the researchers also determined the percentage difference (before minus after dust collection) in efficiency of the solar panel. Data (monthly averages) for the 10 months where there was no rain are listed in the table.

**SOLARCELL**

| Month | Efficiency (% change) | Average Dust Thickness (mm) |
|---|---|---|
| January | 1.5666 | 0.00024 |
| February | 1.9574 | 0.00105 |
| March | 1.3707 | 0.00075 |
| April | 1.9563 | 0.00070 |
| May | 1.6332 | 0.00142 |
| June | 1.8172 | 0.00055 |
| July | 0.9202 | 0.00039 |
| October | 1.8790 | 0.00095 |
| November | 1.5544 | 0.00064 |
| December | 2.0198 | 0.00065 |

*Source:* Siddiqui, R. & Bajpai, U. "Correlation between thicknesses of dust collected on photovoltaic module and difference in efficiencies in composite climate", *International Journal of Energy and Environmental Engineering,* Vol. 4, No. 1, December 2012 (Table 1).

a. Fit the linear model, $E(y) = \beta_0 + \beta_1 x$, to the data where $y$ = Efficiency and $x$ = average dust thickness.
b. Find an estimate of $\sigma$, the true standard deviation of the error, $\epsilon$.
c. Give a practical interpretation of the result, part b.

10.25 *New iron-making process.* An innovative new iron-making technology (called ITmk3) produces high-quality iron nuggets directly from raw iron ore and coal. *Mining Engineering* (Oct. 2004) published the results of pilot tests conducted on the new process. For one phase of the study, the carbon content produced in a pilot plant test was compared to that from laboratory furnace tests. The data for 25 pilot tests are listed in the table on p. 506.

### SPSS Output for Exercise 10.22

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .819[a] | .670 | .587 | 2.35944 |

a. Predictors: (Constant), ABConc

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 45.201 | 1 | 45.201 | 8.119 | .046[a] |
| | Residual | 22.268 | 4 | 5.567 | | |
| | Total | 67.469 | 5 | | | |

a. Predictors: (Constant), ABConc

b. Dependent Variable: WickLength

a. Plot the data points on a scattergram.
b. Fit a simple linear model relating carbon content in a pilot test, $y$, to the carbon content in a lab furnace, $x$. Interpret the estimates of the model parameters.
c. Compute SSE and $s^2$.
d. Compute $s$ and interpret its value.

**CARBON**

| Carbon Content (%) | | Carbon Content (%) | |
|---|---|---|---|
| Pilot Plant | Lab Furnace | Pilot Plant | Lab Furnace |
| 1.7 | 1.6 | 3.4 | 4.3 |
| 3.1 | 2.4 | 3.2 | 3.6 |
| 3.3 | 2.8 | 3.3 | 3.4 |
| 3.6 | 2.9 | 3.1 | 3.3 |
| 3.4 | 3.0 | 3.0 | 3.2 |
| 3.5 | 3.1 | 2.9 | 3.2 |
| 3.8 | 3.2 | 2.6 | 3.4 |
| 3.7 | 3.2 | 2.5 | 3.3 |
| 3.5 | 3.3 | 2.6 | 3.2 |
| 3.4 | 3.3 | 2.6 | 3.1 |
| 3.6 | 3.4 | 2.4 | 3.0 |
| 3.5 | 3.4 | 2.6 | 2.7 |
| 3.9 | 3.8 | | |

*Source*: Hoffman, G., and Tsuge, O. "ITmk3—Application of a new ironmaking technology for the iron ore mining industry." *Mining Engineering*, Vol. 56, No. 9, October 2004 (Figure 8).

10.26 *Drug controlled-release rate study.* Researchers at Dow Chemical Co. investigated the effect of tablet surface area and volume on the rate at which a drug is released in a controlled-release dosage. (*Drug Development and Industrial Pharmacy*, Vol. 28, 2002.) Six similarly shaped tablets were prepared with different weights and thicknesses and the ratio of surface area to volume was measured for each. Using a dissolution apparatus, each tablet was placed in 900 milliliters of de-ionized water and the diffusional drug release rate (percentage of drug released divided by the square root of time) determined. The experimental data are listed in the accompanying table.

a. Fit the simple linear model, $E(y) = \beta_0 + \beta_1 x$, where $y$ = drug release rate and $x$ = surface-area-to-volume ratio.
b. Compute SSE, $s^2$, and $s$.
c. Interpret the value of $s$.

**DOWDRUG**

| Drug Release Rate (% *released*/$\sqrt{time}$) | Surface Area to Volume ($mm^2/mm^3$) |
|---|---|
| 60 | 1.50 |
| 48 | 1.05 |
| 39 | .90 |
| 33 | .75 |
| 30 | .60 |
| 29 | .65 |

*Source*: Reynolds, T., Mitchell, S., and Balwinski, K. "Investigation of the effect of tablet surface area/volume on drug release from Hydroxypropylmethylcellulose controlled-release matrix tablets." *Drug Development and Industrial Pharmacy*, Vol. 28, No. 4, 2002 (Figure 3).

10.27 *Single machine batch scheduling.* In a manufacturing process that involves a single machine, decisions must be made on whether to deliver the product to the customer immediately upon completion of the job or to hold the finished product in a batch with other jobs to be delivered at a later time. A computerized mathematical model for solving the batch scheduling problem was proposed in the *Asian Journal of Industrial Engineering* (Vol. 4, 2012). The performance of the model was graded using a variable called Value of Object Function (VOF). Simulation yielded the following data on VOF and run time (in seconds) for six software runs, each run with a different number of held batches.

**SWRUN**

| Software Run | Number of Batches | VOF | Run Time (seconds) |
|---|---|---|---|
| 1 | 3 | 86.68 | 27 |
| 2 | 4 | 232.87 | 14 |
| 3 | 5 | 372.36 | 12 |
| 4 | 6 | 496.51 | 18 |
| 5 | 7 | 838.82 | 42 |
| 6 | 8 | 1183.00 | 33 |

*Source:* Karimi-Nasab, M., Haddad, H., & Ghanbari, P. "A simulated annealing for the single machine batch scheduling deterioration and precedence constraints", *Asian Journal of Industrial Engineering,* Vol. 4, No. 1, 2012 (Table 2).

a. Use simple linear regression to estimate the equation, $y = \beta_0 + \beta_1 x + \epsilon$, where $y$ = VOF and $x$ = number of batches.
b. Find an estimate of $\sigma^2 = V(\epsilon)$, and $\sigma = \sqrt{V(\epsilon)}$.
c. Which of the two estimates, part b, can be practically interpreted? Give the interpretation.

## Theoretical Exercises

10.28 Show that $V(s^2) = 2\sigma^4/(n-2)$. [*Hint*: The result follows from Theorem 10.1 and the fact that $V(\chi^2) = 2\nu$.]

10.29 Verify that $\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy}$.

## 10.6 Assessing the Utility of the Model: Making Inferences About the Slope $\beta_1$

Refer again to the data of Table 10.1 and suppose that the compression of the insulation material is *completely unrelated* to the pressure. What could be said about the values of $\beta_0$ and $\beta_1$ in the hypothesized probabilistic model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

if $x$ contributes no information for the prediction of $y$? The implication is that the mean of $y$, i.e., the deterministic part of the model $E(y) = \beta_0 + \beta_1 x$, does not change as $x$ changes. Regardless of the value of $x$, you always predict the same value of $y$. In the straight-line model, this means that the true slope, $\beta_1$, is equal to 0. Therefore, to test the null hypothesis that $x$ contributes no information for the prediction of $y$ against the alternative hypothesis that these variables are linearly related with a slope differing from 0, we test

$$H_0: \quad \beta_1 = 0$$
$$H_a: \quad \beta_1 \neq 0$$

If the data support the alternative hypothesis, we will conclude that $x$ does contribute information for the prediction of $y$ using the straight-line model [although the true relationship between $E(y)$ and $x$ could be more complex than a straight line]. Thus, to some extent, this is a test of the utility of the hypothesized model.

The appropriate test statistic is found by considering the sampling distribution of $\hat{\beta}_1$, the least-squares estimator of the slope $\beta_1$. The sampling distribution of this statistic (discussed in Section 9.4) is described in the box.

### Sampling Distribution of $\hat{\beta}_1$

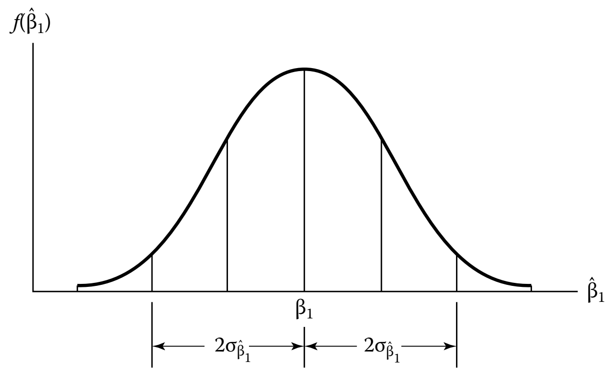
If we make the four assumptions about $\varepsilon$ (see Section 10.2), then the sampling distribution of $\hat{\beta}_1$, the least-squares estimator of slope, will be a normal distribution with mean $\beta_1$ (the true slope) and standard error

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\mathrm{SS}_{xx}}} \approx \frac{s}{\sqrt{\mathrm{SS}_{xx}}} \qquad \text{(see Figure 10.8)}$$

Since $\sigma$ will usually be unknown, the appropriate test statistic will generally be a Student's $T$ statistic formed as follows:

$$T = \frac{\hat{\beta}_1 - \text{Hypothesized value of } \beta_1}{s_{\hat{\beta}_1}} \qquad \text{where } s_{\hat{\beta}_1} = \frac{s}{\sqrt{\mathrm{SS}_{xx}}}$$
$$= \frac{\hat{\beta}_1 - 0}{s/\sqrt{\mathrm{SS}_{xx}}}$$

**FIGURE 10.8**
Sampling distribution of $\hat{\beta}_1$



Note that we have substituted the estimator $s$ for $\sigma$ and then formed $s_{\hat{\beta}_1}$ by dividing $s$ by $\sqrt{SS_{xx}}$. The number of degrees of freedom associated with this $t$ statistic is the same as the number of degrees of freedom associated with $s$. Recall that this will be $(n - 2)$ df when the hypothesized model is a straight line (see Section 10.5).

The setup of our test of the utility of the model is summarized in the box.*

### A Test of Model Utility: Simple Linear Regression

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\beta_1 = 0$ | $H_0$: $\beta_1 = 0$ |
| $H_a$: $\beta_1 < 0$ (or $H_a$: $\beta_1 > 0$) | $H_a$: $\beta_1 \neq 0$ |

*Test statistic*: $T_c = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$

| | |
|---|---|
| *Rejection region:* $T_c < -t_\alpha$ (or $T_c > t_\alpha$) | *Rejection region:* $\lvert T_c \rvert > t_{\alpha/2}$ |
| $p$-value: $P(T < T_c)$ [or $P(T > T_c)$] | $p$-value: $2P(T > \lvert T_c \rvert)$ |

where $t_\alpha$ and $t_{\alpha/2}$ are based on $(n - 2)$ df and obtained from Table 7 of Appendix B.

*Assumptions:* The four assumptions about $\varepsilon$ are listed in Section 10.2.

**Example 10.4**

Testing the slope, $\beta_1$

Refer to Examples 10.1 and 10.3, and test the hypothesis that $\beta_1 = 0$.

Solution

For the insulation compression example, we will choose $\alpha = .05$ and, since $n = 5$, df $= (n - 2) = 5 - 2 = 3$. Then the rejection region for the two-tailed test is

$$T < -t_{.025} \quad \text{or} \quad T > t_{.025}$$

where $t_{.025}$, given in Table 7 of Appendix B, is $t_{.025} = 3.182$. We previously calculated $\hat{\beta}_1 = .7$, $s = .606$, and $SS_{xx} = 10$. Thus,

$$T = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} = \frac{.7}{.606/\sqrt{10}} = 3.66$$

Since this calculated $t$ value falls in the upper-tail rejection region (see Figure 10.9 on page 509), we reject the null hypothesis and conclude that the slope $\beta_1$ is not 0. The sample evidence indicates that $x$ contributes information for the prediction of $y$ using a linear model for the relationship between compression and pressure.

*Note*: We can reach the same conclusion by using the observed significance level ($p$-value) of the test obtained from a computer printout. The SAS printout for the simple linear regression is reproduced in Figure 10.10. Both the test statistic and two-tailed $p$-value are highlighted on the printout. Since $p$-value $= .0354$ is smaller than $\alpha = .05$, we will reject $H_0$.

*A test of hypothesis for $\beta_0$ is rarely of practical importance in simple linear regression. For the sake of completeness, the test statistic is

$$T = \frac{\hat{\beta}_0 - \text{Hypothesized value of } \beta_0}{s\sqrt{(1/n) + (\bar{x})^2/SS_{xx}}}$$

which, given the standard assumption on $\varepsilon$, follows a Student's $T$ distribution with $(n - 2)$ df.

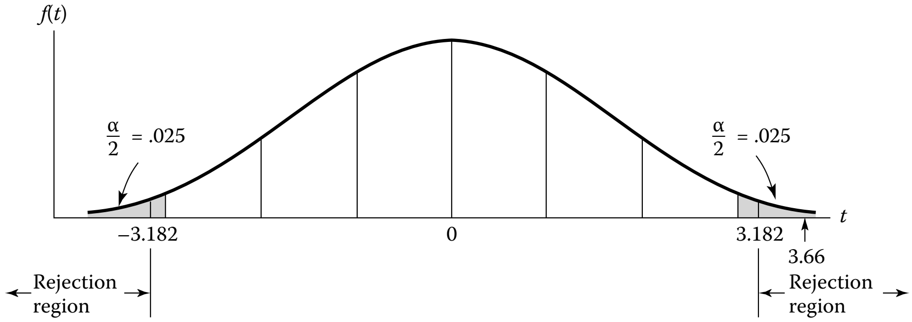



**FIGURE 10.9**
Rejection region and calculated $t$ value for testing whether the slope $\beta_1 = 0$

```
                          The REG Procedure
                            Model: MODEL1
                      Dependent Variable: COMP_Y

               Number of Observations Read          5
               Number of Observations Used          5

                         Analysis of Variance

                                    Sum of          Mean
Source                    DF       Squares        Square    F Value    Pr > F

Model                      1       4.90000       4.90000      13.36    0.0354
Error                      3       1.10000       0.36667
Corrected Total            4       6.00000

              Root MSE              0.60553    R-Square     0.8167
              Dependent Mean        2.00000    Adj R-Sq     0.7556
              Coeff Var            30.27650

                         Parameter Estimates

                    Parameter     Standard
Variable    DF       Estimate        Error   t Value   Pr > |t|     95% Confidence Limits

Intercept    1       -0.10000      0.63509     -0.16     0.8849     -2.12112      1.92112
PRESS_X      1        0.70000      0.19149      3.66     0.0354      0.09061      1.30939
```

**FIGURE 10.10**
SAS simple linear regression of data in Table 10.1

Another way to make inferences about the slope $\beta_1$ is to estimate it using a confidence interval. This interval is formed as shown in the box.

### A $(1-\alpha)100\%$ Confidence Interval for the Slope $\beta_1$

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} \qquad \text{where} \qquad s_{\hat{\beta}_1} = \frac{s}{\sqrt{\mathrm{SS}_{xx}}}$$

and $t_{\alpha/2}$ is based on $(n-2)$ df

**Example 10.5**

Confidence Internal for the Slope, $\beta_1$

Find a 95% confidence interval for $\beta_1$ in Example 10.1. Interpret the result.

Solution

For the insulation compression example, a 95% confidence interval for the slope $\beta_1$ is

$$\hat{\beta}_1 \pm t_{.025} s_{\hat{\beta}_1} = .7 \pm 3.182\left(\frac{s}{\sqrt{\mathrm{SS}_{xx}}}\right)$$

$$= .7 \pm 3.182\left(\frac{.61}{\sqrt{10}}\right) = .7 \pm .61 = (.09, 1.31)$$

(*Note*: This confidence interval is highlighted on the SAS printout, Figure 10.10.) Thus, we estimate that the interval from .09 to 1.31 includes the slope parameter $\beta_1$. Remembering that $y$ is recorded in units of .1 inch and $x$ in units of 10 pounds per square inch, we can say, with 95% confidence, that the mean compression, $E(y)$, will increase between .009 and .131 inch for every 10-pound-per-square-inch increase in pressure, $x$.

Since all the values in this interval are positive, it appears that $\beta_1$ is positive and that the mean of $y$, $E(y)$, increases as $x$ increases. However, the rather large width of the confidence interval reflects the small number of data points (and, consequently, a lack of information) in the experiment. We would expect a narrower interval if the sample size were increased.

Before concluding this section, we call your attention to the similarity between the $t$ statistic for testing hypotheses about $\beta_1$ and the $t$ statistic for testing hypotheses about the means of normal populations in Chapter 8. Also note the similarity of the corresponding confidence intervals. In each case, the general form of the test statistic is

$$T = \frac{\hat{\theta} - \theta_0}{s_{\hat{\theta}}}$$

and the general form of the confidence interval is

$$\hat{\theta} \pm (t_{\alpha/2}) s_{\hat{\theta}}$$

where $\hat{\theta}$ is the estimator of the population parameter $\theta$, $\theta_0$ is the hypothesized value of $\theta$, and $s_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$.

In the optional exercises of this section, we outline the procedure for acquiring the $T$ statistic for testing hypotheses about and constructing confidence intervals for $\beta_1$.

## Applied Exercises

**BLOODTYPE**

10.30 *New method for blood typing.* Refer to the *Analytical Chemistry* (May 2010) study in which medical researchers tested a new method of typing blood using lost cost paper, Exercises 10.6 and 10.22 (p. 505). Recall that the data was used to fit the straight-line model relating $y$ = wicking length (millimeters) to $x$ = antibody concentration. A portion of the SPSS printout not previously shown is displayed on p. 511. Use the information on the printout to find a 95% confidence interval for the slope of the line. Give a practical interpretation of the interval.

## SPSS Output for Exercise 10.30

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 19.393 | 1.708 | | 11.357 | .000 | 14.652 | 24.134 |
| | ABConc | -8.036 | 2.820 | -.819 | -2.849 | .046 | -15.865 | -.206 |

a. Dependent Variable: WickLength

**WATERPIPE**

10.31 *Estimating repair and replacement costs of water pipes.* Refer to the *IHS Journal of Hydraulic Engineering* (September 2012) study of water pipes susceptible to breakage, Exercise 10.8 (p. 497). Recall that civil engineers used simple linear regression to model $y$ = the ratio of repair to replacement cost of commercial pipe as a function of $x$ = the diameter (in millimeters) of the pipe. Are the engineers able to conclude (at $\alpha = .05$) that the cost ratio increases linearly with pipe diameter? If so, provide a 95% confidence interval for the increase in cost ratio for every 1 millimeter increase in pipe diameter.

**SMELTPOT**

10.32 *Extending the life of an aluminum smelter pot.* Refer to *The American Ceramic Society Bulletin* (Feb. 2005) evaluation of commercial bricks used in smelter pots, Exercise 10.9 (p. 497). You fit a straight-line model relating the apparent porosity $y$ (in percent) of the brick to mean pore diameter $x$ (in micrometers) using the data provided.

a. Find a 95% confidence interval for the true slope of the line. Interpret the result.
b. Conduct a test (at $\alpha = .05$) to determine if the true slope of the line differs from 0.
c. Demonstrate that the two inferences, parts **a** and **b**, give the same information on the utility of the straight-line model.

**LIQUIDSPILL**

10.33 *Spreading rate of spilled liquid.* Refer to the *Chemical Engineering Progress* (Jan. 2005) study of the rate at which a spilled volatile liquid will spread across a surface, Exercise 10.10 (p. 497). You fit a straight-line model relating the mass $y$ (in pounds) of the spill to elapsed time $x$ (in minutes) using the data provided.

a. Find a 90% confidence interval for the true slope of the line. Interpret the result.
b. Conduct a test (at $\alpha = .10$) to determine if the true slope of the line differs from 0.
c. Demonstrate that the two inferences, parts **a** and **b**, give the same information on the utility of the straight-line model.

**RAINFALL**

10.34 *New method of estimating rainfall.* Refer to the *Journal of Data Science* (Apr. 2004) comparison of methods for estimating rainfall. Exercises 10.11 (p. 498). Consider the simple linear regression relating rain gauge amount ($y$) to the artificial neural network rain estimate ($x$).

a. Test whether $y$ is positively related to $x$. Use $\alpha = .10$.
b. Construct a 90% confidence interval for $\beta_1$. Practically interpret the result.

**OJUICE**

10.35 *Sweetness of orange juice.* Refer to Exercise 10.12 (p. 498) and the simple linear regression relating the sweetness index ($y$) of an orange juice sample to the amount of water-soluble pectin ($x$) in the juice. Find a 90% confidence interval for the true slope of the line. Interpret the result.

**FINTUBES**

10.36 *Thermal performance of copper tubes.* Refer to the *Journal of Heat Transfer* study of the straight-line relationship between heat transfer enhancement, $y$, and unflooded area ratio, $x$, Exercise 10.14 (p. 499). Construct a 95% confidence interval for $\beta_1$, the slope of the line. Interpret the result.

10.37 *Planning an ecological network.* A new method of planning an ecological network was presented in *Landscape Ecology Engineering* (Jan. 2013). The methodology protects linear green areas that act as ecological corridors for potential movement paths of wild animals (e.g., birds). This requires a prediction of the bird density in the green area. In a sample of 21 bird habitats in China, the researchers determined the bird density (number of birds per hectare) and the percentage of the habitat covered by vegetation (i.e., a green area). Data similar to the data reported in the journal article are listed in the table on p. 512. The researchers used the data to fit the model, $E(y) = \beta_0 + \beta_1 x$, where $y$ = bird density and $x$ =vegetation coverage (percentage).

a. Graph the points in a scatterplot. What type of linear relationship (positive or negative) appears to exist?
b. Fit the straight-line model to the data and obtain the least squares prediction equation.
c. Is there sufficient evidence to indicate that bird density increases linearly as percent vegetation coverage increases? Test using $\alpha = .01$.

### Data for Exercise 10.37

**BIRDDEN**

| HABITAT | DENSITY (birds/hectare) | COVER (%) |
|---|---|---|
| 1 | 0.3 | 0 |
| 2 | 0.25 | 2 |
| 3 | 2 | 4 |
| 4 | 1 | 6 |
| 5 | 0.5 | 9 |
| 6 | 0 | 10 |
| 7 | 3 | 12 |
| 8 | 5 | 17 |
| 9 | 5 | 20 |
| 10 | 1 | 25 |
| 11 | 6 | 30 |
| 12 | 5 | 37 |
| 13 | 8 | 40 |
| 14 | 2 | 45 |
| 15 | 7 | 50 |
| 16 | 16 | 58 |
| 17 | 5 | 60 |
| 18 | 20 | 71 |
| 19 | 5 | 80 |
| 20 | 37 | 90 |
| 21 | 6 | 100 |

10.38 *Pressure stabilization of fresh concrete.* *Engineering Structures* (July 2013) published a study of the characteristics of fresh concrete. One key variable studied was the time (in hours) needed for pressure stabilization of the concrete. The researchers investigated the effect of time needed for pressure stabilization ($x$) on each of three different dependent variables: $y_1$ = initial setting time (hours), $y_2$ = final setting time (hours), and $y_3$ = maturity index (°C-hours). The data on these variables for $n = 8$ fresh concrete lateral pressure tests are listed in the next table.

a. Construct scattergrams to aid the researchers in determining whether pressure stabilization can be used as a reliable predictor of any of the three dependent variables.

b. Support your answer to part a by running three simple linear regression analyses. Are any of the slopes significantly different from 0? (Test using $\alpha = .05$.) Which one?

**CONCRET2**

| Test | $Y_1$ | $Y_2$ | $Y_3$ | X |
|---|---|---|---|---|
| A1 | 4.63 | 7.17 | 385.81 | 12.03 |
| A2 | 4.32 | 6.52 | 358.44 | 11.32 |
| A3 | 4.54 | 6.31 | 292.71 | 9.51 |
| A4 | 4.09 | 6.19 | 253.16 | 8.25 |
| A5 | 4.56 | 6.81 | 279.82 | 9.02 |
| A6 | 4.48 | 6.98 | 318.74 | 9.97 |
| A7 | 4.35 | 6.45 | 262.14 | 8.42 |
| A8 | 4.23 | 6.69 | 244.97 | 7.53 |

*Source:* Santilli, A., Puente, I., & Tanco, M. "Fresh concrete lateral pressure decay: Kinetics and factorial design to determine significant parameters", *Engineering Structures,* Vol. 52, July 2013 (Table 4).

10.39 *Forest fragmentation study.* Ecologists classify the cause of forest fragmentation as either anthropogenic (i.e., due to human development activities such as road construction or logging) or natural in origin (e.g., due to wetlands or wildfire). *Conservation Ecology* (Dec. 2003) published an article on the causes of fragmentation for 54 South American forests. Using advanced high-resolution satellite imagery, the researchers developed two fragmentation indices for each forest—one index for anthropogenic fragmentation and one for fragmentation from natural causes. The values of these two indices (where higher values indicate more fragmentation) for 5 of the forests in the sample are shown in the accompanying table. The data for all 54 forests are saved in the **FORFRAG** file.

**FORFRAG**

(*First 5 observations listed*)

| Ecoregion (forest) | Anthropogenic Index, $y$ | Natural Origin Index, $x$ |
|---|---|---|
| Araucaria moist forests | 34.09 | 30.08 |
| Atlantic Coast restingas | 40.87 | 27.60 |
| Bahia coastal forests | 44.75 | 28.16 |
| Bahia interior forests | 37.58 | 27.44 |
| Bolivian Yungas | 12.40 | 16.75 |

*Source*: Wade, T.G., et al. "Distribution and causes of global forest fragmentation." *Conservation Ecology*, Vol. 72, No. 2, Dec. 2003 (Table 6).

a. Ecologists theorize that an approximately linear (straight-line) relationship exists between the two fragmentation indices. Graph the data for all 54 forests. Does the graph support the theory?

b. Delete the data for the three forests with the largest anthropogenic indices and reconstruct the graph, part **a.** Comment on the ecologists' theory.

c. Fit the straight-line model to the subset **FORFRAG** data file using the method of least-squares. Give the equation of the least-squares prediction equation.
d. Interpret the estimates of $\beta_0$ and $\beta_1$ in the context of the problem.
e. Is there sufficient evidence to indicate that natural origin index ($x$) and anthropogenic index ($y$) are positively linearly related? Test using $\alpha = .05$.
f. Find and interpret a 95% confidence interval for the change in the anthropogenic index ($y$) for every 1-point increase in the natural origin index ($x$).

## Theoretical Exercises

10.40 Explain why

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\mathrm{SS}_{xx}}}$$

is normally distributed with mean 0 and variance 1 when the four assumptions of Section 10.2 are satisfied.

10.41 It can be shown (proof omitted) that the least-squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are independent (in a probabilistic sense) of $s^2$. Use this fact, in conjunction with Theorem 10.1 and the result of Exercise 10.40, to show that

$$T = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{\mathrm{SS}_{xx}}}$$

has a Student's $T$ distribution with $\nu = (n - 2)$ df.

10.42 Use the $T$ statistic in Exercise 10.41 as a pivotal statistic to derive a $(1 - \alpha)100\%$ confidence interval for $\beta_1$.

# 10.7 The Coefficients of Correlation and Determination

In this section, we introduce two statistics that describe the adequacy of the linear regression model: the *coefficient of correlation* and the *coefficient of determination*.

## Coefficient of Correlation

In Section 10.6, we discovered that the least-squares slope, $\hat{\beta}_1$, provides useful information on the linear relationship, or "association," between two variables $y$ and $x$. Another way to measure association is to compute the **Pearson product moment correlation coefficient $r$**. The correlation coefficient, defined in the box, provides a quantitative measure of the strength of the linear relationship between $x$ and $y$ in the sample, as does the least-squares slope $\hat{\beta}_1$. However, unlike the slope, the correlation coefficient $r$ is *scaleless*. The value of $r$ is always between $-1$ and $+1$, no matter what the units of $x$ and $y$ are.

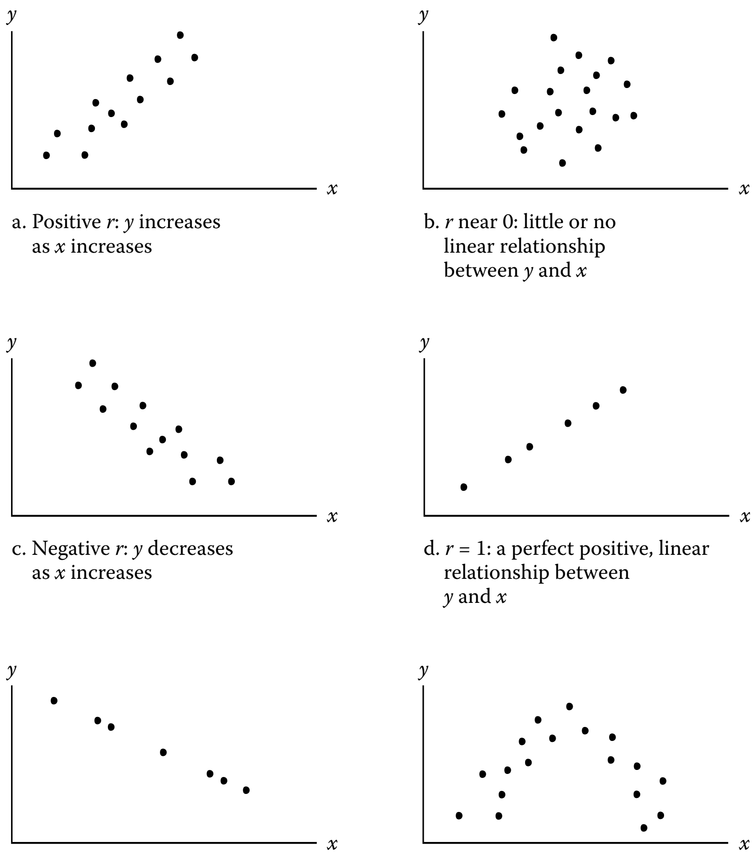
**Definition 10.5**

The **Pearson product moment coefficient of correlation $r$** is a measure of the strength of the linear relationship between two variables $x$ and $y$ in the sample. It is computed (for a sample of $n$ measurements on $x$ and $y$) as follows:

$$r = \frac{\mathrm{SS}_{xy}}{\sqrt{\mathrm{SS}_{xx}\mathrm{SS}_{yy}}}$$

Since both $r$ and $\hat{\beta}_1$ provide information about the utility of the model, it is not surprising that there is a similarity in their computational formulas. In particular, note that $\mathrm{SS}_{xy}$ appears in the numerators of both expressions and, since both denominators are always positive, $r$ and $\hat{\beta}_1$ will always be of the same sign (either both positive or both negative). A value of $r$ near or equal to 0 implies little or no linear relationship between $y$ and $x$. In contrast, the closer $r$ is to 1 or $-1$, the stronger the linear relationship between $y$ and $x$. And, if $r = 1$ or $r = -1$, all the points fall exactly on the least-squares line. Positive values of $r$ imply that $y$ increases as $x$ increases; negative values imply that $y$ decreases as $x$ increases. See Figure 10.11.

**FIGURE 10.11**
Values of $r$ and their implications




a. Positive $r$: $y$ increases as $x$ increases

b. $r$ near 0: little or no linear relationship between $y$ and $x$

c. Negative $r$: $y$ decreases as $x$ increases

d. $r = 1$: a perfect positive, linear relationship between $y$ and $x$

e. $r = -1$: a perfect negative, linear relationship between $y$ and $x$

f. $r$ near 0: little or no linear relationship between $y$ and $x$

## Example 10.6

Finding the Correlation Coefficient, $r$

The data for Example 10.1 are reproduced in Table 10.4. Calculate the coefficient of correlation $r$ between pressure $x$ and compression $y$.

**TABLE 10.4 Compression Versus Pressure for an Insulation Material**

INSULATION

| Pressure $x$, 10 pounds per square inch | Compression $y$, .1 inch |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 2 |
| 4 | 2 |
| 5 | 4 |

Solution

From previous calculations (see Example 10.1), we found $SS_{xy} = 7$, $SS_{xx} = 10$, $\sum y_i = 10$, and $\sum y_i^2 = 26$. Then,

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 26 - \frac{(10)^2}{5} = 26 - 20 = 6$$

and the coefficient of correlation is

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{7}{\sqrt{(10)(6)}} = \frac{7}{7.746} = .904$$

Thus, the pressure and amount of compression are very highly correlated—at least for this sample of five pieces of insulation material. The implication is that a strong positive linear relationship exists between these variables. We must be careful, however, not to jump to any unwarranted conclusions. For instance, the developer of the new insulation material may be tempted to conclude that increasing pressure will always lead to a higher amount of compression. The implication of such a conclusion is that there is a **causal** relationship between the two variables. However, **high correlation does not imply causality**. Many other factors, such as temperature and humidity, may contribute to the increase in the amount of compression produced on the specimens.

### Warning

High correlation does not imply causality. If a large positive or negative value of the sample correlation coefficient $r$ is observed, it is incorrect to conclude that a change in $x$ *causes* a change in $y$. The only valid conclusion is that a linear trend *may* exist between $x$ and $y$.

Keep in mind that the correlation coefficient $r$ measures the correlation between $x$ values and $y$ values in the sample, and that a similar linear coefficient of correlation exists for the population from which the data points were selected. The **population correlation coefficient** is denoted by the symbol $\rho$ (rho). As you might expect, $\rho$ is estimated by the corresponding sample statistic, $r$. Or, rather than estimating $\rho$, we might want to test the hypothesis $H_0$: $\rho = 0$ against $H_a$: $\rho \neq 0$, i.e., test the hypothesis that $x$ contributes no information for the prediction of $y$ using the straight-line model against the alternative that the two variables are at least linearly related. However, we have already performed this identical test in Section 10.6 when we tested $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 \neq 0$.

It is easy to show that $r = \hat{\beta}_1 \cdot \sqrt{SS_{xx}/SS_{yy}}$. Thus, $\hat{\beta}_1 = 0$ implies $r = 0$, and vice versa. Consequently, the null hypothesis $H_0$: $\rho = 0$ is equivalent to the hypothesis $H_0$: $\beta_1 = 0$. When we tested the null hypothesis $H_0$: $\beta_1 = 0$ in connection with the insulation compression example, the data led to a rejection of the hypothesis for $\alpha = .05$. This implies that the null hypothesis of a zero linear correlation between the two variables (pressure and compression) can also be rejected at $\alpha = .05$. The only real difference between the least-squares slope $\hat{\beta}_1$ and the coefficient of correlation $r$ is the measurement scale. Therefore, the information they provide about the utility of the least-squares model is to some extent redundant. Furthermore, the slope $\beta_1$ gives us additional information on the amount of increase (or decrease) in $y$ for every 1-unit increase in $x$. For this reason, we recommend using the slope to make inferences about the existence of a positive or negative linear relationship between two variables. For those who prefer to test for a linear relationship between two variables using the coefficient of correlation $r$, we outline the procedure in the box.

### Test of Hypothesis for Linear Correlation

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\rho = 0$ | $H_0$: $\rho = 0$ |
| $H_a$: $\rho > 0$ (or $\rho < 0$) | $H_a$: $\rho \neq 0$ |

*Test statistic:* $T_c = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

| | |
|---|---|
| *Rejection region:* $T_c > t_\alpha$ (or $T_c < -t_\alpha$) | *Rejection region:* $\lvert T_c \rvert > t_{\alpha/2}$ |
| $p$-value: $P(T > T_c)$ $[\text{or } P(T < T_c)]$ | $p$-value: $2P(T > \lvert T_c \rvert)$ |

where $t_\alpha$ and $t_{\alpha/2}$ are the critical values based on $(n - 2)$df obtained from Table 7 of Appendix B.

*Assumptions:* The sample of $(x, y)$ values is randomly selected from a (bivariate) normal population.*

*The joint probability distribution of $x$ and $y$ will be bivariate normal if the marginal distributions of $x$ and $y$ are both normal.

The next example illustrates how the correlation coefficient $r$ may be a misleading measure of the strength of the association between $x$ and $y$ in situations where the true relationship is nonlinear.

**Example 10.7**

Testing the Correlation Coefficient

Underinflated or overinflated tires can increase tire wear and decrease gas mileage. A manufacturer of a new tire tested the tire for wear at different pressures with the results shown in Table 10.5. Calculate the coefficient of correlation $r$ for the data. Interpret the result.

**TIRES**

**TABLE 10.5 Data for Example 10.7**

| Pressure $x$, pounds per sq. inch | Mileage $y$, thousands | Pressure $x$, pounds per sq. inch | Mileage $y$, thousands |
|---|---|---|---|
| 30 | 29.5 | 33 | 37.6 |
| 30 | 30.2 | 34 | 37.7 |
| 31 | 32.1 | 34 | 36.1 |
| 31 | 34.5 | 35 | 33.6 |
| 32 | 36.3 | 35 | 34.2 |
| 32 | 35.0 | 36 | 26.8 |
| 33 | 38.2 | 36 | 27.4 |

Solution

Rather than perform the calculations by hand, we resort to the use of a computer to find the value of $r$. A SAS printout of the correlation analysis is shown in Figure 10.12. The value of $r$, shaded on the printout, is $r = -.114$. This relatively small value for $r$ describes a weak linear relationship between pressure ($x$) and mileage ($y$). The manufacturer, however, would be remiss in concluding that tire pressure has little or no impact on wear of the tire. On the contrary, the relationship between pressure and wear is fairly strong, as the MINITAB scattergram in Figure 10.13 illustrates. Note that the relationship is not linear, but curvilinear; the underinflated tires (low pressure values) and overinflated tires (high pressure values) both lead to low mileages.

```
Pearson Correlation Coefficients, N = 14
        Prob > |r| under H0: Rho=0

                   PRESS_X      MILEAGE_Y

 PRESS_X           1.00000       -0.11371
                                   0.6987

 MILEAGE_Y        -0.11371        1.00000
                    0.6987
```

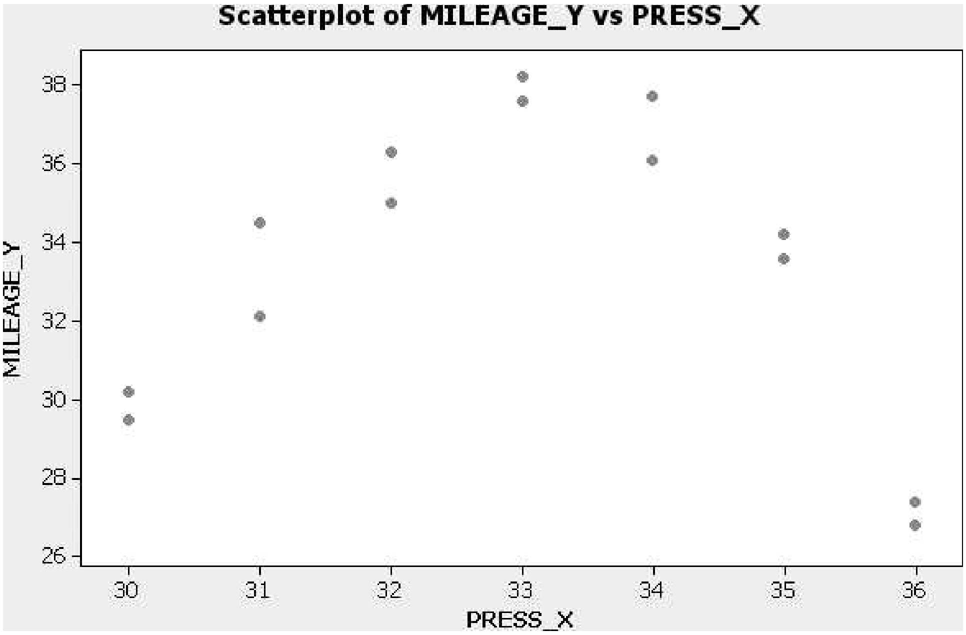
**FIGURE 10.12**
SAS correlation analysis of data in Table 10.5

Example 10.7 points out the danger of using $r$ to determine how well $x$ predicts $y$: The correlation coefficient $r$ describes only the *linear* relationship between $x$ and $y$. For nonlinear relationships, the value of $r$ may be misleading, and we need to resort to other methods for describing and testing such a relationship. Regression models for curvilinear relationships are presented in Chapter 11.

## Coefficient of Determination

Another way to measure the contribution of $x$ in predicting $y$ is to consider how much the errors of prediction of $y$ can be reduced by using the information provided by $x$.

**FIGURE 10.13**
MINITAB scatterplot of data in Table 10.5




To illustrate, suppose a sample of data has the scattergram shown in Figure 10.14a. If we assume that $x$ contributes no information for the prediction of $y$, the best prediction for a value of $y$ is the sample mean $\bar{y}$, which graphs as the horizontal line shown in Figure 10.14b. The vertical line segments in Figure 10.14b are the deviations of the points about the mean $\bar{y}$. Note that the sum of squares of deviations for the model $\hat{y} = \bar{y}$ is $SS_{yy} = \sum(y_i - \bar{y})^2$.

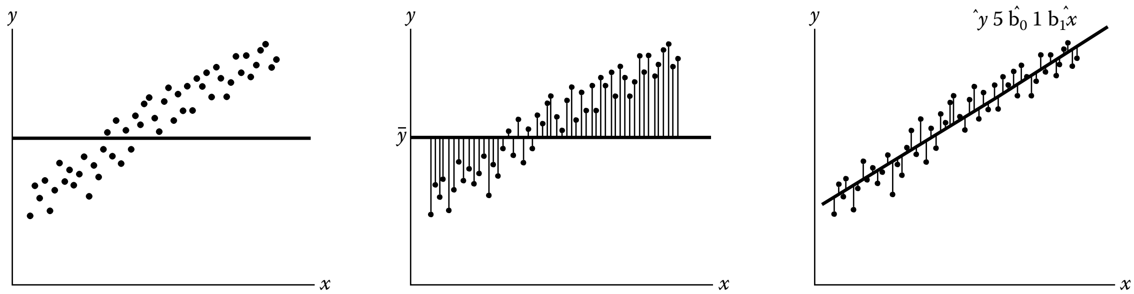
Now suppose that you fit a least-squares line to the same set of data and locate the deviations of the points about the line as shown in Figure 10.14c. Compare the deviations about the prediction lines in parts **b** and **c** in Figure 10.14. You can see that:

1. If $x$ contributes little or no information for the prediction of $y$, then the sums of squares of deviations for the two lines,

$$SS_{yy} = \sum(y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum(y_i - \hat{y}_i)^2$$

will be nearly equal.

2. If $x$ does contribute information for the prediction of $y$, then SSE will be smaller than $SS_{yy}$. In fact, if all the points fall on the least-squares line, then $SSE = 0$.




a. Scattergram of data

b. Assumption: $x$ contributes no information for predicting $y$; $\hat{y} = \bar{y}$

c. Assumption: $x$ contributes information for predicting $y$; $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

**FIGURE 10.14**
A comparison of the sum of squares of deviations for two models

A convenient way of measuring how well the least-squares equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ performs as a predictor of $y$ is to compute the reduction in the sum of squares of deviations that can be attributed to $x$, expressed as a proportion of $SS_{yy}$. This quantity, called the **coefficient of determination**, is

$$\frac{SS_{yy} - SSE}{SS_{yy}}$$

In simple linear regression, it can be shown that this quantity is equal to the square of the simple linear coefficient of correlation $r$.

**Definition 10.6**

The **coefficient of determination** is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

It represents the proportion of the sum of squares of deviations of the $y$ values about their predicted values ($\hat{y}$) that can be attributed to a linear relation between $y$ and $x$. (In simple linear regression, it may also be computed as the square of the coefficient of correlation $r$.)

Note that $r^2$ is always between 0 and 1, because $r$ is between $-1$ and $+1$. Thus, $r^2 = .60$ means that the sum of squares of deviations of the $y$ values about their predicted values has been reduced 60% by the use of $\hat{y}$, instead of $\bar{y}$, to predict $y$. Or, more practically, $r^2 = .60$ implies that the straight-line model relating $y$ to $x$ can explain (or account for) 60% of the variation present in the sample of $y$ values.

### Example 10.8

Finding the Coefficient of Determination, $r^2$

Calculate the coefficient of determination for the insulation compression example. The data are repeated in Table 10.6.

**TABLE 10.6 Data for Example 10.8**

**INSULATION**

| Pressure $x$, 10 pounds per square inch | Compression $y$, .1 inch |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 2 |
| 4 | 2 |
| 5 | 4 |

Solution

We first calculate

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{5} = 26 - \frac{(10)^2}{5} = 26 - 20 = 6$$

From previous calculations, we have

$$SSE = \sum (y_i - \hat{y}_i)^2 = 1.10$$

Then, the coefficient of determination is given by

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{6.0 - 1.1}{6.0} = \frac{4.9}{6.0} = .817$$

*Note:* This value could also be obtained by squaring the correlation coefficient $r = .904$ found in Example 10.6 or directly from a computer printout. The value is highlighted on the SPSS printout, Figure 10.15.

So we know that by using the pressure $x$ to predict compression $y$ with the least-squares line $\hat{y} = -.1 + .7x$, the total sum of squares of deviations of the five sample $y$ values about their predicted values has been reduced 82% by the use of the linear predictor $\hat{y}$. That is, 82% of the sample variation in compression values can be explained by the least-squares line.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .904[a] | .817 | .756 | .606 |

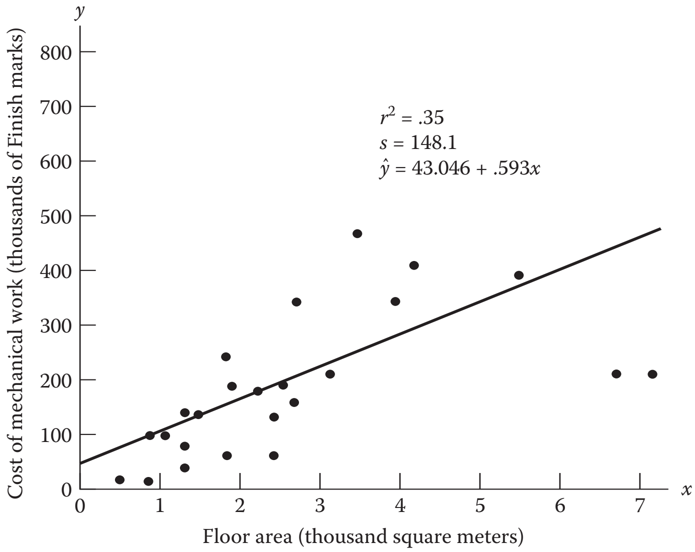
a. Predictors: (Constant), PRESS_X

**FIGURE 10.15**
SPSS printout showing coefficient of determination

### Practical Interpretation of the Coefficient of Determination, $r^2$

About $100(r^2)\%$ of the total sum of squares of deviations of the sample $y$ values about their mean $\bar{y}$ can be explained by (or attributed to) using $x$ to predict $y$ in the straight-line model.

**FIGURE 10.16**
Simple linear model relating cost to floor area




In situations where a straight-line regression model is found to be a statistically adequate predictor of $y$, the value of $r^2$ can help guide the regression analyst in the search for better, more useful models. For example, design engineers used a simple linear model to relate cost of mechanical work (heating, ventilating, and plumbing) in construction to floor area. Based on the data associated with 26 factory and warehouse buildings, the least-squares prediction equation given in Figure 10.16 was found. It was concluded that floor area and mechanical cost are linearly related, since the $T$ statistic (for testing $H_0$: $\beta_1 = 0$) was found to equal 3.61, which is significant with an $\alpha$ as small as .002. Thus, floor area should be useful when predicting the mechanical cost of a factory or warehouse. However, the value of the coefficient of determination $r^2$ was found to be .35. This tells us that only 35% of the variation among mechanical costs is accounted for by the differences in floor areas. This relatively small $r^2$ value led the engineers to include other independent variables (e.g., volume, amount of glass) in the model in an attempt to account for a significant portion of the remaining 65% of the variation in mechanical cost not explained by floor area. In the next chapter, we discuss this important aspect of relating a response to more than one independent variable.

## Applied Exercises

10.43 *Redshifts of Quasi-Stellar Objects.* Refer to the *Journal of Astrophysics & Astronomy* (Mar./Jun. 2003) study of redshifts in Quasi-Stellar Objects (QSOs), Exercise 10.4 (p. 495). Recall that simple linear regression was used to model the magnitude ($y$) of a QSO as a function of redshift level ($x$). In addition to the least-squares line, $\hat{y} = 18.13 + 6.21x$, the coefficient of correlation was determined as $r = .84$.

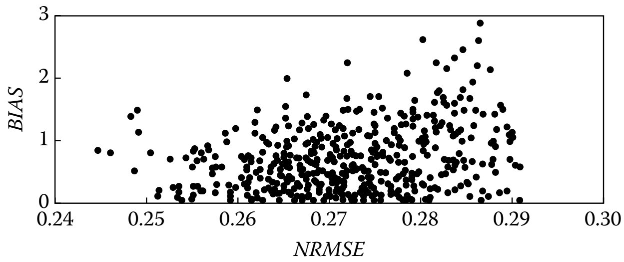
a. Interpret the value of $r$ in the words of the problem.
b. What is the relationship between $r$ and the estimated slope of the line?
c. Find the value of $r^2$ and interpret its value.

10.44 *Find r and $r^2$.* Find the correlation coefficient and the coefficient of determination for the sample data of each of the following exercises. Interpret your results.

a. Exercise 10.7
b. Exercise 10.8
c. Exercise 10.9
d. Exercise 10.10
e. Exercise 10.11
f. Exercise 10.12
g. Exercise 10.13
h. Exercise 10.14
i. Exercise 10.15

10.45 *Evaluation of an imputation method for missing data.* When analyzing large data sets with many variables, researchers often encounter the problem of missing data (e.g., non-response). Typically, an imputation method will be used to substitute in reasonable values (e.g., the mean of the variable) for the missing data. An imputation method that uses "nearest neighbors" as substitutes for the missing data was evaluated in *Data & Knowledge Engineering* (March 2013). Two quantitative assessment measures of the imputation algorithm are normalized root mean square error (NRMSE) and classification bias. The researchers applied the imputation method to a sample of 3600 data sets with missing values and determined the NRMSE and classification bias for each data set. The correlation coefficient between the two variables was reported as $r = .2838$.

a. Conduct a test to determine if the true population correlation coefficient relating NRMSE and bias is positive. Interpret this result practically.
b. A scatterplot for the data (extracted from the journal article) is shown below. Based on the graph, would you recommend using NRMSE as a linear predictor of bias? Explain why your answer does not contradict the result in part a.




10.46 *Fitts' Law.* A robust and highly adopted model of human movement is Fitts' Law. According to Fitts' Law, the time $T$ required to move and select a target of width $W$ that lies at a distance (or amplitude) $A$ is

$$T = a + b\log_2(2A/W)$$

where $a$ and $b$ are constants estimated using simple linear regression. The quantity $\log_2(2A/W)$ is termed the Index of Difficulty (ID) and represents the independent variable (measured in "bits") in the model. Research reported in the *Special Interest Group on Computer-Human Interaction Bulletin* (July 1993) used Fitts' Law to model the time (in milliseconds) required to perform a certain task on a computer. Based on data collected for $n = 160$ trials (using different values of $A$ and $W$), the following least-squares prediction was obtained:

$$\hat{T} = 175.4 + 133.2(\text{ID})$$

a. Interpret the estimates, 175.4 and 133.2.
b. The coefficient of correlation for the analysis is $r = .951$. Interpret this value.
c. Conduct a test to determine whether the Fitts' Law model is statistically adequate for predicting performance time. Use $\alpha = .05$.
d. Calculate the coefficient of determination, $r^2$. Interpret the result.

10.47 *Removing metal from water.* In the *Electronic Journal of Biotechnology* (Apr. 15, 2004), Egyptian scientists studied a new method for removing heavy metals from water. Metal solutions were prepared in glass vessels, then biosorption was used to remove the metal ions. Two variables were measured for each test vessel: $y =$ the metal uptake (milligrams of metal per gram of biosorbent) and $x =$ final concentration of metal in the solution (milligrams per liter).

a. Write a simple linear regression model relating $y$ to $x$.
b. For one metal, simple linear regression analysis yielded $r^2 = .92$. Interpret this result.

10.48 *Wind turbine blade stress.* Refer to the *Wind Engineering* (Jan. 2004) study of two types of timber—radiata pine and hoop pine—used in high-efficiency small wind turbine blades, Exercise 10.16 (p. 520). Data on stress ($y$) and the natural logarithm of number of blade cycles ($x$) for each timber type were analyzed using simple linear regression. The results are reproduced here, with additional information on the coefficient of determination. Interpret the value of $r^2$ for each type of timber.

**Radiata Pine**: $\hat{y} = 97.37 - 2.50x, \quad r^2 = .84$

**Hoop Pine**: $\hat{y} = 122.03 - 2.36x, \quad r^2 = .90$

10.49 *Water content of soil.* The standard method of determining the water content of soil involves the removal of the soil and estimation of soil volume. *Forest Engineering* (July 1999) presented a method that does not require soil removal, i.e., an indirect method, called the radiation method. The new method utilizes the fact that water content is proportional to the number of thermalized hydrogen neutrons emitted from the soil. A sample of 56 soil cores were collected at a depth of 10 feet; for each core, the water content was determined using the standard method and the count of hydrogen neutrons determined by radiation. A simple linear regression of the data yielded the following results:

$$\hat{y} = .088 + .136x, \qquad r^2 = .84$$

where $y =$ water content (grams per cubic centimeter) and $x =$ count ratio (number of hydrogen neutrons divided by the standard count).

a. Interpret the estimated $y$-intercept of the least-squares line.
b. Interpret the estimated slope of the least-squares line.
c. The $p$-value for testing whether the slope is 0 was determined to be .0001. Interpret this result.
d. Interpret the value of $r^2$.

10.50 *Prices of recycled materials.* Prices of recycled materials (e.g., plastics, paper, and glass) are highly volatile due to the fact that supply is constant, rather than tied to demand.

An exploratory study of the prices of recycled products in the United Kingdom was published in *Resources, Conservation, and Recycling* (Vol. 60, 2012). The researchers employed simple linear regression to model $y$ = the monthly price of recycled colored plastic bottles as a function of $x$ = the monthly price of naphtha (a primary material in plastics). The following results were obtained for monthly data collected over a recent 10-year period ($n = 120$ months):

$\hat{y} = -32.35 + 4.82x$, $r = .83$, $r^2 = .69$

$t$-value (for testing $H_0$: $\beta_1 = 0$) $= 16.60$,

Interpret these results. Give your conclusion about the adequacy of the model in the words of the problem.

## Theoretical Exercises

10.51 Verify that

$$\hat{\beta}_1 = r\sqrt{\frac{SS_{yy}}{SS_{xx}}} \quad \text{and} \quad SSE = SS_{yy}(1 - r^2)$$

10.52 Use the result of Exercise 10.51 to show that

$$\frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

## 10.8 Using the Model for Estimation and Prediction

If we are satisfied that a useful model has been found to describe the relationship between the compression of the insulation material and compressive pressure, we are ready to accomplish the original objectives for building the model: using it to estimate or to predict the amount of compression for a particular level of compressive pressure.

The most common uses of a probabilistic model can be divided into two categories. **The first is the use of the model for estimating the mean value of $y$, $E(y)$, for a specific value of $x$.** For our example, we may want to estimate the mean amount of compression for all specimens of insulation subjected to a compressive pressure of 40 $(x = 4)$ pounds per square inch. **The second use of the model entails predicting a particular $y$ value for a given $x$.** That is, if we decide to install the insulation in a particular piece of equipment in which we think it will be subjected to a pressure of 40 pounds per square inch, we will want to predict the insulation compression for this particular specimen of insulation material.

In the case of estimating a mean value of $y$, we are attempting to estimate the mean result of a very large number of experiments at the given $x$ value. In the second case, we are trying to predict the outcome of a single experiment at the given $x$ value. In which of these model uses do you expect to have more success, i.e., which value—the mean or individual value of $y$—can we estimate (or predict) with greater accuracy?

Before answering this question, we first consider the problem of choosing an estimator (or predictor) of the mean (or individual) $y$ value. We will use the least-squares model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

both to estimate the mean value of $y$ and to predict a particular value of $y$ for a given value of $x$. For our example, we found

$$\hat{y} = -.1 + .7x$$

so that the estimated mean compression of all specimens of insulation when $x = 4$ (compressive pressure of 40 pounds per square inch) is

$$\hat{y} = -.1 + .7(4) = 2.7$$

or .27 inch (the units of $y$ are tenths of an inch). The identical value is used to predict the $y$ value when $x = 4$. That is, both the estimated mean value and the predicted value of $y$ equal $\hat{y} = 2.7$ when $x = 4$, as shown in Figure 10.17.

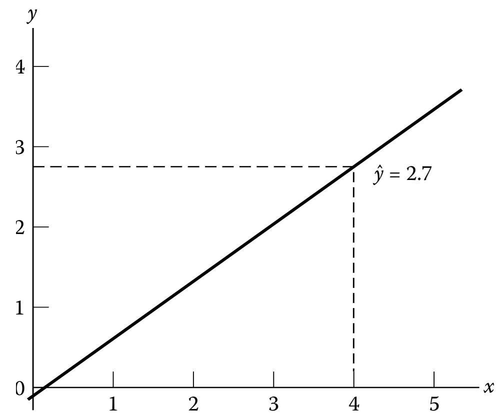



**FIGURE 10.17**
Estimated mean value and predicted individual value of compression $y$ for $x = 4$

The difference in these two model uses lies in the relative accuracy of the estimate and the prediction. These accuracies are best measured by the repeated sampling errors of the least-squares line when it is used as an estimator and as a predictor, respectively. These errors are given in the following box.

### Sampling Errors for the Estimator of the Mean of *y*, *E(y)*, and the Predictor for an Individual *y*

1. The standard deviation of the sampling distribution of the estimator $\hat{y}$ of $E(y)$ at a particular value of $x$, say, $x_p$, is

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $\sigma$ is the standard deviation of the random error $\varepsilon$.

2. The standard deviation of the prediction error for the predictor $\hat{y}$ of an individual $y$ value for $x = x_p$ is

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $\sigma$ is the standard deviation of the random error $\varepsilon$.

The true value of $\sigma$ will rarely be known. Thus, we estimate $\sigma$ by $s$ and calculate the confidence and prediction intervals as shown in the following boxes. (See Figure 10.19 for a comparison of the widths of these intervals.)

### A $(1-\alpha)100\%$ Confidence Interval for the Mean Value of *y*, *E(y)*, for $x = x_p$

$$\hat{y} \pm t_{\alpha/2}(\text{Estimated standard deviation of } \hat{y})$$

or

$$\hat{y} \pm t_{\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n - 2)$ df

A $(1-\alpha)100\%$ Prediction Interval for an Individual $y$ for $x = x_p$

$$\hat{y} \pm t_{\alpha/2}[\text{Estimated standard deviation of } (y - \hat{y})]$$

or

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n-2)$ df

**Example 10.9**

Finding a 95% Confidence Interval for $E(y)$

Find a 95% confidence interval for the mean insulation compression when the pressure is 40 pounds per square inch.

Solution

For a compressive pressure of 40 pounds per square inch, $x_p = 4$ and, since $n = 5$, df $= n - 2 = 3$. Then the confidence interval for the mean value of $y$ is

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

or

$$\hat{y} \pm t_{.025} s \sqrt{\frac{1}{5} + \frac{(4 - \bar{x})^2}{SS_{xx}}}$$

Recall that $\hat{y} = 2.7$, $s = .61$, $\bar{x} = 3$, and $SS_{xx} = 10$. From Table 7 of Appendix B, $t_{.025} = 3.182$. Thus, we have

$$\begin{aligned} 2.7 \pm (3.182)(.61)\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} &= 2.7 \pm (3.182)(.61)(.55) \\ &= 2.7 \pm 1.1 = (1.6, 3.8) \end{aligned}$$

Remembering that compression $(y)$ is measured in units of .1 inch, we estimate that the interval from .16 inch to .38 inch encloses the mean amount of compression when the insulation is subjected to a compressive pressure of 40 pounds per square inch. Note that we used a small amount of data for purposes of illustration in fitting the least-squares line and that the width of the interval could be decreased by using a larger number of data points.

**Example 10.10**

Finding a 95% Prediction Interval for $y$

Predict the amount of compression for an individual piece of insulation material subjected to a compressive pressure of 40 pounds per square inch. Use a 95% prediction interval.

Solution

To predict the compression for a particular piece of insulation material for which $x_p = 4$, we calculate the 95% prediction interval as

$$\begin{aligned} \hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 2.7 \pm (3.182)(.61)\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}} \\ &= 2.7 \pm (3.182)(.61)(1.14) \\ &= 2.7 \pm 2.2 = (.5, 4.9) \end{aligned}$$

Therefore, we predict that the compression for the piece of insulation material will fall in the interval from .05 inch to .49 inch. As in the case for the confidence interval for

**FIGURE 10.18**
MINITAB printout showing confidence interval for $E(y)$ and prediction interval for $y$

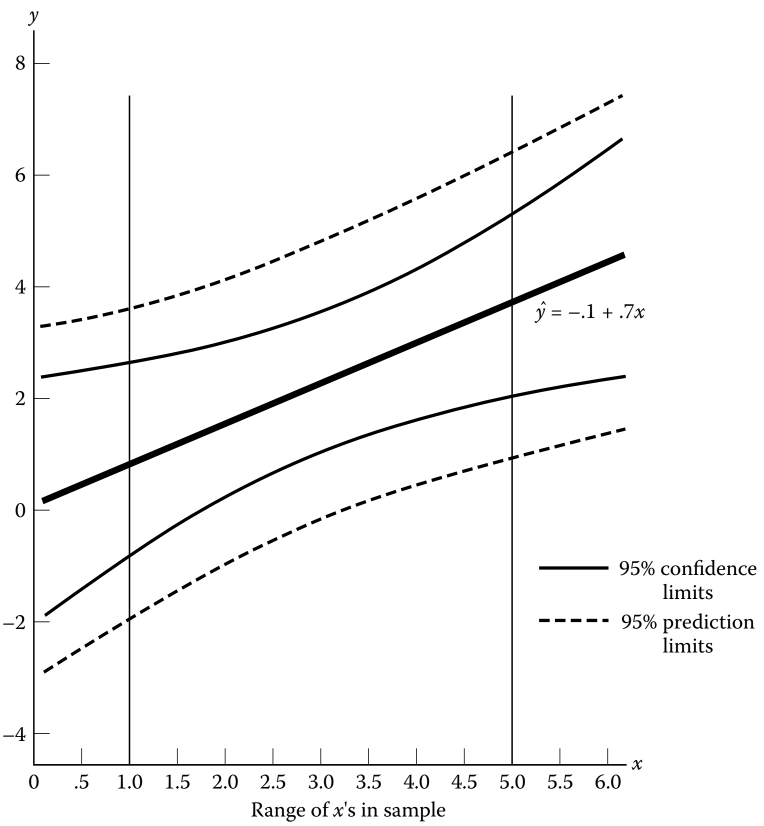
```
Predicted Values for New Observations

New
Obs    Fit  SE Fit       95% CI            95% PI
  1  2.700   0.332  (1.645, 3.755)  (0.503, 4.897)


Values of Predictors for New Observations

New
Obs  PRESS_X
  1     4.00
```

the mean value of $y$, the prediction interval for $y$ is quite large. This is because we have chosen a simple example (only five data points) to fit the least squares line. The width of the prediction interval could be reduced by using a larger number of data points.

Both the confidence interval for $E(y)$ and the prediction interval for $y$ can be obtained using statistical software. Figure 10.18 is a MINITAB printout that gives the confidence and prediction intervals. These intervals (highlighted) agree, except for rounding, with our calculated intervals.

A comparison of the confidence limits for the mean value of $y$ and the prediction limits for some future value of $y$ for various values of compressive pressure $x$ is illustrated in Figure 10.19. It is important to note that the prediction interval for an individual value of $y$ will always be wider than the confidence interval for a mean value of $y$. You can see this by examining the formulas for the two intervals and you can see it in Figure 10.19.

**FIGURE 10.19**
Comparison of widths of 95% confidence and prediction intervals



Additionally, over the range of the sample data, the widths of both intervals increase as the value of $x$ gets farther from $\bar{x}$. (See Figure 10.19.) Thus, the more $x$ deviates from $\bar{x}$, the less useful the interval will be in practice. In fact, when $x$ is selected far enough away from $\bar{x}$ so that it falls outside the range of the sample data, it is dangerous to make any inferences about $E(y)$ or $y$, as explained in the following box.

### Warning

Using the least-squares prediction equation to estimate the mean value of $y$ or to predict a particular value of $y$ for values of $x$ that fall *outside* the range of values of $x$ contained in your sample data may lead to errors of estimation or prediction that are much larger than expected. Although the least-squares model may provide a very good fit to the data over the range of $x$ values contained in the sample, **it could give a poor representation of the true model for values of $x$ outside this region**.

To conclude this section, we will find the variance of the value of $\hat{y}$ when $x = x_p$. This variance plays an important role in developing the confidence interval for $E(y)$ when $x = x_p$ and the prediction interval for a particular value of $y$ when $x = x_p$.

## Example 10.11

**Deriving $V(\hat{y})$**

Find the variance of $\hat{y}$ when $x = x_p$.

**Solution**

When $x = x_p$, we have $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Substituting this value of $\hat{\beta}_0$ into the expression for $\hat{y}$, we obtain

$$\begin{aligned}\hat{y} &= (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1(x_p)\\ &= \bar{y} + \hat{\beta}_1(x_p - \bar{x})\end{aligned}$$

The next step is to express $\hat{y}$ as a linear function of the random $y$ values, $y_1, y_2, \ldots, y_n$ so that we can obtain $V(\hat{y})$ as the variance of a linear function of independent random variables. We now write

$$\begin{aligned}\hat{y} &= \bar{y} + \hat{\beta}_1(x_p - \bar{x})\\ &= \sum \frac{y_i}{n} + \frac{(x_p - \bar{x})}{\mathrm{SS}_{xx}}\sum (x_i - \bar{x})y_i\\ &= \sum \frac{y_i}{n} + \sum \frac{(x_p - \bar{x})(x_i - \bar{x})}{\mathrm{SS}_{xx}} y_i\end{aligned}$$

We can now express $\hat{y}$ as a single summation:

$$\hat{y} = \sum \left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{\mathrm{SS}_{xx}}\right] y_i$$

i.e., $\hat{y}$ is a linear function of the independent random variables, $y_1, y_2, \ldots, y_n$, where the coefficient of $y_i$ is

$$\left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{\mathrm{SS}_{xx}}\right]$$

Then, by Theorem 6.8,

$$V(\hat{y}) = \sum \left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{\mathrm{SS}_{xx}}\right]^2 V(y_i)$$

where $V(y_i) = \sigma^2, i = 1, 2, \ldots, n$. Therefore,

$$V(\hat{y}) = \sum\left[\frac{1}{n^2} + \frac{2}{n}\frac{(x_p - \bar{x})(x_i - \bar{x})}{\mathrm{SS}_{xx}} + \frac{(x_p - \bar{x})^2(x_i - \bar{x})^2}{(\mathrm{SS}_{xx})^2}\right]\sigma^2$$

$$= \left[\frac{n}{n^2} + \frac{2}{n}\frac{(x_p - \bar{x})}{\mathrm{SS}_{xx}}\sum(x_i - \bar{x}) + \frac{(x_p - \bar{x})^2}{(\mathrm{SS}_{xx})^2}\sum(x_i - \bar{x})^2\right]\sigma^2$$

$$= \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{(\mathrm{SS}_{xx})^2}\mathrm{SS}_{xx}\right]\sigma^2 \qquad \text{since } \sum(x_i - \bar{x}) = 0$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\mathrm{SS}_{xx}}\right]$$

You can see that this agrees with the formula for $V(\hat{y})$ given previously in this section.

## Applied Exercises

10.53 *Removing nitrogen from toxic wastewater.* Highly toxic wastewater is produced during the manufacturing of dry-spun acrylic fiber. One way to lessen toxicity is to remove the nitrogen from the wastewater. A group of environmental engineers investigated a promising method—called anaerobic ammonium oxidation—for nitrogen removal and reported the results in the *Chemical Engineering Journal* (April 2013). A sample of 120 specimens of toxic wastewater was collected and each treated with the nitrogen removal method. The amount of nitrogen removed (measured in milligrams per liter) was determined as well as the amount of ammonium (milligrams per liter) used in the removal process. These data (simulated from information provided in the journal article) are saved in the **NITRO** file. The data for the first 5 specimens are shown below. A simple linear regression analysis, where $y =$ amount of nitrogen removed and $x =$ amount of ammonium used, is also shown in the SAS printout on p. 527.

a. Assess, statistically, the adequacy of the fit of the linear model. Do you recommend using the model for predicting nitrogen amount?

b. On the SAS printout, locate a 95% prediction interval for nitrogen amount when amount of ammonium used is 100 milligrams per liter. Practically interpret the result.

c. Will a 95% confidence interval for the mean nitrogen amount when amount of ammonium used is 100 milligrams per liter be wider or narrower than the interval, part b? Explain.

**NITRO** *(first 5 observations of 120 shown)*

| Nitrogen | Ammonium |
|---|---|
| 18.87 | 67.40 |
| 17.01 | 12.49 |
| 23.88 | 61.96 |
| 10.45 | 15.63 |
| 36.03 | 83.66 |

**BBALL**

10.54 *Sound waves from a basketball.* Refer to the *American Journal of Physics* (June 2010) study of sound waves in a spherical cavity, Exercise 10.7 (p. 496). You fit a straight-line model relating frequency of sound waves ($y$) to number of resonances ($x$) using the data provided in Exercise 10.7.

a. Evaluate the adequacy of the model for predicting frequency of sound waves.

b. Use the model to predict the sound wave frequency for the 10th resonance.

c. Form a 90% confidence interval for the prediction, part a. Interpret the result.

d. Suppose you want to predict the sound wave frequency for the 30th resonance. What are the dangers in making this prediction with the fitted model?

**SMELTPOT**

10.55 *Extending the life of an aluminum smelter pot.* Refer to *The American Ceramic Society Bulletin* (Feb. 2005) evaluation of commercial bricks used in smelter pots, Exercise 10.9 (p. 497).

a. Find a 95% prediction interval for the apparent porosity percentage $y$ of a brick with a mean pore diameter of $x = 10$ micrometers. Interpret the result.

**SAS Output for Exercise 10.53**

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: NITRO_Y

| | |
|---|---|
| Number of Observations Read | 120 |
| Number of Observations Used | 120 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 88104 | 88104 | 1075.57 | <.0001 |
| Error | 118 | 9665.80325 | 81.91359 | | |
| Corrected Total | 119 | 97770 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 9.05061 | R-Square | 0.9011 |
| Dependent Mean | 67.91083 | Adj R-Sq | 0.9003 |
| Coeff Var | 13.32720 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 2.22255 | 2.16665 | 1.03 | 0.3071 |
| AMMON_X | 1 | 0.57637 | 0.01757 | 32.80 | <.0001 |

| Output Statistics | | | | | | | |
|---|---|---|---|---|---|---|---|
| Obs | AMMON_X | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Predict | | Residual |
| 1 | 100.0 | . | 59.8596 | 0.8619 | 41.8558 | 77.8634 | . |
| 2 | 67.4 | 18.8700 | 41.0699 | 1.1629 | 22.9999 | 59.1400 | -22.1999 |
| 3 | 12.5 | 17.0100 | 9.4214 | 1.9655 | -8.9190 | 27.7619 | 7.5886 |
| 4 | 62.0 | 23.8800 | 37.9345 | 1.2321 | 19.8465 | 56.0225 | -14.0545 |
| 5 | 15.6 | 10.4500 | 11.2312 | 1.9156 | -7.0885 | 29.5509 | -0.7812 |
| 6 | 83.7 | 36.0300 | 50.4417 | 0.9830 | 32.4136 | 68.4698 | -14.4117 |

b. Will a 95% confidence interval for the mean porosity percentage, $E(y)$, when $x = 10$ micrometers be wider or narrower than the prediction interval, part **a**? Explain.

LIQUIDSPILL

10.56 *Spreading rate of spilled liquid.* Refer to the *Chemical Engineering Progress* (Jan. 2005) study of the rate at which a spilled volatile liquid will spread across a surface, Exercise 10.10 (p. 497).

a. Find a 90% confidence interval for the mean mass, $E(y)$, of all spills with an elapsed time of $x = 8$ minutes. Interpret the result.

b. Will a 90% confidence interval for mean mass, $E(y)$, when $x = \bar{x}$ be wider or narrower than the confidence interval, part **a**. Explain.

**RAINFALL**

10.57 *New method of estimating rainfall.* Refer to the *Journal of Data Science* (Apr. 2004) evaluation of methods for estimating rainfall, Exercise 10.11 (p. 498). Find a 99% prediction interval for the rain gauge amount $y$ when the neural network estimate is $x = 3$ millimeters. Interpret the result.

**OJUICE**

10.58 *Sweetness of orange juice.* Refer to the simple linear regression of sweetness index $y$ and amount of pectin $x$ for $n = 24$ orange juice samples, Exercise 10.12 (p. 498). The SPSS printout of the analysis is shown below. A 90% confidence interval for the mean sweetness index, $E(y)$, for each value of $x$ is shown below on the SPSS spreadsheet. Select an observation and interpret this interval.

**FINTUBES**

10.59 *Thermal performance of copper fin-tubes.* Refer to the *Journal of Heat Transfer* (Aug. 1990) study of copper integral fin-tubes, Exercise 10.14 (p. 499). Find a 90% confidence interval for the mean heat transfer coefficient, $E(y)$, when unflooded area ratio is $x = 1.95$. Interpret the result.

10.60 *Predicting tree heights.* In forestry, the diameter of a tree at breast height (which is fairly easy to measure) is used to predict the height of the tree (a difficult measurement to obtain). Silviculturists working in British Columbia's boreal forest conducted a series of spacing trials to predict the heights of several species of trees. The data in the table on p. 529 are the breast height diameters (in centimeters) and heights (in meters) for a sample of 36 white spruce trees.

a. Construct a scattergram for the data.
b. Assuming the relationship between the variables is best described by a straight line, use the method of least squares to estimate the $y$-intercept and slope of the line.
c. Plot the least-squares line on your scattergram.
d. Do the data provide sufficient evidence to indicate that the breast height diameter $x$ contributes information for the prediction of tree height $y$? Test using $\alpha = .05$.
e. Use your least-squares line to find a 90% confidence interval for the average height of white spruce trees with a breast height diameter of 20 cm. Interpret the interval.

**SPSS Output for Exercise 10.58**

| | run | sweet | pectin | lower90m | upper90m |
|---|---|---|---|---|---|
| 1 | 1 | 5.2 | 220 | 5.64898 | 5.83848 |
| 2 | 2 | 5.5 | 227 | 5.63898 | 5.81613 |
| 3 | 3 | 6.0 | 259 | 5.57819 | 5.72904 |
| 4 | 4 | 5.9 | 210 | 5.66194 | 5.87173 |
| 5 | 5 | 5.8 | 224 | 5.64337 | 5.82560 |
| 6 | 6 | 6.0 | 215 | 5.65564 | 5.85493 |
| 7 | 7 | 5.8 | 231 | 5.63284 | 5.80379 |
| 8 | 8 | 5.6 | 268 | 5.55553 | 5.71011 |
| 9 | 9 | 5.6 | 239 | 5.61947 | 5.78019 |
| 10 | 10 | 5.9 | 212 | 5.65946 | 5.86497 |
| 11 | 11 | 5.4 | 410 | 5.05526 | 5.55416 |
| 12 | 12 | 5.6 | 256 | 5.58517 | 5.73592 |
| 13 | 13 | 5.8 | 306 | 5.43785 | 5.65219 |
| 14 | 14 | 5.5 | 259 | 5.57819 | 5.72904 |
| 15 | 15 | 5.3 | 284 | 5.50957 | 5.68213 |
| 16 | 16 | 5.3 | 383 | 5.15725 | 5.57694 |
| 17 | 17 | 5.7 | 271 | 5.54743 | 5.70434 |
| 18 | 18 | 5.5 | 264 | 5.56591 | 5.71821 |
| 19 | 19 | 5.7 | 227 | 5.63898 | 5.81613 |
| 20 | 20 | 5.3 | 263 | 5.56843 | 5.72031 |
| 21 | 21 | 5.9 | 232 | 5.63125 | 5.80075 |
| 22 | 22 | 5.8 | 220 | 5.64898 | 5.83848 |
| 23 | 23 | 5.8 | 246 | 5.60640 | 5.76091 |
| 24 | 24 | 5.9 | 241 | 5.61587 | 5.77454 |

### Data for Exercise 10.60

**SPRUCE**

| Breast Height Diameter $x$, cm | Height $y$, m | Breast Height Diameter $x$, cm | Height $y$, m |
|---|---|---|---|
| 18.9 | 20.0 | 16.6 | 18.8 |
| 15.5 | 16.8 | 15.5 | 16.9 |
| 19.4 | 20.2 | 13.7 | 16.3 |
| 20.0 | 20.0 | 27.5 | 21.4 |
| 29.8 | 20.2 | 20.3 | 19.2 |
| 19.8 | 18.0 | 22.9 | 19.8 |
| 20.3 | 17.8 | 14.1 | 18.5 |
| 20.0 | 19.2 | 10.1 | 12.1 |
| 22.0 | 22.3 | 5.8 | 8.0 |
| 23.6 | 18.9 | 20.7 | 17.4 |
| 14.8 | 13.3 | 17.8 | 18.4 |
| 22.7 | 20.6 | 11.4 | 17.3 |
| 18.5 | 19.0 | 14.4 | 16.6 |
| 21.5 | 19.2 | 13.4 | 12.9 |
| 14.8 | 16.1 | 17.8 | 17.5 |
| 17.7 | 19.9 | 20.7 | 19.4 |
| 21.0 | 20.4 | 13.3 | 15.5 |
| 15.9 | 17.6 | 22.9 | 19.2 |

*Source:* Scholz, H., Northern Lights College, British Columbia.

10.61 *Life tests of cutting tools.* To improve the quality of the output of any production process, it is necessary first to understand the capabilities of the process (Deming, *Out of the Crisis*, 1982). In a particular manufacturing process, the useful life of a cutting tool is linearly related to the speed at which the tool is operated. The data in the accompanying table were derived from life tests for the two different brands of cutting tools currently used in the production process.

a. Fit the model, $E(y) = \beta_0 + \beta_1 x$, to the data for brand A, where $y$ = useful life and $x$ = cutting speed.

b. Repeat part **a** for brand B.

c. Use a 90% confidence interval to estimate the mean useful life of a brand A cutting tool when the cutting speed is 45 meters per minute. Repeat for brand B. Compare the widths of the two intervals and comment on the reasons for any difference.

d. Use a 90% prediction interval to predict the useful life of a brand A cutting tool when the cutting speed is 45 meters per minute. Repeat for brand B. Compare the widths of the two intervals to each other and to the two intervals you calculated in part **c.** Comment on the reasons for any differences.

e. Note that the estimation and prediction you performed in parts **c** and **d** were for a value of $x$ that was not included in the original sample. That is, the value $x = 45$ was not part of the sample. However, the value is within the range of $x$ values in the sample, so that the regression model spans the $x$ value for which the estimation and prediction were made. In such situations, estimation and prediction represent **interpolations**. Suppose you were asked to predict the useful life of a brand A cutting tool for a cutting speed of $x = 100$ meters per minute. Since the given value of $x$ is outside the range of the sample $x$ values, the prediction is an example of **extrapolation**. Predict the useful life of a brand A cutting tool that is operated at 100 meters per minute, and construct a 95% confidence interval for the actual useful life of the tool. What additional assumption do you have to make in order to ensure the validity of an extrapolation?

**CUTTOOL**

| Cutting Speed (meters per minute) | Useful Life (hours) Brand A | Brand B |
|---|---|---|
| 30 | 4.5 | 6.0 |
| 30 | 3.5 | 6.5 |
| 30 | 5.2 | 5.0 |
| 40 | 5.2 | 6.0 |
| 40 | 4.0 | 4.5 |
| 40 | 2.5 | 5.0 |
| 50 | 4.4 | 4.5 |
| 50 | 2.8 | 4.0 |
| 50 | 1.0 | 3.7 |
| 60 | 4.0 | 3.8 |
| 60 | 2.0 | 3.0 |
| 60 | 1.1 | 2.4 |
| 70 | 1.1 | 1.5 |
| 70 | .5 | 2.0 |
| 70 | 3.0 | 1.0 |

## Theoretical Exercises

10.62 Suppose you want to predict some future value of $y$ when $x = x_p$ using the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The error of prediction will be the difference between the actual value of $y_p$ and the predicted value $\hat{y}$, i.e.,

$$\text{Error of prediction} = y_p - \hat{y}$$

a. Explain why the error of prediction will be normally distributed.

b. Find the expected value and the variance of the error of prediction.

10.63 Explain why

$$Z = \frac{\text{Error or prediction}}{\text{Standard deviation of the error}}$$

$$= \frac{y_p - \hat{y}}{\sigma_{(y_p - \hat{y})}}$$

$$= \frac{y_p - \hat{y}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}}$$

is a standard normal random variable.

10.64 Show that

$$T = \frac{\text{Error of prediction}}{\text{Estimated standard deviation of the error}}$$

$$= \frac{y_p - \hat{y}}{s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}}$$

has a Student's $T$ distribution with $\nu = (n - 2)$ df. Then use the $T$ statistic as a pivotal statistic to derive a $(1 - \alpha)100\%$ prediction interval for $y_p$.

## 10.9 Checking Assumptions: Residual Analysis

When we apply a simple linear regression analysis to the data, we never know for certain whether the assumptions of Section 10.2 are satisfied. How far can we deviate from the assumptions and still expect regression analysis to yield results that will have the reliability stated in this chapter? How can we detect departures (if they exist) from the assumptions, and what can we do about them? We provide some answers to these questions in this section.

Recall (Section 10.2) that the assumptions concern the probability distribution of the random error ($\varepsilon$). To be valid, regression analysis requires (1) $E(\varepsilon) = 0$, (2) $V(\varepsilon) = \sigma^2$ constant, (3) $\varepsilon$ has a normal distribution, and (4) $\varepsilon$'s are independent. It is unlikely that these assumptions are ever satisfied exactly in a practical application of simple linear regression. Fortunately, experience has shown that least-squares regression produces reliable statistical tests, confidence intervals, and prediction intervals as long as the departures from the assumptions are not too great. However, gross violations will lead to unreliable results. Consequently, it is important to check the validity of the assumptions before making model inferences.

In Section 10.3, we defined a regression *residual* as the difference between the actual value of $y$ and its corresponding predicted value, i.e., residual $= (y - \hat{y})$. A residual is an estimate of the true error of prediction for a particular observation; consequently, residuals provide information on the validity of the assumptions on $\varepsilon$. In this section, we illustrate several graphical methods of residual analysis that can be applied to check the assumptions. Not only do these graphs help us determine if a particular assumption is reasonably satisfied, but they also guide the researcher on how to modify the regression model if an assumption is violated.

### Checking Assumption #1: Mean $\varepsilon = 0$

Typically, the assumption of $E(\varepsilon) = 0$ is violated when the deterministic portion of the model is misspecified. In simple linear regression, we hypothesize the straight-line model, $E(y) = \beta_0 + \beta_1 x$. However, suppose the relationship between $y$ and $x$ is nonlinear (i.e., a curvilinear relationship). We will learn in Chapter 12 that a possible deterministic relationship for a nonlinear model is $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Fitting a straight-line model to data that follow a curvilinear relationship leads to a violation of the first assumption.

To see this, suppose the true deterministic relationship is nonlinear, i.e.,

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

but we hypothesize the straight-line relationship,

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Now, for our (misspecified) model, we can write

$$\varepsilon = y - (\beta_0 + \beta_1 x)$$

Then it can be shown that

$$E(\varepsilon) = E(y) - (\beta_0 + \beta_1 x)$$

Substituting the expression for the true $E(y)$, we have

$$E(\varepsilon) = (\beta_0 + \beta_1 x + \beta_2 x^2) - (\beta_0 + \beta_1 x) = \beta_2 x^2$$

Note that the expected value will not equal 0 unless $\beta_2 = 0$ (i.e., when the true deterministic relationship is linear). Consequently, for the misspecified model the assumption that $E(\varepsilon) = 0$ will be violated.

A graphical method for detecting model misspecification in simple linear regression analysis is to construct a plot with the regression residuals on the vertical axis and the values of the independent variable $x$ on the horizontal axis. If the plot reveals a random pattern of points (no trends), then it is likely that the model is specified correctly and the assumption of mean error of 0 is reasonably satisfied. However, if a strong pattern emerges on the residual plot, it is an indication of a misspecified model, violating the first assumption. We illustrate with an example.

**Example 10.12**

Detecting model misspecification

In all-electric homes, the amount of electricity expended is of interest to consumers, builders, and groups involved with energy conservation. Suppose we wish to investigate the July electrical usage, $y$, in all-electric homes and its relationship to the size, $x$, of the home. Moreover, suppose we think that July electrical usage in all electric homes is related to the size of the home by the straight-line model $E(y) = \beta_0 + \beta_1 x$. Data collected for a sample of 15 homes are shown in Table 10.7.

a. Fit the model to the data and assess model adequacy.

b. Plot the regression residuals versus home size ($x$). Do you detect a trend? What is the implication of this plot?

**KWHRS**

**TABLE 10.7 Home Size–Electrical Usage Data**

| Size of Home, $x$ (sq. ft.) | Monthly Usage, $y$ (kilowatt-hours) |
|---|---|
| 1,290 | 1,182 |
| 1,350 | 1,172 |
| 1,470 | 1,264 |
| 1,600 | 1,493 |
| 1,710 | 1,571 |
| 1,840 | 1,711 |
| 1,980 | 1,804 |
| 2,230 | 1,840 |
| 2,400 | 1,986 |
| 2,710 | 2,007 |
| 2,930 | 1,984 |
| 3,000 | 1,960 |
| 3,210 | 2,001 |
| 3,240 | 1,928 |
| 3,520 | 1,945 |

**FIGURE 10.20**
MINITAB Simple Linear Regression Printout, Example 10.12

**Regression Analysis: USAGE versus SIZE**

```
The regression equation is
USAGE = 903 + 0.356 SIZE

Predictor      Coef  SE Coef     T      P
Constant      903.0    132.1  6.83  0.000
SIZE        0.35594  0.05477  6.50  0.000

S = 155.251   R-Sq = 76.5%   R-Sq(adj) = 74.7%

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       1  1017803  1017803  42.23  0.000
Residual Error  13   313338    24103
Total           14  1331140

Obs  SIZE   USAGE     Fit  SE Fit  Residual  St Resid
  1  1290  1182.0  1362.2    68.3    -180.2     -1.29
  2  1350  1172.0  1383.5    65.6    -211.5     -1.50
  3  1470  1264.0  1426.2    60.6    -162.2     -1.13
  4  1600  1493.0  1472.5    55.4      20.5      0.14
  5  1710  1571.0  1511.7    51.4      59.3      0.41
  6  1840  1711.0  1557.9    47.3     153.1      1.04
  7  1980  1804.0  1607.8    43.7     196.2      1.32
  8  2230  1840.0  1696.8    40.3     143.2      0.96
  9  2400  1956.0  1757.3    40.5     198.7      1.33
 10  2710  2007.0  1867.6    46.0     139.4      0.94
 11  2930  1984.0  1945.9    52.9      38.1      0.26
 12  3000  1960.0  1970.8    55.5     -10.8     -0.07
 13  3210  2001.0  2045.6    64.0     -44.6     -0.32
 14  3240  1928.0  2056.3    65.3    -128.3     -0.91
 15  3520  1945.0  2155.9    78.0    -210.9     -1.57
```

Solution

a. A MINITAB printout of the simple linear regression is shown in Figure 10.20. Key statistics for assessing model adequacy are highlighted on the printout. First, note that the test of $H_0$: $\beta_1 = 0$ versus $H_a$: $\beta_1 \neq 0$ results in a $p$-value of .000. Thus, there is sufficient evidence (at any reasonably chosen $\alpha$) of a statistically useful model.

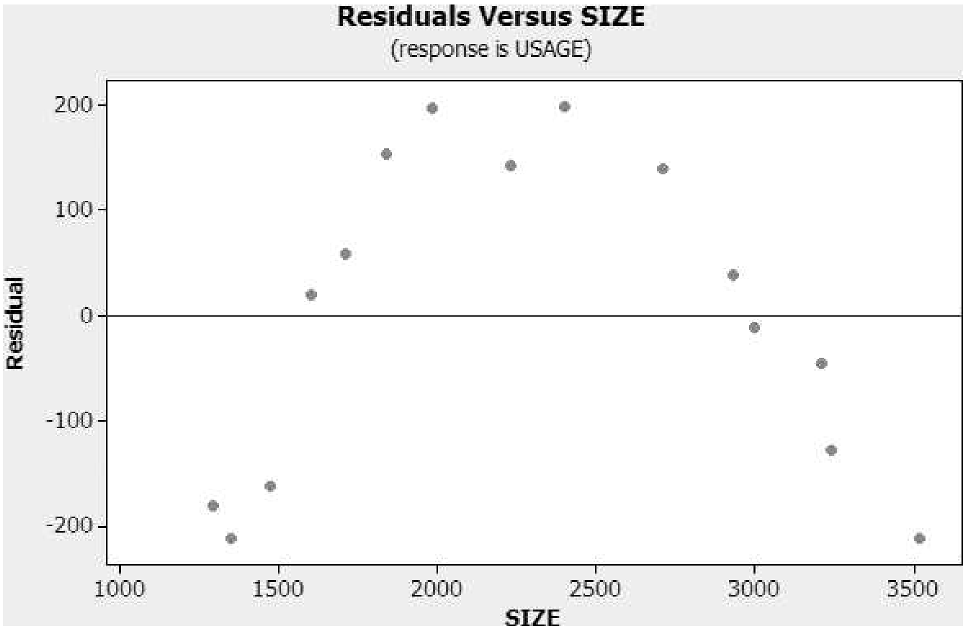
Next, the coefficient of determination is $r^2 = .76$. This implies that about 76% of the sample variation in electrical usage ($y$) can be explained by the linear model.

Finally, the estimated standard deviation of the error term is $s = 155.25$. We can say that about 95% of the actual July electrical usage values fall within $2s = 2(155.25) = 310.5$ kilowatt-hours of their respective predicted values.

Since the model is statistically useful, with a reasonably high $r^2$ value and a reasonably small 2s value, a researcher may choose to use the model for predicting future electrical usage values.

b. The regression residuals for the simple linear regression are also highlighted on Figure 10.20. We used MINITAB to plot these residuals against home size ($x$). The plot is displayed in Figure 10.21. Note the nonrandom pattern of points in the graph. In fact, the residuals exhibit a clear curvilinear trend, with the residuals for the small values of $x$ below the horizontal 0 (mean of the residuals) line, the residuals corresponding to the middle values of $x$ above the 0 line, and the residuals for the largest values of $x$ again below the 0 line. The indication is that the mean value of

**FIGURE 10.21**
MINITAB Residual Plot, Example 10.12




the random error $\varepsilon$ *within* each of these ranges of $x$ (small, medium, large) may not be equal to 0. Such a pattern typically indicates that the model is misspecified, violating the assumption of $E(\varepsilon) = 0$.

Not only does the residual plot indicate a violation of this assumption, the plot guides the researcher as to what modifications to make to the deterministic portion of the model. The curvilinear trend in the plot implies that curvature should be added to the model. We will see in the next chapter that a better model for electrical usage is $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$.

## Checking Assumption #2: Constant Error Variance

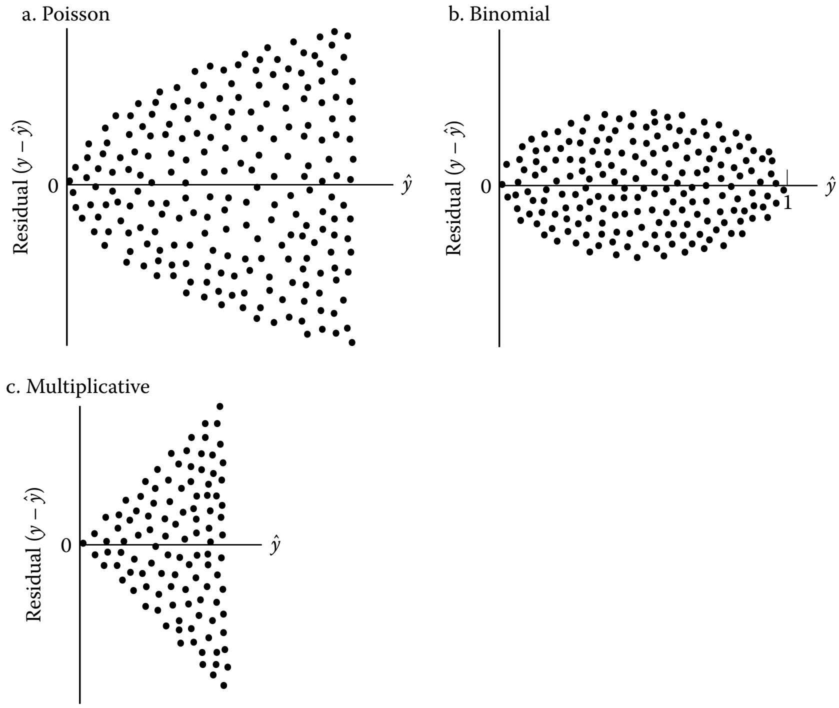
A residual plot can also be used to check the assumption of a constant error variance. Here, the appropriate graph is a plot of the residuals against the predicted value, $\hat{y}$. Like with the previous residual plot, if the graph reveals a random pattern of points (no trends), then it is likely that the assumption of a constant error variance is reasonably satisfied. However, if a strong pattern emerges on the residual plot, it is an indication of a violation of the assumption. For example, a plot of the residuals versus the predicted value $\hat{y}$ may display one of the patterns shown in Figure 10.22. In these figures, the range in values of the residuals increases (or decreases) as $\hat{y}$ increases, thus indicating that the variance of the random error, $\varepsilon$, becomes larger (or smaller) as the estimate of $E(y)$ increases in value. Because $E(y)$ depends on the $x$ values in the model, this implies that the variance of $\varepsilon$ is not constant for all settings of the $x$'s. Errors with a nonconstant variance are said to be **heteroscedastic** in nature.

**Definition 10.7**

In regression, random errors with a *nonconstant variance* are **heteroscedastic** errors. Random errors with a *constant variance* are **homoscedastic** errors.

In the next example we show how to use this plot to detect a nonconstant variance and suggest a useful remedy.

**FIGURE 10.22**
Residual plots showing changes in the variance of $\varepsilon$




## Example 10.13

**Detecting a Nonconstant Variance**

CIVILSAL

The data in Table 10.9 are the salaries, $y$, and years of experience, $x$, for a sample of 50 civil engineers. The first-order model $E(y) = \beta_0 + \beta_1 x$ was fit to the data using MINITAB. The MINITAB printout is shown in Figure 10.23, followed by a plot of the residuals versus $\hat{y}$ in Figure 10.24. Interpret the results. Is there evidence of a violation of the constant error variance assumption?

**TABLE 10.9 Salary Data for Example 10.13**

| Years of Experience $x$ | Salary $y$ | Years of Experience $x$ | Salary $y$ | Years of Experience $x$ | Salary $y$ |
|---|---|---|---|---|---|
| 7 | \$26,075 | 21 | \$43,628 | 28 | \$99,139 |
| 28 | 79,370 | 4 | 16,105 | 23 | 52,624 |
| 23 | 65,726 | 24 | 65,644 | 17 | 50,594 |
| 18 | 41,983 | 20 | 63,022 | 25 | 53,272 |
| 19 | 62,308 | 20 | 47,780 | 26 | 65,343 |
| 15 | 41,154 | 15 | 38,853 | 19 | 46,216 |
| 24 | 53,610 | 25 | 66,537 | 16 | 54,288 |
| 13 | 33,697 | 25 | 67,447 | 3 | 20,844 |
| 2 | 22,444 | 28 | 64,785 | 12 | 32,586 |
| 8 | 32,562 | 26 | 61,581 | 23 | 71,235 |
| 20 | 43,076 | 27 | 70,678 | 20 | 36,530 |
| 21 | 56,000 | 20 | 51,301 | 19 | 52,745 |
| 18 | 58,667 | 18 | 39,346 | 27 | 67,282 |
| 7 | 22,210 | 1 | 24,833 | 25 | 80,931 |
| 2 | 20,521 | 26 | 65,929 | 12 | 32,303 |
| 18 | 49,727 | 20 | 41,721 | 11 | 38,371 |
| 11 | 33,233 | 26 | 82,641 | | |

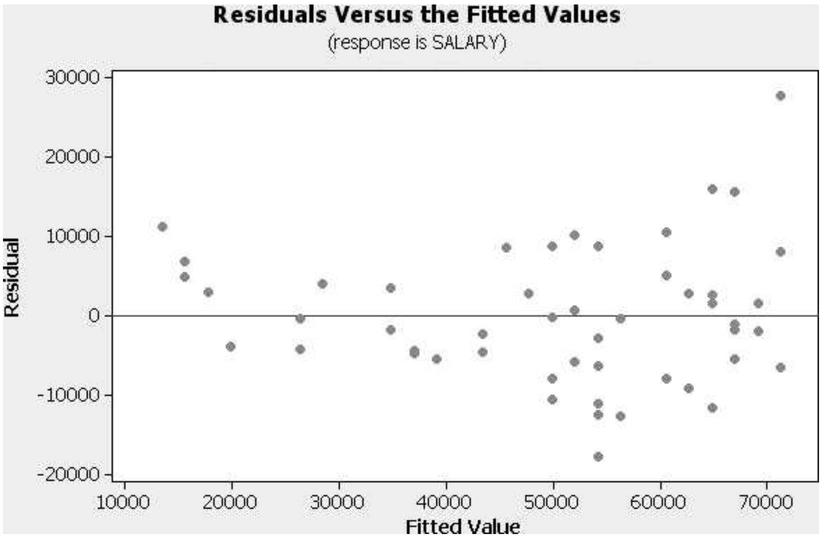
**FIGURE 10.23**
MINITAB regression output for Example 10.13

```
The regression equation is
SALARY = 11369 + 2141 EXP


Predictor     Coef  SE Coef       T      P
Constant     11369     3160    3.60  0.001
EXP         2141.4    160.8   13.31  0.000


S = 8642.44   R-Sq = 78.7%   R-Sq(adj) = 78.2%


Analysis of Variance

Source          DF           SS           MS       F      P
Regression       1  13239655469  13239655469  177.26  0.000
Residual Error  48   3585206077     74691793
Total           49  16824861546
```

**Solution** The MINITAB printout, Figure 10.23, suggests that the first-order model provides an adequate fit to the data. The $r^2$ value, .787, indicates that the model explains 78.7% of the sample variation in salaries. The $T$ value for testing $\beta_1$, 13.31, is highly significant ($p$-value $\approx 0$) and indicates that the model contributes information for the prediction of $y$. However, an examination of the residuals plotted against $\hat{y}$ (Figure 10.24) reveals a potential problem. Note the "cone" shape of the residual variability; the size of the residuals increases as the estimated mean salary increases.

This residual plot indicates that the assumption of a constant error variance is likely to be violated.

**FIGURE 10.24**
MINITAB residual plot for Example 11.13




Regression errors tend to be heteroscedastic when the variance of the dependent variable $y$ depends on the mean of $y$. Variables that represent counts per unit of area, volume, time, etc. (i.e., Poisson random variables) are cases in point. For a Poisson random variable (Section 4.10), we know that $E(y) = V(y)$. Since $\hat{y}$ is an estimate of $E(y)$, a plot of the residuals versus $\hat{y}$ for a Poisson dependent variable will reveal a pattern similar to the one shown in Figure 10.22a.

The remedy for this problem is to utilize a **variance-stabilizing transformation** on $y$. For a Poisson variable, the appropriate transformation is $y^* = \sqrt{y}$. In a simple linear regression application, we fit the model $\sqrt{y} = \beta_0 + \beta_1 x + \varepsilon$. Residuals from this transformed model will no longer exhibit the pattern shown in Figure 10.22a. Rather, these residuals will be randomly distributed.

Another random variable that typically violates the constant variance assumption is the binomial proportion, $y = \hat{p}$. For example, the dependent variable might be $y$ = proportion of fuses in a shipment that are defective. Recall that for a binomial variable, $E(\hat{p}) = p$ and $V(\hat{p}) = \sqrt{(p)(1 - p)}$. Note that the variance is a function of the mean. A plot of the residuals for a binomial dependent variable will often look like that in Figure 10.22b. Here, the residuals tend to have a small variance when the predicted proportion is near 0 or 1, and a large variance when the predicted proportion is near .5. To stabilize the variance for this type of data, use the transformation $y^* = \sin^{-1}\sqrt{y}$.

A third situation that typically requires a variance-stabilizing transformation is when the response $y$ does not follow the form of an *additive model*, $y = E(y) + \varepsilon$, but rather is better represented by a *multiplicative model* of the form $y = E(y) \cdot \varepsilon$. For multiplicative models, the variance of the response will grow proportionally to the square of the mean, i.e., $V(y) = [E(y)]^2 \cdot \sigma^2$. A plot of the residuals for a multiplicative dependent variable will often look like that in Figure 10.22c. To stabilize the variance for this type of data, use the natural logarithm transformation $y^* = \ln(y)$.

The three **variance-stabilizing transformations** we have discussed are summarized in Table 10.8.

**TABLE 10.8 Transformations to Stabilize the Variance of a Response**

| Residual Plot | Type of Data | Characteristics | Transformation |
|---|---|---|---|
| As shown in Figure 10.22a | Poisson | Counts per unit of time, distance, volume, etc. | $y^* = \sqrt{y}$ |
| As shown in Figure 10.22b | Binomial | Proportions, percentages, or numbers of successes for a fixed number $n$ of trials | $y^* = \sin^{-1}\sqrt{y}$ where $y$ is a proportion |
| As shown in Figure 10.22c | Multiplicative | Economic and scientific data | $y^* = \ln(y)$ |

**Example 10.14**

Stabilizing the error variance

Refer to Example 10.13 and the data on salary and experience in Table 10.9. Use the natural logarithmic transformation on the dependent variable and relate ln($y$) to years of experience $x$ with the linear model

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$$

a. Evaluate the adequacy of the model.
b. Interpret the value of $\hat{\beta}_1$.

Solution

a. The MINITAB printout in Figure 10.25 gives the regression analysis for the $n = 50$ measurements. The prediction equation

$$\widehat{\ln y} = 9.84 + .05x$$

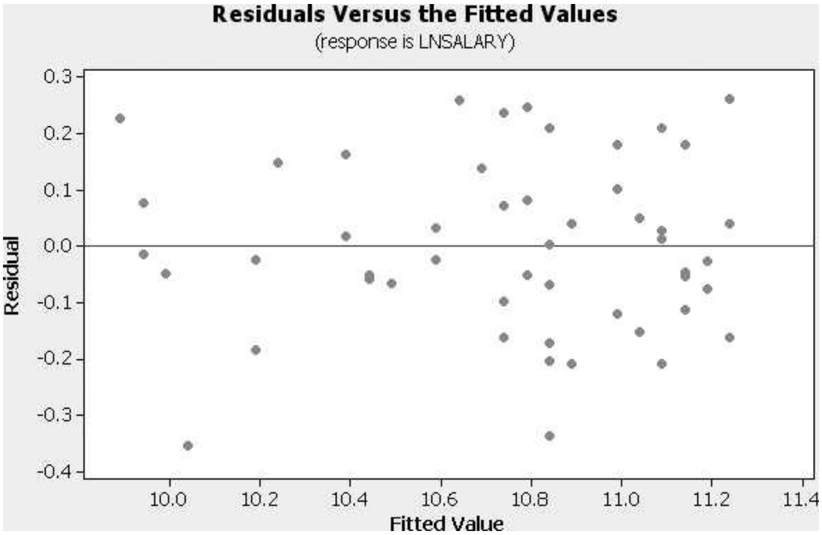
with $r^2 = .864$ and $T = 17.43$ for testing $H_0$: $\beta_1 = 0$, is highly significant ($p$-value $\approx 0$). Both imply that the model contributes significantly to the prediction of ln($y$).

The residual plot, shown in Figure 10.26, indicates that the logarithmic transformation has stabilized the error variances. Note that the cone shape is gone; there is no apparent tendency of the residual variance to increase as mean salary increases. Therefore, we are confident that inferences using the logarithmic model are more reliable than those using the untransformed model.

**FIGURE 10.25**
MINITAB regression output for log model of Example 10.14

```
The regression equation is
LNSALARY = 9.84 + 0.0500 EXP


Predictor       Coef    SE Coef       T      P
Constant     9.84131    0.05636  174.63  0.000
EXP         0.049979   0.002868   17.43  0.000


S = 0.154113   R-Sq = 86.4%   R-Sq(adj) = 86.1%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  7.2122  7.2122  303.66  0.000
Residual Error  48  1.1400  0.0238
Total           49  8.3522
```

**FIGURE 10.26**
MINITAB residual plot for log model of Example 10.14




b. Because we are using the natural logarithm of salary as the dependent variable, the $\beta$ estimates have slightly different interpretations than previously discussed. In general, a parameter $\beta$ in a log-transformed model represents the percentage increase (or decrease) in the dependent variable for a 1-unit increase in the corresponding independent variable. The percentage change is calculated by taking the antilogarithm of the $\beta$ estimate and subtracting 1, i.e., $e^{\hat{\beta}} - 1$.* For example, the percentage change in an engineer's salary associated with a 1-unit (i.e., 1-year) increase in years of experience $x$ is $(e^{\hat{\beta}_1} - 1) = (e^{.05} - 1) = .051$. Thus, we estimate an engineer's salary to increase 5.1% for each additional year of experience.

### Checking Assumption #3: Errors Normally Distributed

Of the four standard regression assumptions about the random error $\varepsilon$, the assumption that $\varepsilon$ is normally

*The result is derived by expressing the percentage change in salary $y$ as $(y_1 - y_0)/y_0$, where $y_1$ = the value of $y$ when, say, $x = 1$, and $y_0$ = the value of $y$ when $x = 0$. Now let $y^* = \ln(y)$ and assume the log model is $y^* = \beta_0 + \beta_1 x$. Then,

$$y = e^{y^*} = e^{\beta_0}e^{\beta_1 x} = \begin{cases} e^{\beta_0} & \text{when } x = 0 \\ e^{\beta_0}e^{\beta_1} & \text{when } x = 1 \end{cases}$$

Substituting, we have

$$\frac{y_1 - y_0}{y_0} = \frac{e^{\beta_0}e^{\beta_1} - e^{\beta_0}}{e^{\beta_0}} = e^{\beta_1} - 1$$

distributed is the least restrictive when we apply regression analysis in practice. That is, moderate departures from the assumption of normality have very little effect on the validity of the statistical tests, confidence intervals, and prediction intervals. In this case, we say that regression is **robust** with respect to nonnormality. However, great departures from normality cast doubt on any inferences derived from the regression analysis.

The methods of Section 5.6 (p. 206) can be used to determine whether the data grossly violate the assumption of normality. To illustrate, a MINITAB stem-and-leaf plot and normal probability plot of the $n = 50$ residuals for the ln(salary) model of Example 10.14 is shown in Figure 10.27. You can see that this distribution is approximately mound-shaped and reasonably symmetric. Consequently, it is unlikely that the normality assumption is grossly violated for this regression analysis.

When nonnormality of the random error term is detected, it can often be rectified by applying one of the transformations listed in Table 10.8. For example, if the relative frequency distribution (or stem-and-leaf display) of the residuals is highly skewed to

**FIGURE 10.27**
MINITAB Graphs for Checking Normality of Residuals for the Model of Example 10.14

**Stem-and-Leaf Display: RESIDUAL**

```
Stem-and-leaf of RESIDUAL  N  = 50
Leaf Unit = 0.010


  1   -3  5
  2   -3  3
  2   -2
  5   -2  000
 10   -1  87665
 12   -1  11
 18   -0  976655
 (8)  -0  44442221
 24   0   0122344
 17   0   5778
 13   1   044
 10   1   688
  7   2   11234
  2   2   66
```

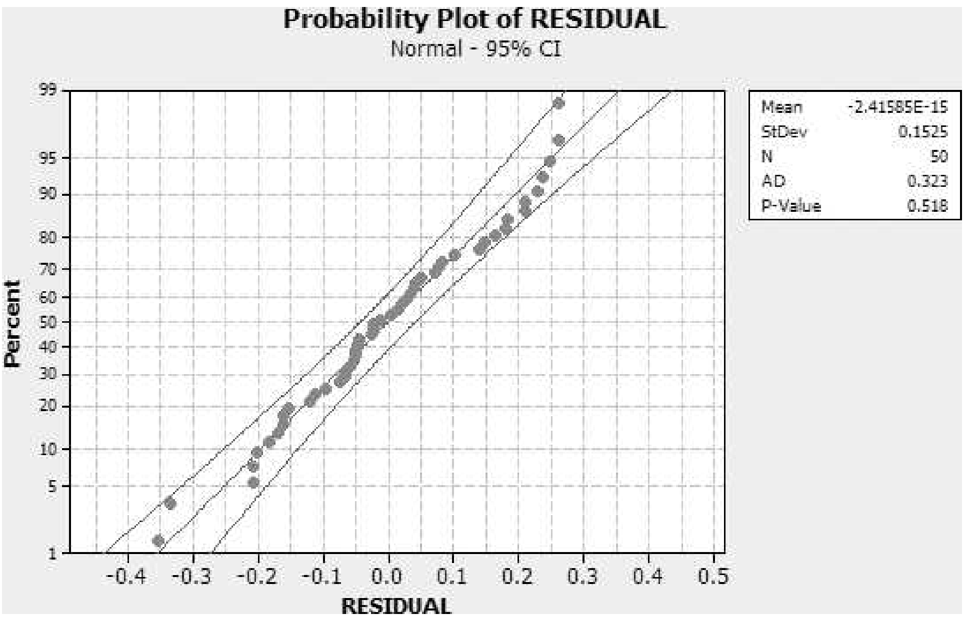
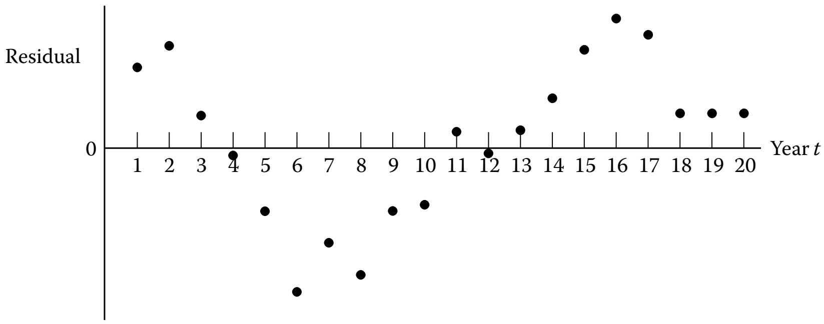


**FIGURE 10.28**
Residual plot for yearly time-series model




the right (as it usually is for Poisson data), the square-root transformation on $y$ will stabilize (approximately) the variance and, at the same time, will reduce skewness in the distribution of residuals.*

**Checking Assumption #4: Independent Errors** The assumption that the random errors are independent (uncorrelated) is most often violated when the data employed in a regression analysis are a **time series**. With time series data, the experimental units in the sample are time periods (e.g., years, months, or days) in consecutive time order.

For most economic and scientific time series, there is a tendency for the regression residuals to have positive and negative runs over time. For example, consider fitting a straight-line regression model to yearly time-series data. The model takes the form

$$E(y) = \beta_0 + \beta_1 t$$

where $y$ is the value of the time series in year $t$. A plot of the yearly residuals may appear as shown in Figure 10.28. Note that if the residual for year $t$ is positive (or negative) there is a tendency for the residual for year $(t + 1)$ to be positive (or negative). That is, neighboring residuals tend to have the same sign and appear to be correlated. Thus, the assumption of independent errors is likely to be violated and any inferences derived from the model are suspect.

Remedial measures for this problem involve proposing complex time-series models that include a model for both the deterministic and the random error components. Time-series models are beyond the scope of this text. Consult the references for this chapter to learn more about these models.

### A Summary of Steps to Follow in a Residual Analysis of the Simple Linear Regression Model

1. Check for a **misspecified model** by plotting the residuals $(y - \hat{y})$ against the independent variable in the model. A curvilinear trend detected in a plot implies that a quadratic term for that particular $x$ variable will probably improve model adequacy (see Chapter 12).
2. Check for **unequal variances** (or, **heteroscedasticity**) by plotting the residuals against the predicted values $(\hat{y})$. If you detect a pattern similar to one of those shown in Figure 10.22, refit the model using the appropriate variance-stabilizing transformation on $y$ (see Table 10.8).
3. Check for **nonnormal errors** by constructing a stem-and-leaf display (or histogram) for the residuals.† If you detect extreme skewness in the data, then apply one of the transformations listed in Table 10.8 (see step 4).

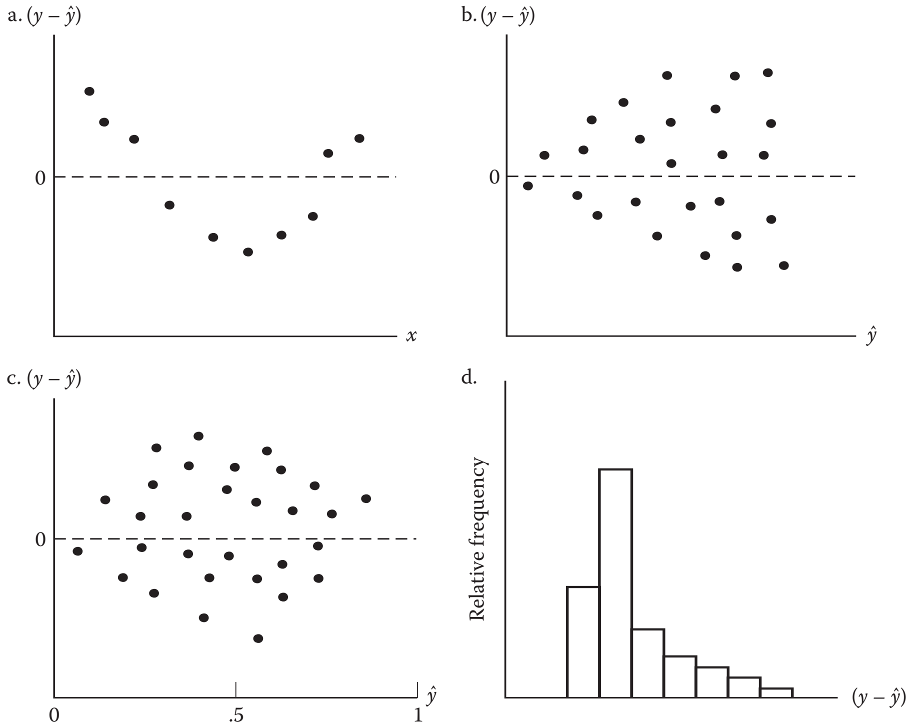
*Nonnormality of residuals may also be due to the presence of one or more unusual observations, called outliers. We discuss the detection of outliers in detail in Chapter 12.

†Hypothesis tests for normality are available (e.g., Shapiro-Wilk test) in most statistical software packages. However, these tests are strict in the sense that data with only a slight departure from normality will usually be deemed nonnormal by the test. Consult the chapter references for more information on these tests. If you do apply them, keep in mind that regression is robust against nonnormal errors.

4. Check for **correlated errors** by plotting the residuals in time order. If you detect runs of positive and negative residuals, propose a time-series model to account for the residual correlation.

## Applied Exercises

10.65 *Interpretation of residual plots.* Identify the problem(s) in each of the following residual plots:




**BIRDDEN**

10.66 *Planning an ecological network.* Refer to the *Landscape Ecology Engineering* (Jan. 2013) study of an ecological network, Exercise 10.37 (p. 511). Recall that the researchers used the data collected for 21 bird habitats to fit the model, $E(y) = \beta_0 + \beta_1 x$, where $y$ = bird density and $x$ = vegetation coverage (percentage).

a. Use the least squares prediction equation to calculate the predicted values of bird density and associated residuals for the model.
b. Plot the residuals against $\hat{y}$. Do you detect a trend?
c. Based on the residual plot, which assumption appears to be violated?
d. What model modification do you recommend?

**BBALL**

10.67 *Sound waves from a basketball.* Refer to the *American Journal of Physics* (June 2010) study of sound waves in a spherical cavity, Exercise 10.54 (p. 526). You fit a straight-line model relating frequency ($y$) of sound waves resulting from striking a basketball with a metal rod to number of resonances ($x$) and determined the model was adequate for predicting $y$.

a. Use the least squares prediction equation to calculate the residuals for the model.
b. Plot the residuals against number of resonances ($x$). Do you detect a trend?
c. Based on the residual plot, which assumption appears to be violated?
d. What model modification do you recommend?

**OJUICE**

10.68 *Sweetness of orange juice.* Refer to the study of the relationship between the "sweetness" of orange juice (measured as an index) and the amount of water soluble pectin (parts per million) used in the manufacturing process, Exercise 10.12 (p. 498). You used simple linear regression to predict sweetness index ($y$) from pectin amount ($x$). Conduct a residual analysis for this model that will provide insight into the validity of the standard regression assumptions on the random error, $\varepsilon$. Do you recommend any model modifications?

**FINTUBES**

10.69 *Thermal performance of copper tubes.* Refer to the *Journal of Heat Transfer* (Aug. 1990) study of the relationship between the amount of heat transferred in a copper tube and the area at the top of the tube that is not flooded by condensed vapor, Exercise 10.14 (p. 499). You used simple linear regression to predict the heat transfer enhancement ratio ($y$) from the unflooded area ratio ($x$). Conduct a residual analysis for this model that will provide insight into the validity of the standard regression assumptions on the random error, $\varepsilon$. Do you recommend any model modifications?

**BOTASH**

10.70 *Cracking in bottom ash waste asphalt.* Refer to the *Journal of Civil Engineering and Construction Technology* (Feb. 2013) study of bottom ash waste asphalt, Exercise 10.15 (p. 499). Recall that the researchers investigated the relationship between the cracking rate $y$ of bottom ash waste asphalt and stress intensity $x$ on data collected for 15 asphalt slabs. In Exercise 10.15 you fit the straight line model relating natural log of crack growth rate to natural log of stress intensity, i.e., $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$. For this model, conduct a residual analysis that will provide insight into the validity of the standard regression assumptions on the random error, $\varepsilon$. Do you recommend any model modifications?

**WATERPIPE**

10.71 *Estimating repair and replacement costs of water pipes.* Refer to the *IHS Journal of Hydraulic Engineering* (September 2012) study of water pipes susceptible to breakage, Exercise 10.31(p. 511). Recall that civil engineers used simple linear regression to model $y$ = the ratio of repair to replacement cost of commercial pipe as a function of $x$ = the diameter (in millimeters) of the pipe. Obtain the regression residuals and construct two graphs: (1) a plot of the residuals against diameter of the pipe, and (2) a normal probability plot. What do these plots suggest about the validity of the assumptions on the random error term?

## 10.10 A Complete Example

**FIREDAM**

**TABLE 10.9 Fire Damage Data**

| Distance from Fire Station $x$, miles | Fire Damage $y$, thousands of dollars |
|---|---|
| 3.4 | 26.2 |
| 1.8 | 17.8 |
| 4.6 | 31.3 |
| 2.3 | 23.1 |
| 3.1 | 27.5 |
| 5.5 | 36.0 |
| .7 | 14.1 |
| 3.0 | 22.3 |
| 2.6 | 19.6 |
| 4.3 | 31.3 |
| 2.1 | 24.0 |
| 1.1 | 17.3 |
| 6.1 | 43.2 |
| 4.8 | 36.4 |
| 3.8 | 26.1 |

In the previous sections, we have presented the basic elements necessary to fit and use a straight-line regression model. In this section, we will assemble these elements by applying them in an example where we use the computer to perform the calculations.

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the residence and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage $y$ and the distance $x$ between the fire and the nearest fire station are recorded for each fire. The results are given in Table 10.9.

**Step 1** First, we hypothesize a model to relate fire damage $y$ to the distance $x$ from the nearest fire station. We will hypothesize a straight-line probabilistic model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

**Step 2** Next, we enter the data into a computer and use a statistical software package to estimate the unknown parameters in the deterministic component of the hypothesized model. The SAS printout for the simple linear regression analysis is shown in Figure 10.29.

The least-squares estimates of $\beta_0$ and $\beta_1$, highlighted on the printout, are

$$\hat{\beta}_0 = 10.277929, \qquad \hat{\beta}_1 = 4.919331$$

Thus, the least squares equation is (after rounding)

$$\hat{y} = 10.278 + 4.919x$$

This prediction equation is shown on the MINITAB scatterplot for the data, Figure 10.30.

Dependent Variable: DAMAGE

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 841.76636 | 841.76636 | 156.89 | <.0001 |
| Error | 13 | 69.75098 | 5.36546 | | |
| Corrected Total | 14 | 911.51733 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 2.31635 | R-Square | 0.9235 |
| Dependent Mean | 26.41333 | Adj R-Sq | 0.9176 |
| Coeff Var | 8.76961 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 10.27793 | 1.42028 | 7.24 | <.0001 | 7.20960 | 13.34625 |
| DISTANCE | 1 | 4.91933 | 0.39275 | 12.53 | <.0001 | 4.07085 | 5.76781 |

Output Statistics

| Obs | DISTANCE | Dep Var DAMAGE | Predicted Value | Std Error Mean Predict | 95% CL Predict | | Residual |
|---|---|---|---|---|---|---|---|
| 1 | 3.4 | 26.2000 | 27.0037 | 0.5999 | 21.8344 | 32.1729 | -0.8037 |
| 2 | 1.8 | 17.8000 | 19.1327 | 0.8340 | 13.8141 | 24.4514 | -1.3327 |
| 3 | 4.6 | 31.3000 | 32.9068 | 0.7915 | 27.6186 | 38.1951 | -1.6068 |
| 4 | 2.3 | 23.1000 | 21.5924 | 0.7112 | 16.3577 | 26.8271 | 1.5076 |
| 5 | 3.1 | 27.5000 | 25.5279 | 0.6022 | 20.3573 | 30.6984 | 1.9721 |
| 6 | 5.5 | 36.0000 | 37.3342 | 1.0573 | 31.8334 | 42.8351 | -1.3342 |
| 7 | 0.7 | 14.1000 | 13.7215 | 1.1766 | 8.1087 | 19.3342 | 0.3785 |
| 8 | 3 | 22.3000 | 25.0359 | 0.6081 | 19.8622 | 30.2097 | -2.7359 |
| 9 | 2.6 | 19.6000 | 23.0682 | 0.6550 | 17.8678 | 28.2686 | -3.4682 |
| 10 | 4.3 | 31.3000 | 31.4311 | 0.7198 | 26.1908 | 36.6713 | -0.1311 |
| 11 | 2.1 | 24.0000 | 20.6085 | 0.7566 | 15.3442 | 25.8729 | 3.3915 |
| 12 | 1.1 | 17.3000 | 15.6892 | 1.0444 | 10.1999 | 21.1785 | 1.6108 |
| 13 | 6.1 | 43.2000 | 40.2858 | 1.2587 | 34.5906 | 45.9811 | 2.9142 |
| 14 | 4.8 | 36.4000 | 33.8907 | 0.8450 | 28.5640 | 39.2175 | 2.5093 |
| 15 | 3.8 | 26.1000 | 28.9714 | 0.6320 | 23.7843 | 34.1585 | -2.8714 |
| 16 | 3.5 | . | 27.4956 | 0.6043 | 22.3239 | 32.6672 | . |

| | |
|---|---|
| Sum of Residuals | 0 |
| Sum of Squared Residuals | 69.75098 |
| Predicted Residual SS (PRESS) | 93.21169 |

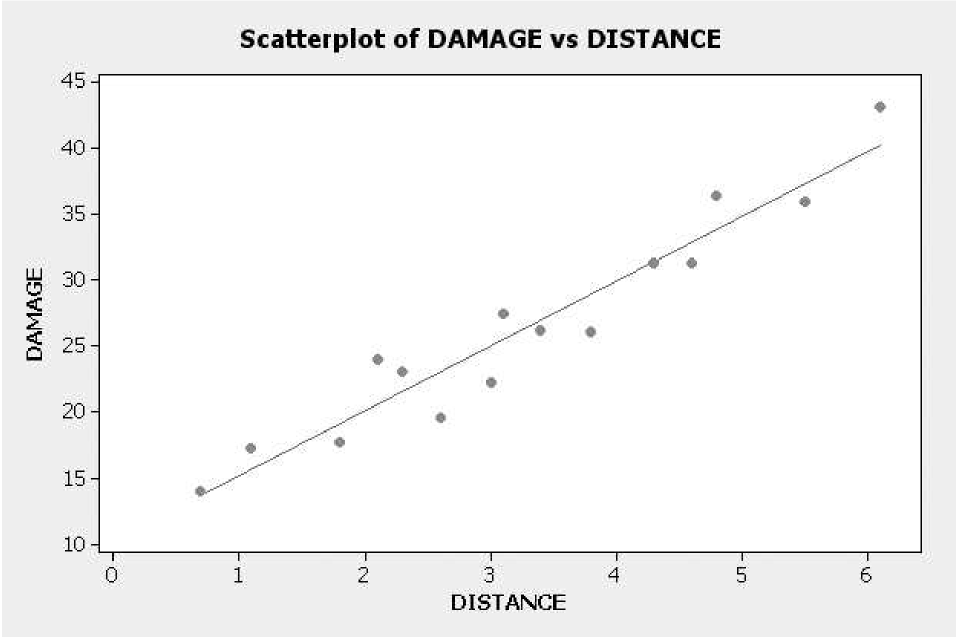
**FIGURE 10.29**
SAS printout for fire damage linear regression

The least-squares estimate of the slope, $\hat{\beta}_1 = 4.92$, implies that the estimated mean damage increases by \$4,920 for each additional mile from the fire station. This interpretation is valid over the range of $x$, or from .7 to 6.1 miles from the station. The estimated $y$-intercept, $\hat{\beta}_0 = 10.28$, has the interpretation that a fire 0 miles from the fire station has an estimated mean damage of \$10,280. Although this would seem to apply to the fire station itself, remember that the $y$-intercept is meaningfully interpretable only if $x = 0$ is within the sampled range of the independent variable. Since $x = 0$ is outside the range, $\hat{\beta}_0$ has no practical interpretation.

**Step 3** Now, we specify the probability distribution of the random error component $\varepsilon$. The assumptions about the distribution will be identical to those listed in Section 10.2.

1. $E(\varepsilon) = 0$.
2. $\text{Var}(\varepsilon) = \sigma^2$ is constant for all $x$ values.
3. $\varepsilon$ has a normal distribution.
4. $\varepsilon$'s are independent.

**FIGURE 10.30**
MINITAB scatterplot of fire damage data with least-squares model




We check the validity of these assumptions by examining residual plots. The residuals of the model (highlighted in Figure 10.29) are plotted against distance ($x$) in Figure 10.31a and against predicted fire damage ($\hat{y}$) in Figure 10.31b. No trends are detected in either graph, indicating that the first two assumptions (mean error of 0 and constant error variance) are likely satisfied. A normal probability plot of the residuals is shown in Figure 10.31c. Since

**FIGURE 10.31**
MINITAB Residual Plots for the Fire Damage Model

a. *Plot of Residuals vs. Distance* ($x$)

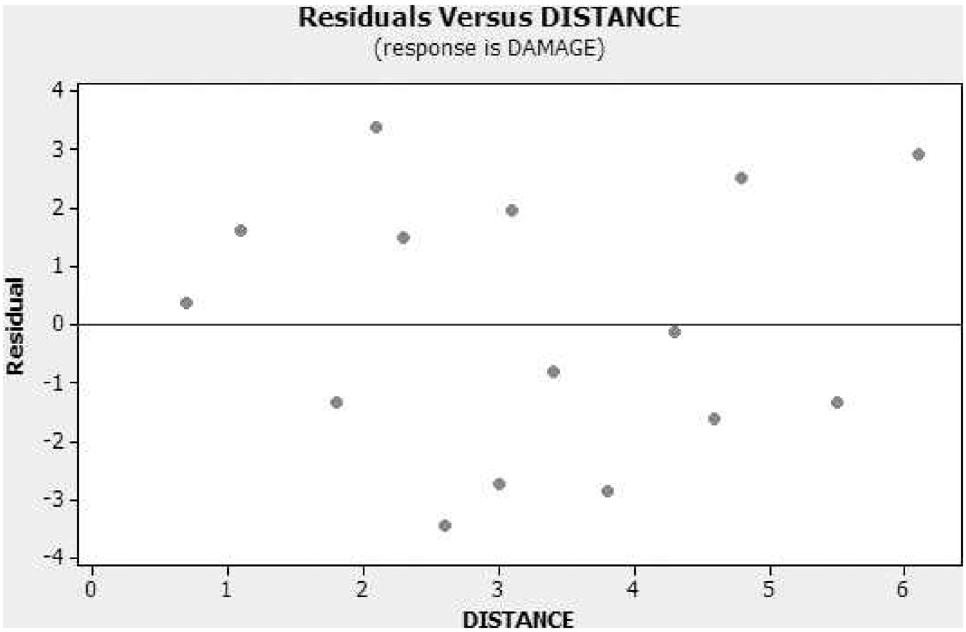
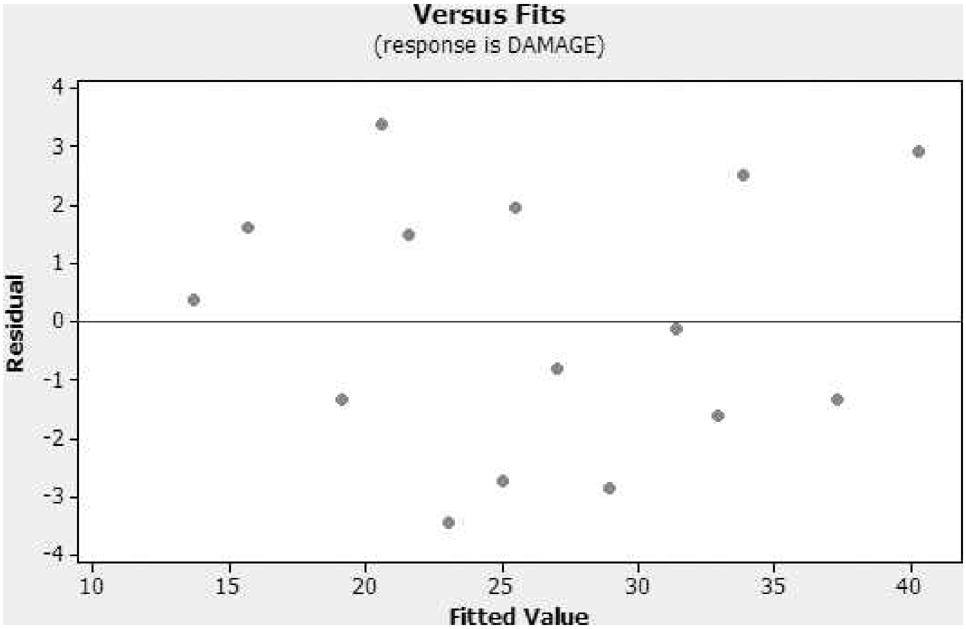
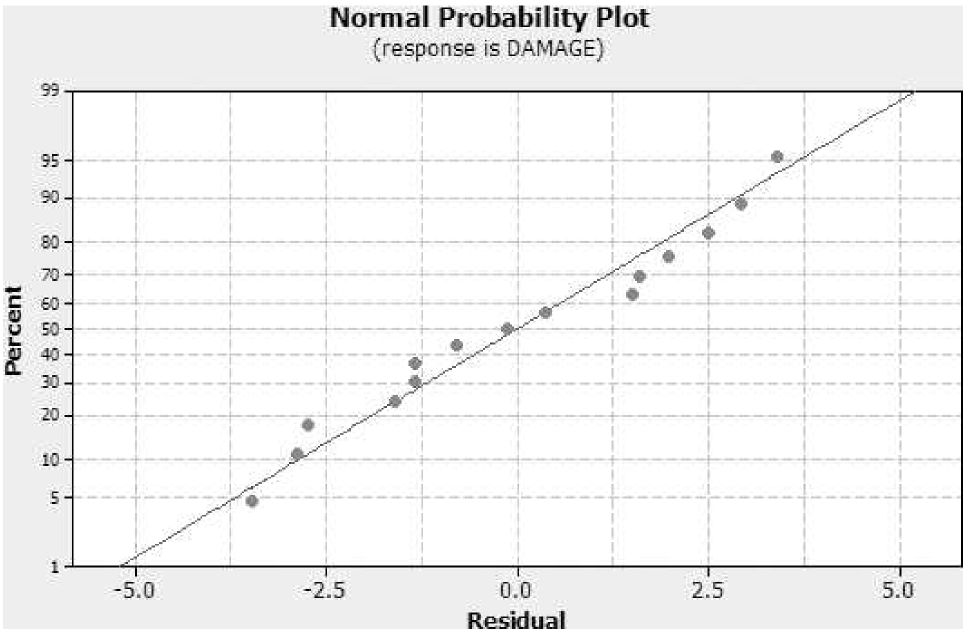


**FIGURE 10.31 (continued):** MINITAB Residual Plots for the Fire Damage Model

b. *Plot of Residuals vs. Predicted Damage* ($\hat{y}$)




c. *Normal Probability Plot of Residuals*




the points fall in nearly a straight line, the assumption of normal errors also appears to be satisfied. Since the data on fire damaged homes were collected independently, the fourth assumption of independent errors is most likely satisfied.

The estimate of the variance $\sigma^2$ of $\varepsilon$, shaded on the printout (Figure 10.29), is $s^2 = 5.36546$. (This value is also called **mean square for error**, or **MSE**.)

The estimated standard deviation of $\varepsilon$, also highlighted on Figure 10.29, is

$$s = 2.31635$$

The value of $s$ implies that most of the observed fire damage ($y$) values will fall within approximately $2s = 4.64$ thousand dollars of their respective predicted values.

**Step 4** We can now check the utility of the hypothesized model, that is, whether $x$ really contributes information for the prediction of $y$ using the straight-line model.

a. *Test of model utility.* First, test the null hypothesis that the slope $\beta_1$ is 0, i.e., that there is no linear relationship between fire damage and the distance from the nearest fire station, against the alternative that $x$ and $y$ are positively linearly related. We test:

$$H_0: \quad \beta_1 = 0$$

$$H_a: \quad \beta_1 > 0$$

The value of the test statistic, shaded on the printout, is $T = 12.525$, and the two-tailed $p$-value (also highlighted) is less than .0001.

Thus, the $p$-value for our one-tailed, upper-tailed test is less than

$$p = \frac{.0001}{2} = .00005$$

Since $\alpha = .05$ exceeds this small $p$-value, there is sufficient evidence to reject $H_0$ and conclude that distance between the fire and the fire station contributes information for the prediction of fire damage and that fire damage increases as the distance increases.

b. *Confidence interval for slope.* We gain additional information about the relationship by forming a confidence interval for the slope $\beta_1$. A 95% confidence interval for $\beta_1$ (highlighted on Figure 10.29) is (4.070, 5.768). We are 95% confident that the interval from \$4,070 to \$5,768 encloses the mean increase ($\beta_1$) in fire damage per additional mile distance from the fire station.

c. *Numerical descriptive measures of model adequacy.* The coefficient of determination (highlighted on the printout) is

$$r^2 = .9235$$

This implies that about 92% of the sample variation in fire damage ($y$) is explained by the distance $x$ between the fire and the fire station.

The coefficient of correlation $r$, which measures the strength of the linear relationship between $y$ and $x$, is not shown on Figure 10.29. Using the facts that $r = \sqrt{r^2}$ in simple linear regression and that $r$ and $\hat{\beta}_1$ have the same sign, we find

$$r = +\sqrt{r^2} = \sqrt{.9235} = .96$$

The high correlation confirms our conclusion that $\beta_1$ differs from 0; it appears that fire damage and distance from the fire station are linearly correlated.

The results of the test for $\beta_1$, the high value for $r^2$, and the relatively small $2s$ value (step 3), all point to a strong linear relationship between $x$ and $y$.

**Step 5** We are now prepared to use the least-squares model. Suppose the insurance company wants to predict the fire damage if a major residential fire were to occur

3.5 miles from the nearest fire station, i.e., $x_p = 3.5$. The predicted value shown (shaded) at the bottom of the SAS printout Figure 10.29, is $\hat{y} = 27.4956$, while the corresponding 95% prediction interval (also highlighted) is (22.3239, 32.6672). Therefore, we predict (with 95% confidence) that the fire damage for a major residential fire 3.5 miles from the nearest fire station will fall between \$22,324 and \$32,667.

*Caution*: We would not use this prediction model to make predictions for homes less than .7 mile or more than 6.1 miles from the nearest fire station. A look at the data in Table 10.9 reveals that all the $x$ values fall between .7 and 6.1. Recall from Section 10.8 that it is dangerous to use the model to make predictions outside the region in which the sample data fall. A straight line might not provide a good model for the relationship between the mean value of $y$ and the value of $x$ when stretched over a wider range of $x$ values.

## 10.11 A Summary of the Steps to Follow in Simple Linear Regression

We have introduced an extremely useful tool in this chapter—the **method of least squares** for fitting a prediction equation to a set of data. This procedure, along with associated statistical tests and estimations, is called a **regression analysis**. In five steps we showed how to use sample data to build a model relating a dependent variable $y$ to a single independent variable $x$.

### Steps to Follow in a Simple Linear Regression Analysis

1. The first step is to hypothesize a **probabilistic model**. In this chapter, we confined our attention to the **straight-line model**, $y = \beta_0 + \beta_1 x + \varepsilon$.
2. The second step is to use the method of least squares to estimate the unknown parameters in the **deterministic component**, $\beta_0 + \beta_1 x$. The least-squares estimates yield a model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with a **sum of squared errors (SSE)** that is smaller than the SSE for any other straight-line model.
3. The third step is to specify the probability distribution of the **random error component** $\boldsymbol{\varepsilon}$. Conduct a **residual analysis** to check the validity of these assumptions.
4. The fourth step is to assess the utility of the hypothesized model. Included here are making inferences about the **slope** $\boldsymbol{\beta_1}$, calculating the **coefficient of correlation** $\boldsymbol{r}$, and calculating the **coefficient of determination** $\boldsymbol{r^2}$.
5. Finally, if we are satisfied with the model, we are prepared to use it. We used the model to **estimate the mean** $\boldsymbol{y}$ **value,** $\boldsymbol{E(y)}$, for a given $x$ value and to **predict an individual** $\boldsymbol{y}$ **value** for a specific value of $x$.

## *STATISTICS IN ACTION REVISITED*

### Can Dowsers Really Detect Water?

We now return to the *Statistics in Action* problem described in the beginning of this chapter—to determine whether, in fact, dowsers can really detect water. Recall that a series of experiments were conducted by a group of university physicists in Munich, Germany. Based on preliminary tests, 43 individuals were selected for the final, carefully controlled, experiment. The researchers set up a 10-meter-long line on the ground floor of a vacant barn and a pipe with running water was located at a random point on the line. On the upper floor of the barn, directly above the experimental line, each of the

43 self-proclaimed dowsers was asked to ascertain (with his or her rod, stick, or other tool) where the pipe with running water on the ground floor was located.

For each trial, two variables were recorded: the actual pipe location (in decimeters from the beginning of the line) and the dowser's guess (also measured in decimeters). Data for the three "best" dowsers (numbered 99, 18, and 108) are saved in the **DOWSING** file (and are listed in Table SIA10.1, p. 484). The German physicists concluded in their final report that the three best dowsers "showed an extraordinarily high rate of success", thus "empirically proving" that dowsers can, in fact, find water.

Professor J.T. Enright of the University of California—San Diego challenged this claim by conducting his own analysis of the data. Let $x$ = dowser's guess (in meters) and $y$ = pipe location (in meters) for each trial. Enright's approach to determining whether the "best" dowsers are effective was to fit the straight-line model, $E(y) = \beta_0 + \beta_1 x$, to the data.

A MINITAB scatterplot of the data is shown in Figure SIA10.1. The least-squares line, obtained from the MINITAB regression printout shown in Figure SIA10.2, is also displayed on the scatterplot. Although the least squares line has a slight upward trend, the variation of the data points around the line is large. It does not appear that a dowser's guess ($x$) will be a very good predictor of actual pipe location ($y$). The two-tailed p-value for testing the null hypothesis, $H_0$: $\beta_1 = 0$, (highlighted on the printout) is $p$-value = .118. Even for an $\alpha$-level as high as $\alpha = .10$, there is insufficient evidence to reject $H_0$. Consequently, the dowsing data provide no statistical support for the German researchers' claim that the three best dowsers have an ability to find underground water with a divining rod.

This lack of support for the "dowsing" theory is made clearer with a confidence interval for the slope of the line. When $n = 26$, $df = (n - 2) = 24$ and $t_{.025} = 2.064$. Substituting this value and the relevant values shown on the MINITAB printout, a 95% confidence interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{0.25}(s_{\hat{\beta}_1}) = .31 \pm (2.064)(.19) = .31 \pm .39, \text{ or, } (-.08, .70)$$

Thus, for every 1-meter increase in a dowser's guess, we estimate (with 95% confidence) that the change in the actual pipe location will range anywhere from a decrease of .08 meter to an increase of .70 meter. In other words, we're not sure whether the pipe location will increase or decrease along the 10-meter pipeline! Keep in mind, also, that the data in Table SIA10.1 represent the "best" performances of the three dowsers, i.e., the outcome of the dowsing experiment in its most favorable light. When the data for all trials is considered and plotted, there is not even a hint of a trend.

**FIGURE SIA10.1**
MINITAB Scatterplot of Dowsing Data

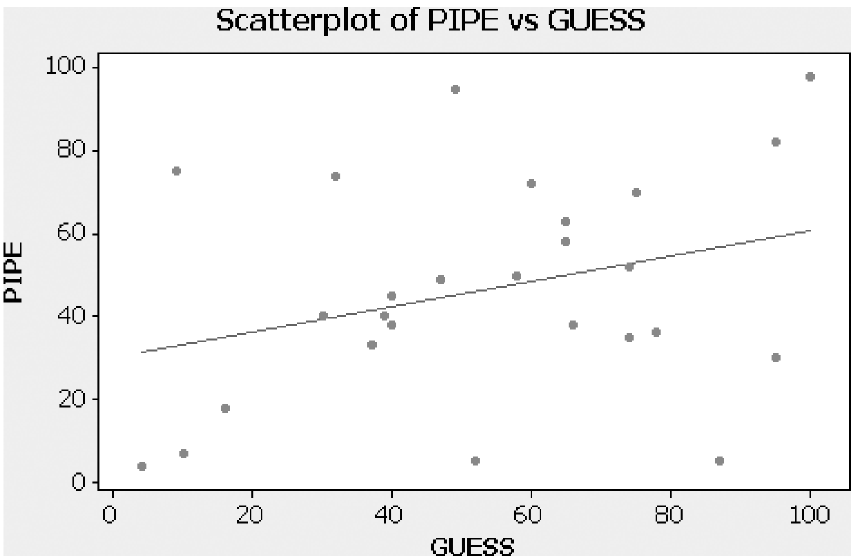


**FIGURE SIA10.2**
MINITAB Simple Linear Regression for Dowsing Data

```
Regression Analysis: PIPE versus GUESS

The regression equation is
PIPE = 30.1 + 0.308 GUESS


Predictor     Coef   SE Coef     T       P
Constant     30.07     11.41  2.63   0.015
GUESS       0.3079    0.1900  1.62   0.118


S = 26.0298   R-Sq = 9.9%   R-Sq(adj) = 6.1%


Analysis of Variance

Source          DF       SS      MS     F      P
Regression       1   1778.9  1778.9  2.63  0.118
Residual Error  24  16261.2   677.6
Total           25  18040.2
```

# Quick Review

## Key Terms

Additive model 536
Coefficient of correlation 546
Coefficient of determination 546
Confidence interval for mean of $y$ 522
Dependent variable 484
Deterministic model 486
Errors of prediction 486
Extrapolation 492
Heteroscedastic errors 533
Homoscedastic errors 533
Independent variable 484
Least-squares equations 490
Least-squares estimates 490
Least-squares line (or prediction equation) 492
Linear statistical models 485
Line of means 486
Method of least squares 488
Multiplicative model 536
Pearson product moment correlation coefficient 513
Population correlation coefficient 515
Prediction equation 484
Prediction interval for $y$ 523
Probabilistic model 546
Random error 546
Residual 546
Robust 538
Regression analysis 546
Regression model 484
Response variable 485
Scattergram 486
Simple linear regression model 486
Slope 487
Variance stabilizing transformation 535
$y$-intercept 487

## Key Formulas

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

where $SS_{xy} = \sum xy - \frac{(\Sigma x)(\Sigma y)}{n}$

$$SS_{xx} = \sum x^2 - \frac{(\Sigma x)^2}{n}$$

Least-squares estimates of $\beta$'s 490

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least-squares line 489

$$SSE = \sum (y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

where $SS_{yy} = \sum y^2 - \frac{(\Sigma y)^2}{n}$

Sum of squared errors 503

$$s^2 = \frac{SSE}{n - 2}$$

Estimated variance of $\sigma^2$ of $\varepsilon$ 503

## Key Formulas (continued)

| | |
|---|---|
| $s_{\hat{\beta}_1} = \dfrac{s}{\sqrt{\mathrm{SS}_{xx}}}$ | Estimated standard error of $\hat{\beta}_1$ 507 |
| $T = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ | Test statistic for $H_0$: $\beta_1 = 0$ 508 |
| $\hat{\beta}_1 \pm (t_{\alpha/2}) s_{\hat{\beta}_1}$ | $(1 - \alpha)100\%$ confidence interval for $\beta_1$ 509 |
| $r = \dfrac{\mathrm{SS}_{xy}}{\sqrt{\mathrm{SS}_{xx}\,\mathrm{SS}_{yy}}} = \pm\sqrt{r^2}$ (same sign as $\hat{\beta}_1$) | Coefficient of correlation 513 |
| $r^2 = \dfrac{\mathrm{SS}_{yy} - \mathrm{SSE}}{\mathrm{SS}_{yy}}$ | Coefficient of determination 518 |
| $\hat{y} \pm (t_{\alpha/2}) s \sqrt{\dfrac{1}{n} + \dfrac{(x_{\mathrm{p}} - \bar{x})^2}{\mathrm{SS}_{xx}}}$ | $(1 - \alpha)100\%$ confidence interval for $E(y)$ when $x = x_{\mathrm{p}}$ 522 |
| $\hat{y} \pm (t_{\alpha/2}) s \sqrt{1 + \dfrac{1}{n} + \dfrac{(x_{\mathrm{p}} - \bar{x})^2}{\mathrm{SS}_{xx}}}$ | $(1 - \alpha)100\%$ prediction interval for $y$ when $x = x_{\mathrm{p}}$ 523 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $y$ | | Dependent variable (variable to be predicted or modeled) |
| $x$ | | Independent (predictor) variable |
| $E(y)$ | | Expected (mean) value of $y$ |
| $\beta_0$ | beta-zero | $y$-intercept of true line |
| $\beta_1$ | beta-one | Slope of true line |
| $\hat{\beta}_0$ | beta-zero hat | Least-squares estimate of $y$-intercept |
| $\hat{\beta}_1$ | beta-one hat | Least-squares estimate of slope |
| $\varepsilon$ | epsilon | Random error |
| $\hat{y}$ | $y$-hat | Predicted value of $y$ |
| $(y - \hat{y})$ | | Error of prediction, or, residual |
| SSE | | Sum of squared errors (will be smallest for least-squares line) |
| $\mathrm{SS}_{xx}$ | | Sum of squares of $x$ values |
| $\mathrm{SS}_{yy}$ | | Sum of squares of $y$ values |
| $\mathrm{SS}_{xy}$ | | Sum of squares of cross-products, $x \cdot y$ |
| $r$ | | Coefficient of correlation |
| $r^2$ | R-squared | Coefficient of determination |
| $x_{\mathrm{p}}$ | | Value of $x$ used to predict $y$ |

## Chapter Summary Notes

- Two quantitative variables in *simple linear regression*: $\boldsymbol{y}$ = **dependent** variable (i.e., the variable to be predicted) and $\boldsymbol{x}$ = **independent** (i.e., predictor) variable.
- General form of a **probabilistic model** for $y$: $y = E(y) + \varepsilon$
- **Simple linear (straight-line) model**: $y = \beta_0 + \beta_1 x + \varepsilon$
- **Slope ($\boldsymbol{\beta_1}$)** represents the change in $y$ for every 1-unit increase in $x$.

- ***y*-intercept ($\beta_0$)** represents the value where the line intercepts the *y*-axis.
- **Steps** in simple linear regression: (1) Hypothesize the model, (2) use the method of least squares to estimate the unknown $\beta$'s, (3) make assumptions on the random error ($\varepsilon$), (4) statistically evaluate the adequacy of the model, and (5) if deemed useful, use the model for estimation and prediction.
- Properties of **method of least squares**: (1) sum of errors of prediction is 0, (2) sum of squared errors of prediction is minimized.
- Estimates of slope and *y*-intercept should only be interpreted *over the range of x-values in the sample*.
- **Four assumptions for** $\varepsilon$: (1) mean of $\varepsilon$ is 0, (2) variance of $\varepsilon$ is constant for all $x$ values, (3) distribution of $\varepsilon$ is normal, (4) values of $\varepsilon$ are independent.
- **Residual analysis** is used to check the validity of the assumptions on $\varepsilon$.
- *Interpretation of estimated standard deviation of* $\varepsilon$: About 95% of the observed $y$ values will lie within $2s$ of the respective predicted values.
- *Statistics used to assess the adequacy of the model*: (1) test of hypothesis for $\beta_1$, (2) confidence interval for $\beta_1$, (3) coefficient of correlation $r$, (4) coefficient of determination, $r^2$.
- Range of correlation coefficient: $-1 \le r \le 1$.
- Range of coefficient of determination: $0 \le r^2 \le 1$.
- **Correlation coefficient** measures the strength of the linear relationship between $x$ and $y$.
- **Coefficient of determination** gives the proportion of the sample variation in $y$ that can be explained by the straight-line model.
- Do not assume that a high correlation implies that $x$ causes $y$.
- For a given *x*-value, a confidence interval for $E(y)$ will be narrower than a prediction interval for $y$.

## Supplementary Applied Exercises

10.72 *Quantum tunneling.* At temperatures approaching absolute zero (273 degrees below zero Celsius), helium exhibits traits that defy many laws of conventional physics. An experiment has been conducted with helium in solid form at various temperatures near absolute zero. The solid helium is placed in a dilution refrigerator along with a solid impure substance, and the proportion (by weight) of the impurity passing through the solid helium is recorded. (This phenomenon of solids passing directly through solids is known as *quantum tunneling*.) The data are given in the table.

**HELIUM**

| Proportion of Impurity Passing Through Helium | Temperature |
|---|---|
| $y$ | $x$, °C |
| .315 | −262 |
| .202 | −265 |
| .204 | −256 |
| .620 | −267 |
| .715 | −270 |
| .935 | −272 |
| .957 | −272 |
| .906 | −272 |
| .985 | −273 |
| .987 | −273 |

a. Construct a scattergram of the data.
b. Find the least-squares line for the data and plot it on your scattergram.
c. Define $\beta_1$ in the context of this problem.
d. Test the hypothesis (at $\alpha = .05$) that temperature contributes no information for the prediction of the proportion of impurity passing through helium when a linear model is used. Draw the appropriate conclusions.
e. Find a 90% confidence interval for $\beta_1$. Interpret your results.
f. Find the coefficient of correlation for the given data.
g. Find the coefficient of determination for the linear model you constructed in part **b.** Interpret your result.
h. Find a 99% prediction interval for the proportion of impurity passing through helium when the temperature is set at −270°C.
i. Estimate the mean proportion of impurity passing through helium when the temperature is set at −270°C. Use a 99% confidence interval.

10.73 *Snow geese feeding trial.* Botanists at the University of Toronto conducted a series of experiments to investigate the feeding habits of baby snow geese. (*Journal of Applied Ecology*, Vol. 32, 1995.) Goslings were deprived of food until their guts were empty, then were allowed to feed for 6 hours on a diet of plants or Purina Duck Chow. For each feeding trial, the change in the weight of the gosling after 2.5 hours was recorded as a percentage of initial weight. Two other variables recorded were digestion efficiency (measured as a percentage) and amount of acid-detergent fiber in the digestive tract (also measured as a percentage). The data for 42 feeding trials are saved in the **SNOWGEESE** file. (The first and last 5 observations are shown in the accompanying table.)

a. The botanists were interested in the correlation between weight change ($y$) and digestion efficiency ($x$). Plot the data for these two variables in a scattergram. Do you observe a trend?
b. Find the coefficient of correlation relating weight change $y$ to digestion efficiency $x$. Interpret this value.
c. Conduct a test to determine whether weight change $y$ is correlated with a digestion efficiency $x$. Use $\alpha = .05$.
d. Repeat parts **b** and **c**, but exclude the data for trials that used duck chow from the analysis. What do you conclude?
e. The botanists were also interested in the correlation between digestion efficiency $y$ and acid-detergent fiber $x$. Repeat parts **a–d** for these two variables.

**SNOWGEESE** (*First and last 5 observations listed*)

| Feeding Trial | Diet | Weight Change (%) | Digestion Efficiency (%) | Acid-Detergent Fibre (%) |
|---|---|---|---|---|
| 1 | Plants | | 0 | 28.5 |
| 2 | Plants | | 2.5 | 27.5 |
| 3 | Plants | | 5 | 27.5 |
| 4 | Plants | 0 | 0 | 32.5 |
| 5 | Plants | 2 | 0 | 32 |
| ⋮ | | | | |
| 38 | Duck Chow | 9 | 59 | 8.5 |
| 39 | Duck Chow | 12 | 52.5 | 8 |
| 40 | Duck Chow | 8.5 | 75 | 6 |
| 41 | Duck Chow | 10.5 | 72.5 | 6.5 |
| 42 | Duck Chow | 14 | 69 | 7 |

*Source*: Gadallah, F. L., and Jefferies, R. L. "Forage quality in brood rearing areas of the lesser snow goose and the growth of captive goslings." *Journal of Applied Biology*, Vol. 32, No. 2, 1995, pp. 281–282 (adapted from Figures 2 and 3).

10.74 *Color analysis of Saturn's moon.* High-resolution images of Iapetus, one of Saturn's largest moons, were recently obtained by the *Cassini* spacecraft and analyzed by NASA (*Science*, Feb. 25, 2005). Using wideband filters, the ratios of ultraviolet to green and infrared to green wavelengths were measured at 24 moon locations. These color ratios are listed in the table below. According to the researchers, "the data's linear trend suggests mixing of two end members: Cassini Regio with a red spectrum and the south polar region with a flat spectrum." Conduct a complete simple linear regression analysis of the data, including a residual analysis. Do the results support the researchers' statement?

10.75 *Vehicle congestion study.* Modern warehouses employ computerized and automated guided vehicles for materials handling. Consequently, the physical layout of the warehouse must be carefully designed to prevent vehicle congestion and optimize response time. Optimal design of an automated warehouse was studied in *The Journal of Engineering for Industry* (Aug. 1993). The layout employed assumes that vehicles do not block each other when they travel within the warehouse, i.e., that there is no congestion. The validity of this assumption was checked by simulating (on a computer) warehouse operations. In each simulation, the number of vehicles was varied and the congestion time (total time one vehicle blocked another) was recorded. The data are shown in the

**SATMOON**

| Region | IR/Green | UV/Green | Region | IR/Green | UV/Green |
|---|---|---|---|---|---|
| Cassini Regio | 1.52 | 0.64 | Bright Terrain | 1.13 | 0.79 |
| | 1.51 | 0.65 | | 1.20 | 0.80 |
| | 1.54 | 0.65 | | 1.22 | 0.80 |
| | 1.53 | 0.66 | | 1.19 | 0.81 |
| Transition Zone | 1.44 | 0.66 | | 1.21 | 0.82 |
| | 1.42 | 0.69 | | 1.16 | 0.83 |
| | 1.42 | 0.70 | | 1.14 | 0.88 |
| | 1.28 | 0.73 | | 1.13 | 0.89 |
| | 1.40 | 0.75 | South Pole | 1.02 | 0.94 |
| | 1.24 | 0.75 | | 0.98 | 0.95 |
| | 1.32 | 0.77 | | 1.01 | 0.99 |
| | 1.26 | 0.77 | | 1.00 | 1.00 |

*Source*: Porco, C. C., et al. "Cassini imaging science: Initial results on Phoebe and Iapetus." *Science*, Vol. 307, No. 5713, Feb. 25, 2005 (Figure 8).

accompanying table. Of interest to the researchers is the relationship between congestion time ($y$) and number of vehicles ($x$). Conduct a complete simple linear regression analysis of the data, including a residual analysis. What conclusions can you draw from the data?

**WAREHOUSE**

| Number of Vehicles | Congestion Time, minutes |
|---|---|
| 1 | 0 |
| 2 | 0 |
| 3 | .02 |
| 4 | .01 |
| 5 | .01 |
| 6 | .01 |
| 7 | .03 |
| 8 | .03 |
| 9 | .02 |
| 10 | .04 |
| 11 | .04 |
| 12 | .04 |
| 13 | .03 |
| 14 | .04 |
| 15 | .05 |

*Source:* Pandit, R., and Palekar, U. S. "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry*, Transactions of the ASME, Vol. 115, Aug. 1993, p. 326 (Table 2).

10.76 Amorphous alloys have been found to have superior corrosion resistance. *Corrosion Science* (Sept. 1993) reported on the resistivity of an amorphous iron–boron–silicon alloy after crystallization. Five alloy specimens were annealed at 700°C, each for a different length of time. The passivation potential—a measure of resistivity of the crystallized alloy—was then measured for each specimen. The experimental data are shown below. Determine whether annealing time ($x$) is a useful linear predictor of passivation potential ($y$). If so, find the expected passivation potential, $E(y)$, when annealing time is set at $x = 30$ minutes, using a 90% confidence interval.

**ALLOY**

| Annealing Time $x$, minutes | Passivation Potential $y$, mV |
|---|---|
| 10 | −408 |
| 20 | −400 |
| 45 | −392 |
| 90 | −379 |
| 120 | −385 |

*Source*: Chattoraj, I., et al. "Polarization and resistivity measurements of post-crystallization changes in amorphous Fe-B-Si alloys." *Corrosion Science*, Vol. 49, No. 9, Sept. 1993, p. 712 (Table 1).

10.77 *Organic chemistry experiment.* Chemists at Kyushu University (Japan) examined the linear relationship between the maximum absorption rate $y$ (in nanomoles) and the Hammett substituent constant $x$ for metacyclophane compounds (*Journal of Organic Chemistry*, July 1995). The data for variants of two compounds are given in the table. The variants of compound 1 are labeled 1a, 1b, 1d, 1e, 1f, 1g, and 1h; the variants of compound 2 are 2a, 2b, 2c, and 2d.

a. Plot the data in a scattergram. Use two different plotting symbols for the two compounds. What do you observe?
b. Using only the data for compound 1, fit the model $E(y) = \beta_0 + \beta_1 x$.
c. Assess the adequacy of the model, part **b.** Use $\alpha = .01$.
d. Repeat parts **b** and **c** using only the data for compound 2.

**ORGCHEM**

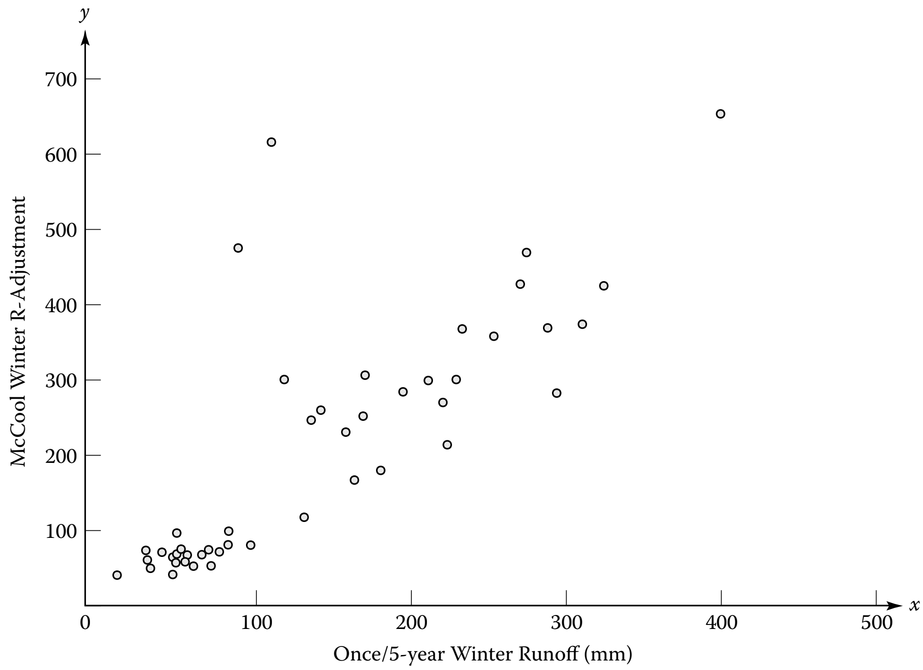
| Compound | Maximum Absorption $y$ | Hammett Constant $x$ |
|---|---|---|
| 1a | 298 | 0.00 |
| 1b | 346 | .75 |
| 1d | 303 | .06 |
| 1e | 314 | −.26 |
| 1f | 302 | .18 |
| 1g | 332 | .42 |
| 1h | 302 | −.19 |
| 2a | 343 | .52 |
| 2b | 367 | 1.01 |
| 2c | 325 | .37 |
| 2d | 331 | .53 |

*Source*: Adapted from Tsuge, A., et al. "Preparation and spectral properties of disubstituted [2-2] metacyclophanes." *Journal of Organic Chemistry*, Vol. 60, No. 15, July 1995, pp. 4390–4391 (Table 1 and Figure 1).

10.78 *Snowmelt runoff erosion.* The U.S. Department of Agriculture has developed and adopted the Universal Soil Loss Equation (USLE) for predicting water erosion of soils. In geographic areas where runoff from melting snow is common, calculating the USLE requires an accurate estimate of snowmelt runoff erosion. An article in the *Journal of Soil and Water Conservation* (Mar.–Apr. 1995) used simple linear regression to develop a snowmelt erosion index. Data for 54 climatological stations in Canada were used to model the McCool winter-adjusted rainfall erosivity index, $y$, as a straight-line function of the once-in-5-year snowmelt runoff amount, $x$ (measured in millimeters).

a. The data points are plotted in the scattergram shown on p. 553. Is there visual evidence of a linear trend?
b. The data for seven stations were removed from the analysis due to lack of snowfall during the study period. Why is this strategy advisable?

**Plot for Exercise 10.78**




c. The simple linear regression on the remaining $n = 47$ data points yielded the following results: $\hat{y} = -6.72 + 1.39x$, $s_{\hat{\beta}_1} = .06$. Use this information to construct a 90% confidence interval for $\beta_1$.

d. Interpret the interval, part **c**.

10.79 *Mercury poisoning in lakes.* In response to a health advisory regarding mercury poisoning in Maine lakes, the Environmental Protection Agency conducted a field study of 120 Maine lakes. (*Statistical Case Studies: A Collaboration Between Academe and Industry*, American Statistical Association, 1998.) In addition to the mercury level (parts per million) of each lake, the EPA measured the lake elevation (feet). The data are saved in the **MAINELAKE** file. (Data for the first 10 lakes are shown in the accompanying table.)

**MAINELAKE** (*Data for first 10 lakes shown*)

| Lake | Mercury Level | Elevation |
|---|---|---|
| Allen Pond | 1.080 | 425 |
| Alligator Pond | 0.025 | 1494 |
| Anasagunticook Lake | 0.570 | 402 |
| Balch & Stump Ponds | 0.770 | 557 |
| Baskahegan Lake | 0.790 | 417 |
| Bauneag Beg Lake | 0.750 | 205 |
| Beaver Pond | 0.270 | 397 |
| Belden Pond | 0.660 | 350 |
| Ben Annis Pond | 0.180 | 122 |
| Bottle Lake | 1.050 | 298 |

a. Consider the simple linear model, $E(y) = \beta_0 + \beta_1 x$, where $y$ = mercury level and $x$ = elevation. Is there evidence that mercury level decreases linearly as elevation increases?

b. Conduct a residual analysis for the linear model. Do the assumptions on $\epsilon$ appear to be satisfied?

10.80 *Rock-drilling experiment.* Two processes for hydraulic drilling of rock are dry drilling and wet drilling. In a dry hole, compressed air is forced down the drill rods to flush the cuttings and drive the hammer; in a wet hole, water is forced down. An experiment was conducted to determine whether the time $y$ it takes to dry drill a distance of 5 feet in rock increases with depth $x$ (*The American Statistician*, Feb. 1991). The results for one portion of the experiment are shown in the table at the top of p. 554. Conduct a complete simple linear regression analysis of the data, including a residual analysis. Interpret the results, practically.

10.81 *Strength of masonry joints.* Civil engineers often use the straight-line equation $E(y) = \beta_0 + \beta_1 x$ to model the relationship between the mean shear strength $E(y)$ of masonry joints and precompression stress $x$. To test this theory, a series of stress tests was performed on solid bricks arranged in triplets and joined with mortar (*Proceedings of*

**TRIPLETS**

| *Triplet Test* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| *Shear Strength, y* | 1.00 | 2.18 | 2.24 | 2.41 | 2.59 | 2.82 | 3.06 |
| *Precompression Stress, x* | 0 | .60 | 1.20 | 1.33 | 1.43 | 1.75 | 1.75 |

*Source*: Riddington, J. R., and Ghazali, M. Z. "Hypothesis for shear failure in masonry joints." *Proceedings of the Institute of Civil Engineers, Part 2*, Mar. 1990. Vol. 89, p. 96 (Fig. 7).

### Data for Exercise 10.80

**DRILLROCK**

| Depth at Which Drilling Begins x, feet | Time to Drill 5 Feet y, minutes |
|---|---|
| 0 | 4.90 |
| 25 | 7.41 |
| 50 | 6.19 |
| 75 | 5.57 |
| 100 | 5.17 |
| 125 | 6.89 |
| 150 | 7.05 |
| 175 | 7.11 |
| 200 | 6.19 |
| 225 | 8.28 |
| 250 | 4.84 |
| 275 | 8.29 |
| 300 | 8.91 |
| 325 | 8.54 |
| 350 | 11.79 |
| 375 | 12.12 |
| 395 | 11.02 |

*Source*: Penner, R., and Watts, D. G. "Mining information." *The American Statistician*, Vol. 45, No. 1, Feb. 1991, p. 6 (Table 1).

*the Institute of Civil Engineers*, Mar. 1990). The precompression stress was varied for each triplet and the ultimate shear load just before failure (called the shear strength) was recorded. The stress results for 7 triplets (measured in N/mm$^2$) are shown in the table on the preceding page.

a. Plot the seven data points in a scattergram. Does the relationship between shear strength and precompression stress appear to be linear?
b. Use the method of least squares to estimate the parameters of the linear model.
c. Interpret the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.
d. Conduct a test to determine if the slope, $\beta_1$, is positive

10.82 *Model of graphic interaction.* The Mixed Arithmetic-Perceptual (MA-P) model is a componential model of graphic interaction that was developed based on analyses of humans interacting with graphical displays on the computer. The assumptions of the MA-P model were tested in a research article reported in the *SIGCHI Bulletin* (July 1993). Using simple linear regression, the researcher modeled response time $y$ (in milliseconds) in a standard graph problem as a function of the number $x$ of processing steps required to solve the problem. A summary of the regression results for $n = 8$ problems follows:

$$\hat{y} = 1{,}346 + 450x, \qquad r^2 = .91$$

a. Interpret the value of $\hat{\beta}_1$.
b. Interpret the value of $r^2$.
c. Conduct a test of model adequacy at $\alpha = .01$. (*Hint*: Base the test on the value of $r$, the correlation coefficient.)

10.83 *High-volume air samplers.* The Environmental Protection Agency establishes industrial and occupational standards for the ambient air quality of total suspended particulates. The high-volume air sampler—the standard device used for sampling total suspended particulates—collects suspended particulates on large filters. The name *high-volume* is derived from the fact that the air sampler has a high sampling flow rate (measured in standard cubic meters per minute). Because of this high flow rate, large quantities of particles are collected over a 24-hour sampling period. However, the flow rate will vary depending on the pressure drop (in inches of water) across the filter medium. An experiment was conducted to investigate the relationship between flow rate and pressure drop. Eight sampling environments in which the high-volume air sampler was implemented yielded the measurements on average flow rate ($y$) and pressure drop across filter ($x$) listed in the table.

**AIRSAMPLER**

| Flow Rate y | Pressure Drop x | Flow Rate y | Pressure Drop x |
|---|---|---|---|
| .92 | 10 | 1.56 | 18 |
| 1.25 | 15 | 1.10 | 13 |
| .60 | 8 | .65 | 9 |
| 1.13 | 12 | 1.33 | 15 |

a. Use the data to develop a simple linear model for predicting the average flow rate of the high-volume air sampler based on the pressure drop of the filter.
b. Is the model of part **a** useful for predicting flow rate? (Use $\alpha = .05$.)
c. Using a 95% prediction interval, predict the flow rate in a sampling environment in which the pressure drop across the filter is 11 inches of water.

10.84 *Solar energy study.* Passive and active solar energy systems are becoming viable options to home builders as installation and operating costs decrease. Laminated solar modules utilize high-quality, single-crystal, silicon solar cells, connected electrically in series, to deliver a specified power output. Research was conducted to investigate the relationship between the solar cell temperature (°C) rise above ambient and the amount of insulation (megawatts per square centimeter). Data collected for six solar cells sampled under identical experimental conditions are recorded in the table on p. 555.

**SOLARCELL2**

| Temperature Rise Above Ambient $y$ | Insulation $x$ |
|---|---|
| 9 | 25 |
| 25 | 70 |
| 20 | 50 |
| 12 | 30 |
| 15 | 45 |
| 22 | 60 |

a. Fit a least-squares line to the data.
b. Plot the data and graph the line as a check on your calculations.
c. Calculate $r$ and $r^2$. Interpret their values.
d. Is the model useful for predicting temperature rise above ambient? (Use $\alpha = .01$.)
e. Estimate the mean temperature rise above ambient for solar cells with insulation of 35 megawatts per square centimeter. Use a 99% confidence interval.

10.85 *Spall damage in bricks.* A recent civil suit revolved around a five-building brick apartment complex located in the Bronx, New York, which began to suffer *spalling* damage (i.e., a separation of some portion of the face of a brick from its body). The owner of the complex alleged that the bricks were defectively manufactured. The brick manufacturer countered that poor design and shoddy management led to the damage. To settle the suit, an estimate of the rate of damage per 1,000 bricks, called the spall rate, was required. (*Chance*, Summer 1994.) The owner estimated the spall rate using several *scaffold-drop* surveys. (With this method, an engineer lowers a scaffold down at selected places on building walls and counts the number of visible spalls for every 1,000 bricks in the observation area.) The brick manufacturer conducted its own survey by dividing the walls of the complex into 83 wall segments and taking a photograph of each wall segment. (The number of spalled bricks that could be made out from each photo was recorded and the sum over all 83 wall segments used as an estimate of total spall damage.) In this court case, the jury was faced with the following dilemma: The scaffold-drop survey provided the most accurate estimate of spall rates in a given wall segment. Unfortunately, the drop areas were not selected at random from the entire complex; rather, drops were made at areas with high spall concentrations, leading to an overestimate of the total damage. On the other hand, the photo survey was complete in that all 83 wall segments in the complex were checked for spall damage. But the spall rate estimated by the photos, at least in areas of high spall concentration, was biased low (spalling damage cannot always be seen from a photo), leading to an underestimate of the total damage.

The data in the table are the spall rates obtained using the two methods at 11 drop locations. Use the data, as did expert statisticians who testified in the case, to help the jury estimate the true spall rate at a given wall segment. Then explain how this information, coupled with the data (not given here) on all 83 wall segments, can provide a reasonable estimate of the total spall damage (i.e., total number of damaged bricks).

**BRICKS**

| Drop Location | Drop Spall Rate (per 1,000 bricks) | Photo Spall Rate (per 1,000 bricks) |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 5.1 | 0 |
| 3 | 6.6 | 0 |
| 4 | 1.1 | .8 |
| 5 | 1.8 | 1.0 |
| 6 | 3.9 | 1.0 |
| 7 | 11.5 | 1.9 |
| 8 | 22.1 | 7.7 |
| 9 | 39.3 | 14.9 |
| 10 | 39.9 | 13.9 |
| 11 | 43.0 | 11.8 |

*Source*: Fairley, W. B., et al. "Bricks, buildings, and the Bronx: Estimating masonry deterioration." *Chance*, Vol. 7. No. 3, Summer 1994, p. 36 (Figure 3). (*Note*: The data points are estimated from the points shown on a scatterplot.)

CHAPTER 11

# Multiple Regression Analysis

## OBJECTIVE

*To extend the methods of Chapter 10 to develop a procedure for predicting a response y based on the values of two or more independent variables; to illustrate the types of practical inferences that can be drawn from this type of analysis*

## CONTENTS



*STATISTICS IN ACTION*
Bid-Rigging in the Highway Construction Industry

## *STATISTICS IN ACTION*

### Bid-Rigging in the Highway Construction Industry

In the United States, commercial contractors bid for the right to construct state highways and roads. A state government agency, usually the Department of Transportation (DOT), notifies various contractors of the state's intent to build a highway. Sealed bids are submitted by the contractors, and the contractor with the lowest bid (building cost) is awarded the road construction contract. The bidding process works extremely well in competitive markets, but has the potential to increase construction costs if the markets are noncompetitive or if collusive practices are present. The latter occurred in the 1980's in Florida. Numerous road contractors either admitted or were found guilty of price-fixing, i.e., setting the cost of construction above the fair, or competitive, cost through bid-rigging or other means.

This Statistics in Action involves data collected by the Florida Attorney General shortly following the price-fixing crisis. The Attorney General's objective is to build a model for the cost ($y$) of a road construction contract awarded using the sealed-bid system. The **FLAG** file contains data for a sample of 235 road contracts. The variables measured for each contract are listed in Table SIA11.1. Ultimately, the Attorney General wants to use the model to predict the costs of future road contracts in the state.

**FLAG**

**TABLE SIA11.1 Variables in the FLAG Data File**

| Variable Name | Type | Description |
|---|---|---|
| CONTRACT | Quantitative | Road contract number |
| COST | Quantitative | Low-bid contract cost (thousands of dollars) |
| DOTEST | Quantitative | DOT engineer's cost estimate (thousands of dollars) |
| STATUS | Qualitative | Bid status (1 = Fixed, 0 = Competitive) |
| B2B1RAT | Quantitative | Ratio of 2nd lowest bid to low bid |
| B3B1RAT | Quantitative | Ratio of 3rd lowest bid to low bid |
| BHB1RAT | Quantitative | Ratio of highest bid to low bid |
| DISTRICT | Qualitative | Location of road (1 = South Florida, 0 = North Florida) |
| BTPRATIO | Quantitative | Ratio of number of bidders to number of plan holders |
| DAYSEST | Quantitative | DOT engineer's estimate of number of workdays required |

We apply the statistical methodology presented in this chapter to the data described in Table SIA11.1. We will learn that two key predictors of contract cost are (1) the DOT engineer's estimate of the contract cost and (2) whether or not any collusion (bid-rigging) was involved in the bidding process. The analysis and results are presented in the *Statistics in Action Revisited* example at the end of this chapter.

## 11.1 General Form of a Multiple Regression Model

Most practical applications of regression analysis utilize models that are more complex than the simple linear (straight-line) model of Chapter 10. For example, a realistic probabilistic model for residential fire damage would include more than just the distance from the fire station discussed in Section 10.10. Factors such as size of the home, building material and extent of flame damage are a few of the many variables that might influence fire damage. Thus, we would want to incorporate these and other potentially important independent variables into the model if we need to make accurate predictions.

Probabilistic models that include more than one independent variable are called **multiple regression models**. The general form of these models is shown in the box. The dependent variable $y$ is now written as a function of $k$ independent variables, $x_1, x_2, \dots, x_k$. The random error term is added to make the model probabilistic rather than deterministic. The value of the coefficient $\beta_i$ determines the contribution of the independent variable $x_i$, given that the other $(k-1)$ independent variables are held constant, and $\beta_0$ is the $y$-intercept. The coefficients $\beta_0, \beta_1, \dots, \beta_k$ will usually be unknown, since they represent population parameters.

### General Form of the Multiple Regression Model*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where

$y$ is the dependent variable
$x_1, x_2, \dots, x_k$ are the independent variables
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the deterministic portion of the model
$\beta_i$ determines the contribution of the independent variable $x_i$

*Note*: The symbols $x_1, x_2, \dots, x_k$ may represent higher-order terms for quantitative predictors or terms that represent qualitative predictors.

At first glance it might appear that the regression model shown here would not allow for anything other than straight-line relationships between $y$ and the independent variables, but this is not true. Actually, $x_1, x_2, \dots, x_k$ can be functions of variables as long as the functions do not contain unknown parameters. For example, the carbon monoxide content $y$ of smoke emitted from a cigarette could be a function of the independent variables

$$x_1 = \text{Tar content}$$

$$x_2 = (\text{Tar content})^2 = x_1^2$$

$$x_3 = 1 \text{ if a filter cigarette, } 0 \text{ if a nonfiltered cigarette}$$

The $x_2$ term is called a **higher-order term** because it is the value of a quantitative variable ($x_1$) squared (i.e., raised to the second power). The $x_3$ term is a **coded variable** representing a **qualitative variable** (filter type). The multiple regression model is quite versatile and can be made to model many different types of response variables.

*The model is also called a **general linear model** because $E(y)$ is a linear function of the unknown parameters, $\beta_0, \beta_1, \beta_2, \dots$. The model

$$E(y) = \beta_0 e^{-\beta_1 x}$$

is *not* a linear model because $E(y)$ is not a *linear* function of the unknown model parameters, $\beta_0$ and $\beta_1$.

Once a model has been chosen to relate $E(y)$ to a set of independent variables, the steps of a multiple regression analysis parallel those of a simple regression analysis. The only differences are that the mathematical theory is beyond the scope of this text and the computations are considerably more complex. In the following sections, we will summarize the assumptions underlying a multiple regression analysis, present the methods for estimating and testing hypotheses about the model parameters, and show how to find a confidence interval for $E(y)$ or a prediction interval for $y$ for specific values of the independent variables. Since most multiple regression analyses are performed on a computer, we demonstrate how to interpret the output produced by statistical software.

### Analyzing a Multiple Regression Model

*Step 1* Hypothesize the deterministic component of the model. This component relates the mean, $E(y)$, to the independent variables $x_1, x_2, \dots, x_k$. This involves the choice of the independent variables to be included in the model.

*Step 2* Use the sample data to estimate the unknown model parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in the model.

*Step 3* Specify the probability distribution of the random error term, $\varepsilon$, and estimate the standard deviation of this distribution, $\sigma$.

*Step 4* Check that the assumptions on $\varepsilon$ are satisfied, and make model modifications if necessary.

*Step 5* Statistically evaluate the usefulness of the model.

*Step 6* When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

## 11.2 Model Assumptions

After we have selected the deterministic portion of a regression model—i.e., a model for $E(y)$—we add a component $\varepsilon$ to compensate for random error.

$$y = E(y) + \varepsilon$$

This component must obey the assumptions of the simple linear regression model—namely, that it is normally distributed with mean 0 and variance equal to $\sigma^2$. Further, we assume that the random errors associated with any pair of $y$ values are independent.

To present formulas for the parameter estimates, we need to write $E(y)$ in a standard form. Thus, we will let

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

be the deterministic component of the model containing $\beta_0$ and $k$ terms involving the predictor variables. The $x$ values that appear in the model are those of Section 11.1. For example, $x_2$ could be $x_1^2$, $x_3$ could be $\sin(x_1)$, etc. *The essential points are that the quantities $x_1, x_2, \dots, x_k$ can be measured without error when a value of $y$* is observed and that they do not involve any unknown parameters.

The linear regression model and associated assumptions are summarized in the box. In Section 11.10, we discuss how to use regression *residuals* to determine whether these assumptions are satisfied. Recall (from Chapter 10) that a residual is the difference between the observed and predicted value of $y$ (i.e., $y - \hat{y}$).

### Assumptions for a Multiple Regression Analysis

1. The mean of $\varepsilon$ is 0, i.e., $E(\varepsilon) = 0$. This implies that the mean of $y$ is equivalent to the deterministic component of the model, i.e.,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

2. For all settings of the independent variables $x_1, x_2, \ldots, x_k$, the variance of $\varepsilon$ is constant, i.e., $V(E_i) = \sigma^2$.
3. The probability distribution of $\varepsilon$ is normal.
4. The random errors are independent (in a probabilistic sense).

## 11.3 Fitting the Model: The Method of Least Squares

The method of fitting a multiple regression model is identical to that of fitting the first-order (straight-line) model. Thus, we will use the method of least squares and choose estimates of $\beta_0, \beta_1, \ldots, \beta_k$ that minimize

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})]^2$$

As in the case of the straight-line model, the sample estimates $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ that minimize SSE will be obtained as solutions to the system of simultaneous linear equations

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0} = 0 \qquad \frac{\partial \text{SSE}}{\partial \hat{\beta}_1} = 0 \qquad \cdots \qquad \frac{\partial \text{SSE}}{\partial \hat{\beta}_k} = 0$$

To illustrate the nature of this system, we will examine the first equation. Taking the partial derivative of SSE with respect to $\hat{\beta}_0$, we obtain

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})]\,(-1)$$

Setting $\partial\text{SSE}/\partial\hat{\beta}_0$ equal to 0 yields

$$\sum y_i - \left(n\hat{\beta}_0 + \sum x_{i1}\hat{\beta}_1 + \sum x_{i2}\hat{\beta}_2 + \cdots + \sum x_{ik}\hat{\beta}_k\right) = 0$$

or

$$n\hat{\beta}_0 + \left(\sum x_{i1}\right)\hat{\beta}_1 + \left(\sum x_{i2}\right)\hat{\beta}_2 + \cdots + \left(\sum x_{ik}\right)\hat{\beta}_k = \sum y_i$$

As in the case of simple linear regression, this is a linear equation in $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The $k$ remaining least-squares equations, all linear equations in $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, are

$$\begin{array}{l}
\left(\sum x_{i1}\right)\hat{\beta}_0 + \left(\sum x_{i1}^2\right)\hat{\beta}_1 + \left(\sum x_{i1}x_{i2}\right)\hat{\beta}_2 + \cdots + \left(\sum x_{i1}x_{ik}\right)\hat{\beta}_k = \sum x_{i1}y_i \\
\left(\sum x_{i2}\right)\hat{\beta}_0 + \left(\sum x_{i1}x_{i2}\right)\hat{\beta}_1 + \left(\sum x_{i2}^2\right)\hat{\beta}_2 + \cdots + \left(\sum x_{i2}x_{ik}\right)\hat{\beta}_k = \sum x_{i2}y_i \\
\qquad \vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad\qquad \vdots \qquad\qquad \vdots \\
\left(\sum x_{ik}\right)\hat{\beta}_0 + \left(\sum x_{i1}x_{ik}\right)\hat{\beta}_1 + \qquad \cdots \qquad + \left(\sum x_{ik}^2\right)\hat{\beta}_k = \sum x_{ik}y_i
\end{array}$$

As you can see, writing the $(k + 1)$ least-squares linear equations is a task; solving them simultaneously by hand is even more difficult. An easy way to express the equations and to solve them is to use matrix algebra, but the inevitable computations are best obtained with statistical software.

In Sections 11.4–11.7, we use matrix algebra to give formulas for the least-squares estimates, SSE, test statistics, confidence intervals, and prediction intervals.

Their use will be illustrated with simple numerical examples. (You may want to review the concepts in Appendix A, "Matrix Algebra," before reading the remainder of this chapter.) In the remaining sections, we present several useful multiple regression models and demonstrate how to analyze each using the printouts for the SAS, SPSS, MINITAB, and EXCEL software.

## 11.4 Computations Using Matrix Algebra: Estimating and Making Inferences About the Individual $\beta$ Parameters

### Least-Squares Equations and Solution

To apply matrix algebra to a regression analysis, we must place the data in matrices in a particular pattern. We will suppose that the model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $x_1, x_2, \ldots$ could actually represent the squares, cubes, cross-products, or other functions of predictor variables, and $\varepsilon$ is a random error. We will assume that we have collected $n$ data points, i.e., $n$ values of $y$ and corresponding values of $x_1, x_2, \ldots, x_k$, and that these are denoted as shown in Table 11.1. Then the two data matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ are as shown in the box.

Notice that the first column in the $\boldsymbol{X}$ matrix is a column of 1's. Thus, we are inserting a value of $x$, namely, $x_0$, as the coefficient of $\beta_0$, where $x_0$ is a variable always equal to 1. Therefore, there is one column in the $\boldsymbol{X}$ matrix for each $\beta$ parameter. Also, remember that a particular data point is identified by specific rows of the $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices. For example, the $y$ value $y_3$ for data point 3 is in the third row of the $\boldsymbol{Y}$ matrix, and the corresponding values of $x_1, x_2, \ldots, x_k$ appear in the third row of the $\boldsymbol{X}$ matrix. Using this notation, the general linear model can be expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

The Data Matrices $Y$ and $X$, the $\hat{\beta}$ Matrix, and the Error Matrix

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The $\hat{\boldsymbol{\beta}}$ matrix shown in the box contains the least-squares estimates (which we are attempting to obtain) of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ of the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

**TABLE 11.1 Notation for Multiple Regression**

| Data Point | $y$-value | $x_1$ | $x_2$ | ... | $x_k$ | Unobservable Random Error |
|---|---|---|---|---|---|---|
| 1 | $y_1$ | $x_{11}$ | $x_{12}$ | ... | $x_{1k}$ | $\varepsilon_1$ |
| 2 | $y_2$ | $x_{21}$ | $x_{22}$ | ... | $x_{2k}$ | $\varepsilon_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| $n$ | $y_n$ | $x_{n1}$ | $x_{n2}$ | ... | $x_{nk}$ | $\varepsilon_n$ |

Using the $\boldsymbol{Y}$ and $\boldsymbol{X}$ data matrices, their transposes, and the $\hat{\boldsymbol{\beta}}$ matrix, we can write the least-squares equations as

**Least-Squares Matrix Equation**

$$(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$$

Thus, $(\boldsymbol{X}'\boldsymbol{X})$ is the coefficient matrix of the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, and $\boldsymbol{X}'\boldsymbol{Y}$ gives the matrix of constants that appear on the right-hand side of the equality signs.

The solution, which follows from Appendix A.3,* is

**Least-Squares Solution**

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Thus, to solve the least-squares matrix equation, the computer calculates $(\boldsymbol{X}'\boldsymbol{X})$, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $\boldsymbol{X}'\boldsymbol{Y}$, and, finally, the product $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. We will illustrate this process using the data for the insulation compression example from Section 10.2.

## Example 11.1

Estimating $\beta$'s using Matrix Algebra: Simple Linear Regression

Find the least-squares line for the insulation compression data repeated in Table 11.2, where $x_1$ = pressure.

**TABLE 11.2 Compression Versus Pressure for an Insulation Material**

**INSULATION**

| Specimen | Pressure $x_1$ | Compression $y$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 3 | 2 |
| 4 | 4 | 2 |
| 5 | 5 | 4 |

**Solution**

The model is the simple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

and the $\boldsymbol{Y}$, $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\varepsilon}$ matrices are

$$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 4 \end{bmatrix} \qquad \boldsymbol{X} = \begin{array}{c} \begin{array}{cc} x_0 & x_1 \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \end{array} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

Then,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 10 \\ 37 \end{bmatrix}$$

*In the notation of Appendix A.3, $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{V} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{Y}$. Then the solution to the equation $\boldsymbol{AV} = \boldsymbol{G}$ is $\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{G}$.

The last matrix that we need is $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. This matrix can be found by using a computer program or by using the method of Appendix A.4. Thus, you would find

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.1 & -.3 \\ -.3 & .1 \end{bmatrix}$$

Then the solution to the least-squares equation is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 1.1 & -.3 \\ -.3 & .1 \end{bmatrix}\begin{bmatrix} 10 \\ 37 \end{bmatrix} = \begin{bmatrix} -.1 \\ .7 \end{bmatrix}$$

Thus, $\hat{\beta}_0 = -.1$, $\hat{\beta}_1 = .7$, and the prediction equation is

$$\hat{y} = -.1 + .7x$$

You can verify that this is the same answer as obtained in Section 10.3.

## Example 11.2

Estimating $\beta$'s Using Matrix Algebra: Multiple Regression

Refer to the insulation compression data in Table 11.2. In addition to compression ($y$) and pressure ($x_1$), temperature ($x_2$) in 10 degrees Centigrade is measured for each specimen. The data are listed in Table 11.3. Now consider the two-variable multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Find the least-squares estimates of $\beta_0$, $\beta_1$, and $\beta_2$.

**TABLE 11.3 Temperature Added to Compression-Pressure Data**

| Spec. | Press. | Temp. | Comp. |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 1 |
| 3 | 3 | 2 | 2 |
| 4 | 4 | 4 | 2 |
| 5 | 5 | 3 | 4 |

Solution

The $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\hat{\boldsymbol{\beta}}$ matrices are shown here:

$$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 4 \end{bmatrix} \qquad \boldsymbol{X} = \begin{array}{c} \begin{array}{ccc} x_0 & x_1 & x_2 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 2 \\ 1 & 4 & 4 \\ 1 & 5 & 3 \end{bmatrix} \end{array}$$

Then:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 5 & 15 & 12 \\ 15 & 55 & 42 \\ 12 & 42 & 34 \end{bmatrix} \qquad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 10 \\ 37 \\ 27 \end{bmatrix}$$

And, using a statistical software package, we obtain

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix}$$

Finally, performing the multiplication, we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix}\begin{bmatrix} 10 \\ 37 \\ 27 \end{bmatrix} \\ &= \begin{bmatrix} .35 \\ 1.15 \\ -.75 \end{bmatrix} \end{aligned}$$

Thus, $\hat{\beta}_0 = .35$, $\hat{\beta}_1 = 1.15$, $\hat{\beta}_2 = -.75$, and the prediction equation is

$$\hat{y} = .35 + 1.15x_1 - .75x_2$$

Recall that the $\hat{\boldsymbol{\beta}}$ matrix is equal to the product of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ matrix and the $\boldsymbol{Y}$ matrix:

$$\hat{\boldsymbol{\beta}} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}$$

Since elements in the $\hat{\boldsymbol{\beta}}$ matrix (i.e., the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$) are obtained by multiplying the rows of $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ by the column matrix $\boldsymbol{Y}$, it follows that $\hat{\beta}_0$ will equal the product of the first row of $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and the $\boldsymbol{Y}$ matrix and, in general, $\hat{\beta}_i$ will equal the product of the $(i+1)$st row of $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{Y}$. Therefore, for $i = 0, 1, 2, \dots, k$, $\hat{\beta}_i$ is a linear function of $n$ normally distributed random variables, $y_1, y_2, \dots, y_n$, and, by Theorem 6.7, $\hat{\beta}_i$ possesses a normal sampling distribution.

Derivation of the means and variances of the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ is beyond the scope of this text. However, it can be shown that the least-squares estimators provide unbiased estimates of $\beta_0, \beta_1, \dots, \beta_k$, i.e.,

$$E(\hat{\beta}_i) = \beta_i \quad \text{for } i = 0, 1, 2, \dots, k$$

The standard errors and covariances of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are determined by the elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix. Thus, if we denote the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix as

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & & \cdots & c_{0k} \\ c_{10} & c_{11} & & \cdots & c_{1k} \\ c_{20} & c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{k0} & \cdot & \cdot & \cdot & c_{kk} \end{bmatrix}$$

then it can be shown (proof omitted) that the standard errors of the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are

$$\sigma_{\hat{\beta}_0} = \sigma\sqrt{c_{00}}$$
$$\sigma_{\hat{\beta}_1} = \sigma\sqrt{c_{11}}$$
$$\sigma_{\hat{\beta}_2} = \sigma\sqrt{c_{22}}$$
$$\vdots$$
$$\sigma_{\hat{\beta}_k} = \sigma\sqrt{c_{kk}}$$

where $\sigma$ is the standard deviation of the random error $\varepsilon$. In other words, the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ give the values of $c_{00}, c_{11}, \dots, c_{kk}$ that are required for finding the standard errors of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$.

The properties of the sampling distributions of the least-squares estimators are summarized in the box.

**THEOREM 11.1**

*Properties of the Sampling Distribution of $\hat{\beta}_i$ $(i = 0, 1, 2, \dots, k)$*

The sampling distribution of $\hat{\beta}_i$ $(i = 0, 1, 2, \dots, k)$ is normal with

$$E(\hat{\beta}_i) = \beta_i \qquad V(\hat{\beta}_i) = c_{ii}\sigma^2 \qquad \sigma_{\hat{\beta}_i} = \sigma\sqrt{c_{ii}}$$

The off-diagonal elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix determine the covariances of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Thus, it can be shown that the covariance of two parameter estimators, say $\hat{\beta}_i$ and $\hat{\beta}_j$ (where $i \neq j$), is equal to

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2 = c_{ji}\sigma^2$$

For example, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) = c_{02}\sigma^2 = c_{20}\sigma^2$ and $\text{Cov}(\hat{\beta}_2, \hat{\beta}_3) = c_{23}\sigma^2 = c_{32}\sigma^2$. These covariances are necessary to determine the variance of the prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

or of any other linear function of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. They will also play a role in finding a confidence interval for $E(y)$ and a prediction interval for $y$.

## Estimating $\sigma^2$

You can see that the variances of the estimators of all of the $\beta$ parameters and of $\hat{y}$ will depend on the value of $\sigma^2$. Since $\sigma^2$ will rarely be known in advance, we must use the sample data to estimate its value.

### Estimator of $\sigma^2$, the Variance of $\varepsilon$ in a Multiple Regression Model

$$s^2 = \frac{\text{SSE}}{n - \text{Number of } \beta \text{ parameters in model}}$$

where

$$\text{SSE} = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$$

**Example 11.3**

Calculating SSE using Matrix Algebra

Find SSE and $s^2$ for the multiple regression model fit to the insulation compression data of Example 11.2.

**Solution** From Example 11.2 we have

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} .35 \\ 1.15 \\ -.75 \end{bmatrix} \quad \text{and} \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 10 \\ 37 \\ 27 \end{bmatrix}$$

Then,

$$\boldsymbol{Y}'\boldsymbol{Y} = [1 \quad 1 \quad 2 \quad 2 \quad 4] \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 4 \end{bmatrix} = 26$$

and $$\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} = [.35 \quad 1.15 \quad -.75] \begin{bmatrix} 10 \\ 37 \\ 27 \end{bmatrix} = 25.8$$

So

$$\text{SSE} = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} = 26 - 25.8 = .2$$

Finally,

$$s^2 = \frac{\text{SSE}}{n - \text{Number of } \beta \text{ parameters in model}} = \frac{.2}{5 - 3} = .1$$

This estimate is needed to construct a confidence interval for an individual $\beta$ parameter, to test a hypothesis concerning its value, or to construct a confidence interval for the mean compression $E(y)$ for a given compressive pressure $x$.

You will not be surprised to learn that the sampling distribution of $s^2$ is related to the chi-square distribution. In fact, Theorems 6.8 and 10.1 are special cases of Theorem 11.2 (proof omitted).

**THEOREM 11.2**

Consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

which contains $(k + 1)$ unknown $\beta$ parameters that must be estimated. If the assumptions of Section 11.2 are satisfied, then the statistic

$$\chi^2 = \frac{\text{SSE}}{\sigma^2} = \frac{[n - (k + 1)]s^2}{\sigma^2}$$

has a chi-square distribution with $\nu = [n - (k + 1)]$ degrees of freedom.

Using Theorem 11.2, we can show that $s^2$ is an unbiased estimator of $\sigma^2$:

$$E(s^2) = E\left\{\frac{\chi^2\sigma^2}{[n - (k + 1)]}\right\} = \frac{\sigma^2}{[n - (k + 1)]}E(\chi^2)$$

where $E(\chi^2) = \nu = [n - (k + 1)]$. Therefore,

$$E(s^2) = \left(\frac{\sigma^2}{[n - (k + 1)]}\right)[n - (k + 1)] = \sigma^2$$

and we conclude that $s^2$ is an unbiased estimator of $\sigma^2$.

## Inferences About the Individual $\beta$ Parameters

A $(1 - \alpha)100\%$ confidence interval for a model parameter $\beta_i$ $(i = 0, 1, 2, \ldots, k)$ can be constructed (see the Theoretical Exercises of this section) using the pivotal method and the $T$ statistic

$$T = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}$$

The quantity $s\sqrt{c_{ii}}$ is the estimated standard error of $\hat{\beta}_i$ and is obtained by replacing $\sigma$ by $s$ in the formula for the standard error. The resulting confidence interval for $\beta_i$ takes the same form as the small-sample confidence interval for a population mean given in Section 7.6.

### A $(1 - \alpha)100\%$ Confidence Interval for $\beta_i$

$$\hat{\beta}_i \pm t_{\alpha/2}(\text{Estimated standard error of } \hat{\beta}_i)$$

or

$$\hat{\beta}_i \pm t_{\alpha/2} s\sqrt{c_{ii}}$$

where $t_{\alpha/2}$ is based on the number of degrees of freedom associated with $s$.

Similarly, the test statistic for testing the null hypothesis $H_0$: $\beta_i = 0$ is

$$T = \frac{\hat{\beta}_i}{\text{Estimated standard error of } \hat{\beta}_i} = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}}$$

The test is summarized in the following box.

**Test of an Individual Parameter Coefficient in the Multiple Regression Model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\beta_i = 0$ | $H_0$: $\beta_i = 0$ |
| $H_a$: $\beta_i > 0$ (or $\beta_i < 0$) | $H_a$: $\beta_i \neq 0$ |

*Test statistic:** $T_c = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}} = \dfrac{\hat{\beta}_i}{s\sqrt{c_{ii}}}$

| | |
|---|---|
| *Rejection region:* $T_c > t_\alpha$ (or $T_c < -t_\alpha$) | *Rejection region:* $|T_c| > t_{\alpha/2}$ |
| $p$-value: $P(T > T_c)$ $[\text{or } P(T < T_c)]$ | $p$-value: $2P(T > |T_c|)$ |

where

$n$ = Number of observations
$k$ = Number of independent variables in the model
and $t_{\alpha/2}$ is based on $[n - (k + 1)]$ df

*Assumptions:* See Section 11.2 for the assumptions about the probability distribution of the random error component $\varepsilon$.

Either the confidence interval or the test can be used to determine whether a term in the model contributes information for the prediction of $y$. We illustrate with examples.

**Example 11.4**

**Using Matrix Algebra to Compute a Confidence Interval for a $\beta$ in Simple Linear Regression**

Refer to the straight-line model, Example 11.1; find the estimated standard error for the sampling distribution of $\hat{\beta}_1$, the estimator of the slope of the line $\beta_1$. Then give a 95% confidence interval for $\beta_1$ and interpret the result.

**Solution**

The $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix for the least-squares solution of Example 11.1 was

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.1 & -3 \\ -.3 & .1 \end{bmatrix}$$

Therefore, $c_{00} = 1.1$ and $c_{11} = .1$. Also, $s^2 = .367$ (from Example 10.3, p. 504). Thus, the estimated standard error for $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = s\sqrt{c_{11}} = \sqrt{.367}\left(\sqrt{.1}\right) = .192$$

A 95% confidence interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{\alpha/2} s\sqrt{c_{11}}$$

$$.7 \pm (3.182)(.192) = (.09, 1.31)$$

The $T$ value, $t_{.025}$, is based on $(n - 2) = 3$ df. Observe that this is the same confidence interval as the one obtained in Example 10.5 (p. 510). With 95% confidence, we say that compression $y$ will increase between .09 and 1.31 units for every 1-unit increase in pressure $x$. Since the slope is significantly different from 0, the implication is that pressure $x$ is a useful linear predictor of compression $y$.

*To test the null hypothesis that a parameter $\beta_i$ equals some value other than zero, say $H_0$: $\beta_1 = \beta_{i0}$, use the test statistic $T = (\hat{\beta}_i - \beta_{i0})/s_{\hat{\beta}_i}$. All other aspects of the test will be described in the box.

**Example 11.5**

**Using Matrix Algebra to Compute a Confidence Interval for a $\beta$ in Multiple Regression**

Refer to the multiple-regression model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, in Examples 11.2 and 11.3.

a. Compute the estimated standard error for $\hat{\beta}_2$.
b. Compute the value of the test statistic for testing $H_0$: $\beta_2 = 0$ against $H_a$: $\beta_2 \neq 0$. State your conclusion at $\alpha = .05$.

Solution

The $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix, obtained in Example 11.2, is

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix}$$

From $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ we see that

$$\begin{aligned} c_{00} &= 1.325 \\ c_{11} &= .325 \\ c_{22} &= .625 \end{aligned}$$

Also, from Example 11.3, $s^2 = .1$; then $s = \sqrt{.1} = .316$.

a. The estimated standard error of $\hat{\beta}_2$ is

$$s_{\hat{\beta}_2} = s\sqrt{c_{22}} = (.316)\sqrt{.625} = .25$$

b. From Example 11.2, $\hat{\beta}_2 = -.75$. The value of the test statistic for testing $H_0$: $\beta_2 = 0$ is

$$T = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{-.75}{.25} = -3.0$$

For this two-tailed $T$ test, we will reject $H_0$ if $|T| > t_{\alpha/2}$. For this example, $T$ has $[n - (k + 1)] = 5 - 3 = 2$ degrees of freedom. Therefore, for $\alpha = .05$, we will reject $H_0$ (see Table 7 of Appendix B) if $|T| > 4.303$. Since the observed value of $T = -3.0$ does not exceed 4.303 in absolute value, there is insufficient evidence to indicate that $\beta_2 \neq 0$. Therefore, the practical implication of the test conclusion is that there is no evidence to indicate temperature($x_2$) is an important predictor in the model.

## 11.5 Assessing Overall Model Adequacy

Conducting $T$ tests on each $\beta$ parameter in a model with a large number of terms is not a good way to determine whether a model is contributing information for the prediction of $y$. If we were to conduct a series of $T$ tests to determine whether the independent variables are contributing to the predictive relationship, it is very likely that we would make one or more errors in deciding which terms to retain in the model and which to exclude. For example, suppose that all the $\beta$ parameters (except $\beta_0$) are in fact equal to 0. Although the probability of concluding that any single parameter differs from 0 is only $\alpha$, the probability of rejecting *at least one* true null hypothesis in a set of $t$ tests is much higher. You can see why this is true by considering the following analogy: The probability of observing a head on a single toss of a coin is only .5, but the probability of observing *at least one* head in five tosses of a coin is .97. Thus, in multiple regression models for which a large number of independent variables are being considered, conducting a series of $T$ tests may include a large number of insignificant variables and exclude some useful ones. If we want to test the overall adequacy of a multiple regression model, we will need a **global test** (one that encompasses all the $\beta$ parameters). We would also like to find some statistical quantity that measures how well the model fits the data.

We begin with the easier problem—finding a measure of how well a multiple regression model fits a set of data. For this we use the multiple regression equivalent of $r^2$, the coefficient of determination for the straight-line model (Chapter 10). Thus, we define the **sample multiple coefficient of determination $R^2$** as

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\text{SSE}}{\text{SS}_{yy}}$$

where $\hat{y}_i$ is the predicted value of $y_i$ for the model. Just as for the simple linear model, $R^2$ is a sample statistic that represents the fraction of the sample variation of the $y$ values (measured by $\text{SS}_{yy}$) that is attributable to the regression model. Thus, $R^2 = 0$ implies a complete lack of fit of the model to the data, and $R^2 = 1$ implies a perfect fit, with the model passing through every data point. In general, the larger the value of $R^2$, the better the model fits the data.

**Definition 11.1**

The **multiple coefficient of determination, $R^2$**, is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}}$$

where $SSE = \Sigma(y_i - \hat{y}_i)^2$, $SS_{yy} = \Sigma(y_i - \bar{y})^2$ and $\hat{y}_i$ is the predicted value of $y_i$ for the multiple regression model.

**Example 11.6**

Computing $R^2$

Refer to the multiple regression model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, for the insulation compression data, Examples 11.2, 11.3, and 11.5. Find $R^2$ for the model and interpret its value.

Solution

From Example 11.3, SSE $= .2$. Now

$$\text{SS}_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2/n$$

$$= 26 - \frac{(10)^2}{5} = 6$$

Therefore,

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}}$$

$$= 1 - \frac{.2}{6} = .967$$

This value of $R^2$ implies that about 97% of the sample variation in compression ($y$) is attributable to, or explained by, one or more of the independent variables pressure ($x_1$) and temperature ($x_2$). Thus, $R^2$ is a sample statistic that tells how well the model fits the data, and thereby represents a measure of the adequacy of the overall model.

The fact that $R^2$ is a sample statistic implies that it can be used to make inferences about the statistical utility of the model for predicting $y$ values for specific settings of the independent variables. For the general linear model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, the test

$H_0$: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

$H_a$: At least one of the above $\beta$ parameters is nonzero would formally test the utility of the overall model. The test statistic used to test this null hypothesis is

$$\textit{Test statistic:}\quad F = \frac{\text{Mean square for model}}{\text{Mean square for error}} = \frac{\text{SS(Model)}/k}{\text{SSE}/[n - (k + 1)]}$$

where $n$ is the number of data points, $k$ is the number of parameters in the model (not including $\beta_0$), and SS(Model) $=$ SS$_{yy}$ $-$ SSE. Under the null hypothesis, this $F$ test statistic has an $F$ probability distribution with $k$ df in the numerator and $[n - (k + 1)]$ df in the denominator. The upper-tail values of the $F$ distribution are given in Tables 9–12 of Appendix B.

It can be shown (proof omitted) that an equivalent form of the test statistic for testing the overall adequacy of the model is

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]}$$

Therefore, the $F$ test statistic becomes large as the coefficient of determination $R^2$ becomes large. To determine how large $F$ must be before we can conclude at a given value of $\alpha$ that the model is useful for predicting $y$, we set up the rejection region as follows:

$$\textit{Rejection region:}\quad F > F_\alpha$$

where

$$\nu_1 = k \text{ df}, \qquad \nu_2 = n - (k + 1) \text{ df}$$

This test procedure is summarized in the box.

### The Analysis of Variance (Global) *F* Test: Testing the Overall Adequacy of the Model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

$H_0$: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_a$: At least one of the parameters, $\beta_1, \beta_2, \ldots, \beta_k$, differs from 0

$$\textit{Test statistic:}\quad F_c = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} = \frac{\text{Mean square for model}}{\text{Mean square for error}} = \frac{\text{SS(Model)}/k}{\text{SSE}/[n - (k + 1)]}$$

*Rejection region:* $F_c > F_\alpha$ where $\nu_1 = k$ and $\nu_2 = [n - (k + 1)]$

$p$-value: $P(F > F_c)$

*Assumptions:* See Section 11.2 for the assumptions about the probability distribution of the random error component $\varepsilon$.

**Example 11.7**

Conducting the Global F Test

Refer to Example 11.6 and the multiple regression model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Test to determine whether the overall model contributes information for the prediction of $y$. Use $\alpha = .10$.

Solution

We want to test $H_0$: $\beta_1 = \beta_2 = 0$. For this example, $n = 5$, $k = 2$ and $n - (k + 1) = 2$. At $\alpha = .10$, we will reject $H_0$: $\beta_1 = \beta_2 = 0$ if $F > F_{.10}$ where $\nu_1 = 2$ and $\nu_2 = 2$. The critical $F$ value from the Appendix is $F_{.10} = 9.0$. Therefore, we will reject $H_0$ if calculated $F > 9.0$ (see Figure 11.1).

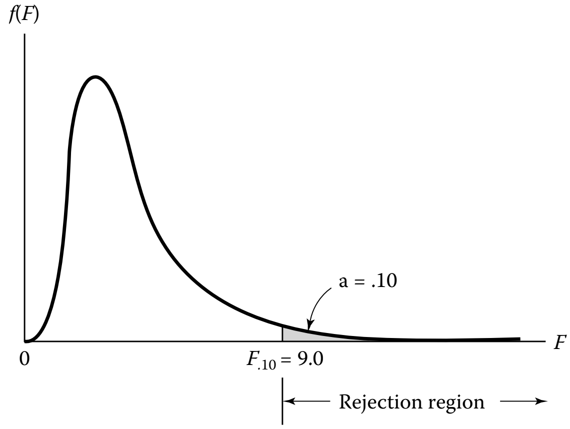



**FIGURE 11.1**
Rejection region for the $F$ statistic with $\nu_1 = 2, \nu_2 = 2$ and $\alpha = .10$

From Example 11.6, we have $R^2 = .967$. Thus, the test statistic is

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} = \frac{.967/2}{(1 - .967)/2} = 29.0$$

Since this value exceeds the tabulated value of 9.0, we conclude that at least one of the model coefficients $\beta_1$ and $\beta_2$, is nonzero. Therefore, this $F$ test indicates that the overall multiple regression model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is useful for predicting compression $y$.

To summarize the discussion so far, the value of $R^2$ is an indicator of how well the prediction equation fits the data. More importantly, it can be used (in the $F$ statistic) to determine whether the data provide sufficient evidence to indicate that the overall model contributes information for the prediction of $y$. However, intuitive evaluations of the contribution of the model based on the computed value of $R^2$ must be examined with care. The value of $R^2$ will increase as more and more variables are added to the model. Consequently, you could force $R^2$ to take a value very close to 1 even though the model contributes no information for the prediction of $y$. In fact, $R^2$ will equal 1 when the number of terms in the model equals the number of data points.

As an alternative to using $R^2$ as a measure of model adequacy, the **adjusted multiple coefficient of determination**, denoted $R_a^2$, is often reported. The formula for $R_a^2$ is shown in the box.

### The Adjusted Multiple Coefficient of Determination

The **adjusted multiple coefficient of determination** is given by

$$R_a^2 = 1 - \frac{(n - 1)}{n - (k + 1)}\left(\frac{\text{SSE}}{\text{SS}_{yy}}\right) = 1 - \frac{n - 1}{n - (k + 1)}(1 - R^2)$$

Unlike $R^2$, $R_a^2$ takes into account ("adjusts for") both the sample size $n$ and the number of $\beta$ parameters in the model. $R_a^2$ will always be smaller than $R^2$, and more importantly, cannot be "forced" to 1 by simply adding more and more independent variables to the model. Consequently, some analysts prefer the more conservative $R_a^2$ when choosing a measure of model adequacy.

From the insulation compression example above,

$$R_a^2 = 1 - \frac{(n-1)}{n-(k+1)}(1 - R^2)$$
$$= 1 - \frac{4}{2}(1 - .967)$$
$$= .934$$

Conservatively, we say that about 93% of the sample variation in compression $y$ can be explained by the model with $x$ and $x^2$. Remember, however, that both $R^2$ and $R_a^2$ are only sample statistics, and **you should not rely solely on their values to tell you whether the model is useful for predicting $y$**. Use the $F$ test (with supporting measure of reliability $\alpha$) to make inferences about the overall adequacy of the multiple regression model.

We conclude this section with some guidelines on how to check the overall utility of a multiple regression model.

### Recommendation for Checking the Utility of a Multiple Regression Model

1. First, conduct a test of overall model adequacy using the $F$ test, that is, test

   $$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

   If the model is deemed adequate (that is, if you reject $H_0$), then proceed to step 2. Otherwise, you should hypothesize and fit another model. The new model may include more independent variables or higher-order terms.
2. Conduct $T$ tests on those $\beta$ parameters in which you are particularly interested (that is, the "most important" $\beta$'s). These usually involve only the $\beta$'s associated with higher-order terms ($x_2$, $x_1$, $x_2$, etc.). However, it is a safe practice to limit the number of $\beta$'s that are tested. Conducting a series of $T$ tests leads to a high overall Type I error rate $\alpha$.
3. Examine the values of $R_a^2$ and $2s$ to evaluate how well, numerically, the model fits the data.

## 11.6 A Confidence Interval for $E(y)$ and a Prediction Interval for a Future Value of $y$

### Confidence Interval for $E(y)$

Suppose we were to postulate that the mean value of the productivity $y$ of a consulting engineering firm is related to the size of the company $x$ and that the relationship could be modeled by the expression

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

A graph of $E(y)$ might appear as shown in Figure 11.2.

We might have several reasons for collecting data on the productivity and size of a set of $n$ engineering firms and finding the least-squares prediction equation,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

For example, we might want to estimate the mean productivity for a company of a given size (say, $x = 2$). That is, we might want to estimate

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$= \beta_0 + 2\beta_1 + 4\beta_2 \qquad \text{where } x = 2$$

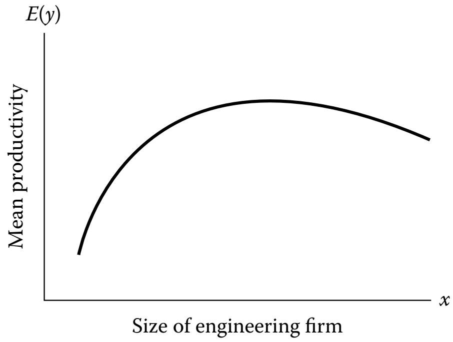



**FIGURE 11.2**
Graph of mean productivity $E(y)$

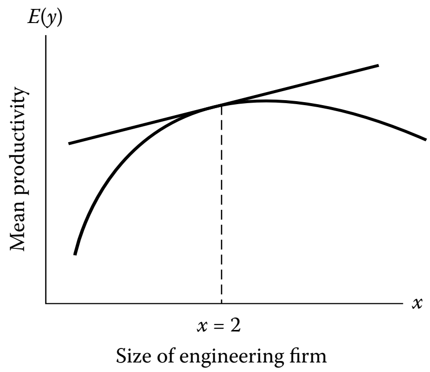



**FIGURE 11.3**
Marginal productivity

Or we might want to estimate the marginal increase in productivity, the slope of a tangent to the curve, when $x = 2$ (see Figure 11.3). The marginal productivity for $y$ when $x = 2$ is the rate of change of $E(y)$ with respect to $x$, evaluated at $x = 2$. For the quadratic model, the marginal productivity for a value of $x$, denoted by the symbol $dE(y)/dx$, is*

$$\frac{dE(y)}{dx} = \beta_1 + 2\beta_2 x$$

Therefore, the marginal productivity at $x = 2$ is

$$\frac{dE(y)}{dx} = \beta_1 + 2\beta_2(2) = \beta_1 + 4\beta_2$$

Note that for $x = 2$, both $E(y)$ and the marginal productivity are *linear* functions of the unknown parameters $\beta_0, \beta_1, \beta_2$ in the model. A problem we pose in this section is that of finding confidence intervals for linear functions of $\beta$ parameters or testing hypotheses concerning their values. We can find these confidence intervals or values of the appropriate test statistics from knowledge of $(\mathbf{X}'\mathbf{X})^{-1}$.

We will suppose that we have a model,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

and that we are interested in making an inference about a linear function of the $\beta$ parameters, say,

$$a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$

where $a_0, a_1, \ldots, a_k$ are known constants. Further, we will use the corresponding linear function of least-squares estimates,

$$\ell = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k$$

as our best estimate of $a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$.

*Note that the marginal productivity for $y$ given $x$ is the first derivative of $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ with respect to $x$.

We recall from Section 11.4 that the least-squares estimators, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, are normally distributed with

$$E(\hat{\beta}_i) = \beta_i$$
$$V(\hat{\beta}_i) = c_{ii}\sigma^2 \qquad (i = 0, 1, 2, \ldots, k)$$

and covariances

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_i) = c_{ij}\sigma^2 \qquad (i \neq j)$$

It then follows by Theorem 6.9 that

$$\ell = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k$$

is normally distributed with mean, variance, and standard deviation as given by Theorem 11.3.

**THEOREM 11.3**

*Properties of the Sampling Distribution of*

$$\ell = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k$$

The sampling distribution of $\ell$ is normal with

$$E(\ell) = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$
$$V(\ell) = [\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}]\sigma^2$$
$$\sigma_\ell = \sqrt{V(\ell)} = \sigma\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}$$

where $\sigma$ is the standard deviation of $\varepsilon$, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the inverse matrix obtained in fitting the least-squares model, and

$$\boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Theorem 11.3 indicates that $\ell$ is an unbiased estimator of

$$E(\ell) = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$

and that its sampling distribution would appear as shown in Figure 11.4.

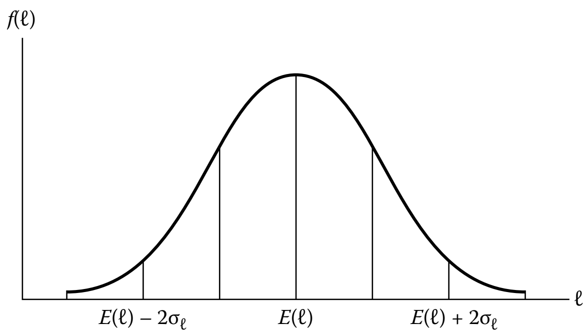



**FIGURE 11.4**
Sampling distribution for $\ell$

Therefore, a $(1 - \alpha)100\%$ confidence interval for $E(\ell)$ is as shown in the following box.

### A $(1 - \alpha)100\%$ Confidence Interval for $E(\ell)$

$$\ell \pm (t_{\alpha/2})s\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}$$

where

$$E(\ell) = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$
$$\ell = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k$$
$$\boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

$s$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are obtained from the least-squares procedure, and $t_{\alpha/2}$ is based on the number of degrees of freedom associated with $s$.

The linear function of the $\beta$ parameters that is most often the focus of our attention is

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

That is, we want to find a confidence interval for $E(y)$ for specific values of $x_1, x_2, \ldots, x_k$. For this special case, $\ell = \hat{y}$ and the $\boldsymbol{a}$ matrix is

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

where the symbols $x_1, x_2, \ldots, x_k$ in the $\boldsymbol{a}$ matrix indicate the specific numerical values assumed by these variables. Thus, the procedure for forming a confidence interval for $E(y)$ is as shown in the box.

### A $(1 - \alpha)100\%$ Confidence Interval for $E(y)$

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}$$

where

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \qquad \boldsymbol{a} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

$s$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are obtained from the least-squares analysis, and $t_{\alpha/2}$ is based on the number of degrees of freedom associated with $s$, namely, $[n - (k + 1)]$.

**Example 11.8**

Using Matrices to Find a Confidence Interval for E(y)

Refer to the data of Examples 11.2 and 11.5 and the multiple regression model for insulation compression $y$, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Find a 95% confidence interval for the mean compression $E(y)$ when the pressure is $x_1 = 5.$ (i.e., 50 psi) and the temperature is $x_2 = 3$ (i.e., 30°C).

Solution

The confidence interval for $E(y)$ for a given value of $x$ is

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}$$

Consequently, we need to find and substitute the values of $\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}$, $t_{\alpha/2}$, $s$, and $\hat{y}$ into this formula. Since we want to estimate

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= \beta_0 + \beta_1(5) + \beta_2(3) \quad \text{when } x_1 = 5 \text{ and } x_2 = 3 \\ &= \beta_0 + 5\beta_1 + 3\beta_2 \end{aligned}$$

it follows that the coefficients of $\beta_0$, $\beta_1$, and $\beta_2$ are $a_0 = 1$, $a_1 = 5$, and $a_2 = 3$ and thus,

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$$

From Examples 11.2 and 11.5, we have $\hat{y} = .35 + 1.15x_1 - .75x_2$, $s^2 = .1$, $s = .316$ and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix}$$

Then,

$$\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} = [1 \quad 5 \quad 3] \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$$

We first calculate

$$\begin{aligned} \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1} &= [1 \quad 5 \quad 3] \begin{bmatrix} 1.325 & -.075 & -.375 \\ -.075 & .325 & -.375 \\ -.375 & -.375 & .625 \end{bmatrix} \\ &= [-.175 \quad .425 \quad -.375] \end{aligned}$$

Then,

$$\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} = [-.175 \quad .425 \quad -.375] \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix} = .825$$

The $T$ value, $t_{.025}$, based on 2 df is 4.303. So, a 95% confidence interval for the mean compression of the insulation material when subjected to a pressure of $x_1 = 5$ and a temperature of $x_2 = 3$ is:

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}$$

Since $\hat{y} = .35 + 1.15x_1 - .75x_2 = .35 + 1.15(5) - .75(3) = 3.85$, the 95% confidence interval for $E(y)$ is

$$3.85 \pm (4.303)(.316)\sqrt{.825} = 3.85 \pm 1.24 = (2.61, 5.09)$$

That is, we are 95% confident that the mean compression, $E(y)$, for all materials subjected to 50 psi $(x_1 = 5)$ of pressure and 30°C $(x_2 = 3)$ of temperature is between 2.61 and 5.09 inches.

### Prediction Interval for a Future Value of *y*

Rather than estimate the mean of $y$, $E(y)$, you may want to predict a new value of $y$, yet unobserved, for specific values of $x_1, x_2, \ldots, x_k$. The difference between these two inferential problems (when each would be pertinent) was explained in Chapter 10, but we will give another example to make certain that the distinction is clear at this point.

Suppose you are the manager of a manufacturing plant and that $y$, the daily profit, is a function of various process variables $x_1, x_2, \ldots, x_k$. Suppose you want to know how much money you would make *in the long run* if the $x$'s are set at specific values. For this case, you would be interested in finding a confidence interval for the mean profit per day, $E(y)$. In contrast, suppose you planned to operate the plant for only one more day! Then you would be interested in predicting the value of $y$, the profit associated with tomorrow's production.

We have indicated that the error of prediction is always larger than the error of estimating $E(y)$. You can see this by comparing the formula for the prediction interval (shown in the next box) with the formula for the confidence interval for $E(y)$ that was given earlier.

A $(1-\alpha)100\%$ Prediction Interval for $y$

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$$

where

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

$s$ and $(\mathbf{X}'\mathbf{X})^{-1}$ are obtained from the least-squares analysis,

$$\mathbf{a} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

contains the numerical values of $x_1, x_2, \ldots, x_k$, and $t_{\alpha/2}$ is based on the number of degrees of freedom associated with $s$, namely, $[n - (k+1)]$.

**Example 11.9**

Using Matrices to Find a Prediction Interval for $y$

Refer to the insulation compression example (Example 11.8). Find a 95% prediction interval for the compression of a particular piece of insulation when it is to be subjected to a pressure of 50 pounds per square inch $(x_1 = 5)$ and a temperature of 30°C $(x_2 = 3)$.

Solution

The 95% prediction interval for the compression of this particular piece of insulation is

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$$

From Example 11.8, when $x_1 = 5$ and $x_2 = 3$, $\hat{y} = 3.85$, $s = .316$, $t_{.025} = 4.303$, and $\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} = .825$. Then the 95% prediction interval for $y$ is

$$3.85 \pm (4.303)(.316)\sqrt{1 + .825}$$
$$3.85 \pm 1.84 = (2.01, 5.69).$$

Thus, we can be 95% confident that the compression $y$ for a particular piece of insulation subjected to a pressure of 50 psi $(x_1 = 5)$ and a temperature of 30°C $(x_2 = 3)$ will fall between 2.01 and 5.69 inches.

## Applied Exercises

11.1 *Extending the life of an aluminum smelter pot.* Refer to *The American Ceramic Society Bulletin* (Feb. 2005) study of aluminum smelter pots, Exercise 10.9 (p. 497). Recall that the life length of a smelter pot depends on the porosity of the brick lining. The apparent porosity of each of six brick specimens, as well as the mean pore diameter of each brick, is reproduced in the accompanying table. Use matrix algebra and the method of least–squares to fit the straight-line model $E(y) = \beta_0 + \beta_1 x$ to the six data points.

**SMELTPOT**

| Brick | Apparent Porosity (%), $y$ | Mean Pore Diameter (micrometers), $x$ |
|---|---|---|
| A | 18.8 | 12.0 |
| B | 18.3 | 9.7 |
| C | 16.3 | 7.3 |
| D | 6.9 | 5.3 |
| E | 17.1 | 10.9 |
| F | 20.4 | 16.8 |

*Source:* Bonadia, P., et al. "Aluminosilicate refractories for aluminum cell linings." *The American Ceramic Society Bulletin*, Vol. 84, No. 2, Feb. 2005 (Table II).

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{Y}$.
c. Verify that $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.50384 & -.12940 \\ -.12940 & .012523 \end{bmatrix}$
d. Find the $\boldsymbol{\beta}$ matrix, then give the least-squares prediction equation.
e. Find SSE and $s^2$.
f. Find the standard error of $\hat{\beta}_1$.
g. Give and interpret a 90% confidence interval for $\beta_1$.
h. Find and interpret the value of $R^2$.
i. Give and interpret a 90% prediction interval for $y$ when $x = 10$.

11.2 *Drug controlled-release rate study.* Refer to the *Drug Development and Industrial Pharmacy* (Vol. 28, 2002.) investigation of the rate at which a drug is released in a controlled-release dosage, Exercise 10.26 (p. 506). Recall that the ratio of surface area to volume and the diffusional release rate (percentage of drug released divided by the square root of time) was measured for each of six drug tablets. The experimental data are reproduced in the table. Use matrix algebra and the method of least–squares to fit the straight-line model $E(y) = \beta_0 + \beta_1 x$ to the six data points.

**DOWDRUG**

| Drug Release Rate, $y$ (% released/$\sqrt{\text{time}}$) | Surface Area to Volume, $x$ (mm$^2$/mm$^3$) |
|---|---|
| 60 | 1.50 |
| 48 | 1.05 |
| 39 | .90 |
| 33 | .75 |
| 30 | .60 |
| 29 | .65 |

*Source:* Reynolds, T., Mitchell, S., and Balwinski, K. "Investigation of the effect of tablet surface area / volume on drug release from Hydroxypropylmethylcellulose controlled-release matrix tablets." *Drug Development and Industrial Pharmacy*, Vol. 28, No. 4, 2002 (Figure 3).

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{Y}$.
c. Verify that $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.64772 & -1.63052 \\ -1.63052 & 1.79506 \end{bmatrix}$
d. Find the $\boldsymbol{\beta}$ matrix, then give the least-squares prediction equation.
e. Find SSE and $s^2$.
f. Find the standard error of $\hat{\beta}_1$.
g. Test $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 \neq 0$, using $\alpha = .05$.
h. Find and interpret the value of $R^2$.
i. Give and interpret a 95% confidence interval for $E(y)$ when $x = 1$.

11.3 *Characterizing bone with fractal geometry.* Refer to the *Medical Engineering & Physics* (May 2013) study involving the use of fractal geometry to characterize human cortical bone, Exercise 10.13 (p. 498). Experimental data on fractal dimension, $x$ (a measure of the variation in the volume of cortical bone tissue) and Young's Modulus, $y$ (a measure of bone tissue's stiffness) for each in a sample of 10 human ribs are reproduced in the table below. Consider the linear model, $E(y) = \beta_0 + \beta_1 x$.

**CORTBONE**

| Young's Modulus (GPa) | Fractal Dimension |
|---|---|
| 18.3 | 2.48 |
| 11.6 | 2.48 |
| 32.2 | 2.39 |
| 30.9 | 2.44 |
| 12.5 | 2.50 |
| 9.1 | 2.58 |
| 11.8 | 2.59 |
| 11.0 | 2.59 |
| 19.7 | 2.51 |
| 12.0 | 2.49 |

*Source:* Sanchez-Molina, D., et al. "Fractal dimension and mechanical properties of human cortical bone", *Medical Engineering & Physics,* Vol. 35, No. 5, May 2013 (Table 1).

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$.
c. Find the least squares estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. (See Theorem A.1 in Appendix A for information on finding $(\boldsymbol{X'X})^{-1}$.
d. Find SSE and $s^2$.
e. Conduct the test, $H_0$: $\beta_1 = 0$ vs. $H_a$: $\beta_1 < 0$ at $\alpha = .01$.
f. Find and interpret $R^2$.
g. Find and interpret a 95% prediction interval for $y$ when $x = 2.50$.

11.4 *Processed straw as thermal insulation.* An article published in *Engineering Structures and Technologies* (Sep. 2012) presented the results of a study on the use of processed straw as thermal insulation for homes. Chopped straw specimens were prepared and tested for thermal conductivity at a temperature of 10º Celcius. Two variables were measured for each of 25 straw specimens: $y$ = thermal conductivity (watts per meter-Kelvin) and $x$ = density (kilograms per cubic meter). The data are provided in the accompanying table. Consider the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$.

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$.

**STRAW**

| Specimen | Thermal Conductivity (y) | Density (x) |
|---|---|---|
| 1 | 0.052 | 49 |
| 2 | 0.045 | 50 |
| 3 | 0.055 | 51 |
| 4 | 0.042 | 56 |
| 5 | 0.048 | 57 |
| 6 | 0.049 | 62 |
| 7 | 0.046 | 64 |
| 8 | 0.047 | 65 |
| 9 | 0.051 | 66 |
| 10 | 0.047 | 68 |
| 11 | 0.049 | 78 |
| 12 | 0.048 | 79 |
| 13 | 0.048 | 82 |
| 14 | 0.052 | 83 |
| 15 | 0.051 | 84 |
| 16 | 0.053 | 98 |
| 17 | 0.054 | 100 |
| 18 | 0.055 | 100 |
| 19 | 0.057 | 101 |
| 20 | 0.055 | 103 |
| 21 | 0.074 | 115 |
| 22 | 0.075 | 116 |
| 23 | 0.077 | 118 |
| 24 | 0.076 | 119 |
| 25 | 0.074 | 120 |

c. Use the technique outlined in Appendix A.4 to find $(\boldsymbol{X'X})^{-1}$. (Be sure to carry out your calculations to six significant digits.)
d. Find the $\hat{\boldsymbol{\beta}}$ matrix and the least-squares prediction equation.
e. Find SSE and $s^2$.
f. Test $H_0$: $\beta_2 = 0$ vs. $H_a$: $\beta_2 \neq 0$ using $\alpha = .05$.
g. Conduct a test of overall model adequacy using $\alpha = .05$.
h. Find and interpret $R_a^2$.
i. Find and interpret a 95% confidence interval for $E(y)$ when $x = 75$.

11.5 *Accelerating hash function computations.* A hash function is an algorithm that maps data of variable length to data of a fixed length. Hash functions are used by engineers responsible for data authentication and data encryption on a large scale. In the *Journal of Cryptographic Engineering* (Nov. 2012), a new algorithm for accelerating hash function computations was proposed. The performance $y$ of the new algorithm (measured as number of CPU cycles per byte), was modeled as a function of message length $x$ (measured in bytes). Data for six different messages submitted for data encryption are listed in the table. Consider the model, $E(y) = \beta_0 + \beta_1 \ln(x)$.

**HASH**

| Message | Performance | Message Length | LN(Length) |
|---|---|---|---|
| 1 | 19.29 | 256 | 5.5452 |
| 2 | 17.09 | 512 | 6.2383 |
| 3 | 15.98 | 1024 | 6.9315 |
| 4 | 15.17 | 4096 | 8.3178 |
| 5 | 14.96 | 20480 | 9.9272 |
| 6 | 14.94 | 102400 | 11.5366 |

*Source:* Gueron, S.. & Krasnov, V. "Parallelizing message schedules to accelerate the computations of hash functions", *Journal of Cryptographic Engineering,* Vol. 2, No. 4, November 2012 (Figure 3).

a. Use matrix algebra to find estimates of $\beta_0$ and $\beta_1$.
b. Conduct the test of overall model adequacy. Use $\alpha = .10$.
c. Find and interpret a 90% confidence interval for the performance level of the new algorithm when applied to a message of length 5,000 bytes.

11.6 *Bubble behavior in subcooled flow boiling.* In industry cooling applications (e.g., cooling of nuclear reactors), a process called subcooled flow boiling is often employed. Subcooled flow boiling is susceptible to small bubbles which occur near the heated surface. The characteristics of these bubbles were investigated in *Heat Transfer Engineering* (Vol. 34, 2013). A series of experiments was conducted to measure two important bubble behaviors—bubble diameter (millimeters) and bubble density (liters per meters squared). The mass flux (kilograms per meters squared per second) and heat flux (megawatts per meters squared) were varied for each experiment. The data obtained at a set pressure are listed in the next table.

a. Consider the multiple regression model, $E(y_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $y_1$ = bubble diameter, $x_1$ = mass flux, and $x_2$ = heat flux. Use matrices to fit the model to the data and test overall adequacy.
b. Consider the multiple regression model, $E(y_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $y_2$ = bubble density, $x_1$ = mass flux, and $x_2$ = heat flux. Use matrices to fit the model to the data and test overall model adequacy.
c. Which of the two dependent variables, diameter ($y_1$) or density ($y_2$), is better predicted by mass flux ($x_1$) and heat flux ($x_2$)? Explain.

**BUBBLE2**

| Bubble | Mass Flux | Heat Flux | Diameter | Density |
|---|---|---|---|---|
| 1 | 406 | 0.15 | 0.64 | 13103 |
| 2 | 406 | 0.29 | 1.02 | 29117 |
| 3 | 406 | 0.37 | 1.15 | 123021 |
| 4 | 406 | 0.62 | 1.26 | 165969 |
| 5 | 406 | 0.86 | 0.91 | 254777 |
| 6 | 406 | 1.00 | 0.68 | 347953 |
| 7 | 811 | 0.15 | 0.58 | 7279 |
| 8 | 811 | 0.29 | 0.98 | 22566 |
| 9 | 811 | 0.37 | 1.02 | 106278 |
| 10 | 811 | 0.62 | 1.17 | 145587 |
| 11 | 811 | 0.86 | 0.86 | 224204 |
| 12 | 811 | 1.00 | 0.59 | 321019 |
| 13 | 1217 | 0.15 | 0.49 | 5096 |
| 14 | 1217 | 0.29 | 0.80 | 18926 |
| 15 | 1217 | 0.37 | 0.93 | 90992 |
| 16 | 1217 | 0.62 | 1.06 | 112102 |
| 17 | 1217 | 0.86 | 0.81 | 192903 |
| 18 | 1217 | 1.00 | 0.43 | 232211 |

*Source:* Puli, U., Rajvanshi, A.K. & Das, S.K. "Investigation of Bubble Behavior in Subcooled Flow Boiling of Water in a Horizontal Annulus Using High-Speed Flow Visualization", *Heat Transfer Engineering,* Vol. 34, No. 10, 2013 (Table 8).

11.7 *Selecting the optimal chemical catalyst.* A study was conducted at Union Carbide to identify the optimal catalyst preparation conditions in the conversion of monoethanolamine (MEA) to ethylenediamine (EDA), a substance used commercially in soaps. For each of 10 preselected catalysts, the following experimental variables were measured:

$y$ = Rate of conversion of MEA to EDA

$x_1$ = Atom ratio of metal used in the experiment

$x_2$ = Reduction temperature

$$x_3 = \begin{cases} 1 & \text{if high acidity support used} \\ 0 & \text{if low acidity support used} \end{cases}$$

The data for the $n = 10$ experiments were used to fit the multiple regression model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. The results are summarized as follows:

| | | | |
|---|---|---|---|
| $\hat{y} = 40.2 - .808x_1 - 6.38x_2 - 4.45x_3$ | | | $R^2 = .899$ |
| $s_{\hat{\beta}_1} = .231$ | $s_{\hat{\beta}_2} = 1.93$ | $s_{\hat{\beta}_2} = .99$ | |

*Source:* Hansen, J. L., and Best, D. C. "How to Pick a Winner." Paper presented at Joint Statistical Meetings, American Statistical Association and Biometric Society, Aug. 1986, Chicago, Illinois.

a. Is there sufficient evidence to indicate that the model is useful for predicting rate of conversion $y$? Test using $\alpha = .01$.
b. Conduct a test to determine whether atom ratio $x_1$ is a useful predictor of rate of conversion $y$. Use $\alpha = .05$.
c. Construct a 95% confidence interval for $\beta_2$. Interpret the interval.

11.8 *Zoning of vacant land.* "Zoning" is defined as the distribution of vacant land to residential and nonresidential uses via policy set by local governments. Although the negative effects of zoning have been studied (e.g., distorting urban property markets, creating barriers to residential mobility, and impeding economic and social integration), little empirical evidence exists identifying the factors that encourage restrictive zoning practices. A study, reported in the *Journal of Urban Economics* (Vol. 21, 1987), developed a series of multiple regression models that hypothesize several determinants of zoning. One of the models studied took the form

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$

where

$y$ = Percentage of vacant land zoned for residential use
$x_1$ = Proportion of existing land in nonresidential use
$x_2$ = Proportion of total tax base derived from nonresidential property

The model was fit to data collected for $n = 185$ municipal communities in northeastern New Jersey, with the following results:

| Independent Variable | Parameter Estimate | Standard Error of Estimate | $T$ value | $p$-value |
|---|---|---|---|---|
| Intercept | 92.26 | 3.07 | 30.05 | $p < .01$ |
| $x_1$ | −96.35 | 46.59 | −2.07 | $p < .05$ |
| $x_1^2$ | 166.80 | 120.88 | 1.38 | $p > .10$ |
| $x_2$ | −75.51 | 13.35 | −5.66 | $p < .01$ |
| Adjusted $R^2 = .25$ $\quad F = 21.86 \quad (p < .01)$ | | | | |

*Source*: Rolleston, B. S. "Determinants of restrictive suburban zoning: An empirical analysis." *Journal of Urban Economics*, Vol. 21, 1987, p. 15, Table 4.

a. Construct a 95% confidence interval for $\beta_3$. Interpret the result.
b. Test the hypothesis that a curvilinear relationship exists between percentage ($y$) of land zoned for residential use and proportion ($x_1$) of existing land in nonresidential use.
c. Is the overall model statistically useful for predicting $y$?
d. Interpret the adjusted $R^2$ value.

11.9 *Usability professionals salary survey.* The Usability Professionals' Association (UPA) supports people who research, design, and evaluate the user experience of products and services (e.g., a design engineer who is evaluating the user interface of new computer technology). Recently, the UPA conducted a salary survey of its members (*UPA Salary Survey,* August 18, 2009). One of the report's authors, Jeff Sauro, investigated how much having a Ph.D. affects salaries in this profession and discussed his analysis on the blog, *www.measuringusability.com*. Sauro fit a first-order multiple regression model for salary ($y$, in dollars) as a function of years of experience ($x_1$), PhD status ($x_2 = 1$ if PhD, 0 if not), and manager status ($x_3 = 1$ if manager, 0 if not). The following prediction equation was obtained:

$$\hat{y} = 52{,}484 + 2{,}941 x_1 + 16{,}880 x_2 + 11{,}108 x_3$$

a. Predict the salary of a UPA member with 10 years of experience, who does not have a PhD, but is a manager.
b. Predict the salary of a UPA member with 10 years of experience, who does have a PhD, but is not a manager.
c. The following coefficient was reported: $R_a^2 = .32$. Give a practical interpretation of this value.
d. A 95% confidence interval for $\beta_1$ was reported as (2700, 3200). Give a practical interpretation of this result.
e. A 95% confidence interval for $\beta_2$ was reported as (11,500, 22,300). Give a practical interpretation of this result.
f. A 95% confidence interval for $\beta_3$ was reported as (7,600, 14,600). Give a practical interpretation of this result.

11.10 *CPU of a supercomputer.* Because the coefficient of determination $R^2$ always increases when a new independent variable is added to the model, it is tempting to include many variables in a model to force $R^2$ to be near 1. However, doing so reduces the degrees of freedom available for estimating $\sigma^2$, which adversely affects our ability to make reliable inferences. As an example, suppose you want to predict the CPU time of a supercomputer job using 18 predictor variables (such as size of job, time of submission, and estimated lines of print). You fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{18} x_{18} + \varepsilon$$

where $y$ = CPU time and $x_1, x_2, \ldots, x_{18}$ are the predictor variables. Using the relevant information on $n = 20$ jobs to fit the model, you obtain $R^2 = .95$. Test to determine whether this value of $R^2$ is large enough for you to infer that this model is useful—i.e., that at least one term in the model is important for predicting CPU time. Use $\alpha = .05$.

## Theoretical Exercises

11.11 If the assumptions of Section 11.2 are satisfied, it can be shown that $s^2$ is independent of $\hat{\beta}_i$, the least-squares estimator of $\beta_i$. Use this fact, along with Theorems 11.1 and 11.2, to show that

$$T = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}$$

has a Student's $T$ distribution with $[n - (k + 1)]$ degrees of freedom.

11.12 Use the $T$ statistic given in Exercise 11.9, in conjunction with the pivotal method, to derive the formula for a $(1 - \alpha)100\%$ confidence interval for $\beta_i$.

11.13 Since $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, are independent of $s^2$, it follows that

$$\ell = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k$$

is independent of $s^2$. Use this fact and Theorems 11.2 and 11.3 to show that

$$T = \frac{\ell - E(\ell)}{s\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

has a Student's $T$ distribution with $[n - (k + 1)]$ degrees of freedom.

11.14 Let $\ell = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$. Use the $T$ statistic of Exercise 11.13, in conjunction with the pivotal method, to derive the formula for a $(1 - \alpha)100\%$ confidence interval for $E(y)$.

11.15 Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$ be the least-squares prediction equation, and let $y$ be some observation to be obtained in the future.

a. Explain why $(\hat{y} - y)$ is normally distributed.

b. Show that

$$E(\hat{y} - y) = 0$$

and

$$V(\hat{y} - y) = [1 + \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}]\sigma^2$$

11.16 Show that

$$T = \frac{\hat{y} - y}{s\sqrt{1 + \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

has a Student's $T$ distribution with $[n - (k + 1)]$ degrees of freedom.

11.17 Use the result of Optional Exercise 11.16 and the pivotal method to derive the formula for a $(1 - \alpha)100\%$ prediction interval for $y$.

## 11.7 A First-Order Model with Quantitative Predictors

Now that we have presented the basic concepts and formulas for multiple regression, we will demonstrate a complete analysis of several common and practical multiple regression models. In this section, we consider a model that includes only terms for *quantitative* independent variables, called a **first-order model**. Note that the first-order model does not include any higher-order terms (such as $x_1^2$). The term *first-order* is derived from the fact that each $x$ in the model is raised to the first power.

### A First-Order Model in Five Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where $x_1, x_2, \ldots, x_5$ are all quantitative variables that *are not* functions of other independent variables.

*Note:* $\beta_i$ represents the slope of the line relating $y$ to $x_i$ when all the other $x$'s are held fixed.

Recall that in the straight-line model (Chapter 10)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

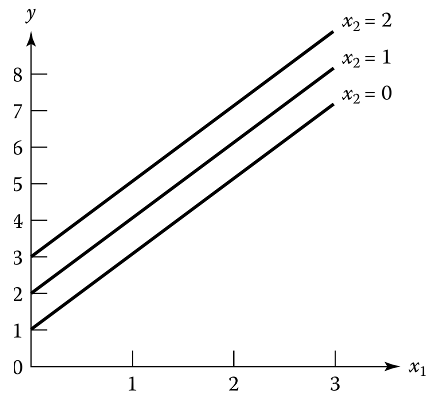
$\beta_0$ represents the $y$-intercept of the line and $\beta_1$ represents the slope of the line. From our discussion in Chapter 10, $\beta_1$ has a practical interpretation—it represents the mean change in $y$ for every 1-unit increase in $x$. When the independent variables are quantitative, the $\beta$ parameters in the first-order model specified in the box have similar interpretations. The difference is that when we interpret the $\beta$ that multiplies one of the variables (e.g., $x_1$), we must be certain to hold the values of the remaining independent variables (e.g., $x_2, x_3$) fixed.

To see this, suppose that the mean $E(y)$ of a response $y$ is related to two quantitative independent variables, $x_1$ and $x_2$, by the first-order model

$$E(y) = 1 + 2x_1 + x_2$$

In other words, $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$.

**FIGURE 11.5**
Graphs of $E(y) = 1 + 2x_1 + x_2$ for $x_2 = 0, 1, 2$




Now, when $x_2 = 0$, the relationship between $E(y)$ and $x_1$ is given by

$$E(y) = 1 + 2x_1 + (0) = 1 + 2x_1$$

A graph of this relationship (a straight line) is shown in Figure 11.5. Similar graphs of the relationship between $E(y)$ and $x_1$ for $x_2 = 1$,

$$E(y) = 1 + 2x_1 + (1) = 2 + 2x_1$$

and for $x_2 = 2$,

$$E(y) = 1 + 2x_1 + (2) = 3 + 2x_1$$

also are shown in Figure 11.5. Note that the slopes of the three lines are all equal to $\beta_1 = 2$, the coefficient that multiplies $x_1$.

Figure 11.5 exhibits a characteristic of all first-order models: If you graph $E(y)$ versus any one variable—say, $x_1$—for fixed values of the other variables, the result will always be a *straight line* with slope equal to $\beta_1$. If you repeat the process for other values of the fixed independent variables, you will obtain a set of *parallel* straight lines. This indicates that the effect of the independent variable $x_i$ on $E(y)$ is independent of all the other independent variables in the model, and this effect is measured by the slope $\beta_i$ (as stated in the box).

The first-order model is the most basic multiple regression model encountered in practice.

## Example 11.10

A First-order Model for Total Production Time

Consider a production process in which one or more workers are engaged in a variety of tasks. For such a process, the total time spent in production varies as a function of the size of the work pool and the level of output of the various activities. At a large metropolitan department store, the number of hours worked ($y$) per day by the clerical staff may depend on the following variables:

$x_1$ = Number of pieces of mail processed (open, sort, etc.)

$x_2$ = Number of gift cards sold

$x_3$ = Number of store charge accounts transacted

$x_4$ = Number of change order transactions or returns processed

$x_5$ = Number of checks cashed

The data for a sample of 52 working days are listed in Table 11.4. The store's production engineer wants to model number of hours worked with the first order model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

a. Use scattergrams to plot the sample data. Interpret the plots.

b. Use the method of least squares to estimate the model parameters. Interpret the $\beta$-estimates.

c. Find the estimate of $\sigma$, the standard deviation of the random error term, and interpret its value.

d. Assess the adequacy of the model by conducting a test of hypothesis at $\alpha = .05$.

CLERICAL

**TABLE 11.4** Daily Data on Workloads of a Clerical Staff

| Obs. | Day of Week | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|---|
| 1 | M | 128.5 | 7,781 | 100 | 886 | 235 | 644 |
| 2 | T | 113.6 | 7,004 | 110 | 962 | 388 | 589 |
| 3 | W | 146.6 | 7,267 | 61 | 1,342 | 398 | 1,081 |
| 4 | Th | 124.3 | 2,129 | 102 | 1,153 | 457 | 891 |
| 5 | F | 100.4 | 4,878 | 45 | 803 | 577 | 537 |
| 6 | S | 119.2 | 3,999 | 144 | 1,127 | 345 | 563 |
| 7 | M | 109.5 | 11,777 | 123 | 627 | 326 | 402 |
| 8 | T | 128.5 | 5,764 | 78 | 748 | 161 | 495 |
| 9 | W | 131.2 | 7,392 | 172 | 876 | 219 | 823 |
| 10 | Th | 112.2 | 8,100 | 126 | 685 | 287 | 555 |
| 11 | F | 95.4 | 4,736 | 115 | 436 | 235 | 456 |
| 12 | S | 124.6 | 4,337 | 110 | 899 | 127 | 573 |
| 13 | M | 103.7 | 3,079 | 96 | 570 | 180 | 428 |
| 14 | T | 103.6 | 7,273 | 51 | 826 | 118 | 463 |
| 15 | W | 133.2 | 4,091 | 116 | 1,060 | 206 | 961 |
| 16 | Th | 111.4 | 3,390 | 70 | 957 | 284 | 745 |
| 17 | F | 97.7 | 6,319 | 58 | 559 | 220 | 539 |
| 18 | S | 132.1 | 7,447 | 83 | 1,050 | 174 | 553 |
| 19 | M | 135.9 | 7,100 | 80 | 568 | 124 | 428 |
| 20 | T | 131.3 | 8,035 | 115 | 709 | 174 | 498 |
| 21 | W | 150.4 | 5,579 | 83 | 568 | 223 | 683 |
| 22 | Th | 124.9 | 4,338 | 78 | 900 | 115 | 556 |
| 23 | F | 97.0 | 6,895 | 18 | 442 | 118 | 479 |
| 24 | S | 114.1 | 3,629 | 133 | 644 | 155 | 505 |
| 25 | M | 88.3 | 5,149 | 92 | 389 | 124 | 405 |
| 26 | T | 117.6 | 5,241 | 110 | 612 | 222 | 477 |
| 27 | W | 128.2 | 2,917 | 69 | 1,057 | 378 | 970 |
| 28 | Th | 138.8 | 4,390 | 70 | 974 | 195 | 1,027 |
| 29 | F | 109.5 | 4,957 | 24 | 783 | 358 | 893 |
| 30 | S | 118.9 | 7,099 | 130 | 1,419 | 374 | 609 |
| 31 | M | 122.2 | 7,337 | 128 | 1,137 | 238 | 461 |
| 32 | T | 142.8 | 8,301 | 115 | 946 | 191 | 771 |
| 33 | W | 133.9 | 4,889 | 86 | 750 | 214 | 513 |
| 34 | Th | 100.2 | 6,308 | 81 | 461 | 132 | 430 |
| 35 | F | 116.8 | 6,908 | 145 | 864 | 164 | 549 |
| 36 | S | 97.3 | 5,345 | 116 | 604 | 127 | 360 |
| 37 | M | 98.0 | 6,994 | 59 | 714 | 107 | 473 |
| 38 | T | 136.5 | 6,781 | 78 | 917 | 171 | 805 |
| 39 | W | 111.7 | 3,142 | 106 | 809 | 335 | 702 |
| 40 | Th | 98.6 | 5,738 | 27 | 546 | 126 | 455 |
| 41 | F | 116.2 | 4,931 | 174 | 891 | 129 | 481 |
| 42 | S | 108.9 | 6,501 | 69 | 643 | 129 | 334 |
| 43 | M | 120.6 | 5,678 | 94 | 828 | 107 | 384 |
| 44 | T | 131.8 | 4,619 | 100 | 777 | 164 | 834 |
| 45 | W | 112.4 | 1,832 | 124 | 626 | 158 | 571 |
| 46 | Th | 92.5 | 5,445 | 52 | 432 | 121 | 458 |
| 47 | F | 120.0 | 4,123 | 84 | 432 | 153 | 544 |
| 48 | S | 112.2 | 5,884 | 89 | 1,061 | 100 | 391 |
| 49 | M | 113.0 | 5,505 | 45 | 562 | 84 | 444 |
| 50 | T | 138.7 | 2,882 | 94 | 601 | 139 | 799 |
| 51 | W | 122.1 | 2,395 | 89 | 637 | 201 | 747 |
| 52 | Th | 86.6 | 6,847 | 14 | 810 | 230 | 547 |

*Source:* Adapted from Smith, G. L. *Work Measurement*. Columbus, OH: Grid Publishing Co., 1978 (Table 3–1).

e. Find a 95% confidence interval for $\beta_2$. Interpret the result.

f. Find the adjusted coefficient of determination, $R_a^2$, and interpret the result.

g. Find a 95% prediction interval for the number of hours worked on a day when $x_1 = 5{,}000$ of pieces of mail are processed, $x_2 = 75$ gift certificates are sold, $x_3 = 900$ store charge accounts transactions are made, $x_4 = 200$ change order transactions are processed, and $x_5 = 650$ checks are cashed. Interpret the result.

Solution

a. MINITAB side-by-side scatterplots for examining the relationships between the dependent variable $y$ and each of the five independent variables are shown in Figure 11.6. Of the five variables, number of checks ($x_5$) appears to have the strongest linear relationship with $y$, and number of pieces of mail processed ($x_1$) appears to have the weakest linear relationship.

b. We fit the model to the data using SAS. The results are shown in the SAS printout, Figure 11.7. The least-squares estimates of the $\beta$ parameters (highlighted on the printout) yield the following prediction equation:

$$\hat{y} = 66.3 + 0.0012x_1 + 0.116x_2 + 0.013x_3 - 0.045x_4 + 0.056x_5$$

We know that with first-order models, $\beta_i$ represents the slope of the line relating $y$ to $x_i$ for fixed values of the other $x$'s in the model. That is, $\beta_i$ measures the change in $E(y)$ for every 1-unit increase in $x_i$ when all the other independent variables in the model are held constant. Consequently, we obtain the following interpretations:

$\hat{\beta}_1 = .0012$: We estimate the mean number of daily work hours, $E(y)$, to *increase* .0012 hour for every additional piece of mail processed (i.e., for every 1-unit increase in $x_1$), when the other $x$'s in the model are held fixed.

$\hat{\beta}_2 = .116$: We estimate the mean number of daily work hours, $E(y)$, to *increase* .116 hour for every additional gift certificate sold (i.e., for every 1-unit increase in $x_2$), when the other $x$'s in the model are held fixed.

$\hat{\beta}_3 = .013$: We estimate the mean number of daily work hours, $E(y)$, to *increase* .013 hour for every additional charge account transaction (i.e., for every 1-unit increase in $x_3$), when the other $x$'s in the model are held fixed.

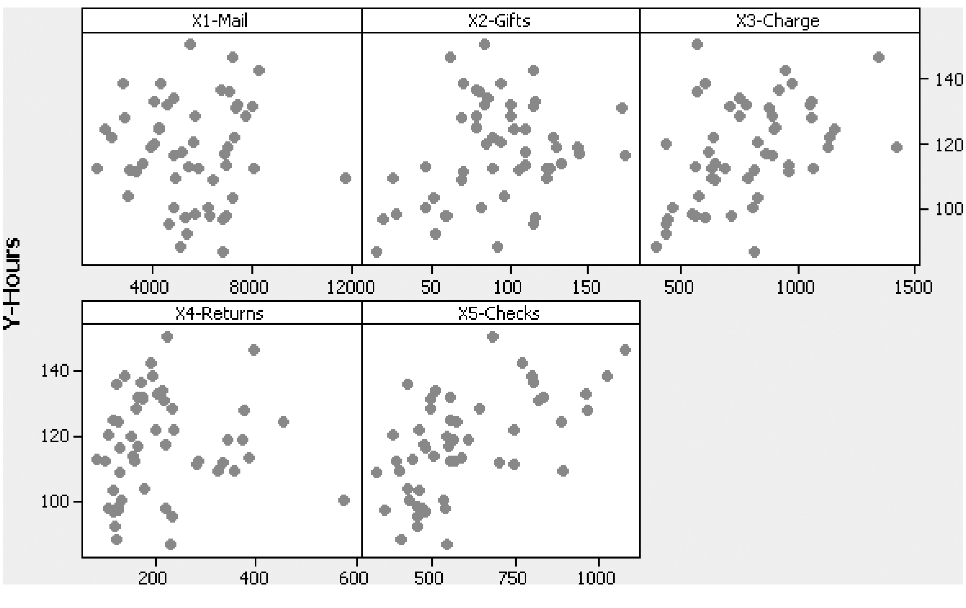



**FIGURE 11.6**
MINITAB scatterplots of data in Table 11.3

The REG Procedure
Model: MODEL1
Dependent Variable: HOURS

| | |
|---|---|
| Number of Observations Read | 52 |
| Number of Observations Used | 52 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 6576.40174 | 1315.28035 | 10.55 | <.0001 |
| Error | 46 | 5736.11057 | 124.69806 | | |
| Corrected Total | 51 | 12313 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 11.16683 | R-Square | 0.5341 |
| Dependent Mean | 117.37692 | Adj R-Sq | 0.4835 |
| Coeff Var | 9.51365 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 66.26491 | 9.10141 | 7.28 | <.0001 | 47.94472 | 84.58510 |
| MAIL | 1 | 0.00124 | 0.00088788 | 1.40 | 0.1686 | -0.00054543 | 0.00303 |
| GIFTS | 1 | 0.11581 | 0.04608 | 2.51 | 0.0155 | 0.02306 | 0.20856 |
| CHARGE | 1 | 0.01269 | 0.00871 | 1.46 | 0.1520 | -0.00484 | 0.03021 |
| RETURNS | 1 | -0.04549 | 0.01755 | -2.59 | 0.0128 | -0.08082 | -0.01016 |
| CHECKS | 1 | 0.05616 | 0.01096 | 5.12 | <.0001 | 0.03410 | 0.07822 |

**FIGURE 11.7**
SAS regression output for first-order model of work hours

$\hat{\beta}_4 = -.045$: We estimate the mean number of daily work hours, $E(y)$, to *decrease* .045 hour for every change order transaction processed (i.e., for every 1-unit increase in $x_4$), when the other $x$'s in the model are held fixed.

$\hat{\beta}_5 = .056$: We estimate the mean number of daily work hours, $E(y)$, to increase .056 hour for every additional check cashed (i.e., for every 1-unit increase in $x_5$), when the other $x$'s in the model are held fixed.

The value $\hat{\beta}_0 = 66.3$ does not have a meaningful interpretation in this example. To see this, note that $\hat{y} = \hat{\beta}_0$ when $x_1 = x_2 = x_3 = x_4 = x_5 = 0$. Thus, $\hat{\beta}_0 = 66.3$ represents the estimated mean number of hours worked on a day when the values of all the independent variables are equal to 0. Since a workday with no mail to process, no gift certificates sold, no charge account or change order transactions, or no checks cashed is impractical, the value of the estimated $y$-intercept has no meaningful interpretation. *In general, $\hat{\beta}_0$ will not have a practical interpretation unless it makes sense to set the values of the x's simultaneously equal to 0.*

c. The estimate for $\sigma$, highlighted on the printout as ROOT MSE, is $s = 11.7$. A useful interpretation of the estimated standard deviation $s$ is that the interval $\pm 2s$ will provide a rough approximation to the accuracy with which the model will predict future values of $y$. Our reasoning is as follows: If the assumptions about the random error $\varepsilon$ in Section 11.2 hold, then $\varepsilon$ is normally distributed with mean 0 and standard deviation $\sigma$. Consequently, about 95% of the errors of prediction will fall within $2\sigma$ of 0, or equivalently, 95% of the actual $y$'s will fall within $2\sigma$ of their corresponding predicted values. Thus, we expect the first-order model to provide predictions of work hours to within about $\pm 2s = \pm 2(11.17) = \pm 22.34$ hours of their true values.

d. To conduct a test of overall model adequacy for this first-order model with five independent variables, we want to test

$H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
$H_a$: At least one of the above $\beta$'s is not equal to 0

The test statistic, $F = 10.55$, is highlighted on Figure 11.7, as is the observed significance level of the test, $p$-value $< .0001$. Since this $p$-value is less than $\alpha = .05$, we reject $H_0$ and conclude that at least one of the model parameters is nonzero—i.e., we conclude that the first-order model is *statistically* useful for predicting number of daily work hours, $y$.

e. A 95% confidence interval for $\beta_2$, highlighted on the SAS printout in the row corresponding to the variable $x_2$ (GIFTS), is (.023, .209). Our interpretation is similar to the one given in part **b**; we are 95% confident that for every 1-unit increase in number of gift certificates sold, daily work hours ($y$) will increase between .023 and .209 hour.

f. The adjusted coefficient of determination, highlighted on Figure 11.7, is $R_a^2 = .4835$. This implies that after adjusting for sample size and terms in the model, the first-order model accounts for about 48% of the sample variation in daily work hours ($y$).

g. We used MINITAB to obtain the 95% prediction interval for the desired future value of $y$. The specified values of the $x$'s as well as the prediction interval (97.09, 142.88) are shaded on Figure 11.8. We can be 95% confident that the number of hours worked on a day when $x_1 = 5{,}000$ of pieces of mail are processed, $x_2 = 75$ gift certificates are sold, $x_3 = 900$ store charge accounts transactions are made, $x_4 = 200$ charge order transactions are processed, and $x_5 = 650$ checks are cashed falls between 97 and 142 hours.

```
Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  119.98    2.17  (115.62, 124.35)  (97.09, 142.88)

Values of Predictors for New Observations

New
Obs  X1-Mail  X2-Gifts  X3-Charge  X4-Returns  X5-Checks
  1     5000      75.0        900         200        650
```

**FIGURE 11.8**
MINITAB prediction interval for $y$ in first-order model of work hours

Although the model in Example 11.10 is statistically useful for predicting $y$ (i.e., the global $F$ test is significant), it may not be *practically* useful. Only about 48% of the sample variation in daily work hours $y$ can be explained by the model, and the model standard deviation indicates that we can predict $y$ to within 22 hours—a value that may lead to larger-than-desired errors of prediction. Consequently, we may want to improve the model before applying it in practice. In the next two sections, we consider some alternative multiple regression models that are more complex than the first-order model.

## Applied Exercises

11.18 *Whales entangled in fishing gear.* Entanglement of marine mammals (e.g., whales) in fishing gear is considered a significant threat to the species. A study published in *Marine Mammal Science* (April 2010) investigated the characteristics of whales entangled in fishing nets in the East Sea of Korea. A sample of 207 entanglements of whales in the area was used to model the body length ($y$, in meters) of the entangled whale. Two independent variables used to predict whale length were water depth of the entanglement ($x_1$, in meters) and distance of the entanglement from land ($x_2$, in miles).

a. Give the equation of a first-order model for length ($y$) as a function of the two independent variables.
b. The marine scientists theorize that the length of an entangled whale will increase linearly as the water depth increases, for entanglements that are a fixed distance from land. Explain how to use the model, part a, to test this theory.
c. The $p$-value for testing $H_0$: $\beta_2 = 0$ in the model, part a, was reported as .013. Interpret this result using $\alpha = .05$.

11.19 *A rubber additive made from cashew nut shells.* Cardanol, an agricultural byproduct of cashew nut shells, is a cheap and abundantly available renewable resource. In *Industrial & Engineering Chemistry Research* (May 2013), researchers investigated the use of cardanol as an additive for natural rubber. Cardanol was grafted onto pieces of natural rubber latex and the chemical properties examined. One property of interest is the grafting efficiency (measured as a percentage) of the chemical process. The researchers manipulated several independent variables in the chemical process—$x_1$ = initiator concentration (parts per hundred resin), $x_2$ = cardanol concentration (parts per hundred resin), $x_3$ = reaction temperature (degrees Centigrade) and $x_4$ = reaction time (hours). Values of these variables, as well as the dependent variable $y$ = grafting efficiency, were recorded for a sample of $n = 9$ chemical runs. The data are provided in the accompanying table. A MINITAB analysis of the first-order model, $E(y_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, is shown at the bottom of the page.

**GRAFTING**

| Run | GE ($y$) | IC ($x_1$) | CC ($x_2$) | RTEMP ($x_3$) | RTIME ($x_4$) |
|---|---|---|---|---|---|
| 1 | 81.94 | 1 | 5 | 35 | 6 |
| 2 | 52.38 | 1 | 10 | 50 | 8 |
| 3 | 54.62 | 1 | 15 | 65 | 10 |
| 4 | 84.92 | 2 | 5 | 50 | 10 |
| 5 | 78.93 | 2 | 10 | 65 | 6 |
| 6 | 36.47 | 2 | 15 | 35 | 8 |
| 7 | 67.79 | 3 | 5 | 65 | 8 |
| 8 | 43.96 | 3 | 10 | 35 | 10 |
| 9 | 42.85 | 3 | 15 | 50 | 6 |

*Source:* Mohapatra, S. & Nando, G.B. "Chemical Modification of Natural Rubber in the Latex Stage by Grafting Cardanol, a Waste from the Cashew Industry and a Renewable Resource", *Industrial & Engineering Chemistry Research,* Vol. 52, No. 17, May 2013 (Tables 2 and 3).

a. Conduct a test of overall model adequacy. Use $\alpha = .10$.
b. Interpret, practically, the value of $R_a^2$.
c. Interpret, practically, the value of $s$.
d. Find and interpret a 90% confidence interval for $\beta_3$.
e. Conduct a test of $H_0$: $\beta_4 = 0$. What do you conclude?

**MINITAB Output for Exercise 11.19**

```
Regression Analysis: GE versus IC, CC, RTEMP, RTIME

The regression equation is
GE = 97.3 - 5.72 IC - 3.36 CC + 0.433 RTEMP - 1.69 RTIME


Predictor      Coef  SE Coef      T      P
Constant      97.28    27.40   3.55  0.024
IC           -5.723    4.581  -1.25  0.280
CC          -3.3570   0.9162  -3.66  0.022
RTEMP        0.4330   0.3054   1.42  0.229
RTIME        -1.685    2.290  -0.74  0.503


S = 11.2206   R-Sq = 81.4%   R-Sq(adj) = 62.9%


Analysis of Variance

Source          DF      SS     MS     F      P
Regression       4  2208.2  552.1  4.38  0.091
Residual Error   4   503.6  125.9
Total            8  2711.8
```

11.20 *Highway crash data analysis.* Civil engineers at Montana State University have written a tutorial on an empirical Bayes method for analyzing before and after highway crash data (*Montana Department of Transportation, Research Report,* May 2004). The initial step in the methodology is to develop a Safety Performance Function (SPF)—a mathematical model that estimates crash occurrence for a given roadway segment. Using data collected for over 100 roadway segments, the engineers fit the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $y$ = number of crashes per 3 years, $x_1$ = roadway length (miles), and $x_2$ = AADT = average annual daily traffic (number of vehicles). The results are shown in the following tables.

**Interstate Highways**

| Variable | Parameter Estimate | Standard Error | $T$ value |
|---|---|---|---|
| Intercept | 1.81231 | .50568 | 3.58 |
| Length ($x_1$) | .10875 | .03166 | 3.44 |
| AADT ($x_2$) | .00017 | .00003 | 5.19 |

**Noninterstate Highways**

| Variable | Parameter Estimate | Standard Error | $T$ value |
|---|---|---|---|
| Intercept | 1.20785 | .28075 | 4.30 |
| Length ($x_1$) | .06343 | .01809 | 3.51 |
| AADT ($x_2$) | .00056 | .00012 | 4.86 |

a. Give the least-squares prediction equation for the interstate highway model.
b. Give practical interpretations of the $\beta$ estimates, part **a**.
c. Refer to part **a**. Find a 99% confidence interval for $\beta_1$ and interpret the results.
d. Refer to part **a**. Find a 99% confidence interval for $\beta_2$ and interpret the results.
e. Repeat parts **a–d** for the noninterstate highway model.

11.21 *Global warming and foreign investments.* Scientists believe that a major cause of global warming is higher levels of carbon dioxide ($CO_2$) in the atmosphere. In the *Journal of World-Systems Research* (Summer 2003), sociologists examined the impact of foreign investment dependence on $CO_2$ emissions in $n = 66$ developing countries. In particular, the researchers modeled the level of $CO_2$ emissions in a year based on foreign investments and other independent variables measured 16 years earlier. The variables and the model results are listed in the next table.

a. Interpret the value of $R^2$.
b. Conduct a test of overall model adequacy. Use $\alpha = .01$.
c. Conduct a test to determine if agricultural production ($x_5$) is a statistically useful predictor of $CO_2$ emissions ($y$). Use $\alpha = .01$.

| $y$ = ln(level of $CO_2$ emissions in current year) | $\beta$ Estimate | $T$ value | $p$-value |
|---|---|---|---|
| $x_1$ = ln(foreign investments) | .79 | 2.52 | $<.05$ |
| $x_2$ = gross domestic investment | .01 | .13 | $>.10$ |
| $x_3$ = trade exports | −.02 | −1.66 | $>.10$ |
| $x_4$ = ln(GNP) | −.44 | −.97 | $>.10$ |
| $x_5$ = agricultural production | −.03 | −.66 | $>.10$ |
| $x_6$ = 1 if African country, 0 if not | −1.19 | −1.52 | $>.10$ |
| $x_7$ = ln(level of $CO_2$ emissions) | .56 | 3.35 | $<.001$ |
| $R^2 = .31$ | | | |

*Source:* Grimes, P., and Kentor, J. "Exporting the greenhouse: Foreign capital penetration and $CO_2$ emissions 1980–1996." *Journal of World-Systems Research*, Vol. IX, No. 2, Summer 2003 (Table 1).

11.22 *Growth of Japanese beetles.* In the *Journal of Insect Behavior* (Nov. 2001), biologists at Eastern Illinois University published the results of their study on Japanese beetles. The biologists collected beetles over a period of $n = 13$ summer days in a soybean field. For one portion of the study, the biologists modeled $y$, the average size (in millimeters) of female beetles as a function of the average daily temperature $x_1$ (degrees) and Julian date $x_2$.

a. Write a first-order model for $E(y)$ as a function of $x_1$ and $x_2$.
b. The model was fit to the data, with the following results. Interpret the estimate of $\beta_1$.

| Variable | Parameter Estimate | $T$ value | $p$-value |
|---|---|---|---|
| Intercept | 6.51 | 26.0 | $<.0001$ |
| Temperature ($x_1$) | −.002 | −0.72 | .49 |
| Date ($x_2$) | −.010 | −3.30 | .008 |

c. Conduct a test to determine whether the average size of female Japanese beetles decreases linearly as temperature increases. Use $\alpha = .05$.

11.23 *Emotional intelligence and team performance.* The *Engineering Project Organizational Journal* (Vol. 3., 2013) published the results of an exploratory study designed to gain a better understanding of how the emotional intelligence of individual team members relates directly to the performance of their team during an engineering project. Undergraduate students enrolled in the course, *Introduction to the Building Industry,* participated in the study. All students completed an emotional intelligence test and received an interpersonal score, stress management score, and mood score. Students were grouped into $n = 23$ teams and assigned a group project. However, each student received an individual project score. These scores were averaged to obtain the dependent variable in the analysis—mean project score ($y$). Three independent variables were determined for each team: range of interpersonal scores ($x_1$), range of stress management scores ($x_2$), and range of mood scores ($x_3$). Data (simulated from information provided in the article) are listed in the table.

**TEAMPERF**

| Team | Intrapersonal Range | Stress Range | Mood Range | Project Average |
|---|---|---|---|---|
| 1 | 14 | 12 | 17 | 88.0 |
| 2 | 21 | 13 | 45 | 86.0 |
| 3 | 26 | 18 | 6 | 83.5 |
| 4 | 30 | 20 | 36 | 85.5 |
| 5 | 28 | 23 | 22 | 90.0 |
| 6 | 27 | 24 | 28 | 90.5 |
| 7 | 21 | 24 | 38 | 94.0 |
| 8 | 20 | 30 | 30 | 85.5 |
| 9 | 14 | 32 | 16 | 88.0 |
| 10 | 18 | 32 | 17 | 91.0 |
| 11 | 10 | 33 | 13 | 91.5 |
| 12 | 28 | 43 | 28 | 91.5 |
| 13 | 19 | 19 | 21 | 86.0 |
| 14 | 26 | 31 | 26 | 83.0 |
| 15 | 25 | 31 | 11 | 85.0 |
| 16 | 40 | 35 | 24 | 84.0 |
| 17 | 27 | 12 | 14 | 85.5 |
| 18 | 30 | 13 | 29 | 85.0 |
| 19 | 31 | 24 | 28 | 84.5 |
| 20 | 25 | 26 | 16 | 83.5 |
| 21 | 23 | 28 | 12 | 85.0 |
| 22 | 20 | 32 | 10 | 92.5 |
| 23 | 35 | 35 | 17 | 89.0 |

a. Hypothesize a first-order model for project score ($y$) as a function of $x_1$, $x_2$, and $x_3$.
b. Fit the model, part a, to the data using statistical software.
c. Is there sufficient evidence to indicate the overall model is statistically useful for predicting $y$? Test using $\alpha = .05$.
d. Evaluate the model statistics $R_a^2$ and $2s$.
e. Find and interpret a 95% prediction interval for $y$ when $x_1 = 20, x_2 = 30$, and $x_3 = 25$.

11.24 *Arsenic in groundwater.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of the reliability of a commercial kit to test for arsenic in groundwater, Exercise 7.59 (p. 330). Recall that a field kit was used to test a sample of 328 groundwater wells in Bangladesh. In addition to the arsenic level (micrograms per liter), the latitude (degrees), longitude (degrees), and depth (feet) of each well was measured. The data are saved in the **ASWELLS** file.

**ASWELLS**

(*Data for first and last five wells shown*)

| Wellid | Latitude | Longitude | Depth | Arsenic |
|---|---|---|---|---|
| 10 | 23.7887 | 90.6522 | 60 | 331 |
| 14 | 23.7886 | 90.6523 | 45 | 302 |
| 30 | 23.7880 | 90.6517 | 45 | 193 |
| 59 | 23.7893 | 90.6525 | 125 | 232 |
| 85 | 23.7920 | 90.6140 | 150 | 19 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 7353 | 23.7949 | 90.6515 | 40 | 48 |
| 7357 | 23.7955 | 90.6515 | 30 | 172 |
| 7890 | 23.7658 | 90.6312 | 60 | 175 |
| 7893 | 23.7656 | 90.6315 | 45 | 624 |
| 7970 | 23.7644 | 90.6303 | 30 | 254 |

a. Write a first-order model for arsenic level ($y$) as a function of latitude, longitude, and depth.
b. Fit the model to the data using the method of least-squares.
c. Give practical interpretations of the $\beta$ estimates.
d. Find the model standard deviation, $s$, and interpret its value.

11.25 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 8.29 (p. 392). Recall that the heat rate (kilojoules per kilowatt per hour) was measured for each in a sample of 67 gas turbines augmented with high-pressure inlet fogging. In addition, several other variables were measured, including cycle speed (revolutions per minute), inlet temperature (C°), exhaust gas temperature (C°), cycle pressure ratio, and air mass flow rate (kilograms per second). The full data set is saved in the **GASTURBINE** file. (First and last 5 observations are shown in the table on p. 591.)

a. Write a first-order model for heat rate ($y$) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.
b. Fit the model to the data using the method of least-squares.
c. Give practical interpretations of the $\beta$ estimates.
d. Find the model standard deviation, $s$, and interpret its value.

## Data for Exercise 11.25

**GASTURBINE**

(*Data for first and last five gas turbines shown*)

| Rpm | Cpratio | Inlet-Temp | Exh-Temp | Airflow | Heatrate |
|---|---|---|---|---|---|
| 27245 | 9.2 | 1134 | 602 | 7 | 14622 |
| 14000 | 12.2 | 950 | 446 | 15 | 13196 |
| 17384 | 14.8 | 1149 | 537 | 20 | 11948 |
| 11085 | 11.8 | 1024 | 478 | 27 | 11289 |
| 14045 | 13.2 | 1149 | 553 | 29 | 11964 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 18910 | 14.0 | 1066 | 532 | 8 | 12766 |
| 3600 | 35.0 | 1288 | 448 | 152 | 8714 |
| 3600 | 20.0 | 1160 | 456 | 84 | 9469 |
| 16000 | 10.6 | 1232 | 560 | 14 | 11948 |
| 14600 | 13.4 | 1077 | 536 | 20 | 12414 |

*Source:* Bhargava, R., and Meher-Homji, C. B. "Parametric analysis of existing gas turbines with inlet evaporative and overspray fogging." *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005.

11.26 *Contamination of fish in the Tennessee River.* Refer to the U.S. Army Corps of Engineers data on fish contaminated from the toxic discharges of a chemical plant located on the banks of the Tennessee River in Alabama. Recall that the engineers measured the length (in centimeters), weight (in grams), and DDT level (in parts per million) for 144 captured fish. In addition, the number of miles upstream from the river was recorded. The data are saved in the **DDT** file. (The first and last five observations are shown in the table.)

**DDT**

| River | Mile | Species | Length | Weight | DDT |
|---|---|---|---|---|---|
| FC | 5 | CHANNELCATFISH | 42.5 | 732 | 10.00 |
| FC | 5 | CHANNELCATFISH | 44.0 | 795 | 16.00 |
| FC | 5 | CHANNELCATFISH | 41.5 | 547 | 23.00 |
| FC | 5 | CHANNELCATFISH | 39.0 | 465 | 21.00 |
| FC | 5 | CHANNELCATFISH | 50.5 | 1252 | 50.00 |
| ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ |
| TR | 345 | LARGEMOUTHBASS | 23.5 | 358 | 2.00 |
| TR | 345 | LARGEMOUTHBASS | 30.0 | 856 | 2.20 |
| TR | 345 | LARGEMOUTHBASS | 29.0 | 793 | 7.40 |
| TR | 345 | LARGEMOUTHBASS | 17.5 | 173 | 0.35 |
| TR | 345 | LARGEMOUTHBASS | 36.0 | 1433 | 1.90 |

a. Fit the first-order model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ to the data, where $y$ = DDT level, $x_1$ = mile, $x_2$ = length, and $x_3$ = weight. Report the least-squares prediction equation.

b. Find the estimate of the standard deviation of $\varepsilon$ for the model and give a practical interpretation of its value.

c. Do the data provide sufficient evidence to conclude that DDT level increases as length increases? Report the observed significance level of the test and reach a conclusion using $\alpha = .05$.

d. Find and interpret a 95% confidence interval for $\beta_3$.

e. Test the overall adequacy of the model using $\alpha = .05$.

f. Predict, with 95% confidence, the DDT level of a fish caught 100 miles upstream with a length of 40 cm and a weight of 800 g. Interpret the result.

11.27 *Extracting water from oil.* In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage ($y$) required to separate the water from the oil. (*Journal of Colloid and Interface Science*, Aug. 1995.) The seven independent variables investigated in the study are listed in the table below. (Each variable was measured at two levels—a "low" level and a "high" level.) Sixteen water/oil mixtures were prepared using different combinations of the independent variables; then each emulsion was exposed to a high electric field. In addition, three mixtures were tested when all independent variables were set to 0. The data for all 19 experiments are saved in the **WATEROIL** file. (The first 5 experiments are listed in the next table.)

a. Propose a first-order model for $y$ as a function of all seven independent variables.

b. Use a statistical software package to fit the model to the data in the table.

c. Fully interpret the $\beta$ estimates.

d. Assess model adequacy by conducting the $F$ test, interpreting $R_a^2$, and interpreting $2s$.

e. Consider the model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_5$. The researchers concluded that "in order to break a water-oil mixture with the lowest possible voltage, the volume fraction of the disperse phase ($x_1$) should be high, while the salinity ($x_2$) and the amount of surfactant ($x_5$) should be low." Use this information to find a 95% prediction interval for this "low" voltage $y$. Interpret the interval.

**Data for Exercise 11.27**

**WATEROIL**

*(Data for 5 of 19 experiments shown.)*

| Experiment Number | Voltage $y$ (kw/cm) | Disperse Phase Volume $x_1$ (%) | Salinity $x_2$ (%) | Temperature $x_3$ (°C) | Time Delay $x_4$ (hours) | Surfactant Concentration $x_5$ (%) | Span: Triton $x_6$ | Solid Particles $x_7$ (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | .64 | 40 | 1 | 4 | .25 | 2 | .25 | .5 |
| 2 | .80 | 80 | 1 | 4 | .25 | 4 | .25 | 2 |
| 3 | 3.20 | 40 | 4 | 4 | .25 | 4 | .75 | .5 |
| 4 | .48 | 80 | 4 | 4 | .25 | 2 | .75 | 2 |
| 5 | 1.72 | 40 | 1 | 23 | .25 | 4 | .75 | 2 |

*Source:* Fordedal, H., et al. "A multivariate analysis of W/O emulsions in high external electric fields as studied by means of dielectric time domain spectroscopy." *Journal of Colloid and Interface Science*, Vol. 173, No. 2, Aug. 1995, p. 398 (Table 2).

## 11.8 An Interaction Model with Quantitative Predictors

In Section 11.7, we demonstrated the relationship between $E(y)$ and the independent variables in a first-order model. When $E(y)$ is graphed against any one variable (say, $x_1$) for fixed values of the other variables, the result is a set of *parallel* straight lines (see Figure 11.5). When this situation occurs (as it always does for a first-order model), we say that the relationship between $E(y)$ and any one independent variable *does not depend* on the values of the other independent variables in the model.

However, if the relationship between $E(y)$ and $x_1$ does, in fact, depend on the values of the remaining $x$'s held fixed, then the first-order model is not appropriate for predicting $y$. In this case, we need another model that will take into account this dependence. Such a model includes the *cross-products* of two or more $x$'s.

For example, suppose that the mean value $E(y)$ of a response $y$ is related to two quantitative independent variables, $x_1$ and $x_2$, by the model

$$E(y) = 1 + 2x_1 - x_2 + x_1x_2$$

A graph of the relationship between $E(y)$ and $x_1$ for $x_2 = 0, 1,$ and 2 is displayed in Figure 11.9. Note that the graph shows three nonparallel straight lines. You can verify that the slopes of the lines differ by substituting each of the values $x_2 = 0, 1,$ and 2 into the equation. For $x_2 = 0$:

$$E(y) = 1 + 2x_1 - (0) + x_1(0) = 1 + 2x_1 \qquad (\text{slope} = 2)$$

For $x_2 = 1$:

$$E(y) = 1 + 2x_1 - (1) + x_1(1) = 3x_1 \qquad (\text{slope} = 3)$$

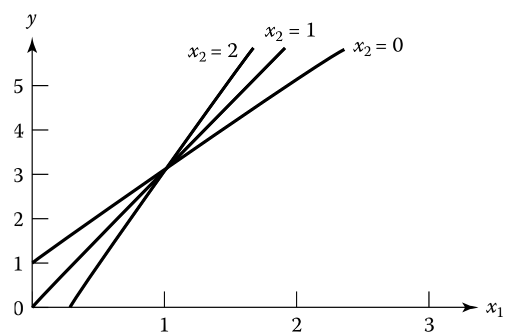



**FIGURE 11.9**
Graphs of $1 + 2x_1 - x_2 + x_1x_2$ for $x_2 = 0, 1, 2$

For $x_2 = 2$:

$$E(y) = 1 + 2x_1 - (2) + x_1(2) = -1 + 4x_1 \qquad \text{(slope} = 4)$$

Note that the slope of each line is represented by $\beta_1 + \beta_3 x_2 = 2 + x_2$. Thus, the effect on $E(y)$ of a change in $x_1$ (i.e., the slope) now *depends* on the value of $x_2$. When this situation occurs, we say that $x_1$ and $x_2$ **interact**. The cross-product term, $x_1 x_2$, is called an **interaction term**, and the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is called an **interaction model** with two quantitative variables.

### An Interaction Model Relating $E(y)$ to Two Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where

$(\beta_1 + \beta_3 x_2)$ represents the change in $E(y)$ for every 1-unit increase in $x_1$, holding $x_2$ fixed

$(\beta_2 + \beta_3 x_1)$ represents the change in $E(y)$ for every 1-unit increase in $x_2$, holding $x_1$ fixed

**Example 11.11**

**An Interaction Model for Production Man-Hours**

In a production facility, an accurate estimate of man-hours needed to complete a task is crucial to management. A manufacturer of boiler drums wants to use regression to predict the number of man-hours needed to erect the drums in future projects. To accomplish this, data for a sample of 35 boilers were collected. In addition to man-hours ($y$), the variables measured were boiler capacity $x_1$ (thousand pounds per hour) and boiler design pressure $x_2$ (pounds per square inch, psi). The data are listed in Table 11.5. The manufacturer believes that the rate at which man-hours increase with boiler capacity will be greater for boilers designed at higher pressures. Hence, the following interaction model is proposed:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

SPSS was used to fit the model to the data. The SPSS printout is shown in Figure 11.10.

a. Test the overall utility of the model using the global $F$ test at $\alpha = .05$.

b. Test the hypothesis (at $\alpha = .05$) that the slope of the relationship between man-hours ($y$) and boiler capacity ($x_1$) increases as the design pressure ($x_2$) increases—that is, that capacity and pressure interact positively.

c. Estimate the change in man-hours ($y$) for every 1-psi increase in design pressure ($x_2$) when boiler capacity ($x_1$) is 750 thousand pounds/hour.

**Solution**

a. The global $F$ test is used to test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

The test statistic and $p$-value of the test (highlighted on the SPSS printout) are $F = 28.652$ and $p$-value $= 0$, respectively. Since $\alpha = .05$ exceeds the $p$-value, there is sufficient evidence to conclude that the model fit is a statistically useful predictor of man-hours, $y$.

b. The hypothesis of interest to the manufacturer concerns the interaction parameter $\beta_3$. Specifically,

$$H_0: \beta_3 = 0$$
$$H_a: \beta_3 > 0$$

Since we are testing an individual $\beta$ parameter, a $T$ test is required. The test statistic and two-tailed $p$-value (highlighted on the printout) are $T = 2.233$ and

**BOILERS**

**TABLE 11.5 Data for Boiler Drum Study**

| Man-Hours $y$ | Boiler Capacity (thousand lbs/hr) $x_1$ | Design Pressure (psi) $x_2$ | Man-Hours $y$ | Boiler Capacity (thousand lbs/hr) $x_1$ | Design Pressure (psi) $x_2$ |
|---|---|---|---|---|---|
| 3,137 | 120.0 | 375 | 14,791 | 1,089.5 | 2,170 |
| 3,590 | 65.0 | 750 | 2,680 | 125.0 | 750 |
| 4,526 | 150.0 | 500 | 2,974 | 120.0 | 375 |
| 10,825 | 1,073.8 | 2,170 | 1,965 | 65.0 | 750 |
| 4,023 | 150.0 | 325 | 2,566 | 150.0 | 500 |
| 7,606 | 610.0 | 1,500 | 1,515 | 150.0 | 250 |
| 3,748 | 88.2 | 399 | 2,000 | 150.0 | 500 |
| 2,972 | 88.2 | 399 | 2,735 | 150.0 | 325 |
| 3,163 | 88.2 | 399 | 3,698 | 610.0 | 1,500 |
| 4,065 | 90.0 | 1,140 | 2,635 | 90.0 | 1,140 |
| 2,048 | 30.0 | 325 | 1,206 | 30.0 | 325 |
| 6,500 | 441.0 | 410 | 3,775 | 441.0 | 410 |
| 5,651 | 441.0 | 410 | 3,120 | 441.0 | 410 |
| 6,565 | 441.0 | 410 | 4,206 | 441.0 | 410 |
| 6,387 | 441.0 | 410 | 4,006 | 441.0 | 410 |
| 6,454 | 627.0 | 1,525 | 3,728 | 627.0 | 1,525 |
| 6,928 | 610.0 | 1,500 | 3,211 | 610.0 | 1,500 |
| 4,268 | 150.0 | 500 | 1,200 | 30.0 | 325 |

*Source:* Kelly Uscategui, former graduate student, University of South Florida

$p$-value $= .033$, respectively. The upper-tailed $p$-value, obtained by dividing the two-tailed $p$-value in half, is $.033/2 = .0165$. Since $\alpha = .05$ exceeds the $p$-value, the manufacturer can reject $H_0$ and conclude that the rate of change of man-hours with capacity increases as the design pressure increases; that is, $x_1$ and $x_2$ interact positively. Thus, it appears that the interaction term should be included in the model.

c. To estimate the change in man-hours, $y$, for every 1-unit increase in design pressure, $x_2$, we need to estimate the slope of the line relating $y$ to $x_2$ when the boiler capacity is at $x_1 = 750$ thousand pounds/hour. An analyst who is not careful may estimate this slope as $\hat{\beta}_2 = -1.53$. Although the coefficient of $x_2$ is negative, this does *not* imply that man-hours decreases as the design pressure increases. Since interaction is present, the rate of change (slope) of man-hours with the design pressure *depends* on $x_1$, the boiler capacity. Thus, the estimated rate of change of $y$ for a unit increase in $x_2$ (1 psi) for $x_1 = 750$ is

$$\text{Estimated } x_2 \text{ slope} = \hat{\beta}_2 + \hat{\beta}_3 x_1 = -1.53 + .003(750) = .72$$

In other words, we estimate that the man-hours required to erect a boiler drum with a capacity of 750 thousand pounds/hour will *increase* by about .72 man-hour for every 1 psi increase in design pressure. Extreme care is needed in interpreting the signs and sizes of coefficients in a multiple regression model.

Example 11.11 illustrates an important point about conducting $T$ tests on the $\beta$ parameters in the interaction model. The "most important" $\beta$ parameter in this model is the interaction $\beta$, $\beta_3$. (Note that this $\beta$ is also the one associated with the highest-order

**FIGURE 11.10**
SPSS regression output for interaction model of man-hours

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .854[a] | .729 | .703 | 1472.227 |

a. Predictors: (Constant), CAP_PRESS, PRESSURE, CAPACITY

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 2E+008 | 3 | 62102055.14 | 28.652 | .000[a] |
| | Residual | 7E+007 | 32 | 2167451.542 | | |
| | Total | 3E+008 | 35 | | | |

a. Predictors: (Constant), CAP_PRESS, PRESSURE, CAPACITY
b. Dependent Variable: MANHRS

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 3012.677 | 732.354 | | 4.114 | .000 |
| | CAPACITY | 3.786 | 2.135 | .394 | 1.774 | .086 |
| | PRESSURE | -1.529 | 1.083 | -.314 | -1.412 | .168 |
| | CAP_PRESS | .003 | .002 | .745 | 2.233 | .033 |

a. Dependent Variable: MANHRS

term in the model, $x_1x_2$.*) Consequently, we will want to test $H_0$: $\beta_3 = 0$ after we have determined that the overall model is useful for predicting $y$. Once interaction is detected (as in Example 11.11), however, tests on the first-order terms $x_1$ and $x_2$ should *not* be conducted since they are meaningless tests; the presence of interaction implies that both $x$'s are important.

### Caution

Once interaction has been deemed important in the model $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$, do not conduct $T$ tests on the $\beta$ coefficients of the first-order terms $x_1$ and $x_2$. These terms should be kept in the model regardless of the magnitude of their associated $p$-values shown on the printout.

*The order of a term is equal to the sum of the exponents of the quantitative variables included in the term. Thus, when $x_1$ and $x_2$ are both quantitative variables, the cross-product, $x_1x_2$, is a second-order term.

## Applied Exercises

11.28 *Whales entangled in fishing gear.* Refer to the *Marine Mammal Science* (April 2010) study of whales entangled in fishing nets, Exercise 11.18 (p. 581). Recall that the body length ($y$, in meters) of an entangled whale was modeled as a function of two independent variables—water depth of the entanglement ($x_1$, in meters) and distance of the entanglement from land ($x_2$, in miles).

a. Give the equation of an interaction model for length ($y$) as a function of the two independent variables.

b. Construct a graph of length vs. water depth that illustrates the interaction between the two independent variables. Use two different values of distance in the graph.
c. Suppose the marine scientists theorize that the length of an entangled whale will increase linearly as the water depth increases, but that the increase will be greater the farther the distance of the entanglement from land. Explain how to use the model, part a, to test this theory.

**BUBBLE2**

11.29 *Bubble behavior in subcooled flow boiling.* Refer to the *Heat Transfer Engineering* (Vol. 34, 2013) study of bubble behavior in subcooled flow boiling, Exercise 11.6 (p. 580). Recall that bubble density (liters per meters squared) was modeled as a function of mass flux (kilograms per meters squared per second) and heat flux (megawatts per meters squared).

a. Write an interaction model for bubble density ($y$) as a function of $x_1$ = mass flux and $x_2$ = heat flux.
b. Fit the interaction model, part a, to the data using statistical software. Give the least-squares prediction equation.
c. Evaluate overall model adequacy by conducting a global $F$ test (at $\alpha = .05$) and interpreting the model statistics, $R_a^2$ and $2s$.
d. Conduct a test (at $\alpha = .05$) to determine whether mass flux and heat flux interact.
e. How much do you expect bubble density to decrease for every 1 kg/m$^2$-sec increase in mass flux, when heat flux is set at .5 megawatta/m$^2$?

11.30 *Predicting thrust force of a metal drill.* In *Frontiers in Automobile and Mechanical Engineering* (Nov. 2010), a model was developed to predict the thrust force when drilling into a hybrid metal composite. Three variables related to thrust force are spindle speed (revolutions per minute), feed rate (millimeters per minute), and fraction weight of silicon carbide in composite (percentage). Experimental data was collected by varying these three variables at two levels each: speed (1000 and 3000 rpm), rate (50 and 150 mm/min), and weight (5 and 15 percent). For each combination of these values, thrust force (Newtons) of the drill was measured. The data (adapted from information in the article) are listed at the top of the next column.

a. Write the equation of an interaction model for thrust force as a function of spindle speed, feed rate, and fraction weight. Include all possible 2-variable interaction terms in the model.
b. Give a function of the model parameters (i.e., the slope) that represents the change in force for every 1% increase in weight when rate is fixed at 50 mm/minute and speed is fixed at 1000 rpm.
c. Give a function of the model parameters (i.e., the slope) that represents the change in force for every 1% increase in weight when rate is fixed at 150 mm/minute and speed is fixed at 1000 rpm.

**DRILLMETAL**

| Experiment | SPEED | RATE | PCTWT | FORCE |
|---|---|---|---|---|
| 1 | 1000 | 50 | 5 | 510 |
| 2 | 3000 | 50 | 5 | 540 |
| 3 | 1000 | 150 | 5 | 710 |
| 4 | 3000 | 150 | 5 | 745 |
| 5 | 1000 | 50 | 15 | 615 |
| 6 | 3000 | 50 | 15 | 635 |
| 7 | 1000 | 150 | 15 | 810 |
| 8 | 3000 | 150 | 15 | 850 |
| 9 | 1000 | 50 | 5 | 500 |
| 10 | 3000 | 50 | 5 | 545 |
| 11 | 1000 | 150 | 5 | 720 |
| 12 | 3000 | 150 | 5 | 730 |
| 13 | 1000 | 50 | 15 | 600 |
| 14 | 3000 | 50 | 15 | 615 |
| 15 | 1000 | 150 | 15 | 825 |
| 16 | 3000 | 150 | 15 | 840 |

d. Fit the interaction model, part a, to the data using statistical software. Give the least-squares prediction equation.
e. Which $\beta$ should you test to determine if the slopes identified in parts b and c are significantly different? Carry out this test and interpret the results.

11.31 *Multivariable testing.* The technique of multivariable testing (MVT) was discussed In *The Journal of the Reliability Analysis Center* (First Quarter, 2004). MVT was shown to improve the quality of carbon-foam rings used in nuclear missile housings. The rings are produced via a casting process that involves mixing ingredients, oven curing, and carving the finished part. One type of defect analyzed was the number $y$ of black streaks in the manufactured ring. Two variables found to impact the number of defects were turntable speed (revolutions per minute), $x_1$, and cutting blade position (inches from center), $x_2$.

a. The researchers discovered "an interaction between blade position and turntable speed." Hypothesize a regression model for $E(y)$ that incorporates this interaction.
b. The researchers reported a positive linear relationship between number of defects ($y$) and turntable speed ($x_1$), but found that the slope of the relationship was much steeper for lower values of cutting blade position ($x_2$). What does this imply about the interaction term in the model, part **a**? Explain.

**ASWELLS**

11.32 *Arsenic in groundwater.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of the reliability of a commercial kit to test for arsenic in groundwater, Exercise 11.24 (p. 590). Recall that you fit a first-order model for arsenic level ($y$) as a function of latitude ($x_1$), longitude ($x_2$), and depth ($x_3$).

a. Write a model for arsenic level ($y$) that includes first-order terms for latitude, longitude, and depth, as well as terms for interaction between latitude and depth and interaction between longitude and depth.
b. Fit the interaction model, part a. Give the least-squares prediction equation.
c. Conduct a test (at $\alpha = .05$) to determine whether latitude and depth interact to affect arsenic level.
d. Conduct a test (at $\alpha = .05$) to determine whether longitude and depth interact to affect arsenic level.
e. Practically interpret the results of the tests, parts **c** and **d**.

**GASTURBINE**

11.33 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 11.25 (p. 590). Recall that you fit a first-order model for heat rate ($y$) as a function of speed ($x_1$), inlet temperature ($x_2$), exhaust temperature ($x_3$), cycle pressure ratio ($x_4$), and air flow rate ($x_5$).

a. Researchers hypothesize that the linear relationship between heat rate ($y$) and temperature (both inlet and exhaust) depends on air flow rate. Write a model for heat rate that incorporates the researchers' theories.
b. Fit the interaction model, part a. Give the least-squares prediction equation.
c. Conduct a test (at $\alpha = .05$) to determine whether inlet temperature and air flow rate interact to affect heat rate.
d. Conduct a test (at $\alpha = .05$) to determine whether exhaust temperature and air flow rate interact to affect heat rate.
e. Practically interpret the results of the tests, parts **c** and **d**.

**DDT**

11.34 *Contamination of fish in the Tennessee River.* Refer to the U.S. Army Corps of Engineers data on contaminated fish, Exercise 11.26 (p. 591). You fit the first-order model relating DDT level ($y$) to miles upstream ($x_1$), fish length ($x_2$), and fish weight ($x_3$).

a. Propose a model for $E(y)$ that hypothesizes that the rate of increase of DDT level with length is greater for heavier contaminated fish.
b. Fit the model, part a, to the data. Give the least-squares prediction equation.
c. Test the theory, part **a**, using $\alpha = .10$. What do you conclude?

**WATEROIL**

11.35 *Extracting water from oil.* Refer to the *Journal of Colloid and Interface Science* (Aug. 1995) study of the factors that influence the voltage level ($y$) required to separate water from oil, Exercise 11.27 (p. 591).

a. Consider using only volume ($x_1$) and salinity ($x_2$) to predict $y$. Write the equation of an interaction model for $E(y)$.
b. Fit the interaction model, part a. Give the least-squares prediction equation.
c. Conduct a test (at $\alpha = .10$) to determine whether volume and salinity interact to affect voltage level.
d. Give the estimated change in voltage ($y$) for every 1% increase in volume ($x_1$) when salinity is set at $x_2 = 4\%$.

## 11.9 A Quadratic (Second-Order) Model with a Quantitative Predictor

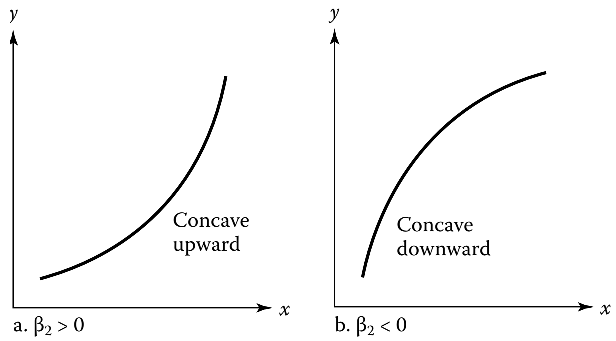
All of the models discussed in the previous sections proposed straight-line relationships between $E(y)$ and each of the independent variables in the model. In this section, we consider a nonlinear model that allows for curvature in the relationship between $y$ and a single quantitative predictor, $x$. This model is a **second-order model** because it will include an $x^2$-term.

The form of this model, called the **quadratic model**, is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

The term involving $x^2$, called a **quadratic term** (or **second-order term**), enables us to hypothesize curvature in the graph of the response model relating $y$ to $x$. Graphs of the quadratic model for two different values of $\beta_2$ are shown in Figure 11.11. When the curve opens upward, the sign of $\beta_2$ is positive (see Figure 11.11a); when the curve opens downward, the sign of $\beta_2$ is negative (see Figure 11.11b).

**FIGURE 11.11**
Graphs for two quadratic models




### A Quadratic (Second-Order) Model in a Single Quantitative Independent Variable

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x^2$$

where

$\beta_0$ is the $y$-intercept of the curve
$\beta_1$ is a shift parameter
$\beta_2$ is the rate of curvature

## Example 11.12

A Quadratic Model for Electrical Usage

Refer to Example 10.12 (p. 531) where we investigated the July electrical usage, $y$, in all-electric homes and its relationship to the size, $x$, of the home. Recall that the 1st order model, $E(y) = \beta_0 + \beta_1 x$, was deemed to be an inadequate fit. Now consider the quadratic model,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

The data for $n = 15$ homes are reproduced in Table 11.6.

a. Construct a scatterplot for the data. Is there evidence to support the use of a quadratic model?
b. Use the method of least-squares to estimate the unknown parameters $\beta_0$, $\beta_1$, and $\beta_2$ in the quadratic model.
c. Graph the prediction equation and assess how well the model fits the data, both visually and numerically.
d. Interpret the $\beta$ estimates.
e. Is the overall model useful (at $\alpha = .01$) for predicting electrical usage $y$?
f. Is there sufficient evidence of concave downward curvature in the electrical usage–home size relationship? Test using $\alpha = .01$.

Solution

a. A scattergram for the data of Table 11.5 produced using MINITAB, is shown in Figure 11.12. The figure illustrates that the electrical usage appears to increase in a curvilinear manner with the size of the home. This provides some support for the inclusion of the quadratic term $x^2$ in the model.
b. We used SAS to fit the model to the data in Table 11.5. Part of the SAS regression output is displayed in Figure 11.13. The least-squares estimates of the $\beta$ parameters (highlighted) are $\hat{\beta}_0 = -806.72$, $\hat{\beta}_1 = 1.962$, and $\hat{\beta}_2 = -.00034$. Therefore, the equation that minimizes the SSE for the data is

$$\hat{y} = -806.72 + 1.962x - .00034x^2$$

c. Figure 11.14 is a graph of the least-squares prediction equation. Note that the graph provides a good fit to the data of Table 11.6. A numerical measure of fit is obtained with the adjusted coefficient of determination, $R_a^2$. From the SAS printout, $R_a^2 = .9735$. This implies that about 97% of the sample variation in electrical usage ($y$) can be explained by the quadratic model (after adjusting for sample size and degrees of freedom).

**KWHRS**

**TABLE 11.6 Home Size–Electrical Usage Data**

| Size of Home $x$ (sq. ft) | Monthly Usage $y$ (kilowatt-hours) |
|---|---|
| 1,290 | 1,182 |
| 1,350 | 1,172 |
| 1,470 | 1,264 |
| 1,600 | 1,493 |
| 1,710 | 1,571 |
| 1,840 | 1,711 |
| 1,980 | 1,804 |
| 2,230 | 1,840 |
| 2,400 | 1,956 |
| 2,930 | 1,954 |
| 2,710 | 2,007 |
| 3,000 | 1,960 |
| 3,210 | 2,001 |
| 3,240 | 1,928 |
| 3,520 | 1,945 |

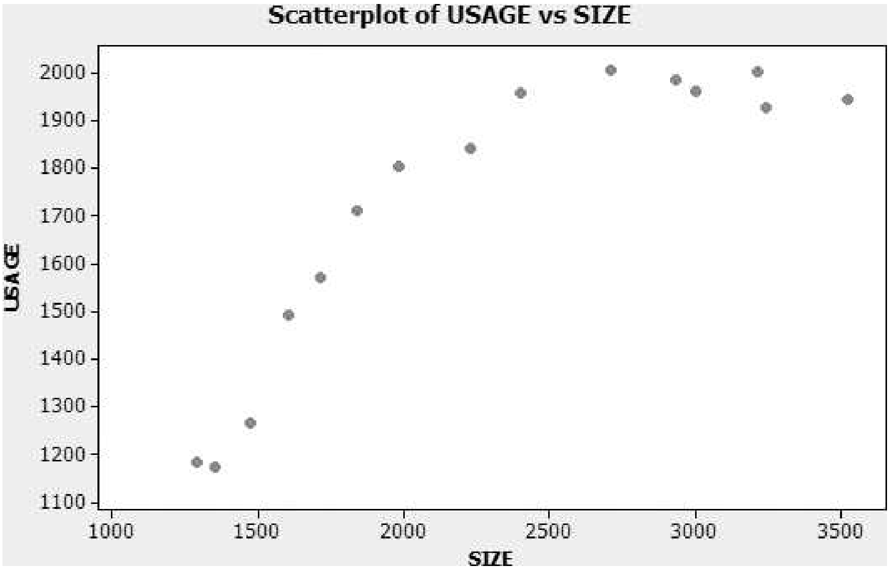



**FIGURE 11.12**
MINITAB scatterplot for electrical usage data

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: USAGE

| | |
|---|---|
| Number of Observations Read | 15 |
| Number of Observations Used | 15 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 1300900 | 650450 | 258.11 | <.0001 |
| Error | 12 | 30240 | 2520.01514 | | |
| Corrected Total | 14 | 1331140 | | | |

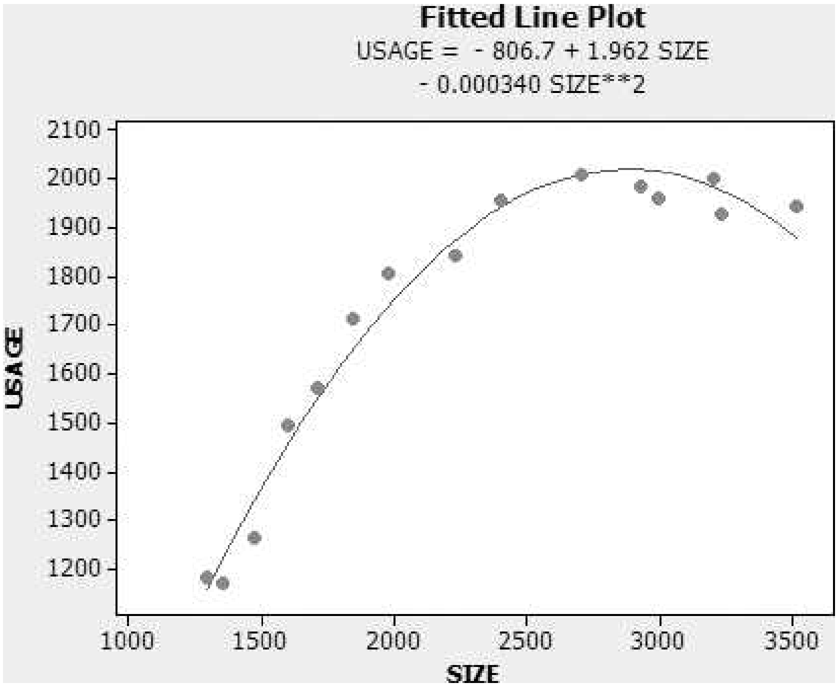
| | | | |
|---|---|---|---|
| Root MSE | 50.19975 | R-Square | 0.9773 |
| Dependent Mean | 1721.20000 | Adj R-Sq | 0.9735 |
| Coeff Var | 2.91656 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -806.71666 | 166.87209 | -4.83 | 0.0004 |
| SIZE | 1 | 1.96162 | 0.15252 | 12.86 | <.0001 |
| SIZESQ | 1 | -0.00034044 | 0.00003212 | -10.60 | <.0001 |

**FIGURE 11.13**
SAS Regression Output for Electrical Usage Data

**FIGURE 11.14**
MINITAB Plot of Least Squares Model for Electrical Usage




d. The interpretation of the estimated coefficients in a quadratic model must be undertaken cautiously. First, the estimated $y$-intercept, $\hat{\beta}_0$, can be meaningfully interpreted only if the range of the independent variable includes zero—that is, if $x = 0$ is included in the sampled range of $x$. Although $\hat{\beta}_0 = -806.72$ seems to imply that the estimated electrical usage is negative when $x = 0$, this zero point is not in the range of the sample (the lowest value of $x$ is 1,290 square feet), and the value is nonsensical (a home with 0 square feet); thus the interpretation of $\hat{\beta}_0$ is not meaningful.

The estimated coefficient of $x$ is $\hat{\beta}_1 = 1.962$, but it no longer represents a slope in the presence of the quadratic term $x^2$.* The estimated coefficient of the first-order term $x$ will not, in general, have a meaningful interpretation in the quadratic model. Consequently, there is no need to conduct any inferential statistical analyses (e.g., confidence interval or test) on $\beta_1$.

The sign of the coefficient, $\hat{\beta}_2 = -.00034$, of the quadratic term, $x^2$, is the indicator of whether the curve is concave downward (mound-shaped) or concave upward (bowl-shaped). A negative $\hat{\beta}_2$ implies downward concavity, as in this example (Figure 11.14), and a positive $\hat{\beta}_2$ implies upward concavity. Rather than interpreting the numerical value of $\hat{\beta}_2$ itself, we utilize a graphical representation of the model, as in Figure 11.14, to describe the model.

Note that Figure 11.14 implies that the estimated electrical usage is leveling off as the home sizes increase beyond 2,500 square feet. In fact, the concavity of the model would lead to decreasing usage estimates if we were to display the model out to 4,000 square feet and beyond (see Figure 11.15). However, model interpretations are not meaningful outside the range of the independent variable, which has a maximum value of 3,520 square feet in this example. Thus, although the model

*For students with knowledge of calculus, note that the slope of the quadratic model is the first derivative $\partial y/\partial x = \beta_1 + 2\beta_2 x$. Thus, the slope varies as a function of $x$, rather than the constant slope associated with the straight-line model.

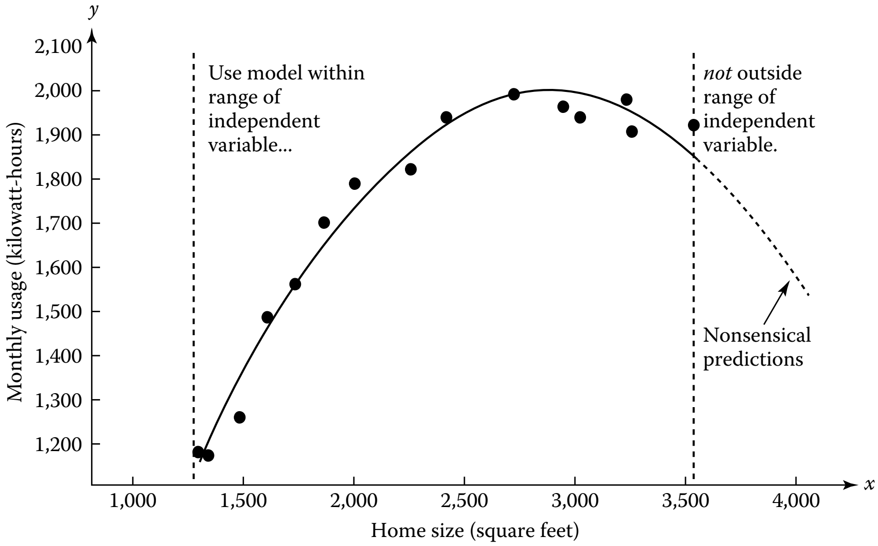



**FIGURE 11.15**
Potential misuse of quadratic model

appears to support the hypothesis that the *rate of increase* per square foot *decreases* for the home sizes near the high end of the sampled values, the conclusion that usage will actually begin to decrease for very large homes would be a *misuse* of the model, since no homes of 4,000 square feet or more were included in the sample.

e. To test whether the quadratic model is statistically useful, we conduct the global $F$ test:

$H_0$: $\beta_1 = \beta_2 = 0$
$H_a$: At least one of the above coefficients is nonzero

From the SAS printout, Figure 11.13, the test statistic is $F = 258.11$ with an associated $p$-value $< .0001$. For any reasonable $\alpha$, we reject $H_0$ and conclude that the overall model is a useful predictor of electrical usage, $y$.

f. Figure 11.14 shows concave downward curvature in the relationship between size of a home and electrical usage in the sample of 10 data points. To determine if this type of curvature exists in the population, we want to test

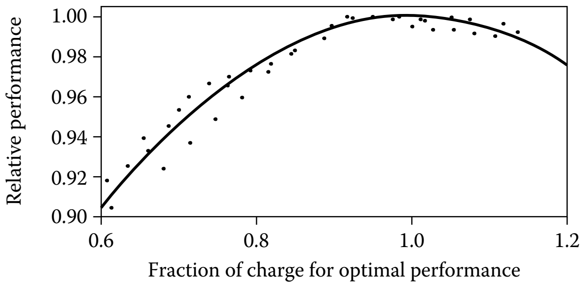
$H_0$: $\beta_2 = 0$ (no curvature in the response curve)
$H_a$: $\beta_2 < 0$ (downward concavity exists in the response curve)

The value of the test statistic for testing $\beta_2$, highlighted on the printout, is $T = -10.60$, and the associated two-tailed $p$-value is .0001. Since this is a one-tailed test, the appropriate $p$-value is $.0001/2 = .00005$. Now $\alpha = .01$ exceeds this $p$-value. Thus, there is very strong evidence of downward curvature in the population: that is, electrical usage increases more slowly per square foot for large homes than for small homes.

(*Note:* The SAS printout in Figure 11.13 also provides the $T$ test statistic and corresponding two-tailed $p$-values for the tests of $H_0$: $\beta_0 = 0$ and $H_0$: $\beta_1 = 0$. Since the interpretation of these parameters is not meaningful for this model, the tests are not of interest.)

## Applied Exercises

11.36 *Commercial refrigeration systems.* The role of maintenance in energy saving in commercial refrigeration was the topic of an article in the *Journal of Quality in Maintenance Engineering* (Vol. 18, 2012). The authors provided the following illustration of data relating the efficiency (relative performance) of a refrigeration system to the fraction of total charges for cooling the system required for optimal performance. Based on the data shown in the graph, hypothesize an appropriate model for relative performance ($y$) as a function of fraction of charge ($x$). What are the hypothesized signs (positive or negative) of the $\beta$-parameters in the model?




11.37 *Estimating repair and replacement costs of water pipes.* Refer to the *IHS Journal of Hydraulic Engineering* (September, 2012) study of the repair and replacement of water pipes, Exercise 10.8 (p. 497). Recall that a team of civil engineers used regression analysis to model $y$ = the ratio of repair to replacement cost of commercial pipe as a function of $x$ = the diameter (in millimeters) of the pipe. Data for a sample of 13 different pipe sizes are reproduced in the accompanying table. In Exercise 10.8, you fit a straight-line model to the data. Now consider the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. A MINITAB printout of the analysis follows.

a. Give the least squares prediction equation relating ratio of repair to replacement cost ($y$) to pipe diameter ($x$).
b. Conduct a global $F$-test for the model using $\alpha = .01$. What do you conclude about overall model adequacy?
c. Evaluate the adjusted coefficient of determination, $R_a^2$ for the model.
d. Give the null and alternative hypotheses for testing if the rate of increase of ratio ($y$) with diameter ($x$) is slower for larger pipe sizes.
e. Carry out the test, part **d**, using $\alpha = .01$.
f. Locate, on the printout, a 95% prediction interval for the ratio of repair to replacement cost for a pipe with a diameter of 240 millimeters. Interpret the result

11.38 *Monitoring impedance to leg movements.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) experiment to monitor the impedance to leg movement, Exercise 2.46 (p. 51). Recall that engineers attached electrodes to the ankles and knees of volunteers and measured the signal-to-noise ratio (SNR) of impedance changes.

**WATERPIPE**

| DIAMETER | RATIO |
|---|---|
| 80 | 6.58 |
| 100 | 6.97 |
| 125 | 7.39 |
| 150 | 7.61 |
| 200 | 7.78 |
| 250 | 7.92 |
| 300 | 8.20 |
| 350 | 8.42 |
| 400 | 8.60 |
| 450 | 8.97 |
| 500 | 9.31 |
| 600 | 9.47 |
| 700 | 9.72 |

*Source:* Suribabu, C.R. & Neelakantan, T.R. "Sizing of water distribution pipes based on performance measure and breakage-repair replacement economics", *IHS Journal of Hydraulic Engineering,* Vol. 18, No. 3, September 2012 (Table 1).

**MINITAB Output for Exercise 11.37**

**Regression Analysis: RATIO versus DIAMETER, DIAMSQ**

```
The regression equation is
RATIO = 6.27 + 0.00791 DIAMETER - 0.000004 DIAMSQ


Predictor          Coef     SE Coef      T      P
Constant         6.2660      0.1554  40.31  0.000
DIAMETER      0.0079145   0.0009974   7.93  0.000
DIAMSQ      -0.00000426  0.00000132  -3.23  0.009


S = 0.162211   R-Sq = 97.7%   R-Sq(adj) = 97.2%


Analysis of Variance

Source          DF       SS      MS       F      P
Regression       2  11.0810  5.5405  210.56  0.000
Residual Error  10   0.2631  0.0263
Total           12  11.3441


Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI              95% PI
      1  7.9785  0.0580  (7.8494, 8.1077)  (7.5947, 8.3624)


Values of Predictors for New Observations

New Obs  DIAMETER  DIAMSQ
      1       250   62500
```

For the optimum ankle-knee electrode pair, the engineers examined the relationship between knee joint angle, $x$ (degrees), and knee impedance change, $y$ (ohms). A quadratic model was fit to the data with the following results:

$$\hat{y} = 24.83 + .041x + .0005x^2, \qquad R^2 = .903$$

a. Sketch the least-squares prediction equation. Identify the nature of the curvature estimated by the model.
b. Predict the knee impedance change ($y$) for an electrode pair with a knee joint angle of $x = 50$ degrees.
c. Interpret the value of $R^2$.

11.39 *Estimating change-point dosage.* A standard method for studying toxic substances and their effects on humans is to observe the responses of rodents exposed to various doses of the substance over time. In the *Journal of Agricultural, Biological, and Environmental Statistics* (June 2005), researchers used least-squares regression to estimate the "change-point" dosage—defined as the largest dose level that has no adverse effects. Data were obtained from a dose-response study of rats exposed to the toxic substance aconiazide. A sample of 50 rats was evenly divided into five dosage groups: 0, 100, 200, 500, and 750 milligrams per kilograms of body weight. The dependent variable $y$ measured was the weight change (in grams) after a 2-week exposure. The researchers fit the quadratic model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $x$ = dosage level, with the following results: $\hat{y} = 10.25 + .0053x - .0000266x^2$.

a. Construct a rough sketch of the least-squares prediction equation. Describe the nature of the curvature in the estimated model.
b. Estimate the weight change ($y$) for a rat given a dosage of 500 mg/kg of aconiazide.
c. Estimate the weight change ($y$) for a rat given a dosage of 0 mg/kg of aconiazide. (This dosage is called the "control" dosage level.)
d. Find the smallest dosage level $x$ that yields an estimated weight change below the estimated weight change for the control group. This value is the "change-point" dosage. [*Hint:* Find the value of $x$ for which $E(y|x) < E(y|x = 0)$.]

11.40 *Failure times of silicon wafer microchips.* Researchers at National Semiconductor experimented with tin-lead solder bumps used to manufacture silicon wafer integrated circuit chips. (International Wafer Level Packaging Conference, Nov. 3-4, 2005.) The failure times of the microchips (in hours) was determined at different solder temperatures (degrees Centigrade). The data for one experiment are given in the table. The researchers want to predict failure time ($y$) based on solder temperature ($x$).

a. Construct a scatterplot for the data. What type of relationship, linear or curvilinear, appears to exist between failure time and solder temperature?
b. Fit the model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, to the data. Give the least squares prediction equation.
c. Conduct a test to determine if there is upward curvature in the relationship between failure time and solder temperature. (Use $\alpha = .05$.)

**WAFER**

| Temperature (°C) | Time to Failure (hours) |
|---|---|
| 165 | 200 |
| 162 | 200 |
| 164 | 1200 |
| 158 | 500 |
| 158 | 600 |
| 159 | 750 |
| 156 | 1200 |
| 157 | 1500 |
| 152 | 500 |
| 147 | 500 |
| 149 | 1100 |
| 149 | 1150 |
| 142 | 3500 |
| 142 | 3600 |
| 143 | 3650 |
| 133 | 4200 |
| 132 | 4800 |
| 132 | 5000 |
| 134 | 5200 |
| 134 | 5400 |
| 125 | 8300 |
| 123 | 9700 |

*Source:* Gee, S. & Nguyen, L. "Mean time to failure in wafer level-CSP packages with SnPb and SnAgCu solder bmps", International Wafer Level Packaging Conference, San Jose, CA, Nov. 3-4, 2005 (adapted from Figure 7).

11.41 *Planning an ecological network.* Refer to the *Landscape Ecology Engineering* (Jan. 2013) study of a new method of planning an ecological network, Exercise 10.37 (p. 511). Based on a sample of 21 bird habitats in China, the researchers modeled $y$ = the bird density (number of birds per hectare) as a linear function of $x$ = the percentage of the habitat covered by vegetation (i.e., a green area). Data similar to the data reported in the journal article are reproduced in the table on p. 604. Suppose the researchers want to know if there is a curvilinear relationship between $y$ and $x$. Specifically, is there evidence to indicate that the rate of increase of bird density with percent vegetation coverage is steeper for greener habitats (i.e., habitats with a greater percentage of vegetation). Conduct the appropriate analysis to answer the researchers' question.

### Data for Exercise 11.41

**BIRDDEN**

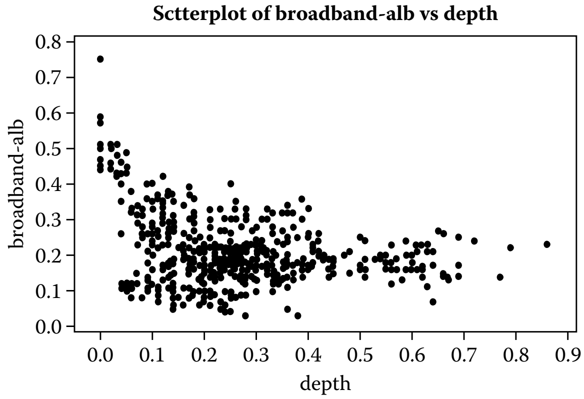
| HABITAT | DENSITY (birds/hectare) | COVER (%) |
|---|---|---|
| 1 | 0.3 | 0 |
| 2 | 0.25 | 2 |
| 3 | 2 | 4 |
| 4 | 1 | 6 |
| 5 | 0.5 | 9 |
| 6 | 0 | 10 |
| 7 | 3 | 12 |
| 8 | 5 | 17 |
| 9 | 5 | 20 |
| 10 | 1 | 25 |
| 11 | 6 | 30 |
| 12 | 5 | 37 |
| 13 | 8 | 40 |
| 14 | 2 | 45 |
| 15 | 7 | 50 |
| 16 | 16 | 58 |
| 17 | 5 | 60 |
| 18 | 20 | 71 |
| 19 | 5 | 80 |
| 20 | 37 | 90 |
| 21 | 6 | 100 |

11.42 *Catalytic converters in cars.* A quadratic model was applied to motor vehicle toxic emissions data collected over a 15-year period in Mexico City. (*Environmental Science & Engineering*. Sept. 1, 2000.) The following equation was used to predict the percentage ($y$) of motor vehicles without catalytic converters in the Mexico City fleet for a given year ($x$): $\hat{y} = 325{,}790 - 321.67x + 0.794x^2$.

a. Explain why the value $\hat{\beta}_0 = 325{,}790$ has no practical interpretation.
b. Explain why the value $\hat{\beta}_1 = -321.67$ should not be interpreted as a slope.
c. Examine the value of $\hat{\beta}_2$ to determine the nature of the curvature (upward or downward) in the sample data.
d. The researchers used the model to estimate "that just after the year 2021 the fleet of cars with catalytic converters will completely disappear." Comment on the danger of using the model to predict $y$ in the year 2021.

**PONDICE**

11.43 *Characteristics of sea ice meltponds.* Surface albedo is defined as the ratio of solar energy directed upward from a surface over energy incident upon the surface. Surface albedo is a critical climatological parameter of sea ice. The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice meltponds in the Canadian Arctic. (See Example 2.1, p. 25). Data for 504 ice meltponds located in the Barrow Strait in the Canadian Arctic are saved in the **PONDICE** file. Environmental engineers want to examine the relationship between the broadband surface albedo level, $y$, of the ice and pond depth, $x$ (meters).




**MINITAB scatterplot for Exercise 11.43**

a. A MINITAB scatterplot for the data is shown above. Based on the scatterplot, hypothesize a model for $E(y)$ as a function of $x$.
b. Fit the model, part **a**, to the data. Give the least-squares prediction equation.
c. Conduct a test of overall model adequacy using $\alpha = .01$.
d. Conduct tests (at $\alpha = .01$) on any important $\beta$ parameters in the model.
e. Find and interpret the values of adjusted $R^2$ and $s$.

11.44 *Forest fragmentation study.* Refer to the *Conservation Ecology* (Dec. 2003) study of the causes of fragmentation for 54 South American forests, Exercise 10.39 (p. 512). Recall that ecologists have developed two fragmentation indices for each forest—one index for anthropogenic fragmentation ($y$) and one for fragmentation from natural causes ($x$). Data on these two indices for all 54 forests are saved in the **FORFRAG** file. (The first five observations are reproduced in the table on p. 605.) In Exercise 10.33 you fit a simple linear model to the data, after removing data for the three forests with the largest anthropogenic indices. Now consider a quadratic model for $E(y)$.

a. Fit the quadratic model to all the data using the method of least–squares. Give the equation of the least-squares prediction equation.
b. Interpret the estimates of $\beta_0$, $\beta_1$, and $\beta_2$ in the context of the problem.
c. Is there sufficient evidence of a curvilinear relationship between natural origin index ($x$) and anthropogenic index ($y$)? Test using $\alpha = .05$.

**Data for Exercise 10.44**

**FORFRAG**

(*First five observations listed*)

| Ecoregion (forest) | Anthropogenic Index, $y$ | Natural Origin Index, $x$ |
|---|---|---|
| Araucaria moist forests | 34.09 | 30.08 |
| Atlantic Coast restingas | 40.87 | 27.60 |
| Bahia coastal forests | 44.75 | 28.16 |
| Bahia interior forests | 37.58 | 27.44 |
| Bolivian Yungas | 12.40 | 16.75 |

*Source:* Wade, T. G., et al. "Distribution and causes of global forest fragmentation." *Conservation Ecology*, Vol. 72, No. 2, Dec. 2003 (Table 6).

11.45 *Kinetics of fluorocarbon plasmas.* Fluorocarbon plasmas are used in the production of semiconductor materials. In the *Journal of Applied Physics* (Dec. 1, 2000), electrical engineers at Nagoya University (Japan) studied the kinetics of fluorocarbon plasmas in order to optimize material processing. In one portion of the study, the surface production rate of fluorocarbon radicals emitted from the production process was measured at various points in time (in milliseconds) after the radio frequency power was turned off. The data are given in the accompanying table. Consider a model relating surface production rate ($y$) to time ($x$).

**RADICALS**

| Rate | Time | Rate | Time |
|---|---|---|---|
| 1.00 | 0.1 | 0.00 | 1.7 |
| 0.80 | 0.3 | −0.10 | 1.9 |
| 0.40 | 0.5 | −0.15 | 2.1 |
| 0.20 | 0.7 | −0.05 | 2.3 |
| 0.05 | 0.9 | −0.13 | 2.5 |
| 0.00 | 1.1 | −0.08 | 2.7 |
| −0.05 | 1.3 | 0.00 | 2.9 |
| −0.02 | 1.5 | | |

*Source:* Takizawa, K., et al. "Characteristics of $C_3$ radicals in high-density $C_4F_8$ plasmas studied by laser-induced fluorescence spectroscopy." *Journal of Applied Physics*, Vol. 88, No. 11, Dec. 1, 2000 (Figure 7).

a. Graph the data in a scattergram. What trend do you observe?
b. Fit a quadratic model to the data. Give the least-squares prediction equation.
c. Is there sufficient evidence of upward curvature in the relationship between surface production rate and time after turn off? Use $\alpha = .05$.

## 11.10 Regression Residuals and Outliers

In Section 10.9 we demonstrated the importance of a residual analysis for checking whether the assumptions about the random error $\varepsilon$ in the straight-line model, $y = \beta_0 + \beta_1 x + \varepsilon$, are reasonably satisfied. Recall that a regression *residual* is defined as the difference between the actual value of $y$ and its corresponding predicted value, i.e., $(y - \hat{y})$. Conducting a residual analysis of a multiple regression model is equally as important. The residual plots and graphs presented in Section 10.9 apply also to a multiple regression model. A checklist of these residual plots and associated assumptions, as well as the recommended model modification if an assumption is violated, is provided in Table 11.7.

**Example 11.13**

**Checking Assumptions for a Multiple Regression Model**

Refer to Example 11.10 (p. 581) and the multiple regression model for $y$ = total number of man-hours worked per day by a member of the clerical staff of a large department store. The independent variables used in the model were: $x_1$ = number of pieces of mail processed, $x_2$ = number of gift cards sold, $x_3$ = number of store charge accounts transacted, $x_4$ = number of change order transactions or returns processed, and $x_5$ = number of checks cashed. The first-order model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$, was fit to data collected for 52 work days. (The data are listed in Table 11.4.) Conduct a residual analysis for the model in order to determine if the least squares assumptions on the random error term are reasonably satisfied.

Solution

We analyzed the model using MINITAB. A MINITAB printout of the model, including a list of residuals (shaded), is shown in Figure 11.16.

*Assumption of mean error of 0:* To check the first assumption, we used MINITAB to plot the residuals against each of the five independent variables in the model. These

**TABLE 11.7 Using Residuals to Check Assumptions in Multiple Regression**

| Assumption on $\varepsilon$ | Residual Plot/Graph | Violation | Model Modification |
|---|---|---|---|
| $E(\varepsilon) = 0$ | Plot residuals vs. each $x$ in model | Pattern in graph (i.e., a curvilinear trend) indicating a misspecified model | Add term(s) to model to account for pattern (e.g., add $x^2$) |
| $V(\varepsilon) = \sigma^2$ constant | Plot residuals vs. $\hat{y}$ | Pattern in graph (i.e., cone shape, football shape) [See Figure 10.22] | Use a variance-stabilizing transformation on $y$, e.g., $\ln(y)$, $\sqrt{y}$, etc. [See Table 10.8] |
| Normal distribution | Histogram, stem-leaf plot, or normal probability plot of residuals | Highly skewed distribution | Use a normalizing transformation on $y$, e.g., $\ln(y)$, $\sqrt{y}$, etc. [Reminder, regression is *robust* with respect to nonnormal data] |
| Independent | Plot residuals vs. time-series variable (if data is recorded sequentially over time) | Long runs of positive residuals followed by long runs of negative residuals [See Figure 10.28] | Use a time-series model that accounts for residual correlation |

graphs are shown in Figure 11.17. There do not appear to be any strong trends (e.g., no curvilinear trends) in the residuals, indicating that the model is not misspecified (at least with respect to curvature). Thus, the assumption of $E(\varepsilon) = 0$ appears to be reasonably satisfied.

**FIGURE 11.16a**
MINITAB Printout for Model of Man-Hours, Example 11.13

**Regression Analysis: Y-Hours versus X1-Mail, X2-Gifts, ...**

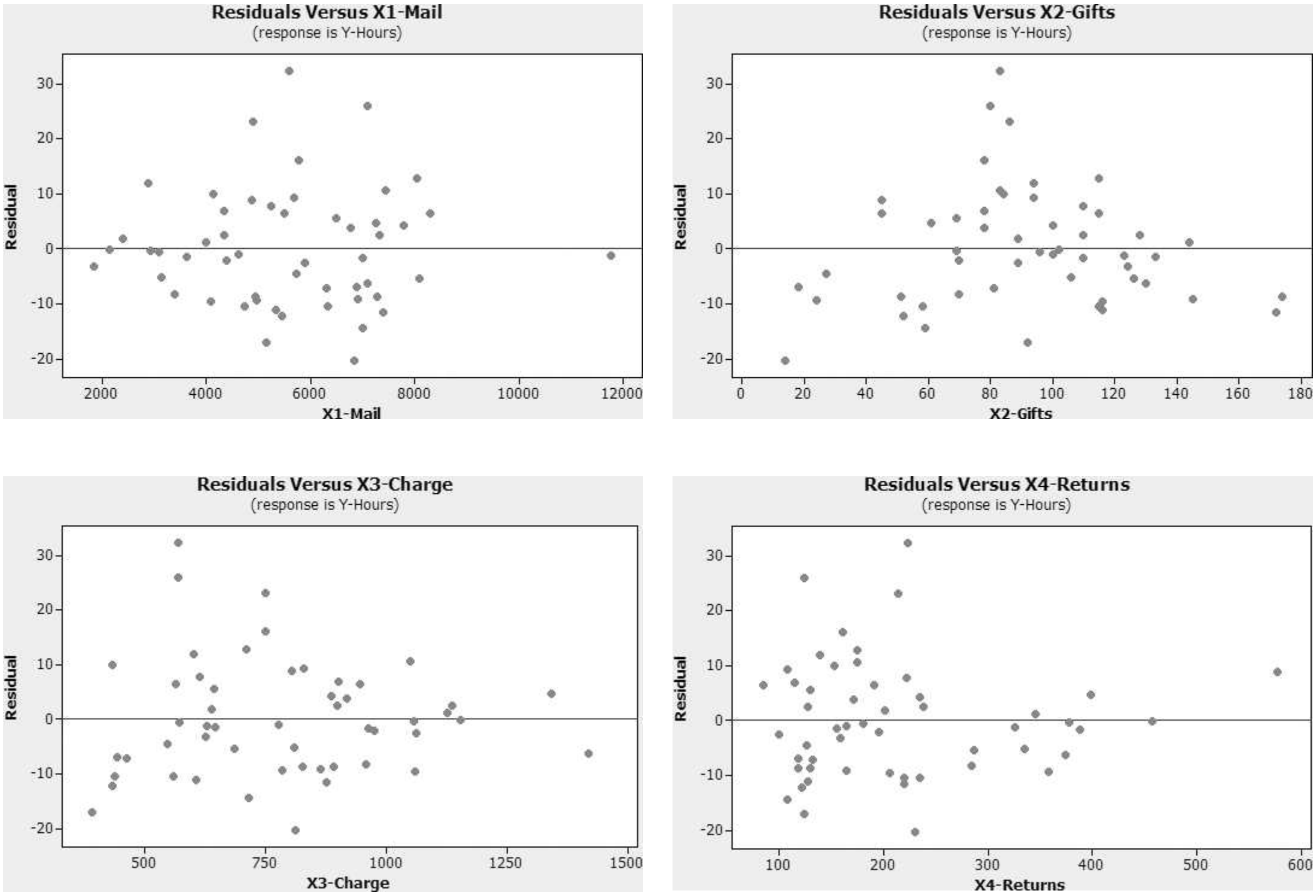
```
The regression equation is
Y-Hours = 66.3 + 0.00124 X1-Mail + 0.116 X2-Gifts + 0.0127 X3-Charge
          - 0.0455 X4-Returns + 0.0562 X5-Checks


Predictor        Coef     SE Coef      T      P
Constant       66.265       9.101   7.28  0.000
X1-Mail     0.0012418   0.0008879   1.40  0.169
X2-Gifts      0.11581     0.04608   2.51  0.016
X3-Charge    0.012685    0.008708   1.46  0.152
X4-Returns   -0.04549     0.01755  -2.59  0.013
X5-Checks     0.05616     0.01096   5.12  0.000


S = 11.1668   R-Sq = 53.4%   R-Sq(adj) = 48.3%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       5   6576.4  1315.3  10.55  0.000
Residual Error  46   5736.1   124.7
Total           51  12312.5
```

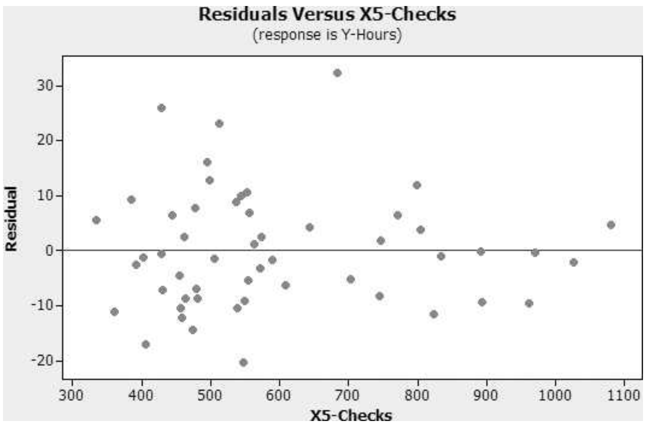
**FIGURE 11.16b**
List of Residuals for Model of Man-Hours, Example 11.13

| Obs | X1-Mail | Y-Hours | Fit | SE Fit | Residual | St Resid |
|---|---|---|---|---|---|---|
| 1 | 7781 | 128.50 | 124.23 | 2.60 | 4.27 | 0.39 |
| 2 | 7004 | 113.60 | 115.33 | 3.52 | -1.73 | -0.16 |
| 3 | 7267 | 146.60 | 141.98 | 5.56 | 4.62 | 0.48 |
| 4 | 2129 | 124.30 | 124.60 | 4.91 | -0.30 | -0.03 |
| 5 | 4878 | 100.40 | 91.63 | 7.07 | 8.77 | 1.01 X |
| 6 | 3999 | 119.20 | 118.13 | 4.42 | 1.07 | 0.10 |
| 7 | 11777 | 109.50 | 110.84 | 6.54 | -1.34 | -0.15 |
| 8 | 5764 | 128.50 | 112.42 | 1.99 | 16.08 | 1.46 |
| 9 | 7392 | 131.20 | 142.74 | 5.31 | -11.54 | -1.17 |
| 10 | 8100 | 112.20 | 117.72 | 3.83 | -5.52 | -0.53 |
| 11 | 4736 | 95.40 | 105.91 | 3.73 | -10.51 | -1.00 |
| 12 | 4337 | 124.60 | 122.20 | 2.91 | 2.40 | 0.22 |
| 13 | 3079 | 103.70 | 104.29 | 3.36 | -0.59 | -0.06 |
| 14 | 7273 | 103.60 | 112.32 | 3.49 | -8.72 | -0.82 |
| 15 | 4091 | 133.20 | 142.83 | 4.07 | -9.63 | -0.93 |
| 16 | 3390 | 111.40 | 119.64 | 2.99 | -8.24 | -0.77 |
| 17 | 6319 | 97.70 | 108.18 | 2.67 | -10.48 | -0.97 |
| 18 | 7447 | 132.10 | 121.59 | 3.37 | 10.51 | 0.99 |
| 19 | 7100 | 135.90 | 109.95 | 2.57 | 25.95 | 2.39R |
| 20 | 8035 | 131.30 | 118.61 | 2.91 | 12.69 | 1.18 |
| 21 | 5579 | 150.40 | 118.23 | 2.87 | 32.17 | 2.98R |
| 22 | 4338 | 124.90 | 118.10 | 3.27 | 6.80 | 0.64 |
| 23 | 6895 | 97.00 | 104.05 | 4.09 | -7.05 | -0.68 |
| 24 | 3629 | 114.10 | 115.65 | 3.26 | -1.55 | -0.15 |
| 25 | 5149 | 88.30 | 105.35 | 3.12 | -17.05 | -1.59 |
| 26 | 5241 | 117.60 | 109.97 | 2.49 | 7.63 | 0.70 |
| 27 | 2917 | 128.20 | 128.57 | 4.16 | -0.37 | -0.04 |
| 28 | 4390 | 138.80 | 140.99 | 4.55 | -2.19 | -0.21 |
| 29 | 4957 | 109.50 | 119.00 | 4.51 | -9.50 | -0.93 |
| 30 | 7099 | 118.90 | 125.33 | 5.18 | -6.43 | -0.65 |
| 31 | 7337 | 122.20 | 119.69 | 3.96 | 2.51 | 0.24 |
| 32 | 8301 | 142.80 | 136.50 | 3.92 | 6.30 | 0.60 |
| 33 | 4889 | 133.90 | 110.89 | 1.96 | 23.01 | 2.09R |
| 34 | 6308 | 100.20 | 107.47 | 2.72 | -7.27 | -0.67 |
| 35 | 6908 | 116.80 | 125.97 | 3.14 | -9.17 | -0.86 |
| 36 | 5345 | 97.30 | 108.44 | 2.87 | -11.14 | -1.03 |
| 37 | 6994 | 98.00 | 112.54 | 2.90 | -14.54 | -1.35 |
| 38 | 6781 | 136.50 | 132.78 | 3.34 | 3.72 | 0.35 |
| 39 | 3142 | 111.70 | 116.89 | 3.25 | -5.19 | -0.49 |
| 40 | 5738 | 98.60 | 103.27 | 3.50 | -4.67 | -0.44 |
| 41 | 4931 | 116.20 | 124.99 | 4.37 | -8.79 | -0.86 |
| 42 | 6501 | 108.90 | 103.38 | 2.95 | 5.52 | 0.51 |
| 43 | 5678 | 120.60 | 111.40 | 3.19 | 9.20 | 0.86 |
| 44 | 4619 | 131.80 | 132.82 | 3.28 | -1.02 | -0.10 |
| 45 | 1832 | 112.40 | 115.72 | 4.02 | -3.32 | -0.32 |
| 46 | 5445 | 92.50 | 104.75 | 3.05 | -12.25 | -1.14 |
| 47 | 4123 | 120.00 | 110.19 | 3.05 | 9.81 | 0.91 |
| 48 | 5884 | 112.20 | 114.75 | 4.73 | -2.55 | -0.25 |
| 49 | 5505 | 113.00 | 106.56 | 3.20 | 6.44 | 0.60 |
| 50 | 2882 | 138.70 | 126.90 | 3.90 | 11.80 | 1.13 |
| 51 | 2395 | 122.10 | 120.44 | 3.42 | 1.66 | 0.16 |
| 52 | 6847 | 86.60 | 106.92 | 4.09 | -20.32 | -1.96 |

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

**FIGURE 11.17**
MINITAB Residual Plots for Checking Assumption #1, Example 11.13



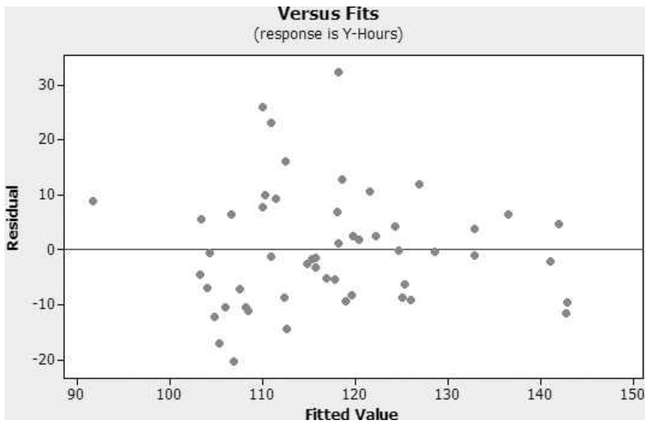
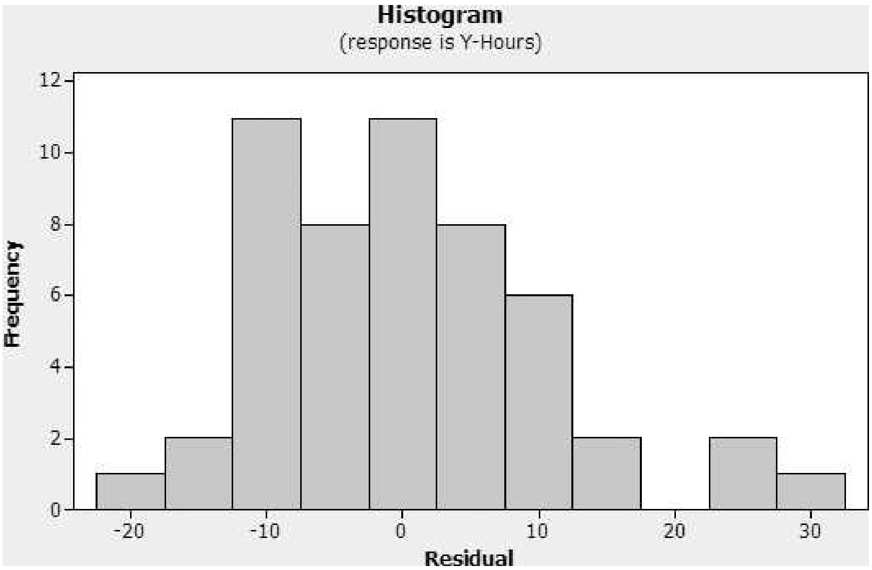



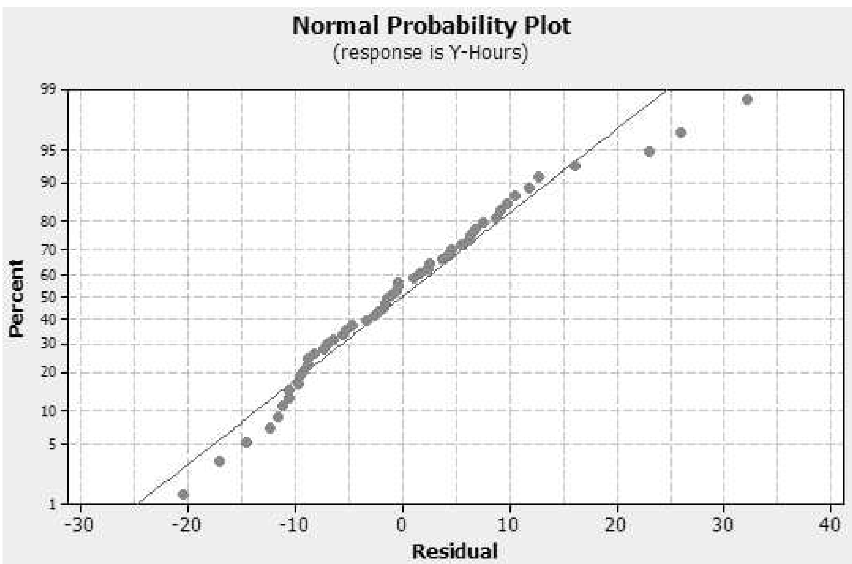

**FIGURE 11.18**
MINITAB Residual Plot for Checking Assumption #2, Example 11.13

*Assumption of constant error variance:* To check the second assumption, we used MINITAB to plot the residuals against predicted man-hours, $\hat{y}$. This graph is shown in Figure 11.18. Since the dependent variable is number of man-hours in a day, it would likely follow a Poisson distribution. Recall from Section 10.8, the residual plot pattern for Poisson data that violate this assumption is a bullet shape, with the variance in the residuals increasing as $\hat{y}$ increases. The graph shows no evidence of this type of pattern. In fact, the points seem to be randomly scattered. Consequently, there is no need for a variance-stabilizing transformation on $y$—the assumption constant error variance appears to be reasonably satisfied.

*Assumption of normal errors:* To check the third assumption, we used MINITAB to generate both a histogram and normal probability plot for the residuals. These graphs are shown in Figure 11.19. The histogram is mound-shaped and the points on the normal probability plot fall nearly in a straight line. These graphs, coupled with the fact that regression is robust for small to moderate departures from normality, lead us to conclude that the assumption of normal errors is reasonably satisfied.

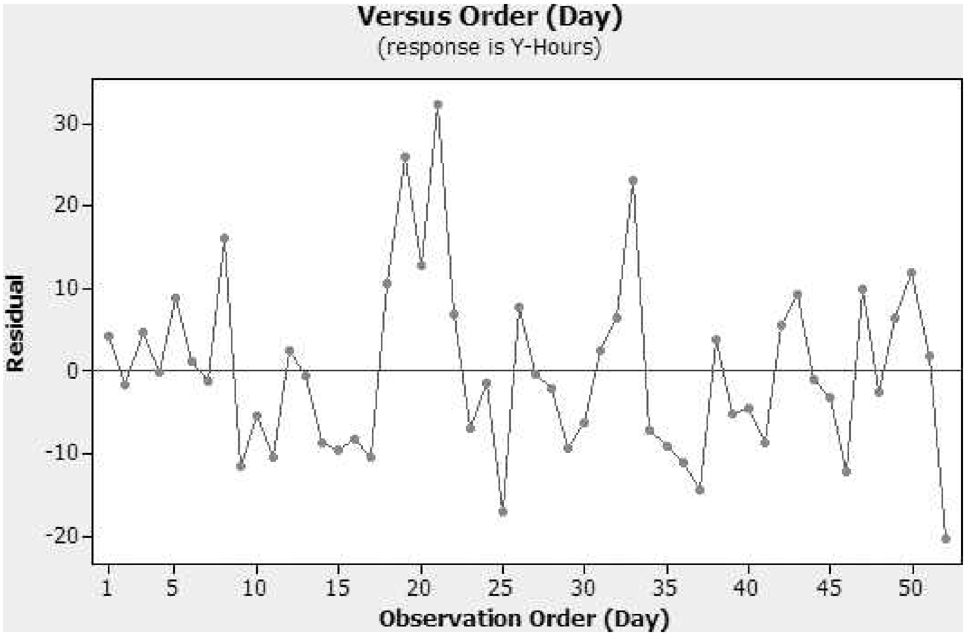
**FIGURE 11.19**
MINITAB Residual Plots for Checking Assumption #3, Example 11.13



*Assumption of independent errors:* To check the fourth assumption, we used MINITAB to plot the residuals in time order. This is possible since the data is collected over 52 consecutive days. The graph is shown in Figure 11.20. If the residuals were strongly positively correlated, we would see residuals for consecutive days that tend to have the same sign (i.e., both positive or both negative). That is, there would be a long run of positive residuals followed by a long run of negative residuals, followed by another long run of positive residuals, etc. There does not appear to be any evidence of this type of residual correlation in the graph. Consequently, the assumptions of independent errors is reasonably satisfied.

**FIGURE 11.20**
MINITAB Residual Plot for Checking Assumption #4, Example 11.13




In addition to checking assumptions, residuals can also be used to detect **outliers** and **influential observations**. Outliers are values of $y$ that appear to be in disagreement with the model. Since almost all values of $y$ should lie within $3\sigma$ of $E(y)$, the mean value of $y$, we would expect most of them to lie within $3s$ of $\hat{y}$. Here, it is helpful to consider **standardized residuals**. A standardized residual is a residual value divided by $s$. If a residual is larger than $3s$ (in absolute value), or, equivalently, a standardized residual is larger than 3 (in absolute value), we consider it an outlier and seek background information that might explain the reason for its large value.

**Definition 11.3**

A standardized residual for the ith observation (denoted $z_i$) is computed by dividing the corresponding residual by $s$, i.e.,

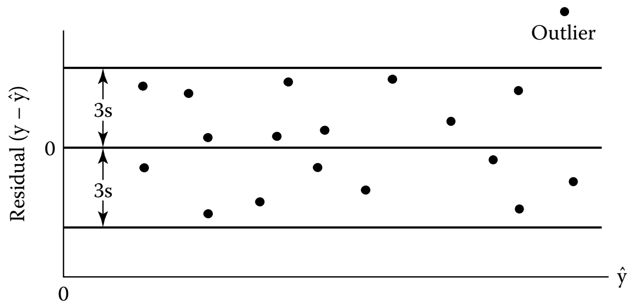
$$z_i = (y_i - \hat{y}_i)/s$$

**Definition 11.4**

A residual that is larger than $3s$ (in absolute value), or a standardized residual that is larger than 3 (in absolute value), is considered to be an **outlier**.

To detect outliers, we can construct horizontal lines located a distance of $3s$ above and below 0 (see Figure 11.21) on a residual plot. Any residual falling outside the

**FIGURE 11.21**
3$s$ lines used to locate outliers




band formed by these lines would be considered an outlier. We would then initiate an investigation to seek the cause of the departure of such observations from expected behavior.

Although some analysts advocate elimination of outliers, regardless of whether cause can be assigned, others encourage the correction of only those outliers that can be traced to specific causes. The best philosophy is probably a compromise between these extremes. For example, before deciding the fate of an outlier you may want to determine how much influence it has on the regression analysis. When an accurate outlier (i.e., an outlier that is not due to recording or measurement error) is found to have a dramatic effect on the regression analysis, it may be the model and not the outlier that is suspect. Omission of important independent variables or higher-order terms could be the reason why the model is not predicting well for the outlying observation.

Several sophisticated numerical techniques are available for identifying outlying **influential** observations. One of these methods requires that you delete observations one at a time, each time refitting the regression model based on only the remaining $n - 1$ observations. This method is based on a statistical procedure, called the **jackknife**,* that is gaining increasing acceptance among practitioners. The basic principle of the jackknife when applied to regression is to compare the regression results using all $n$ observations to the results with the $i$th observation deleted, to ascertain how much influence a particular observation has on the analysis. Using the jackknife, several alternative influence measures can be calculated.

The **deleted residual**, $d_i = y_i - \hat{y}_{(i)}$, measures the difference between the observed value $y_i$ and the predicted value $\hat{y}_{(i)}$, based on the model with the $i$th observation deleted. [The notation $(i)$ is generally used to indicate that the observed value $y_i$ was deleted from the regression analysis.] An observation with an unusually large (in absolute value) deleted residual is considered to have large influence on the fitted model.

**Definition 11.5**

A **deleted residual** (denoted $d_i$) is the difference between the observed response $y_i$ and the predicted value $\hat{y}_{(i)}$, obtained when the data for the ith observation is deleted from the analysis, i.e.,

$$d_i = y_i - \hat{y}_{(i)}$$

**Definition 11.6**

An observation with an unusually large (in absolute value) deleted residual is considered to be an **influential observation**. [Note: Deleted residuals larger than 3$s$ in absolute value are considered "unusually large".]

*The procedure derives its name from the Boy Scout jackknife, which serves as a handy tool in a variety of situations when specialized techniques may not be applicable. [See Belsley, Kuh, and Welsch (1980).]

A measure closely related to the deleted residual is the difference between the predicted value based on the model fit to all $n$ observations and the predicted value obtained when $y_i$ is deleted, i.e., $\hat{y}_i - \hat{y}_{(i)}$. When the difference $\hat{y}_i - \hat{y}_{(i)}$ is large relative to the predicted value $\hat{y}_i$, the observation $y_i$ is said to influence the regression fit.

A third way to identify an influential observation using the jackknife is to calculate, for each $\beta$ parameter in the model, the difference between the parameter estimate based on all $n$ observations and the estimate based on only $n - 1$ observations (with the observation in question deleted). Consider, for example, the straight-line model $E(y) = \beta_0 + \beta_1 x$. The differences $\hat{\beta}_0 - \hat{\beta}_0^{(i)}$ and $\hat{\beta}_1 - \hat{\beta}_1^{(i)}$ measure how influential the $i$th observation $y_i$ is on the parameter estimates. [Using the ($i$) notation defined earlier, $\hat{\beta}_1^{(i)}$ represents the estimate of the $\beta_i$ coefficient when the $i$th observation is omitted from the analysis.] If the parameter estimates change drastically, i.e., if the absolute differences are large, $y_i$ is deemed an influential observation.

### Recommendation

After performing a multiple regression analysis, it is important to check for *outliers* by locating residuals that lie a distance of $3s$ or more above or below 0 on a residual plot versus $\hat{y}$. Before eliminating an outlier from the analysis, you should conduct an investigation to determine its cause. If the outlier is found to be the result of a coding or recording error, fix it or remove it. Otherwise, you may want to determine how *influential* the outlier is before deciding its fate. Several measures of influence are available, including *deleted residuals*.

## Applied Exercises

11.46 *Failure times of silicon wafer microchips.* Refer to the National Semiconductor study of manufactured silicon wafer integrated circuit chips, Exercise 11.40 (p. 603). Recall that the failure times of the microchips (in hours) was determined at different solder temperatures (degrees Centigrade). The data are repeated in the table.

a. Fit the straight-line model $E(y) = \beta_0 + \beta_1 x$ to the data, where $y$ = failure time and $x$ = solder temperature.
b. Compute the residual for a microchip manufactured at a temperature of 152°C.
c. Plot the residuals against solder temperature ($x$). Do you detect a trend?
d. In Exercise 11.40c, you determined that failure time ($y$) and solder temperature ($x$) were curvilinearly related. Does the residual plot, part c, support this conclusion?

**WAFER**

| Temperature (°C) | Time to Failure (hours) | Temperature (°C) | Time to Failure (hours) |
|---|---|---|---|
| 165 | 200 | 149 | 1150 |
| 162 | 200 | 142 | 3500 |
| 164 | 1200 | 142 | 3600 |
| 158 | 500 | 143 | 3650 |
| 158 | 600 | 133 | 4200 |
| 159 | 750 | 132 | 4800 |
| 156 | 1200 | 132 | 5000 |
| 157 | 1500 | 134 | 5200 |
| 152 | 500 | 134 | 5400 |
| 147 | 500 | 125 | 8300 |
| 149 | 1100 | 123 | 9700 |

*Source:* Gee, S. & Nguyen, L. "Mean time to failure in wafer level-CSP packages with SnPb and SnAgCu solder bmps", International Wafer Level Packaging Conference, San Jose, CA, Nov. 3-4, 2005 (adapted from Figure 7).

11.47 *Modeling PCB Concentration.* PCBs make up a family of hazardous chemicals that are often dumped, illegally, by industrial plants into the surrounding streams, rivers, or bays. The table below reports the annual concentrations of PCBs (measured in parts per billion) in water samples collected for two consecutive years from 37 U.S. bays and estuaries. An official from the Environmental Protection Agency wants to model the year 2 PCB concentration ($y$) of a bay as a function of the previous year's PCB concentration ($x$).

a. Fit the first-order model, $E(y) = \beta_0 + \beta_1 x$, to the data. Give the least-squares prediction equation.
b. Is the model adequate for predicting $y$? Explain.
c. Construct a residual plot for the data. Do you detect any outliers? If so, identify them.
d. Refer to part **c**. Although the residual for Boston Harbor is not, by definition, an outlier, the EPA believes that it has strong influence on the regression because of its large $y$ value. Remove the observation for Boston Harbor from the data and refit the model. Has model adequacy improved?
e. An alternative approach is to use the natural log transformations $y^* = \ln(y + 1)$ and $x^* = \ln(x + 1)$, and fit the model $E(y^*) = \beta_0 + \beta_1 x^*$. Fit this model, then conduct a test for model adequacy and perform a residual analysis. Interpret the results. In particular, comment on the residual value for Boston Harbor.

**DDT**

11.48 *Contamination of fish in the Tennessee River.* Refer to the U.S. Army Corps of Engineers data on fish contaminated from the toxic discharges of a chemical plant located on the banks of the Tennessee River in Alabama. In Exercise 11.26 (p. 591) you fit the first-order model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where $y$ = DDT level in captured fish, $x_1$ = miles captured upstream, $x_2$ = fish length, and $x_3$ = fish weight. Conduct a complete residual analysis for the model. Do you recommend that any model modifications be made? Explain.

**GASTURBINE**

11.49 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 11.25 (p. 590). Consider, again, the first-order model for heat rate ($y$) as a function of

**BAYPCB**

| | | PCB Concentration | | | | PCB Concentration | |
|---|---|---|---|---|---|---|---|
| Bay | State | Year 1 | Year 2 | Bay | State | Year 1 | Year 2 |
| Casco Bay | ME | 95.28 | 77.55 | Mississippi River Delta | LA | 34 | 30.14 |
| Merrimack River | MA | 52.97 | 29.23 | Barataria Bay | LA | 0 | 0 |
| Salem Harbor | MA | 533.58 | 403.1 | San Antonio Bay | TX | 0 | 0 |
| Boston Harbor | MA | 17,104.86 | 736 | Corpus Christi Bay | TX | 0 | 0 |
| Buzzards Bay | MA | 308.46 | 192.15 | San Diego Harbor | CA | 422.1 | 531.67 |
| Narragansett Bay | RI | 159.96 | 220.6 | San Diego Bay | CA | 6.74 | 9.3 |
| East Long Island Sound | NY | 10 | 8.62 | Dana Point | CA | 7.06 | 5.74 |
| West Long Island Sound | NY | 234.43 | 174.31 | Seal Beach | CA | 46.71 | 46.47 |
| Raritan Bay | NJ | 443.89 | 529.28 | San Pedro Canyon | CA | 159.56 | 176.9 |
| Delaware Bay | DE | 2.5 | 130.67 | Santa Monica Bay | CA | 14 | 13.69 |
| Lower Chesapeake Bay | VA | 51 | 39.74 | Bodega Bay | CA | 4.18 | 4.89 |
| Pamlico Sound | NC | 0 | 0 | Coos Bay | OR | 3.19 | 6.6 |
| Charleston Harbor | SC | 9.1 | 8.43 | Columbia River Mouth | OR | 8.77 | 6.73 |
| Sapelo Sound | GA | 0 | 0 | Nisqually Beach | WA | 4.23 | 4.28 |
| St. Johns River | FL | 140 | 120.04 | Commencement Bay | WA | 20.6 | 20.5 |
| Tampa Bay | FL | 0 | 0 | Elliott Bay | WA | 329.97 | 414.5 |
| Apalachicola Bay | FL | 12 | 11.93 | Lutak Inlet | AK | 5.5 | 5.8 |
| Mobile Bay | AL | 0 | 0 | Nahku Bay | AK | 6.6 | 5.08 |
| Round Island | MS | 0 | 0 | | | | |

*Source: Environmental Quality*, 1987–1988.

## SAS Output for Exercise 11.49

| | HEATRATE | _PRED | _RESID | _H | _RSTUDENT | _DFFITS |
|---|---|---|---|---|---|---|
| 1 | 14622 | 14262.043775 | 359.95622475 | 0.1855911418 | 0.8675525828 | 0.4141457267 |
| 2 | 13196 | 12527.866128 | 668.1338724 | 0.095127268 | 1.5482334503 | 0.5019901643 |
| 3 | 11948 | 12303.859631 | -355.8596307 | 0.0519439992 | -0.794131791 | -0.185884543 |
| 4 | 11289 | 12038.477619 | -749.4776187 | 0.0521260193 | -1.703738164 | -0.399535001 |
| 5 | 11964 | 12229.510458 | -265.5104583 | 0.0495908163 | -0.59039758 | -0.13486209 |
| 6 | 10526 | 10680.392768 | -154.3927685 | 0.0247893211 | -0.338260917 | -0.053930558 |
| 7 | 10387 | 10519.725593 | -132.7255935 | 0.0287913306 | -0.291316918 | -0.05015801 |
| 8 | 10592 | 10367.641131 | 224.35886877 | 0.0626537807 | 0.5019535902 | 0.1297738503 |
| 9 | 10460 | 10566.32604 | -106.3260395 | 0.0403612504 | -0.234717477 | -0.048136426 |
| 10 | 10086 | 10232.796663 | -146.7966633 | 0.0638701019 | -0.328244987 | -0.085739086 |
| 11 | 14628 | 13214.015064 | 1413.9849358 | 0.0660538381 | 3.4645251575 | 0.9213652471 |
| 12 | 13396 | 12563.017765 | 832.98223507 | 0.0569040646 | 1.9095424681 | 0.4690542002 |
| 13 | 11726 | 11963.551915 | -237.5519147 | 0.0807474376 | -0.536836461 | -0.159106878 |
| 14 | 11252 | 11310.001708 | -58.00170759 | 0.0457271404 | -0.12835846 | -0.028097989 |
| 15 | 12449 | 12087.066769 | 361.93323148 | 0.0979474326 | 0.8284045229 | 0.2729751205 |
| 16 | 11030 | 11313.295597 | -283.2955965 | 0.0625950725 | -0.63458402 | -0.163981777 |
| 17 | 10787 | 11134.121831 | -347.1218314 | 0.0530684167 | -0.774899499 | -0.183444267 |
| 18 | 10603 | 10982.776871 | -379.7768706 | 0.0603576289 | -0.851959048 | -0.215925289 |
| 19 | 10144 | 10331.066242 | -187.0662422 | 0.0480282163 | -0.415017328 | -0.093218567 |
| 20 | 11674 | 11509.933417 | 164.06658345 | 0.0509387171 | 0.3644294875 | 0.0844287547 |
| 21 | 11510 | 11025.244807 | 484.75519289 | 0.0994655353 | 1.115553658 | 0.3707461139 |
| 22 | 10946 | 10780.076994 | 165.92300575 | 0.0278163559 | 0.3641431984 | 0.0615953836 |
| 23 | 10508 | 10705.096399 | -197.0963995 | 0.0352719536 | -0.434428773 | -0.083067452 |
| 24 | 10604 | 10896.236335 | -292.2363354 | 0.0513484991 | -0.650831939 | -0.151418655 |
| 25 | 10270 | 10345.90378 | -75.90377957 | 0.0360856225 | -0.167149718 | -0.032341006 |
| 26 | 10529 | 10593.014477 | -64.01447678 | 0.0644165255 | -0.143077515 | -0.037542982 |
| 27 | 10360 | 10618.441095 | -258.4410951 | 0.0815098668 | -0.584544836 | -0.17413484 |
| 28 | 14796 | 14708.018722 | 87.981278035 | 0.1196385964 | 0.2027537434 | 0.0747436047 |
| 29 | 12913 | 12891.983127 | 21.016872703 | 0.1248270761 | 0.0485612744 | 0.0183399243 |
| 30 | 12270 | 12673.365781 | -403.3657807 | 0.0365598218 | -0.894173101 | -0.174185203 |
| 31 | 11842 | 12254.869917 | -412.8699173 | 0.0444623577 | -0.919363731 | -0.198316893 |
| 32 | 10656 | 11663.002778 | -1007.002778 | 0.0833329646 | -2.378187485 | -0.717048776 |
| 33 | 11360 | 11218.425759 | 141.57424065 | 0.0616396938 | 0.3161704598 | 0.0810338947 |
| 34 | 11136 | 11197.799982 | -61.79998173 | 0.0564808302 | -0.137544055 | -0.033652474 |
| 35 | 10814 | 11356.639343 | -542.6393433 | 0.0424252255 | -1.213244202 | -0.255372399 |
| 36 | 13523 | 12489.538273 | 1033.4617267 | 0.1808345787 | 2.6038547325 | 1.2234090394 |
| 37 | 11289 | 11159.830604 | 129.16939627 | 0.0537139524 | 0.2872150154 | 0.0684288872 |
| 38 | 11183 | 11020.912821 | 162.08717923 | 0.0922202447 | 0.368136372 | 0.1173360912 |
| 39 | 10951 | 11690.398363 | -739.3983628 | 0.043185789 | -1.671507351 | -0.355111643 |
| 40 | 9722 | 10114.049783 | -392.0497828 | 0.0745533047 | -0.886650757 | -0.251657764 |
| 41 | 10481 | 10541.142037 | -60.14203715 | 0.0880224842 | -0.136148709 | -0.042297848 |
| 42 | 9812 | 9972.0182951 | -160.0182951 | 0.0503975323 | -0.355316634 | -0.081855763 |
| 43 | 9669 | 9959.4134861 | -290.4134861 | 0.1395029404 | -0.679307097 | -0.273516253 |
| 44 | 9643 | 9748.8383688 | -105.8383688 | 0.0783958802 | -0.238416847 | -0.069536265 |
| 45 | 9115 | 8325.4058325 | 789.59416751 | 0.1527764312 | 1.9097622284 | 0.8109768659 |
| 46 | 9115 | 9117.0677296 | -2.067729619 | 0.1988637759 | -0.004993454 | -0.002487859 |
| 47 | 11588 | 10532.976758 | 1055.0232421 | 0.1136884675 | 2.5502108124 | 0.9133581089 |
| 48 | 10888 | 10500.011954 | 387.9880461 | 0.0913012389 | 0.8854991793 | 0.2806832718 |
| 49 | 9738 | 9935.8913064 | -197.8913064 | 0.0602504304 | -0.441963917 | -0.11190792 |
| 50 | 9295 | 8909.4259476 | 385.57405239 | 0.0765052121 | 0.8727508315 | 0.2511993774 |
| 51 | 9421 | 9412.8050086 | 8.194991357 | 0.0739287913 | 0.018407208 | 0.0052008285 |
| 52 | 9105 | 8582.8087155 | 522.19128447 | 0.1360909184 | 1.2295887723 | 0.48802317 |
| 53 | 10233 | 9732.5653104 | 500.43468958 | 0.1101355734 | 1.1594647716 | 0.4079053895 |
| 54 | 10186 | 10365.872705 | -179.8727048 | 0.090892499 | -0.408338798 | -0.129115103 |
| 55 | 9918 | 9659.626414 | 258.37358604 | 0.0480920098 | 0.5739844118 | 0.1290146639 |
| 56 | 9209 | 9515.4158984 | -306.4158984 | 0.0445866186 | -0.680211734 | -0.146943618 |
| 57 | 9532 | 9638.1825427 | -106.1825427 | 0.0647075576 | -0.237434446 | -0.062452168 |
| 58 | 9933 | 10280.590042 | -347.5900416 | 0.0512493612 | -0.775203502 | -0.180170595 |
| 59 | 9152 | 8894.7720107 | 257.22798934 | 0.0812541489 | 0.5817040777 | 0.1729924644 |
| 60 | 9295 | 9640.0725037 | -345.0725037 | 0.0695095373 | -0.777122543 | -0.212400602 |
| 61 | 16243 | 15757.951033 | 485.04896726 | 0.288675505 | 1.259441918 | 0.8023229754 |
| 62 | 14628 | 15114.912042 | -486.9120421 | 0.2045952179 | -1.194032875 | -0.60557777 |
| 63 | 12766 | 13140.94411 | -374.9441104 | 0.0823489349 | -0.85112664 | -0.254967245 |
| 64 | 8714 | 8415.2383967 | 298.76160334 | 0.5520390072 | 0.9724353283 | 1.0795071223 |
| 65 | 9469 | 9760.4835859 | -291.4835859 | 0.1177328552 | -0.673300499 | -0.24595617 |
| 66 | 11948 | 11751.975827 | 196.02417311 | 0.1803903839 | 0.4688789032 | 0.2199702059 |
| 67 | 12414 | 12704.256015 | -290.2560146 | 0.067064899 | -0.651850908 | -0.17477133 |

speed ($x_1$), inlet temperature ($x_2$), exhaust temperature ($x_3$), cycle pressure ratio ($x_4$), and air flow rate ($x_5$). A SAS printout with influence diagnostics for the 67 observations in the **GASTURBINE** file is shown above. Interpret these results. Do you detect any influential observations? (*Note:* "Studentized" deleted residuals, i.e., the deleted residuals divided by their standard errors, are given under the heading _RSTUDENT; the difference between fits, $\hat{y}_i - \hat{y}_{(i)}$, is given under _DFFITS.)

**GRAFTING**

11.50 *A rubber additive made from cashew nut shells.* Refer to the *Industrial & Engineering Chemistry Research* (May 2013) study of the use of cardanol as an additive for natural rubber, Exercise 11.19 (p. 588). Recall that you analyzed a first-order model for $y$ = grafting efficiency as a function of $x_1$ = initiator concentration (parts per hundred resin), $x_2$ = cardanol concentration (parts per hundred resin), $x_3$ = reaction temperature (degrees Centigrade) and $x_4$ = reaction time (hours). Conduct a complete residual analysis for the model. Do you recommend any model modifications?

**BUBBLE2**

11.51 *Bubble behavior in subcooled flow boiling.* Refer to the *Heat Transfer Engineering* (Vol. 34, 2013) study of bubble behavior in subcooled flow boiling, Exercises 11.6 and 11.29 (p. 596). You fit an interaction model for bubble density (y) as a function of $x_1$ = mass flux and $x_2$ = heat flux. Conduct a complete residual analysis for the model. Do you recommend any model modifications?

11.52 *Home-improvement sales.* The data in the table below are sales, $y$, in thousands of dollars per week, for home-improvement outlets in each of four cities. The objective is to model sales, $y$, as a function of traffic flow, adjusting for city-to-city variations that might be due to size or other market conditions. The model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where

$$x_1 = \begin{cases} 1 & \text{if city 1} \\ 0 & \text{if other} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if city 2} \\ 0 & \text{if other} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if city 3} \\ 0 & \text{if other} \end{cases} \qquad x_4 = \text{traffic flow}$$

A SAS printout for the regression analysis is provided on pages 616–617.

a. Is the model statistically useful for predicting $y$? Explain.
b. Do you detect any outliers?
c. Refer to part **b**. Are the outliers detected in the residual plot influential?
d. Note that the value of sales ($y$) for the 13th observation was incorrectly entered into the computer as 82.0; the correct value is 8.2. Make the correction and rerun the regression analysis. Interpret the results.

11.53 *Assembly line breakdowns.* Breakdowns of machines that produce steel cans are very costly. The more breakdowns, the fewer cans produced, and the smaller the company's profits. To help anticipate profit loss, the owners of a can company would like to find a model that will predict the number of breakdowns on the assembly line. The model proposed by the company's statisticians is the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where $y$ is the number of breakdowns per 8-hour shift,

$$x_1 = \begin{cases} 1 & \text{if afternoon shift} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if midnight} \\ 0 & \text{otherwise} \end{cases}$$

**HOMEIMPROVE**

| City | Traffic Flow thousands of cars | Weekly Sales $y$, thousands of dollars | City | Traffic Flow thousands of cars | Weekly Sales $y$, thousands of dollars |
|---|---|---|---|---|---|
| 1 | 59.3 | 6.3 | 3 | 75.8 | 8.2 |
| 1 | 60.3 | 6.6 | 3 | 48.3 | 5.0 |
| 1 | 82.1 | 7.6 | 3 | 41.4 | 3.9 |
| 1 | 32.3 | 3.0 | 3 | 52.5 | 5.4 |
| 1 | 98.0 | 9.5 | 3 | 41.0 | 4.1 |
| 1 | 54.1 | 5.9 | 3 | 29.6 | 3.1 |
| 1 | 54.4 | 6.1 | 3 | 49.5 | 5.4 |
| 1 | 51.3 | 5.0 | 4 | 73.1 | 8.4 |
| 1 | 36.7 | 3.6 | 4 | 81.3 | 9.5 |
| 2 | 23.6 | 2.8 | 4 | 72.4 | 8.7 |
| 2 | 57.6 | 6.7 | 4 | 88.4 | 10.6 |
| 2 | 44.6 | 5.2 | 4 | 23.2 | 3.3 |

## SAS Output for Exercise 11.52

The REG Procedure
Model: MODEL1
Dependent Variable: y

| | |
|---|---|
| Number of Observations Read | 24 |
| Number of Observations Used | 24 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 1469.76287 | 367.44072 | 1.66 | 0.1996 |
| Error | 19 | 4194.22671 | 220.74877 | | |
| Corrected Total | 23 | 5663.98958 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 14.85762 | R-Square | 0.2595 |
| Dependent Mean | 9.07083 | Adj R-Sq | 0.1036 |
| Coeff Var | 163.79550 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > ¦t¦ |
|---|---|---|---|---|---|
| Intercept | 1 | -16.45925 | 13.16400 | -1.25 | 0.2264 |
| x1 | 1 | 1.10609 | 8.42257 | 0.13 | 0.8969 |
| x2 | 1 | 6.14277 | 11.67997 | 0.53 | 0.6050 |
| x3 | 1 | 14.48962 | 9.28839 | 1.56 | 0.1353 |
| x4 | 1 | 0.36287 | 0.16791 | 2.16 | 0.0437 |

Output Statistics

| Obs | Dependent Variable | Predicted Value | Std Error Mean Predict | Residual | Std Error Residual | Student Residual | -2-1 0 1 2 | Cook's D |
|---|---|---|---|---|---|---|---|---|
| 1 | 6.3000 | 6.1652 | 4.9535 | 0.1348 | 14.008 | 0.00962 | ¦ ¦ ¦ | 0.000 |
| 2 | 6.6000 | 6.5281 | 4.9596 | 0.0719 | 14.005 | 0.00513 | ¦ ¦ ¦ | 0.000 |
| 3 | 7.6000 | 14.4387 | 6.3195 | -6.8387 | 13.447 | -0.509 | ¦ *¦ ¦ | 0.011 |
| 4 | 3.0000 | -3.6324 | 6.6491 | 6.6324 | 13.287 | 0.499 | ¦ ¦ ¦ | 0.012 |
| 5 | 9.5000 | 20.2084 | 8.2476 | -10.7084 | 12.358 | -0.866 | ¦ *¦ ¦ | 0.067 |
| 6 | 5.9000 | 4.2783 | 5.0130 | 1.6217 | 13.986 | 0.116 | ¦ ¦ ¦ | 0.000 |
| 7 | 6.1000 | 4.3871 | 5.0054 | 1.7129 | 13.989 | 0.122 | ¦ ¦ ¦ | 0.000 |
| 8 | 5.0000 | 3.2622 | 5.1069 | 1.7378 | 13.952 | 0.125 | ¦ ¦ ¦ | 0.000 |
| 9 | 3.6000 | -2.0357 | 6.1807 | 5.6357 | 13.511 | 0.417 | ¦ ¦ ¦ | 0.007 |
| 10 | 2.8000 | -1.7527 | 9.1137 | 4.5527 | 11.734 | 0.388 | ¦ ¦ ¦ | 0.018 |
| 11 | 6.7000 | 10.5850 | 8.9723 | -3.8850 | 11.843 | -0.328 | ¦ ¦ ¦ | 0.012 |
| 12 | 5.2000 | 5.8677 | 8.5897 | -0.6677 | 12.123 | -0.0551 | ¦ ¦ ¦ | 0.000 |
| 13 | 82.0000 | 25.5362 | 7.2703 | 56.4638 | 12.957 | 4.358 | ¦ ¦******¦ | 1.196 |
| 14 | 5.0000 | 15.5571 | 5.6157 | -10.5571 | 13.755 | -0.767 | ¦ *¦ ¦ | 0.020 |
| 15 | 3.9000 | 13.0533 | 5.7339 | -9.1533 | 13.707 | -0.668 | ¦ *¦ ¦ | 0.016 |
| 16 | 5.4000 | 17.0812 | 5.6598 | -11.6812 | 13.737 | -0.850 | ¦ *¦ ¦ | 0.025 |
| 17 | 4.1000 | 12.9082 | 5.7479 | -8.8082 | 13.701 | -0.643 | ¦ *¦ ¦ | 0.015 |
| 18 | 3.1000 | 8.7714 | 6.4338 | -5.6714 | 13.392 | -0.423 | ¦ ¦ ¦ | 0.008 |
| 19 | 5.4000 | 15.9926 | 5.6193 | -10.5926 | 13.754 | -0.770 | ¦ *¦ ¦ | 0.020 |
| 20 | 8.4000 | 10.0668 | 6.7066 | -1.6668 | 13.258 | -0.126 | ¦ ¦ ¦ | 0.001 |
| 21 | 9.5000 | 13.0423 | 7.0271 | -3.5423 | 13.091 | -0.271 | ¦ ¦ ¦ | 0.004 |
| 22 | 8.7000 | 9.8128 | 6.6916 | -1.1128 | 13.265 | -0.0839 | ¦ ¦ ¦ | 0.000 |
| 23 | 10.6000 | 15.6187 | 7.5002 | -5.0187 | 12.826 | -0.391 | ¦ ¦ ¦ | 0.010 |
| 24 | 3.3000 | -8.0406 | 9.9965 | 11.3406 | 10.992 | 1.032 | ¦ ¦** ¦ | 0.176 |

**SAS Output for Exercise 11.52 *(Continued)***

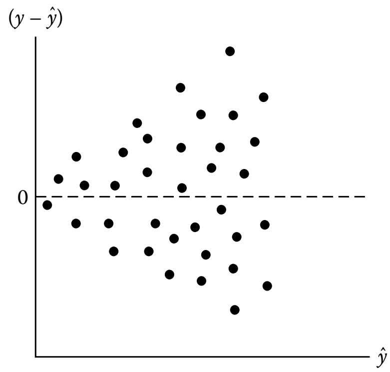
Output Statistics

| Obs | RStudent | Hat Diag H | Cov Ratio | DFFITS | DFBETAS Intercept | x1 | x2 | x3 | x4 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.009366 | 0.1112 | 1.4742 | 0.0033 | -0.0001 | 0.0020 | 0.0000 | 0.0000 | 0.0001 |
| 2 | 0.004998 | 0.1114 | 1.4747 | 0.0018 | -0.0001 | 0.0011 | 0.0000 | 0.0000 | 0.0001 |
| 3 | -0.4984 | 0.1809 | 1.4939 | -0.2342 | 0.1256 | -0.1339 | -0.0539 | -0.0510 | -0.1455 |
| 4 | 0.4891 | 0.2003 | 1.5339 | 0.2447 | 0.1410 | 0.0780 | -0.0604 | -0.0572 | -0.1633 |
| 5 | -0.8606 | 0.3081 | 1.5482 | -0.5743 | 0.3965 | -0.2848 | -0.1700 | -0.1609 | -0.4592 |
| 6 | 0.1129 | 0.1138 | 1.4735 | 0.0405 | 0.0054 | 0.0224 | -0.0023 | -0.0022 | -0.0063 |
| 7 | 0.1192 | 0.1135 | 1.4724 | 0.0427 | 0.0053 | 0.0237 | -0.0023 | -0.0022 | -0.0062 |
| 8 | 0.1213 | 0.1181 | 1.4799 | 0.0444 | 0.0094 | 0.0234 | -0.0040 | -0.0038 | -0.0108 |
| 9 | 0.4079 | 0.1731 | 1.5134 | 0.1866 | 0.0964 | 0.0680 | -0.0413 | -0.0391 | -0.1116 |
| 10 | 0.3791 | 0.3763 | 2.0190 | 0.2945 | 0.0859 | -0.0178 | 0.1667 | -0.0348 | -0.0995 |
| 11 | -0.3202 | 0.3647 | 2.0048 | -0.2426 | 0.0614 | -0.0127 | -0.1967 | -0.0249 | -0.0711 |
| 12 | -0.0536 | 0.3342 | 1.9667 | -0.0380 | 0.0017 | -0.0004 | -0.0286 | -0.0007 | -0.0020 |
| 13 | 179.3101 | 0.2394 | 0.0000 | 100.6096 | -55.1618 | 11.4109 | 23.6508 | 69.3701 | 63.8990 |
| 14 | -0.7589 | 0.1429 | 1.3061 | -0.3098 | -0.0000 | -0.0000 | 0.0000 | -0.1873 | 0.0000 |
| 15 | -0.6578 | 0.1489 | 1.3673 | -0.2752 | -0.0480 | 0.0099 | 0.0206 | -0.1435 | 0.0556 |
| 16 | -0.8439 | 0.1451 | 1.2625 | -0.3477 | 0.0374 | -0.0077 | -0.0160 | -0.2237 | -0.0433 |
| 17 | -0.6327 | 0.1497 | 1.3806 | -0.2654 | -0.0489 | 0.0101 | 0.0209 | -0.1370 | 0.0566 |
| 18 | -0.4141 | 0.1875 | 1.5381 | -0.1990 | -0.0838 | 0.0173 | 0.0359 | -0.0710 | 0.0971 |
| 19 | -0.7616 | 0.1430 | 1.3049 | -0.3111 | 0.0096 | -0.0020 | -0.0041 | -0.1919 | -0.0112 |
| 20 | -0.1224 | 0.2038 | 1.6389 | -0.0619 | -0.0237 | 0.0469 | 0.0318 | 0.0409 | -0.0084 |
| 21 | -0.2639 | 0.2237 | 1.6557 | -0.1417 | -0.0278 | 0.0974 | 0.0591 | 0.0797 | -0.0461 |
| 22 | -0.0817 | 0.2028 | 1.6408 | -0.0412 | -0.0164 | 0.0314 | 0.0215 | 0.0276 | -0.0049 |
| 23 | -0.3824 | 0.2548 | 1.6888 | -0.2236 | -0.0104 | 0.1378 | 0.0743 | 0.1054 | -0.1037 |
| 24 | 1.0336 | 0.4527 | 1.7946 | 0.9400 | 0.9216 | -0.6183 | -0.6154 | -0.6930 | -0.7023 |

$x_3$ is the temperature of the plant (F°), and $x_4$ is the number of inexperienced personnel working on the assembly line. After the model is fit using the least-squares procedure, the residuals are plotted against $\hat{y}$, as shown in the accompanying figure.

a. Do you detect a pattern in the residual plot? What does this suggest about the least-squares assumptions?

b. Given the nature of the response variable $y$ and the pattern detected in part **a**, what model adjustments would you recommend?



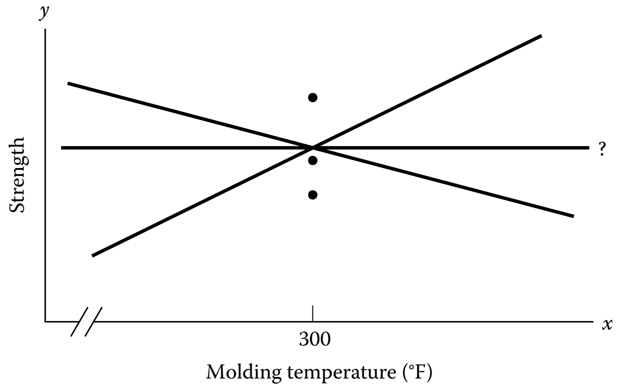

## 11.11 Some Pitfalls: Estimability, Multicollinearity, and Extrapolation

There are several problems you should be aware of when constructing a prediction model for some response $y$. A few of the most important will be discussed in this section.

**Problem 1: Parameter Estimability** Suppose you want to fit a model relating the strength $y$ of a new type of plastic fitting to molding temperature $x$. We propose the first-order model

$$E(y) = \beta_0 + \beta_1 x$$

**FIGURE 11.22**
Plastic strength and molding temperature data




Now, suppose we mold a sample of three plastic fittings, each at a temperature of 300°F. The data are graphed in Figure 11.22. You can see the problem: The parameters of the line cannot be estimated when all the data are concentrated at a single $x$ value. Recall that it takes two points ($x$ values) to fit a straight line. Thus, the parameters are not estimated when only one $x$ value is observed.

A similar problem would occur if we attempted to fit the second-order model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

to a set of data for which only one or two different $x$ values were observed (see Figure 11.23). At least three different $x$ values must be observed before a second-order model can be fit to a set of data (that is, before all three parameters are estimable). In general, the number of levels of $x$ must be at least one more than the order of the polynomial in $x$ that you want to fit. Remember, also, that the sample size $n$ must be sufficiently large to allow degrees of freedom for estimating $\sigma^2$.

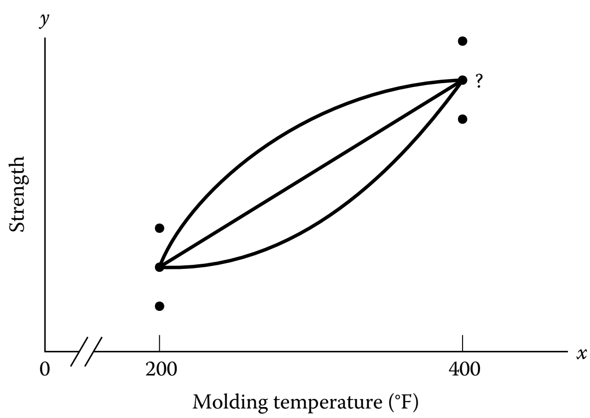
### Requirements for Fitting a $p$th-Order Polynomial Regression Model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

1. The number of levels of $x$ must be greater than or equal to $(p + 1)$.
2. The sample size $n$ must be greater than $(p + 1)$ to allow sufficient degrees of freedom for estimating $\sigma^2$.

Since many variables observed in nature cannot be controlled by the researcher, the independent variables will almost always be observed at a sufficient number of levels to permit estimation of the model parameters. However, when the computer program you use suddenly refuses to fit a model, the problem is probably inestimable parameters.

**FIGURE 11.23**
Only two $x$ values observed—second-order model is not estimable



**Problem 2: Parameter Interpretation** Given that the parameters of the model are estimable, it is important to interpret the parameter estimates correctly. A typical misconception is that $\hat{\beta}_i$ *always* measures the effect of $x_i$ on $E(y)$, *independent* of the other $x$ variables in the model. This may be true for some models, but is not true in general (e.g., the interaction model of Section 11.8). Generally, the interpretation of an individual $\beta$ parameter becomes increasingly more difficult as the model becomes more complex. In Chapter 12, we give the $\beta$ interpretations for a number of different multiple regression models.

Another misconception about parameter estimates is that a statistically significant $\hat{\beta}_i$ value establishes a *cause-and-effect* relationship between $E(y)$ and $x_i$. That is, if $\hat{\beta}_i$ is found to be significantly greater than 0, then some practitioners would infer that an increase in $x_i$ *causes* an increase in the mean response, $E(y)$. However, we warned in Section 10.7 about the dangers of inferring a causal relationship between two variables. There may be many other independent variables (some of which we may have included in our model, some of which we may have omitted) that affect the mean response. Unless we can control the values of these other variables, we are uncertain about what is actually causing the observed increase in $y$. In Chapter 13, we introduce the notion of **designed experiments**, where the values of the independent variables are set in advance before the value of $y$ is observed. Only with such an experiment can a cause-and-effect relationship be established.

**Problem 3: Multicollinearity** Often, two or more of the independent variables used in the model for $E(y)$ will contribute redundant information. That is, the independent variables will be correlated with each other. For example, suppose we want to construct a model to predict the gasoline mileage rating, $y$, of a truck as a function of its load, $x_1$, and the horsepower, $x_2$, of its engine. In general, you would expect heavier loads to require greater horsepower and to result in lower mileage ratings. Thus, although both $x_1$ and $x_2$ contribute information for the prediction of mileage rating, some of the information is overlapping, because $x_1$ and $x_2$ are correlated.

When the independent variables are correlated, we say that multicollinearity exists. In practice, it is not uncommon to observe correlations among the independent variables. However, a few problems arise when serious multicollinearity is present in the regression analysis.

**DEFINITION 11.7**

**Multicollinearity** exists when two or more independent variables used in regression are correlated.

First, high correlations among the independent variables increase the likelihood of rounding errors in the calculations of the $\beta$ estimates, standard errors, and so forth.* Second, the regression results may be confusing and misleading.

To illustrate, if the gasoline mileage rating model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

were fit to a set of data, we might find that the $t$ values for both $\hat{\beta}_1$ and $\hat{\beta}_2$ (the least-squares estimates) are nonsignificant. However, the $F$ test for $H_0$: $\beta_1 = \beta_2 = 0$ would probably be highly significant. The tests may seem to be contradictory, but really they are not. The $T$ tests indicate that the contribution of one variable, say, $x_1$ = load, is not significant after the effect of $x_2$ = horsepower has been discounted (because $x_2$ is also in the model). The significant $F$ test, on the other hand, tells us that at least one of the two variables is making a contribution to the prediction of $y$ (i.e., either $\beta_1$, $\beta_2$, or both differ from 0). In fact, both are probably contributing, but the contribution of one overlaps with that of the other.

*The result is due to the fact that, in the presence of severe multicollinearity, the computer has difficulty inverting the $(\boldsymbol{X}'\boldsymbol{X})$ matrix.

Multicollinearity can also have an effect on the signs of the parameter estimates. More specifically, a value of $\hat{\beta}_i$ may have the opposite sign from what is expected. For example, we expect the signs of both of the parameter estimates for the gasoline mileage rating model to be negative, yet the regression analysis for the model might yield the estimates $\hat{\beta}_1 = .2$ and $\hat{\beta}_2 = -.7$. The positive value of $\hat{\beta}_1$ seems to contradict our expectation that heavy loads will result in lower mileage ratings. However, it is dangerous to interpret a $\beta$ coefficient when the independent variables are correlated. Because the variables contribute redundant information, the effect of load ($x_1$) on mileage rating is measured only partially by $\hat{\beta}_1$. Also, we warned in the discussion of Problem 2 that we cannot establish a cause-and-effect relationship between $y$ and the predictor variables based on **observational data** (data for which the values of the independent variables are uncontrolled). By attempting to interpret the value $\hat{\beta}_1$, we are really trying to establish a cause-and-effect relationship between $y$ and $x_1$ (by suggesting that a heavy load $x_1$ will *cause* a lower mileage rating $y$).

How can you avoid the problems of multicollinearity in regression analysis? One way is to conduct a designed experiment so that the levels of the $x$ variables are uncorrelated. Unfortunately, time and cost constraints may prevent you from collecting data in this manner. For these and other reasons, most data collected in scientific studies are observational. Since observational data frequently consist of correlated independent variables, you will need to recognize when multicollinearity is present and, if necessary, make modifications in the regression analysis.

Several methods are available for detecting multicollinearity in regression. A simple technique is to calculate the coefficient of correlation $r$ between each pair of independent variables in the model and use the procedure outlined in Section 10.7 to test for evidence of positive or negative correlation. If one or more of the $r$ values is statistically different from 0, the variables in question are correlated and a severe multicollinearity problem may exist.* Other indications of the presence of multicollinearity include those mentioned above—namely, nonsignificant $T$ tests for the individual $\beta$ parameters when the $F$ test for overall model adequacy is significant, and parameter estimates with opposite signs from what is expected.†

The methods for detecting multicollinearity are summarized in the box. We illustrate the use of these statistics in Example 11.16

### Detecting Multicollinearity in the Regression Model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The following are indicators of multicollinearity:

1. Significant correlations between pairs of independent variables in the model
2. Nonsignificant $T$ tests for the individual $\beta$ parameters when the $F$ test for overall model adequacy $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is significant
3. Opposite signs (from what is expected) in the estimated parameters

*Remember that $r$ measures only the pairwise correlation between $x$ values. Three variables, $x_1$, $x_2$, and $x_3$, may be highly correlated as a group but may not exhibit large pairwise correlations. Thus, multicollinearity may be present even when all pairwise correlations are not significantly different from 0.

†More formal methods for detecting multicollinearity, such as variance-inflation factors (VIFs), are available. Independent variables with a VIF of 10 or above are usually considered to be highly correlated with one or more of the other independent variables in the model. Calculation of VIFs are beyond the scope of this introductory text. Consult the chapter references for a discussion of VIFs and other formal methods of detecting multicollinearity.

**Example 11.14**

**Detecting Signs of Multicollinearity**

The Federal Trade Commission (FTC) annually ranks varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide contents. The U.S. Surgeon General considers each of these three substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine contents of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. Table 11.8 presents data on tar, nicotine, and carbon monoxide contents (in milligrams) and weight (in grams) for a sample of 25 (filter) brands tested in a recent year. Suppose we want to model carbon monoxide content, $y$, as a function of tar content, $x_1$, nicotine content, $x_2$, and weight, $x_3$, using the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The model is fit to the 25 data points in Table 11.8, and a portion of the SAS printout is shown in Figure 11.24. Examine the printout. Do you detect any signs of multicollinearity?

Solution

First, note that the $F$ test for overall model utility is highly significant. The test statistic ($F = 78.98$) and observed significance level ($p$-value $< .0001$) are highlighted on the SAS printout, Figure 11.24. Therefore, we can conclude at, say, $\alpha = .01$, that at least one of the parameters, $\beta_1$, $\beta_2$, or $\beta_3$, in the model is nonzero. The $T$ tests for two of three individual $\beta$'s, however, are nonsignificant. (The $p$-values for these tests are

FTC2

**TABLE 11.8 FTC Cigarette Data for Example 11.14**

| Tar ($x_1$) | Nicotine ($x_2$) | Weight ($x_3$) | Carbon Monoxide ($y$) |
|---|---|---|---|
| 14.1 | .86 | .9853 | 13.6 |
| 16.0 | 1.06 | 1.0938 | 16.6 |
| 29.8 | 2.03 | 1.1650 | 23.5 |
| 8.0 | .67 | .9280 | 10.2 |
| 4.1 | .40 | .9462 | 5.4 |
| 15.0 | 1.04 | .8885 | 15.0 |
| 8.8 | .76 | 1.0267 | 9.0 |
| 12.4 | .95 | .9225 | 12.3 |
| 16.6 | 1.12 | .9372 | 16.3 |
| 14.9 | 1.02 | .8858 | 15.4 |
| 13.7 | 1.01 | .9643 | 13.0 |
| 15.1 | .90 | .9316 | 14.4 |
| 7.8 | .57 | .9705 | 10.0 |
| 11.4 | .78 | 1.1240 | 10.2 |
| 9.0 | .74 | .8517 | 9.5 |
| 1.0 | .13 | .7851 | 1.5 |
| 17.0 | 1.26 | .9186 | 18.5 |
| 12.8 | 1.08 | 1.0395 | 12.6 |
| 15.8 | .96 | .9573 | 17.5 |
| 4.5 | .42 | .9106 | 4.9 |
| 14.5 | 1.01 | 1.0070 | 15.9 |
| 7.3 | .61 | .9806 | 8.5 |
| 8.6 | .69 | .9693 | 10.6 |
| 15.2 | 1.02 | .9496 | 13.9 |
| 12.0 | .82 | 1.1184 | 14.9 |

*Source*: Federal Trade Commission

```
                              Dependent Variable: CO

                    Number of Observations Read          25
                    Number of Observations Used          25

                              Analysis of Variance

                                     Sum of          Mean
  Source                   DF       Squares        Square    F Value    Pr > F

  Model                     3     495.25781     165.08594      78.98    <.0001
  Error                    21      43.89259       2.09012
  Corrected Total          24     539.15040

             Root MSE               1.44573    R-Square     0.9186
             Dependent Mean        12.52800    Adj R-Sq     0.9070
             Coeff Var             11.53996

                              Parameter Estimates

                      Parameter       Standard                           Variance
Variable      DF       Estimate          Error    t Value    Pr > |t|   Inflation

Intercept      1        3.20219        3.46175       0.93      0.3655           0
TAR            1        0.96257        0.24224       3.97      0.0007    21.63071
NICOTINE       1       -2.63166        3.90056      -0.67      0.5072    21.89992
WEIGHT         1       -0.13048        3.88534      -0.03      0.9735     1.33386


               Pearson Correlation Coefficients, N = 25
                      Prob > |r| under H0: Rho=0

                              TAR        NICOTINE         WEIGHT

           TAR            1.00000         0.97661        0.49077
                                           <.0001         0.0127

           NICOTINE       0.97661         1.00000        0.50018
                           <.0001                         0.0109

           WEIGHT         0.49077         0.50018        1.00000
                           0.0127          0.0109
```

**FIGURE 11.24**
SAS printout for model of CO content, Example 11.14

highlighted on the printout.) Unless tar ($x_1$) is the only one of the three variables useful for predicting carbon monoxide content, these results are the first indication of a potential multicollinearity problem.

The negative values for $\hat{\beta}_2$ and $\hat{\beta}_3$ (highlighted on the printout) are a second clue to the presence of multicollinearity. From past studies, the FTC expects carbon monoxide content ($y$) to increase when either nicotine content ($x_2$) or weight ($x_3$) increases—that is, the FTC expects *positive* relationships between $y$ and $x_2$ and between $y$ and $x_3$, not negative ones.

All signs indicate that a serious multicollinearity problem exists.*

To confirm our suspicions, we had SAS produce the coefficient of correlation, $r$, for each of the three pairs of independent variables in the model. The resulting output is shown (highlighted) at the bottom of Figure 11.24. You can see that tar ($x_1$) and nicotine ($x_2$) are highly correlated $(r = .9766)$ while weight ($x_3$) is moderately correlated with the other two $x$'s $(r \approx .5)$ All three correlations have $p$-values less than .05; consequently, all three are significantly different from 0 at $\alpha = .05$.

*Note also that the variance-inflation factors (VIFs) for both tar and nicotine, given on the SAS printout, Figure 11.24, exceed 10.

Once you have detected that a multicollinearity problem exists, there are several alternative measures available for solving the problem. The appropriate measure to take depends on the severity of the multicollinearity and the ultimate goal of the regression analysis.

Some researchers, when confronted with highly correlated independent variables, choose to include only one of the correlated variables in the final model. One way of deciding which variable to include is by using **stepwise regression**, a topic discussed in Chapter 12. Generally, only one (or a small number) of a set of multicollinear independent variables will be included in the regression model by the stepwise regression procedure. This procedure tests the parameter associated with each variable in the presence of all the variables already in the model. For example, in fitting the gasoline mileage rating model introduced earlier, if at one step the variable representing truck load is included as a significant variable in the prediction of the mileage rating, the variable representing horsepower will probably never be added in a future step. Thus, if a set of independent variables is thought to be multicollinear, some screening by stepwise regression may be helpful.

If you are interested in using the model for estimation and prediction, you may decide not to drop any of the independent variables from the model. In the presence of multicollinearity, we have seen that it is dangerous to interpret the individual $\beta$'s for the purpose of establishing cause and effect. However, confidence intervals for $E(y)$ and prediction intervals for $y$ generally remain unaffected **as long as the values of the independent variables used to predict $y$ follow the same pattern of multicollinearity exhibited in the sample data**. That is, you must take strict care to ensure that the values of the $x$ variables fall within the experimental region. (We will discuss this problem in further detail in Problem 4.) Alternatively, if your goal is to establish a cause-and-effect relationship between $y$ and the independent variables, you will need to conduct a designed experiment to break up the pattern of multicollinearity.

### Solutions to Some Problems Created by Multicollinearity*

1. Drop one or more of the correlated independent variables from the final model. A screening procedure such as stepwise regression (see Chapter 12) is helpful in determining which variables to drop.
2. If you decide to keep all the independent variables in the model:
   a. Avoid making inferences about the individual $\beta$ parameters (such as establishing a cause-and-effect relationship between $y$ and the predictor variables).
   b. Restrict inferences about $E(y)$ and future $y$ values to values of the independent variables that fall within the experimental region (see Problem 4).
3. If your ultimate objective is to establish a cause-and-effect relationship between $y$ and the predictor variables, use a designed experiment (see Chapter 13).
4. To reduce rounding errors in polynomial regression models, code the independent variables so that first-, second-, and higher-order terms for a particular $x$ variable are not highly correlated (see Chapter 12).

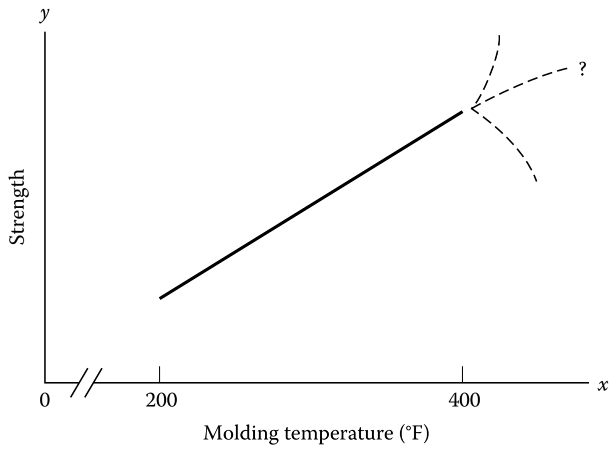
When fitting a polynomial regression model [for example, the second-order model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$], the independent variables $x_1 = x$ and $x_2 = x^2$ will often be correlated. If the correlation is high, the computer solution may result in

*Several other solutions are available. For example, in the case where higher-order regression models are fit, the analyst may want to code the independent variables so that higher-order terms (e.g., $x^2$) for a particular $x$-variable are not highly correlated with $x$. One transformation that works is $Z = (x - \bar{x})/s$ (See Optional Section 12.5). Other, more sophisticated procedures for addressing multicollinearity (such as **ridge regression**) are beyond the scope of this text. Consult the references.

extreme rounding errors. For this model, the solution is not to drop one of the independent variables, but to transform the $x$ variable in such a way that the correlation between the coded $x$ and $x^2$ values is substantially reduced. Coding the independent quantitative variables in polynomial regression models is discussed in Chapter 12.

**Problem 4: Prediction Outside the Experimental Region** The fitted regression model enables us to construct a confidence interval for $E(y)$ and a prediction interval for $y$ for values of the independent variable only within the region of experimentation, i.e., within the range of values of the independent variables used in the experiment. For example, suppose that you conduct experiments on the mean strength of the plastic fittings (see Figure 11.23) at several different temperatures in the interval 200°F to 400°F. The regression model that you fit to the data is valid for estimating $E(y)$ or for predicting values of $y$ for values of $x$ in the range 200°F $\leq x \leq$ 400°F. However, if you attempt to **extrapolate** beyond the experimental region, you risk the possibility that the fitted model is no longer a good approximation to the mean strength of the plastic (see Figure 11.25). For example, the plastic may become too brittle when formed at 500°F and possess no strength at all. Estimating and predicting outside of the experimental region is sometimes necessary. If you do so, keep in mind the possibility of a large extrapolation error.

**FIGURE 11.25**
Using a regression model outside the experimental region




## Applied Exercises

**GRAFTING**

11.54 *A rubber additive made from cashew nut shells.* Refer to the *Industrial & Engineering Chemistry Research* (May 2013) study of the use of cardanol as an additive for natural rubber, Exercises 11.19 and 11.50 (p. 588, 615). In both exercises, you analyzed a first-order model for $y$ = grafting efficiency as a function of $x_1$ = initiator concentration (parts per hundred resin), $x_2$ = cardanol concentration (parts per hundred resin), $x_3$ = reaction temperature (degrees Centigrade) and $x_4$ = reaction time (hours).

a. Suppose an engineer wants to predict the grafting efficiency of chemical run with initiator concentration set at 5 parts per hundred resin, cardanol concentration at 20 parts per hundred resin, reaction temperature at 30 degrees, and reaction time at 5 hours. Would you recommend using the prediction equation from Exercise 11.19 to obtain this prediction? Explain.

b. Examine the data and determine if there is any evidence of multicollinearity. (Note: This result is due to the design of the experiment.)

c. Does the data allow the researchers to investigate whether grafting efficiency ($y$) is curvilinearly related to reaction time ($x_4$)? Explain.

**TEAMPERF**

11.55 *Emotional intelligence and team performance.* Refer to the *Engineering Project Organizational Journal* (Vol. 3., 2013) study of the relationship between emotional intelligence of individual team members and their performance during an engineering project, Exercise 11.23 (p. 589). Using data on $n$ = 23 teams you fit a first-order model for mean project score ($y$) as a function of range of interpersonal scores ($x_1$), range of stress management scores ($x_2$), and range of mood scores ($x_3$). Do you detect any signs of multicollinearity in the data?

11.56 *Global warming and foreign investments.* The *Journal of World-Systems Research* (Summer 2003) reported on a study of the link between foreign investments made 16 years earlier and carbon dioxide emissions. The researchers modeled the annual level ($y$) of $CO_2$ emissions as a function of seven independent variables for $n = 66$ developing countries. A matrix given the correlation ($r$) for each pair of independent variables is shown at the bottom of the page. Identify the independent variables that are highly correlated. What problems may result from including these highly correlated variables in the regression model?

11.57 *Accuracy of software effort estimates.* Periodically, software engineers must provide estimates of their effort in developing new software. In the *Journal of Empirical Software Engineering* (Vol. 9, 2004), multiple regression was used to predict the accuracy of these effort estimates. The dependent variable, defined as the relative error in estimating effort,

$$y = (\text{Actual effort} - \text{Estimated effort})/(\text{Actual effort})$$

was determined for each in a sample of $n = 49$ software development tasks. Two independent variables used in the model for relative error in estimating effort were company role of estimator ($x_1 = 1$ if developer, 0 if project leader) and previous accuracy ($x_2 = 1$ if more than 20% accurate, 0 if less than 20% accurate). The multiple regression yielded the prediction equation $\hat{y} = .12 - .28x_1 + .27x_2$. The researcher is concerned that the sign of the estimated $\beta$ multiplied by $x_1$ is the opposite from what is expected. (The researcher expects a project leader to have a smaller relative error of estimation than a developer.) Give at least one reason why this phenomenon occurred.

11.58 *Steam processing of peat.* A bioengineer wants to model the amount ($y$) of carbohydrate solubilized during steam processing of peat as a function of temperature ($x_1$), exposure time ($x_2$), and pH value ($x_3$). Data collected for each of 15 peat samples were used to fit the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

A summary of the regression results follows:

$$\hat{y} = -3{,}000 + 3.2x_1 - .4x_2 - 1.1x_3 \qquad R^2 = .93$$

$$s_{\hat{\beta}_1} = 2.4 \qquad s_{\hat{\beta}_2} = .6 \qquad s_{\hat{\beta}_3} = .8$$

$$r_{12} = .92 \qquad r_{13} = .87 \qquad r_{23} = .81$$

Based on these results, the bioengineer concludes that none of the three independent variables, $x_1$, $x_2$, and $x_3$, is a useful predictor of carbohydrate amount, $y$. Do you agree with this statement? Explain.

11.59 *Engineering market research.* The management of an engineering consultant firm is considering the possibility of setting up its own market research department rather than continuing to use the services of a market research firm. Management wants to know what salary should be paid to a market researcher, based on years of experience. An independent consultant has proposed the quadratic model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where

$y$ = Annual salary (thousands of dollars)
$x$ = Years of experience

To fit the model, the consultant randomly sampled three market researchers at other firms and recorded the information given in the accompanying table. Give your opinion regarding the adequacy of the proposed model.

| | $y$ | $x$ |
|---|---|---|
| *Researcher 1* | 40 | 2 |
| *Researcher 2* | 25 | 1 |
| *Researcher 3* | 42 | 3 |

**Correlation Matrix for Exercise 11.56**

| Independent Variable | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ = ln(level of $CO_2$ emissions) |
|---|---|---|---|---|---|---|
| $x_1$ = ln foreign investments | .13 | .57 | .30 | −.38 | .14 | −.14 |
| $x_2$ = gross domestic investment | | .49 | .36 | −.47 | −.14 | .25 |
| $x_3$ = trade exports | | | .43 | −.47 | −.06 | −.07 |
| $x_4$ = ln(GNP) | | | | −.84 | −.53 | .42 |
| $x_5$ = agricultural production | | | | | .45 | −.50 |
| $x_6$ = 1 if African Country, 0 if not | | | | | | −.47 |

*Source:* Grimes, P., and Kentor, J. "Exporting the greenhouse: Foreign capital penetration and $CO_2$ emissions 1980–1996." *Journal of World-Systems Research*, Vol. IX, No. 2, Summer 2003 (Appendix B).

11.60 *Redundancy of correlated variables.* In a classic research paper, Hamilton (1987) illustrated the multicollinearity problem with an example using the data shown in the next table. The values of $x_1$, $x_2$, and $y$ in the table represent appraised land value, appraised improvements value, and sale price, respectively, of a randomly selected residential property. (All measurements are in thousands of dollars.)

**MCDATA**

| $x_1$ | $x_2$ | $y$ | $x_1$ | $x_2$ | $y$ |
|---|---|---|---|---|---|
| 22.3 | 96.6 | 123.7 | 30.4 | 77.1 | 128.6 |
| 25.7 | 89.4 | 126.6 | 32.6 | 51.1 | 108.4 |
| 38.7 | 44.0 | 120.0 | 33.9 | 50.5 | 112.0 |
| 31.0 | 66.4 | 119.3 | 23.5 | 85.1 | 115.6 |
| 33.9 | 49.1 | 110.6 | 27.6 | 65.9 | 108.3 |
| 28.3 | 85.2 | 130.3 | 39.0 | 49.0 | 126.3 |
| 30.2 | 80.4 | 131.3 | 31.6 | 69.6 | 124.6 |
| 21.4 | 90.5 | 114.4 | | | |

*Source:* Hamilton, D. "Sometimes $R^2 > r^2_{yx1} + r^2_{yx2}$: Correlated variables are not always redundant," *The American Statistician*, Vol. 41, No. 2, May 1987, pp. 129–132.

a. Calculate the coefficient of correlation between $y$ and $x_1$. Is there evidence of a linear relationship between sale price and appraised land value?
b. Calculate the coefficient of correlation between $y$ and $x_2$. Is there evidence of a linear relationship between sale price and appraised improvements?
c. Based on the results in parts **a** and **b**, do you think the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ will be useful for predicting sale price?
d. Fit the model in part **c**, and conduct a test of model adequacy. In particular, note the value of $R^2$. Does the result agree with your answer to part **c**?
e. Calculate the coefficient of correlation between $x_1$ and $x_2$. What does the result imply?
f. Many researchers avoid the problems of multicollinearity by always omitting all but one of the "redundant" variables from the model. Would you recommend this strategy for this example? Explain. (Hamilton notes that in this case, such a strategy "can amount to throwing out the baby with the bathwater.")

**FTC2**

11.61 *Analysis of cigarette data.* Refer to the FTC cigarette data of Example 11.14 (p. 582). Recall that the data are saved in the **FTC2** file.

a. Fit the model $E(y) = \beta_0 + \beta_1 x_1$ to the data. Is there evidence that tar content ($x_1$) is useful for predicting carbon monoxide content ($y$)?
b. Fit the model $E(y) = \beta_0 + \beta_2 x_2$ to the data. Is there evidence that nicotine content ($x_2$) is useful for predicting carbon monoxide content ($y$)?
c. Fit the model $E(y) = \beta_0 + \beta_3 x_3$ to the data. Is there evidence that weight ($x_3$) is useful for predicting carbon monoxide content ($y$)?
d. Compare the signs of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ in the models of parts **a**, **b**, and **c**, respectively, to the signs of the $\hat{\beta}$'s in the multiple regression model fit in Example 11.14. The fact that the $\hat{\beta}$'s change dramatically when the independent variables are removed from the model is another indication of a serious multicollinearity problem.

## 11.12 A Summary of the Steps to Follow in a Multiple Regression Analysis

We have discussed some of the methodology of **multiple regression analysis**, a technique for modeling a dependent variable $y$ as a function of several independent variables $x_1, x_2, \ldots, x_k$. The steps we follow in constructing and using multiple regression models are much the same as those for the simple straight-line models:

1. The form of the probabilistic model is hypothesized.
2. The model coefficients are estimated using least squares.
3. The probability distribution of $\varepsilon$ is specified and $\sigma^2$ is estimated.
4. The utility of the model is checked using the analysis of variance $F$ test and the multiple coefficient of determination $R^2$. The $T$ tests on individual $\beta$ parameters aid in deciding the final form of the model.
5. An analysis of residuals is conducted to determine if the data comply with the assumptions in step 3.
6. If the model is deemed useful and the assumptions are satisfied, it may be used to make estimates and to predict values of $y$ to be observed in the future.

We have covered steps 2–6 in this chapter, assuming that the model was specified. Chapter 12 is devoted to step 1—model construction.

## *STATISTICS IN ACTION REVISITED*
### Building a Model for Road Construction Costs in a Sealed Bid Market

We now return to the *Statistics in Action* problem described in the beginning of this chapter—to build a model for the cost (y) of a road construction contract awarded using the sealed-bid system. Recall that the **FLAG** file contains data for a sample of 235 road contracts. (The variables measured for each contract are listed in Table SIA11.1, p. 557.)

We begin our analysis by constructing scatterplots of the data, with the dependent variable cost ($y$) plotted against each of the potential predictors. These are shown in the MINITAB printout, Figure SIA11.1. It is also prudent to construct a matrix of pairwise correlations for the potential independent variables to check for multicollinearity. The MINITAB correlation matrix is shown in Figure SIA11.2. From the scatterplots, it appears that several of the independent variables—the DOT engineer's cost estimate (DOTEST), estimate of workdays (DAYSEST), and fixed or competitive bid status (STATUS)—would be good linear predictors of contract cost. However, Figure SIA11.2 shows that the correlation between DOTEST and DAYSEST is $r = .798$—a fairly high value. To avoid the problems that occur when multicollinearity is present in the data, we will fit a regression model using the two independent variables, $x_1 =$ DOT cost estimate and $x_2 =$ bid status. (Note that bid status is a qualitative variable. We learn in Chapter 12 to create the "dummy" variable,

$$x_2 = \begin{cases} 1 & \text{if fixed contract} \\ 0 & \text{if competitive contract} \end{cases}$$

for a two-level qualitative independent variable.)

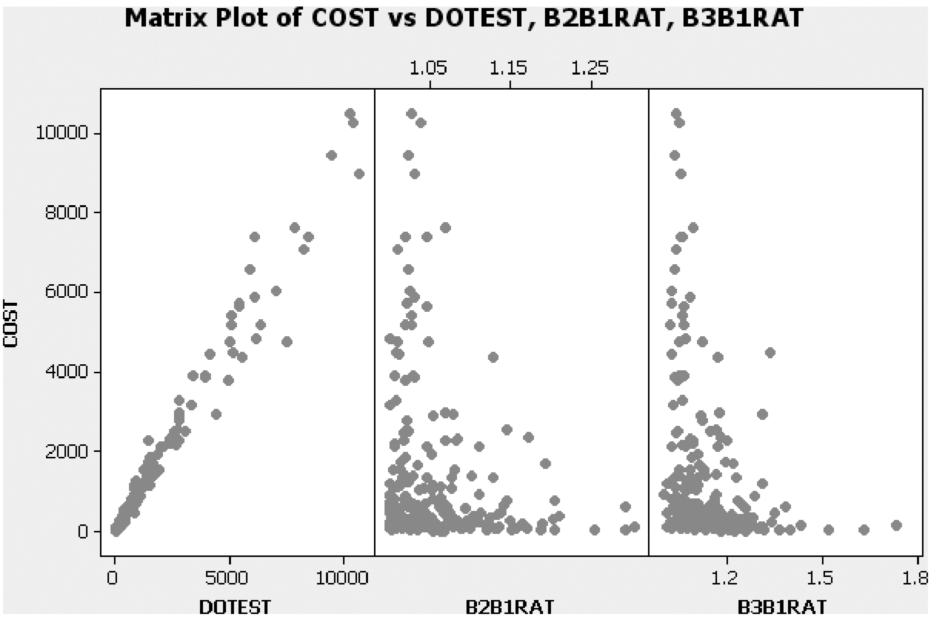
Initially, we consider an interaction model for contract cost ($y$). The model is given by the equation

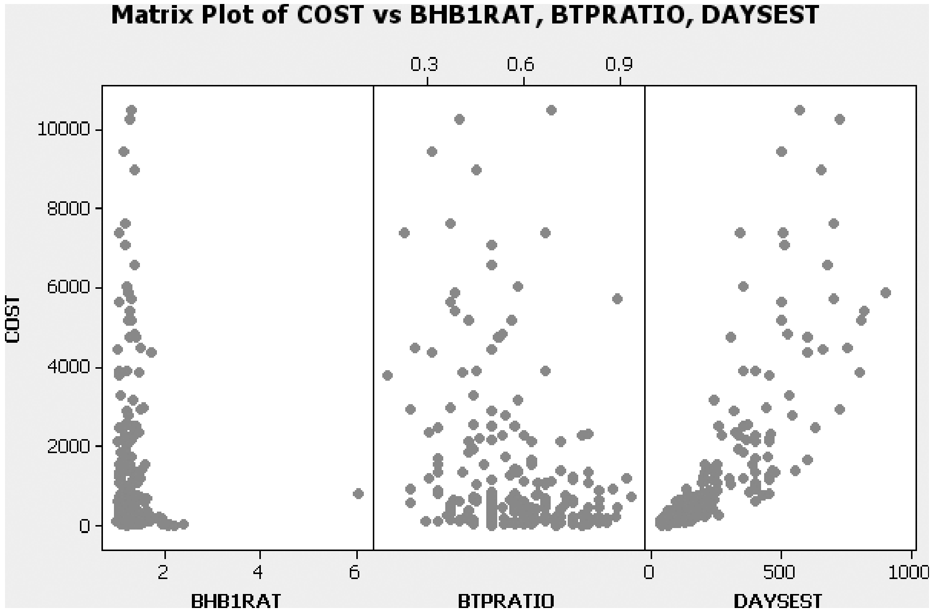
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

The MINITAB printout for this model is shown in Figure SIA11.3. The global $F$ statistic ($F = 3281.22$) and associated $p$-value (.000) shown on the printout indicate that the overall model is statistically useful for predicting construction cost. The value of $R_a^2$ indicates that the model can explain 97.7% of the sample variation in contract cost. Also, the $T$ test for the interaction term, $\beta_3 x_1 x_2$, is significant ($p$-value $= .000$), implying that the relationship between contract cost ($y$) and DOT cost estimate ($x_1$) depends on bid status (fixed or competitive). These results provide strong statistical support for using the model for estimation and prediction.

The nature of the interaction is illustrated in the MINITAB graph of the least-squares prediction equation for the reduced model, Figure SIA11.4. You can see that the rate of increase of contract cost ($y$) with

**FIGURE SIA11.1**
MINITAB scatterplots for **FLAG** data





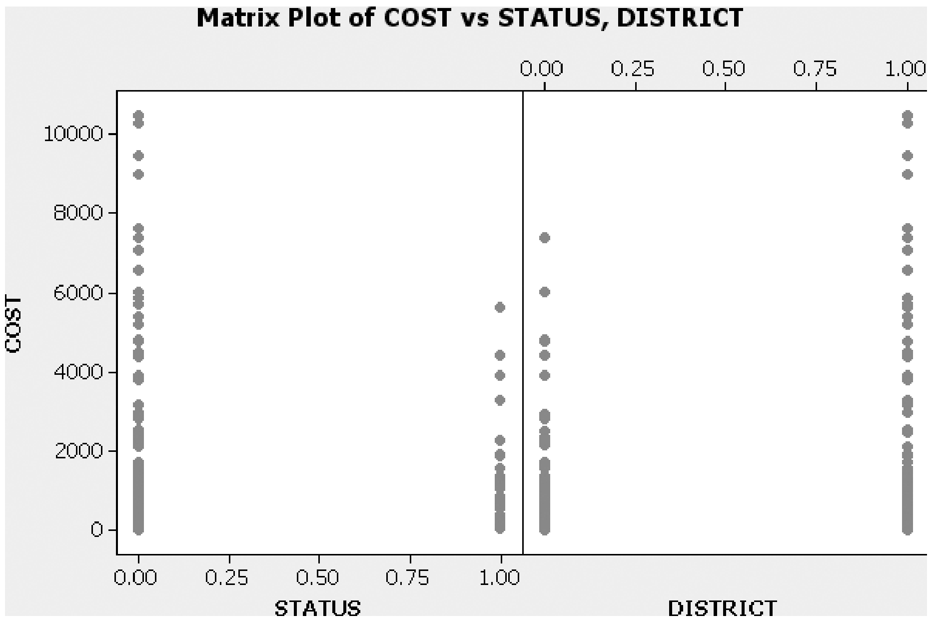




**FIGURE SA11.1 (*Continued*)**

| | DOTEST | B2B1RAT | B3B1RAT | BHB1RAT | STATUS | DISTRICT | BTPRATIO |
|---|---|---|---|---|---|---|---|
| B2B1RAT | -0.153 | | | | | | |
| B3B1RAT | -0.242 | 0.535 | | | | | |
| BHB1RAT | -0.054 | 0.199 | 0.268 | | | | |
| STATUS | -0.137 | -0.117 | -0.195 | -0.149 | | | |
| DISTRICT | 0.251 | 0.081 | -0.058 | -0.132 | 0.097 | | |
| BTPRATIO | -0.329 | -0.139 | -0.142 | -0.022 | 0.001 | -0.153 | |
| DAYSEST | 0.798 | -0.113 | -0.213 | 0.013 | -0.113 | 0.193 | -0.362 |

Cell Contents: Pearson correlation

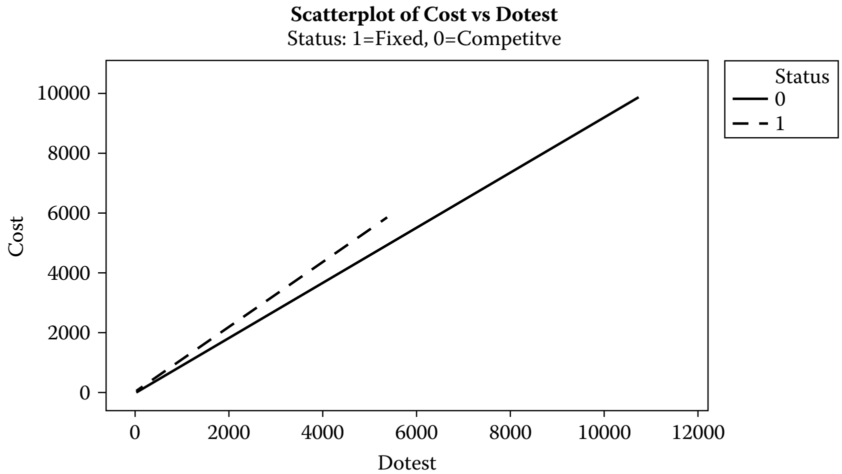
**FIGURE SIA11.2**

MINITAB correlation matrix for potential predictors of road construction cost

**FIGURE SIA11.3**
MINITAB regression printout for interaction model of cost

```
The regression equation is
COST = - 6.4 + 0.921 DOTEST + 28.7 STATUS + 0.163 STA_DOT


Predictor       Coef   SE Coef      T      P
Constant       -6.43     26.21  -0.25  0.806
DOTEST      0.921336  0.009723  94.75  0.000
STATUS         28.67     58.66   0.49  0.625
STA_DOT      0.16328   0.04043   4.04  0.000


S = 296.699   R-Sq = 97.7%   R-Sq(adj) = 97.7%


Analysis of Variance

Source           DF         SS         MS        F      P
Regression        3  866540004  288846668  3281.22  0.000
Residual Error  231   20334968      88030
Total           234  886874973
```

**FIGURE SIA11.4**
MINITAB plot of least-squares prediction equation for interaction model




the DOT engineer's estimate of cost ($x_1$) is steeper for fixed contracts than for competitive contracts. Before actually using the model in practice, we need to examine the residuals to be sure that the standard regression assumptions are reasonably satisfied.

Figures SIA11.5 and SIA11.6 are MINITAB graphs of the residuals from the interaction model. The histogram shown in Figure SIA11.5 appears to be approximately normally distributed; consequently, the assumption of normal errors is reasonably satisfied. The scatterplot of the residuals against $y$ shown in Figure SIA11.6, however, shows a distinct "funnel" pattern; this indicates that the assumption of a constant error variance is likely to be violated. One way to modify the model to satisfy this assumption is to use a variance-stabilizing transformation (such as the natural log) on cost ($y$). When both the $y$ and $x$ variables in a regression equation are economic variables (prices, costs, salaries, etc.), it is often advantageous to transform the $x$ variable also. Consequently, we'll modify the model by making a log transform on both cost ($y$) and DOTEST ($x_1$).

Our modified (log-log) interaction model takes the form

$$E(y^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 (x_1^*) x_2$$

where $y^* = \ln(\text{COST})$ and $x_1^* = \ln(\text{DOTEST})$. The MINITAB printout for this model is shown in Figure SIA11.7, followed by graphs of the residuals in Figures SIA11.8 and SIA11.9. The histogram shown in Figure SIA11.8 is approximately normal, and more importantly, the scatterplot of the residuals shown in Figure SIA11.9 has no distinct trend. It appears that the log transformations successfully stabilized the error variance. Note, however, that the $T$ test for the interaction term in the model (highlighted in Figure SIA11.7) is no longer

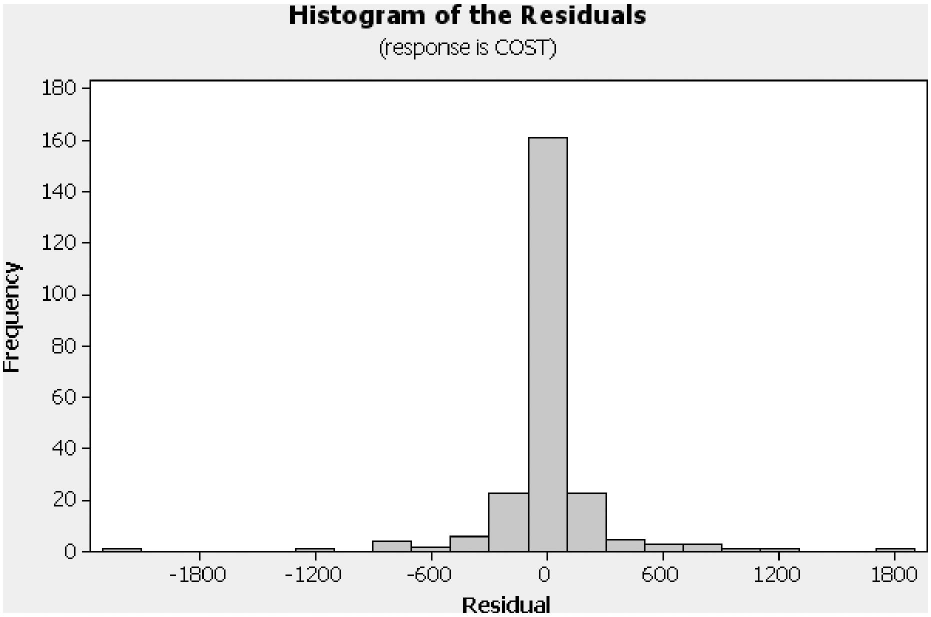



**FIGURE SIA11.5**
MINITAB histogram of residuals from interaction model

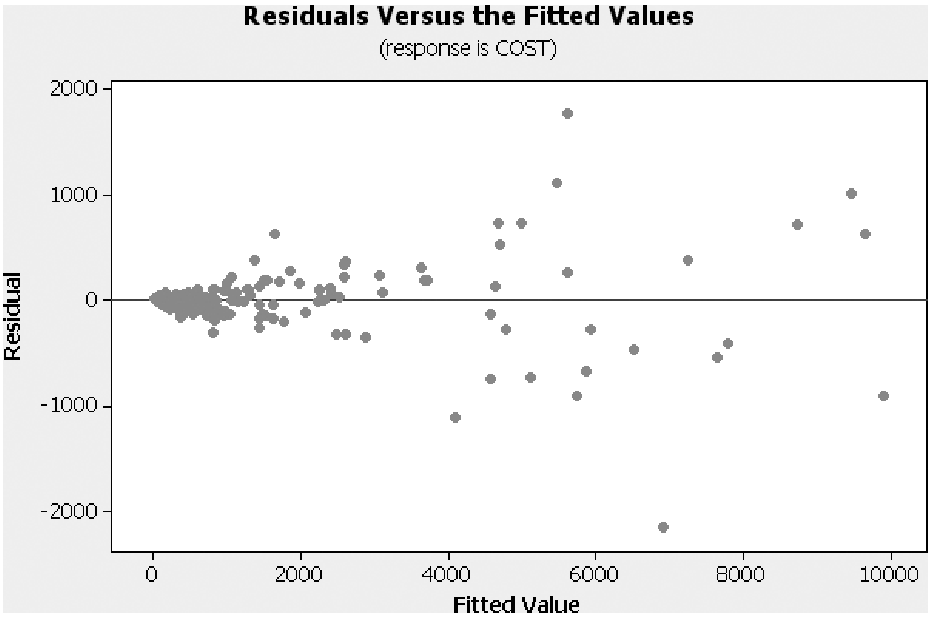



**FIGURE SIA11.6**
MINITAB plot of residuals for interaction model

statistically significant ($p$-value = .420). Consequently, we will drop the interaction term from the model and use the simpler modified model

$$E(y^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2$$

to predict road contract cost.

The MINITAB printout for the modified, no-interaction model is shown in Figure SIA11.10. The least-squares prediction equation is

$$\widehat{\ln(y)} = -0.147 + 1.01 \ln(x_1) + 0.217x_2$$

Suppose we want to predict road construction cost when the DOT estimate is \$370,000 (i.e., $x_1 = 370$) and the contract is fixed (i.e., $x_2 = 1$). Now $\ln(370) = 5.91$. Substituting these values into the prediction equation, we have

$$\widehat{\ln(y)} = -0.147 + 1.01 \ln(x_1) + 0.217x_2 = -0.147 + 1.01(5.91) + 0.217(1) = 6.02$$

**FIGURE SIA11.7**
MINITAB regression printout for modified (log-log) interaction model of road construction cost

```
The regression equation is
LNCOST = - 0.162 + 1.01 LNDOTEST + 0.324 STATUS - 0.0176 STA_LNDOT


Predictor       Coef   SE Coef       T      P
Constant    -0.16188   0.05193   -3.12  0.002
LNDOTEST     1.00780   0.00798  126.23  0.000
STATUS        0.3243    0.1356    2.39  0.018
STA_LNDOT   -0.01762   0.02181   -0.81  0.420


S = 0.154922   R-Sq = 98.8%   R-Sq(adj) = 98.7%


Analysis of Variance

Source           DF      SS      MS        F      P
Regression        3  439.64  146.55  6105.87  0.000
Residual Error  231    5.54    0.02
Total           234  445.18
```

**FIGURE SIA11.8**
MINITAB histogram of residuals from modified (log-log) model

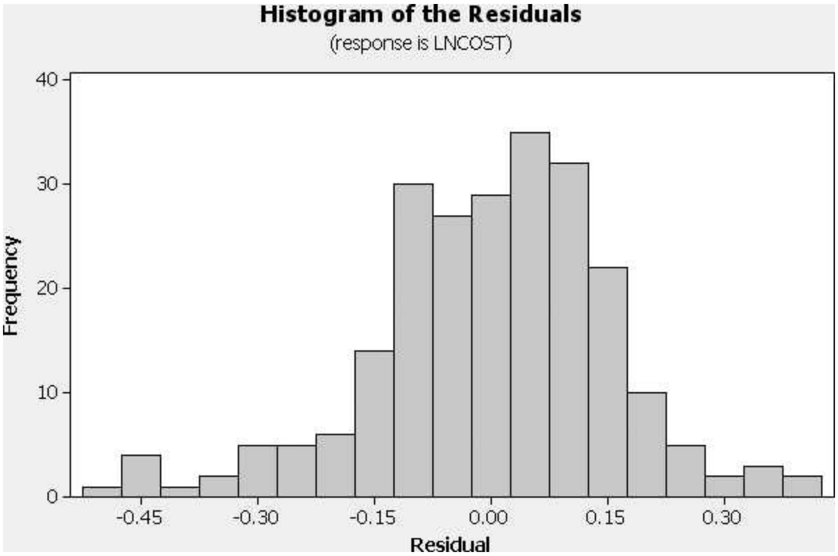



**FIGURE SIA11.9**
MINITAB plot of residuals from modified (log-log) model

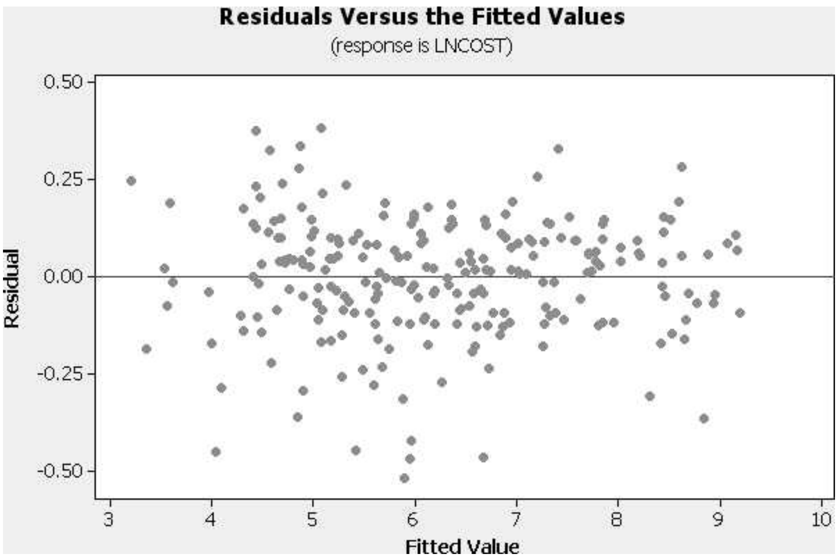


**FIGURE SIA11.10**
MINITAB regression printout for simpler, modified (log-log) model of road construction cost

```
The regression equation is
LNCOST = - 0.147 + 1.01 LNDOTEST + 0.217 STATUS


Predictor       Coef   SE Coef       T      P
Constant    -0.14684   0.04846   -3.03  0.003
LNDOTEST     1.00543   0.00742  135.45  0.000
STATUS       0.21656   0.02475    8.75  0.000


S = 0.154773   R-Sq = 98.8%   R-Sq(adj) = 98.7%


Analysis of Variance

Source           DF      SS      MS        F      P
Regression        2  439.61  219.80  9175.83  0.000
Residual Error  232    5.56    0.02
Total           234  445.16
Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  6.0152  0.0219  (5.9720, 6.0584)  (5.7072, 6.3232)


Values of Predictors for New Observations

New
Obs  LNDOTEST  STATUS
  1      5.91    1.00
```

To convert this predicted natural log value back into thousands of dollars, we take the antilogarithm, $e^{6.02} = 411.6$. Thus, for a fixed road contract with a DOT estimate of $370,000, the model predicts the cost to be $411.6 thousand.

Both the predicted ln($y$) and corresponding 95% prediction interval (5.71,6.32) are shown (highlighted) at the bottom of the MINITAB printout, Figure SIA 11.10. Taking antilogs of the interval endpoints, we obtain, $e^{5.71} = 301.9$ and $e^{6.32} = 555.6$. Consequently, the model predicts (with 95% confidence) that the cost for a fixed road contract with a DOT estimate of $370,000 will fall between $301.9 thousand and $555.6 thousand. Note the wide range of the prediction interval. This is due to the relatively large magnitude of the model standard deviation, $s$. Although the model has been deemed statistically useful for predicting contract cost, it may not be "practically" useful. To reduce the magnitude of $s$, the Florida attorney general will need to improve the model's predictive ability.

# Supplementary Review

## Key Terms

Adjusted multiple coefficient of determination 571
Analysis of variance $F$ test 626
Coded variable 558
Deleted residual 611
Designed experiments 619
Extrapolation 617
First-order model 582
Global $F$ test 570
Heteroscedastic errors 533
Higher-order term 558
Influential observation 609
Interaction 592
Interaction model 592
Interaction term 593
Jackknife 611
Multicollinearity 619
Multiple coefficient of determination 569
Multiple regression model 572
Multiplicative model 559
Observational data 620
Outlier 609
Parameter estimability 617
Quadratic model 597
Quadratic term 597
Qualitative variable 558
Residual 559
Residual analysis 559
Robust method 606
Second-order model 597
Second-order term 597
Standardized residual 607
Stepwise regression 623
Time series 606
Variance-stabilizing transformation 609

## Key Formulas

| | |
|---|---|
| $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ | Matrix representation of least-squares solution 562 |
| $\text{SSE} = \boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}$ | Matrix representation of sum of squared errors 565 |
| $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-(k+1)}$ | Estimator of $\sigma^2$ for a model with $k$ independent variables 565 |
| $\text{Var}(\hat{\beta}_j) = c_{jj}\sigma^2$, where $c_{jj}$ is $j$th diagonal element of $(\boldsymbol{X'X})^{-1}$ | Variance of a beta estimate 564 |
| $T = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$ | Test statistic for testing $H_0$: $\beta_i = 0$ 567 |
| $\hat{\beta}_i \pm (t_{\alpha/2})s_{\hat{\beta}_i}$ where $t$ depends on $n-(k+1)$ df | $100(1-\alpha)\%$ confidence interval for $\beta_i = 0$ 566 |
| $R^2 = \dfrac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}}$ | Multiple coefficient of determination 569 |
| $R_a^2 = 1 - \left[\dfrac{(n-1)}{n-(k+1)}\right](1-R^2)$ | Adjusted multiple coefficient of determination 571 |
| $F = \dfrac{\text{MS(Model)}}{\text{MSE}} = \dfrac{R^2/k}{(1-R^2)/[n-(k+1)]}$ | Test statistic for testing $H_0$: $\beta_i = \beta_2 = \cdots = \beta_k = 0$ 570 |
| $\hat{y} \pm t_{\alpha/2}(s)\sqrt{\boldsymbol{a}'(\boldsymbol{X'X})^{-1}\boldsymbol{a}}$ | $100(1-\alpha)\%$ confidence interval for $E(y)$ when $\boldsymbol{a}' = [1\ x_1\ x_2\ \ldots\ x_k]$ 575 |
| $\hat{y} \pm t_{\alpha/2}(s)\sqrt{1 + \boldsymbol{a}'(\boldsymbol{X'X})^{-1}\boldsymbol{a}}$ | $100(1-\alpha)\%$ prediction interval for $y$ when $\boldsymbol{a}' = [1\ x_1\ x_2\ \ldots\ x_k]$ 575 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ | First-order model with two quantitative independent variables 582 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ | Interaction model with two quantitative independent variables 593 |
| $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ | Quadratic model 598 |
| $y - \hat{y} = \hat{\varepsilon}$ | Regression residual 605 |
| $(y - \hat{y})/s$ | Standardized residual 605 |
| $y_i - \hat{y}_{(i)}$ | Deleted residual 605 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $x_1^2$ | $x$-1 squared | Quadratic term that allows for curvature in the relationship between $y$ and $x$ |
| $x_1x_2$ | $x$-1 $x$-2 | Interaction term |
| MSE | M-S-E | Mean square for error (estimates $\sigma^2$) |
| $\beta_i$ | beta-$i$ | Coefficient of $x_i$ in the model |
| $\hat{\beta}_i$ | beta-$i$-hat | Least-squares estimate of $\beta_i$ |
| $s_{\hat{\beta}_i}$ | $s$ of beta-$i$-hat | Estimated standard error of $\hat{\beta}_i$ |
| $R^2$ | $R$-squared | Multiple coefficient of determination |
| $R_a^2$ | $R$-squared adjusted | Adjusted multiple coefficient of determination |
| $F$ | | Test statistic for testing global usefulness of model |
| $\hat{\varepsilon}$ | epsilon-hat | Estimated random error, or residual |
| $\ln(y)$ | Natural log of $y$ | Natural logarithm of dependent variable |

## Chapter Summary Notes

- **Steps in multiple regression**: (1) Hypothesize the deterministic form of the model, (2) use the method of least squares to estimate the unknown $\beta$'s, (3) make assumptions on the random error ($\varepsilon$), (4) check the assumptions and make model modifications, (5) statistically evaluate the adequacy of the model, (6) if deemed useful, use the model for estimation and prediction.
- **Four assumptions for $\varepsilon$**: (1) mean of $\varepsilon$ is 0, (2) variance of $\varepsilon$ is constant for all $x$ values, (3) distribution of $\varepsilon$ is normal, (4) values of $\varepsilon$ are independent.
- **First-order model in $k$ quantitative $x$'s:** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$, where each $\beta_i$ represents the change in $y$ for every 1-unit increase in $x_i$, holding the other $x$'s fixed.
- **Adjusted coefficient of determination ($R_a^2$)** cannot be "forced" to 1 by adding independent variables to the model.
- **Recommendation for checking statistical utility of the model:** (1) Conduct the **global $F$ test**, (2) if test is significant, conduct $T$ tests on the "most important" $\beta$'s only, (3) interpret the value of $2s$, (4) interpret the value of $R_a^2$.
- **Interaction model in 2 quantitative $x$'s:** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, where $(\beta_1 + \beta_3 x_2)$ represents the change in $y$ for every 1-unit increase in $x_1$ for fixed $x_2$, and $(\beta_2 + \beta_3 x_1)$ represents the change in $y$ for every 1-unit increase in $x_2$ for fixed $x_1$.
- Once interaction is tested and deemed important, avoid conducting $T$ tests on the first-order terms in the model.
- **Quadratic model in 1 quantitative $x$:** $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ where $\beta_2 > 0$ implies *upward* curvature and $\beta_2 < 0$ implies *downward* curvature.
- Once curvature is tested and deemed important, avoid conducting a $t$ test on the first-order term in the model.
- Properties of **regression residuals**: (1) sum of residuals $= 0$, (2) sum of squared residuals $=$ SSE.
- To detect a **misspecified model**: Plot residuals against each quantitative $x$ in the model—look for trends (e.g., curvilinear trend).
- To identify **outliers**: Find residuals that are greater than $3s$ in absolute value.
- To identify **influential observations**: Find **deleted residuals** that are greater than $3s$ in absolute value.
- To detect **non-normal errors**: Graph residuals in a histogram, stem-and-leaf plot, or normal probability plot—look for strong departures from normality.
- To detect a **nonconstant error variance** (e.g., *heteroscedasticity*): Plot residuals against $\hat{y}$—look for patterns (e.g., cone-shaped pattern).
- **Multicollinearity** occurs when two or more of the $x$'s in the model are correlated.
- Indicators of multicollinearity: (1) highly correlated $x$'s, (2) significant global $F$ test but all $T$ tests nonsignificant, (3) signs on the $\beta$ estimates opposite from expected.
- **Extrapolation** occurs when you predict $y$ for values of the $x$'s that are outside of the range of the sample data.

# Supplementary Applied Exercises

11.62 *Vehicle congestion study.* Refer to the *Journal of Engineering for Industry* study of vehicle congestion in an automated warehouse, Exercise 10.73. (p. 550). The data on number of vehicles ($x$) and congestion time ($y$) are reproduced in the table. Consider the straight-line model $E(y) = \beta_0 + \beta_1 x$.

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$.
c. Find the least-squares estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. [*Note*: See Theorem A.1 in Appendix A for information on finding $(\boldsymbol{X'X})^{-1}$.]
d. Find SSE and $s^2$.
e. Conduct the test $H_0$: $\beta_1 = 0$ vs. $H_a$: $\beta_1 > 0$ at $\alpha = .01$.
f. Find and interpret $R^2$.
g. Find and interpret a 99% prediction interval for $y$ when $x = 5$.

**WAREHOUSE**

| Number of Vehicles, $x$ | Congestion Time, $y$ (minutes, hundredths) | Number of Vehicles, $x$ | Congestion Time, $y$ (minutes, hundredths) |
|---|---|---|---|
| 1 | 0 | 9 | 2 |
| 2 | 0 | 10 | 4 |
| 3 | 2 | 11 | 4 |
| 4 | 1 | 12 | 4 |
| 5 | 1 | 13 | 3 |
| 6 | 1 | 14 | 4 |
| 7 | 3 | 15 | 5 |
| 8 | 3 | | |

*Source:* Pandit, R., and U. S. Palekat, "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry*, Transactions of the ASME, Vol. 115, Aug. 1993, p. 326 (Table 2).

11.63 *Optical density of a liquid.* Poly (perfluoropropylene-oxide), i.e., PPFPO, is a viscous liquid used extensively in the electronics industry as a lubricant. In a study reported in *Applied Spectroscopy* (Jan. 1986), the infrared reflectance spectra properties of PPFPO were examined. The optical density ($y$) for the prominent infrared absorption of PPFPO was recorded for different experimental settings of band frequency ($x_1$) and film thickness ($x_2$) in a Perkin–Elmer Model 621 infrared spectrometer. The results are given in the table below. Consider the first-order model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

**PPFPO**

| Optical Density $y$ | Band Frequency $x_1$, cm$^{-1}$ | Film Thickness $x_2$, milligrams |
|---|---|---|
| .231 | 740 | 1.1 |
| .107 | 740 | .62 |
| .053 | 740 | .31 |
| .129 | 805 | 1.1 |
| .069 | 805 | .62 |
| .030 | 805 | .31 |
| 1.005 | 980 | 1.1 |
| .559 | 980 | .62 |
| .321 | 980 | .31 |
| 2.948 | 1,235 | 1.1 |
| 1.633 | 1,235 | .62 |
| .934 | 1,235 | .31 |

*Source:* Pacansky, J., England, C. D., and Waltman, R. "Infrared spectroscopic studies of poly (perfluoropropyleneoxide) on gold substrates: A classical dispersion analysis for the refractive index." *Applied Spectroscopy*, Vol. 40, No. 1, Jan. 1986, p. 9 (Table 1).

a. Construct $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices for the data.
b. Find $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$.
c. Use the technique outlined in Appendix A.4 to find $(\boldsymbol{X'X})^{-1}$. (Be sure to carry out your calculations to six significant digits.)
d. Find the $\hat{\boldsymbol{\beta}}$ matrix and the least-squares prediction equation.
e. Find $s$ and interpret its value.
f. Find $R_a^2$ and interpret its value.
g. Conduct a test of overall model adequacy using $\alpha = .10$.
h. Find and interpret a 90% confidence interval for $\beta_1$.
i. Find and interpret a 90% confidence interval for $\beta_2$.
j. Find and interpret a 90% prediction interval for $y$ when $x_1 = 950$ and $x_2 = .62$.

**PPFPO**

11.64 *Optical density of a liquid.* Refer to the *Applied Spectroscopy* study of the optical density ($y$) for infrared absorption of the liquid PPFPO, Exercise 11.63 (p. 635). In addition to optical density, band frequency ($x_1$) and film thickness ($x_2$) were measured for 12 experiments. (The data are saved in the **PPFPO** file.)

a. Write an interaction model for optical density ($y$) as a function of band frequency ($x_1$) and film thickness ($x_2$).
b. Give a practical explanation of the statement, "band frequency ($x_1$) and film thickness ($x_2$) interact."
c. A MINITAB printout for the interaction model, part **a**, is shown at the bottom of the page. Give the least-squares prediction equation.
d. Is there sufficient evidence (at $\alpha = .01$) of interaction between band frequency and film thickness?
e. For each level of film thickness ($x_2$), use the $\beta$ estimates of the model to sketch the relationship between optical density ($y$) band frequency ($x_1$).

11.65 *Chemical yield study.* An experiment was conducted to investigate the effect of temperature ($T$) and pressure ($P$) on the yield $y$ of a chemical. Each of the two factors, temperature and pressure, was held constant at two levels—$T$ at 50° and 70°, $P$ at 10 pounds per square inch and 20 pounds per square inch—and the yield of each of the four combinations was measured.

**MINITAB Output for Exercise 11.64**

```
The regression equation is
DENSITY = - 0.214 + 0.000257 FREQ - 3.72 THICK + 0.00497 FREQ_THICK


Predictor          Coef     SE Coef      T      P
Constant        -0.2143      0.6866  -0.31  0.763
FREQ          0.0002567   0.0007157   0.36  0.729
THICK           -3.7200      0.9147  -4.07  0.004
FREQ_THICK    0.0049655   0.0009535   5.21  0.001


S = 0.205676   R-Sq = 96.0%   R-Sq(adj) = 94.5%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        3  8.0481  2.6827  63.42  0.000
Residual Error    8  0.3384  0.0423
Total            11  8.3865
```

The results are shown in the accompanying table.

**CHEMYLD**

| Temperature | Pressure | Yield |
|---|---|---|
| 50 | 10 | 24.5 |
| 50 | 10 | 26.0 |
| 50 | 20 | 28.4 |
| 50 | 20 | 28.1 |
| 70 | 10 | 22.1 |
| 70 | 10 | 20.8 |
| 70 | 20 | 16.7 |
| 70 | 20 | 15.3 |

a. Fit the linear model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

to the data.

b. Conduct a test of overall model adequacy. Use $\alpha = .01$.

c. Is there evidence of interaction between temperature and pressure? Test using $\alpha = .01$.

11.66 *Permeability of concrete.* J. Vuorinen carried out a series of experiments to gather information on the coefficient of permeability of concrete (*Magazine of Concrete Research*, Sept. 1985). In one experiment, the outflow of water from the pores of a concrete specimen after it had been under saturating water pressure for a period of time was recorded for different combinations of concrete permeability and porosity. The resulting water quantities after different lapses of time for one permeability–porosity combination are given in the table.

**CONPERM1**

| Time $t$, seconds | Water Outflow $w$, grams per cylinder |
|---|---|
| 201 | 3.88 |
| 325 | 4.93 |
| 525 | 6.42 |
| 775 | 7.80 |
| 975 | 8.72 |
| 1,200 | 9.60 |

*Source*: Vuorinen, J. "Applications of diffusion theory to permeability tests on concrete, Part II: Pressure-saturation test on concrete and coefficient of permeability." *Magazine of Concrete Research*, Vol. 37, No. 132, Sept. 1985, p. 156 (Table II.1).

a. According to Vuorinen, "the quantity of water discharged is approximately in linear relationship with the square root of time" for most of the permeability–porosity combinations. Fit the following model to the data in the table:

$$E(w) = \alpha_0 + \alpha_1 \sqrt{t}$$

b. Is there sufficient evidence to indicate that quantity of water outflow and the square root of time are linearly related? Test using $\alpha = .10$.

11.67 *Permeability of concrete (continued).* Refer to Exercise 11.66. Vuorinen fit the water outflow–time model for each of nine permeability–porosity combinations and used the results to develop a model for the coefficient of permeability of concrete, $y$. Specifically, he fit the model*

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$x_1$ = Porosity of the cement

$x_2$ = Estimated slope coefficient ($\hat{\alpha}_1$) of the corresponding water outflow–time regression line

The data are reproduced in the next table.

**CONPERM2**

| Coefficient of Permeability $y$, (meters per second) $\times 10^{-11}$ | Porosity $x_1$ | Estimated Water Outflow Time Slope Coefficient $x_2$ |
|---|---|---|
| 1.00 | .050 | .903 |
| 1.00 | .035 | .722 |
| 1.00 | .025 | .590 |
| .10 | .050 | .345 |
| .10 | .035 | .282 |
| .10 | .025 | .233 |
| .01 | .050 | .103 |
| .01 | .035 | .091 |
| .01 | .025 | .078 |

*Source:* Vuorinen, J. "Applications of diffusion theory to permeability tests on concrete, Part II: Pressure-saturation test on concrete and coefficient of permeability." *Magazine of Concrete Research*, Vol. 37, No. 132, Sept. 1985, p. 156 (Table II.1).

a. Find the least-squares prediction equation and interpret the $\beta$ estimates.

b. Conduct a test of overall model utility. Interpret the $p$-value of the test.

c. Is there evidence that concrete porosity $x_1$ is a useful predictor of coefficient of permeability $y$? Test using $\alpha = .05$.

d. Is there evidence that the estimated water outflow–time slope is a useful predictor of coefficient of permeability $y$? Test using $\alpha = .05$.

*In actuality, Vuorinen fit the logarithmic model $\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \varepsilon$

e. Find $R^2$ and interpret its value.
f. Find the estimate of $\sigma$ and interpret its value.
g. Find a 95% prediction interval for $y$ when $x_1 = .05$ and $x_2 = .30$. Interpret the result.

11.68 *Snow geese feeding trial.* The *Journal of Applied Ecology* (Vol. 32, 1995) published a study of the feeding habits of baby snow geese. The data on gosling weight change, digestion efficiency, acid-detergent fiber (all measured as percentages) and diet (plants or duck chow) for 42 feeding trials are saved in the **SNOWGEESE** file. Selected observations are shown in the following table. The botanists were interested in predicting weight change ($y$) as a function of the other variables. Consider the first-order model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $x_1$ is digestion efficiency and $x_2$ is acid-detergent fiber.

a. Find the least-squares prediction equation for weight change, $y$.
b. Interpret the $\beta$-estimates in the equation, part **a**.
c. Conduct a test to determine if digestion efficiency, $x_1$, is a useful linear predictor of weight change. Use $\alpha = .01$.
d. Form a 99% confidence interval for $\beta_2$. Interpret the result.
e. Find and interpret $R^2$ and $R_a^2$. Which statistic is the preferred measure of model fit? Explain.
f. Is the overall model statistically useful for predicting weight change? Test using $\alpha = .05$.

**SNOWGEESE**

(*First and last five trials shown*)

| Feeding Trial | Diet | Weight Change (%) | Digestion Efficiency (%) | Acid-Detergent Fiber (%) |
|---|---|---|---|---|
| 1 | Plants | −6 | 0 | 28.5 |
| 2 | Plants | −5 | 2.5 | 27.5 |
| 3 | Plants | −4.5 | 5 | 27.5 |
| 4 | Plants | 0 | 0 | 32.5 |
| 5 | Plants | 2 | 0 | 32 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 38 | Duck Chow | 9 | 59 | 8.5 |
| 39 | Duck Chow | 12 | 52.5 | 8 |
| 40 | Duck Chow | 8.5 | 75 | 6 |
| 41 | Duck Chow | 10.5 | 72.5 | 6.5 |
| 42 | Duck Chow | 14 | 69 | 7 |

*Source:* Gadallah, F. L., and Jefferies, R. L., "Forage quality in brood rearing areas of the lesser snow goose and the growth of captive goslings." *Journal of Applied Biology*, Vol. 32, No. 2, 1995, pp. 281–282 (adapted from Figures 2 and 3).

11.69 *Solar lighting with lasers.* Engineers at the University of Massachusetts studied the viability of using semiconductor lasers for solar lighting in spaceborne applications (*Journal of Applied Physics*, Sept. 1993). A series of $n = 8$ experiments with quantum-well lasers yielded the following observations on solar pumping threshold current ($y$) and waveguide Al mole fraction ($x$):

**SOLAR2**

| Threshold Current $y$, A/cm$^{-2}$ | Waveguide Al Mole Fraction $x$ |
|---|---|
| 273 | .15 |
| 175 | .20 |
| 146 | .25 |
| 166 | .30 |
| 162 | .35 |
| 165 | .40 |
| 245 | .50 |
| 314 | .60 |

*Source:* Unnikrishnan, S., and Anderson, N. G. "Quantum-well lasers for direct solar photopumping." *Journal of Applied Physics*, Vol. 74, No. 6, Sept. 15, 1993, p. 4226 (data adapted from Figure 2).

a. The researchers theorize that the relationship between threshold current ($y$) and waveguide Al composition ($x$) will be represented by a U-shaded curve. Hypothesize a model that corresponds to this theory.
b. Plot the data points in a scattergram. Comment on the researchers' theory.
c. Fit the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, to the data using the method of least squares.
d. Conduct a formal test of the researchers' theory. Use $\alpha = .10$.

11.70 *Plastic extrusion experiment.* An experiment was conducted to investigate the effect of extrusion pressure $P$ and temperature at extrusion $T$ on the strength $y$ of a new type of plastic. Two plastic specimens were prepared for each of five combinations of pressure and temperature. The specimens were then tested in random order, and the breaking strength for each specimen was recorded. The independent variables were coded as follows to simplify computations:

$$x_1 = \frac{P - 200}{10}$$
$$x_2 = \frac{T - 400}{25}$$

The $n = 10$ data points are listed in the table.

**PLASTIC**

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| 5.2; 5.0 | −2 | 2 |
| .3; −.1 | −1 | −1 |
| −1.2; −1.1 | 0 | −2 |
| 2.2; 2.0 | 1 | −1 |
| 6.2; 6.1 | 2 | 2 |

a. Give the $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices needed to fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.
b. Find the least-squares prediction equation. Interpret the $\beta$ estimates.
c. Find SSE, $s^2$, and $s$. Interpret the value of $s$.
d. Does the model contribute information for the prediction of $y$? Test using $\alpha = .05$.
e. Find $R^2$ and interpret its value.
f. Test the null hypothesis that $\beta_1 = 0$. Use $\alpha = .05$. What is the practical implication of the test?
g. Find a 90% confidence interval for the mean strength of the plastic for $x_1 = -2$ and $x_2 = 2$.
h. Suppose a single specimen of the plastic is to be installed in the engine mount of a Douglas DC-10 aircraft. Find a 90% prediction interval for the strength of this specimen if $x_1 = -2$ and $x_2 = 2$.

11.71 *Walking study.* The *American Scientist* (July–Aug. 1998) published a study of the relationship between self-avoiding and unrooted walks. In a self-avoiding walk you never retrace or cross your own path, whereas an unrooted walk is a path in which the starting and ending points are impossible to distinguish. The possible number of walks of each type of various lengths are recorded in the accompanying table. Suppose you want to model the number of unrooted walks ($y$) as a function of walk length ($x$). Consider the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Is there sufficient evidence of an upward concave curvilinear relationship between $y$ and $x$? Test at $\alpha = .10$.

**WALK**

| Walk Length (Number of steps) | Unrooted Walks | Self-Avoiding Walks |
|---|---|---|
| 1 | 1 | 4 |
| 2 | 2 | 12 |
| 3 | 4 | 36 |
| 4 | 9 | 100 |
| 5 | 22 | 284 |
| 6 | 56 | 780 |
| 7 | 147 | 2,172 |
| 8 | 388 | 5,916 |

*Source:* Hayes, B. "How to avoid yourself." *American Scientist*, Vol. 86, No. 4, July–Aug. 1998, p. 317 (Figure 5).

11.72 *Deep space survey of quasars.* A quasar is a distant celestial object (at least 4 billion light-years away) that provides a powerful source of radio energy. *The Astronomical Journal* (July 1995) reported on a study of 90 quasars detected by a deep space survey. The survey enabled astronomers to measure several different quantitative characteristics of each quasar, including redshift range, line flux (erg/cm$^2$·s), line luminosity (erg/s), $AB_{1450}$ magnitude, absolute magnitude, and rest frame equivalent width. The data for a sample of 25 large (redshift) quasars are saved in the **QUASAR** file. (Several quasars are listed in the table on the next page.)

a. Hypothesize a first-order model for equivalent width, $y$, as a function of the first four variables in the table.
b. Fit the first-order model to the data. Give the least-squares prediction equation.
c. Interpret the $\beta$ estimates in the model.
d. Test the overall adequacy of the model using $\alpha = .05$.
e. Test to determine whether redshift ($x_1$) is a useful linear predictor of equivalent width ($y$), using $\alpha = .05$.
f. Find and interpret a 95% prediction interval for equivalent width ($y$) for the first quasar listed in the **QUASAR** file.

11.73 *Urban air analysis.* Chemical engineers at Tokyo Metropolitan University analyzed urban air specimens for the presence of low molecular weight dicarboxylic acid (*Environmental Science & Engineering*, Oct. 1993). The dicarboxylic acid (as a percent of total carbon) and oxidant concentrations for 19 air specimens collected from urban Tokyo are listed in the table on below. Consider a straight-line model relating dicarboxylic acid percentage ($y$) to oxidant concentration ($x$). Conduct a complete residual analysis.

**URBANAIR**

| Dicarboxylic Acid % | Oxidant ppm | Dicarboxylic Acid % | Oxidant ppm |
|---|---|---|---|
| .85 | 78 | .50 | 32 |
| 1.45 | 80 | .38 | 28 |
| 1.80 | 74 | .30 | 25 |
| 1.80 | 78 | .70 | 45 |
| 1.60 | 60 | .80 | 40 |
| 1.20 | 62 | .90 | 45 |
| 1.30 | 57 | 1.22 | 41 |
| .20 | 49 | 1.00 | 34 |
| .22 | 34 | 1.00 | 25 |
| .40 | 36 | | |

*Source*: Kawamura, K., and Ikushima, K. "Seasonal changes in the distribution of dicarboxylic acids in the urban atmosphere." *Environmental Science & Technology*, Vol. 27, No. 10, Oct. 1993, p. 2232 (data extracted from Figure 4).

11.74 *Elastic properties of moissanite.* Moissanite is a popular abrasive material because of its extreme hardness. Another important property of moissanite is elasticity. The elastic properties of the material were investigated in the *Journal of Applied Physics* (Sept. 1993). A diamond anvil cell was used to compress a mixture of moissanite, sodium chloride, and gold in a ratio of 33:99:1 by volume. The

## Data for Exercise 11.72

**QUASAR**

(*First five quasars shown.*)

| Quasar | Redshift ($x_1$) | Line Flux ($x_2$) | Line Luminosity ($x_3$) | $AB_{1450}$ $x_4$ | Absolute Magnitude ($x_5$) | Rest Frame Equivalent Width ($y$) |
|---|---|---|---|---|---|---|
| 1 | 2.81 | −13.48 | 45.29 | 19.50 | −26.27 | 117 |
| 2 | 3.07 | −13.73 | 45.13 | 19.65 | −26.26 | 82 |
| 3 | 3.45 | −13.87 | 45.11 | 18.93 | −27.17 | 33 |
| 4 | 3.19 | −13.27 | 45.63 | 18.59 | −27.39 | 92 |
| 5 | 3.07 | −13.56 | 45.30 | 19.59 | −26.32 | 114 |

*Source:* Schmidt, M., Schneider, D. P., and Gunn, J. E. "Spectroscopic CCD surveys for quasars at large redshift." *The Astronomical Journal*, Vol. 110, No. 1, July 1995, p. 70 (Table 1).

compressed volume, $y$, of the mixture (relative to the zero-pressure volume) was measured at each of 11 different pressures (GPa). The results are displayed in the table below.

a. Fit the straight-line regression model $E(y) = \beta_0 + \beta_1 x$ to the data.
b. Calculate the regression residuals for this model.
c. Plot the residuals against $x$. Do you detect a trend?
d. Propose an alternative model based on the plot, part **c**.
e. Fit and analyze the model, part **d**.

**ELASTICITY**

| Compressed Volume $y$, % | Pressure $x$, GPa | Compressed Volume $y$, % | Pressure $x$, GPa |
|---|---|---|---|
| 100 | 0 | 85.2 | 51.6 |
| 96 | 9.4 | 83.3 | 60.1 |
| 93.8 | 15.8 | 82.9 | 62.6 |
| 90.2 | 30.4 | 82.9 | 62.6 |
| 87.7 | 41.6 | 81.7 | 68.4 |
| 86.2 | 46.9 | | |

*Source*: Bassett, W. A., Weathers, M. S., and Wu, T. G. "Compressibility of SiC up to 68.4 GPa." *Journal of Applied Physics*, Vol. 74, No. 6, Sept. 15, 1993, p. 3825 (Table 1).

11.75 *Soil loss during rainfall.* Phosphorus used in soil fertilizers can contaminate freshwater sources during rainfall runoff. Consequently, it is important for water-quality engineers to estimate the amount of dissolved phosphorus in the water. *Geoderma* (June 1995) presented an investigation of the relationship between soil loss and percentage of dissolved phosphorus in water samples collected at 20 fertilized watersheds in Oklahoma. The data are given in the table.

a. Plot the data in a scattergram. Do you detect a linear or curvilinear trend?
b. Fit the quadratic model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ to the data.
c. Conduct a test to determine if a curvilinear relationship exists between dissolved phosphorus percentage ($y$) and soil loss ($x$). Test using $\alpha = .05$.

**PHOSPHOR**

| Watershed | Soil Loss $x$ (kilometers per half-acre) | Dissolved Phosphorus Percentage $y$ |
|---|---|---|
| 1 | 18 | 42.3 |
| 2 | 17 | 50.2 |
| 3 | 35 | 52.7 |
| 4 | 16 | 77.1 |
| 5 | 14 | 36.8 |
| 6 | 54 | 17.5 |
| 7 | 153 | 66.4 |
| 8 | 81 | 67.5 |
| 9 | 183 | 28.9 |
| 10 | 284 | 15.1 |
| 11 | 767 | 20.1 |
| 12 | 148 | 38.3 |
| 13 | 649 | 5.6 |
| 14 | 479 | 8.6 |
| 15 | 1,371 | 5.5 |
| 16 | 9,150 | 4.6 |
| 17 | 15,022 | 2.2 |
| 18 | 69 | 77.9 |
| 19 | 4,392 | 7.8 |
| 20 | 312 | 42.9 |

*Source:* Sharpley, A. N., Robinson, J. S., and Smith S. J. "Bioavailable phosphorus dynamics in agricultural soils and effects on water quality." *Geoderma*, Vol. 67, No. 1–2, June 1995, p. 11 (Table 4).

**ASWELLS**

11.76 *Arsenic in groundwater.* Refer to the *Environmental Science & Technology* (Jan. 2005) study of the reliability of a commercial kit to test for arsenic in groundwater, Exercise 11.24 (p. 590). Recall that you fit a first-order model for arsenic level ($y$) as a function of latitude, longitude, and depth to the data saved in the **ASWELLS** file. Conduct a complete residual analysis of the model. Do you recommend any model modifications?

**PONDICE**

11.77 *Characteristics of sea ice meltponds.* Refer to the surface albedo study of pond ice, Exercise 11.43 (p. 604). Recall that you fit a second-order model for broadband surface albedo level ($y$) as a function of pond depth ($x$) to data saved in the **PONDICE** file. Conduct a complete residual analysis of the model. Do you recommend any model modifications?

11.78 *Prototyping in information systems.* To meet the increasing demand for new software products, many systems development experts have adopted a prototyping methodology. The effect of prototyping on the system development life cycle (SDLC) was investigated in the *Journal of Computer Information Systems* (Spring 1993). A survey of 500 randomly selected corporate-level management information systems (MIS) managers was conducted. Three potential independent variables were: (1) *importance* of prototyping to each phase of the SDLC; (2) degree of *support* prototyping provides for the SDLC; and (3) degree to which prototyping *replaces* each phase of the SDLC. The table (next column) gives the pairwise correlations of the three variables in the survey data for one particular phase of the SDLC. Use this information to assess the degree of multicollinearity in the survey data. Would you recommend using all three independent variables in a regression analysis? Explain.

| Variable Pairs | Correlation Coefficient, $r$ |
|---|---|
| Importance–Replace | .2682 |
| Importance–Support | .6991 |
| Replace–Support | $-.0531$ |

*Source:* Hardgrave, B. C., Doke, E. R., and Swanson, N. E, "Prototyping effects of the system development life cycle: An empirical study." *Journal of Computer Information Systems*, Vol. 33, No. 3, Spring 1993, p. 16 (Table 1).

11.79 *Sintering experiment. Sintering*, one of the most important techniques of materials science, is used to convert a powdered material into a porous solid body. The following two measures characterize the final product:

$$V_v = \text{Percentage of total volume of final product that is solid}$$
$$= \left(\frac{\text{Solid volume}}{\text{Porous volume} + \text{Solid volume}}\right) \cdot 100$$
$$S_v = \text{Solid–pore interface area per unit volume of the product}$$

When $V_v = 100\%$, the product is completely solid—i.e., it contains no pores. Both $V_v$ and $S_v$ are estimated by a microscopic examination of polished cross sections of sintered material. Generally, the longer a powdered material is sintered, the more solid will be the product. Thus, we would expect $S_v$ to decrease and $V_v$ to increase as the sintering time is increased. The table at the bottom of the page gives the mean and standard deviation of the values of $S_v$ (in squared centimeters per cubic centimeter) and $V_v$ (percentage) for 100 specimens of sintered nickel for six different sintering times.*

a. Plot the sample means of the $S_v$ measurements versus sintering time. Hypothesize a linear model relating mean $S_v$ to sintering time $x$.
b. Plot the sample means of the $V_v$ measurements versus sintering time. Hypothesize a linear model relating mean $V_v$ to sintering time $x$.

**SINTERING**

| | | $S_v$ | | $V_v$ | |
|---|---|---|---|---|---|
| Sample | Time minutes | Mean | Standard Deviation | Mean | Standard Deviation |
| 1 | 1.0 | 1,076.5 | 295.0 | 95.83 | 1.2 |
| 2 | 10.0 | 736.0 | 181.9 | 96.73 | 2.1 |
| 3 | 28.5 | 509.4 | 154.7 | 97.38 | 2.1 |
| 4 | 150.0 | 299.5 | 161.0 | 97.82 | 1.5 |
| 5 | 450.0 | 165.0 | 110.4 | 99.03 | 1.3 |
| 6 | 1,000.0 | 72.9 | 76.6 | 99.49 | 1.1 |

*Data and experimental information provided by Guoquan Liu while visiting at the University of Florida.

c. Fit a linear model relating $E(S_v)$ to sintering time $x$. Show that the data may violate the assumptions of Section 11.2. What model modifications do you suggest?

d. Consider second-order model relating $V_v$ to sintering time $x$. Fit the model $E(V_v) = \beta_0 + \beta_1 x + \beta_2 x^2$ to the data and conduct a complete regression analysis. Ultimately, you want to predict the value of $V_v$ at sintering time 150 minutes.

e. The unstable values of the standard deviations for $S_v$ shown in the table indicate a strong possibility that the standard regression assumption of equal variance is violated for the model of part **c**. We can satisfy this assumption by transforming the response to a new response that has a constant variance. Consider the natural log transform* $S_v^* = \ln(S_v)$. Fit the model $E(S_v^*) = \beta_0 + \beta_1 x$ to the data and give the least-squares prediction equation.

f. Is the model in part **e** adequate for predicting $\ln(S_v)$? Test using $\alpha = .05$.

g. Refer to the model, part **e**. The predicted value of $S_v$ is the antilog,

$$\hat{S}_v = e^{\widehat{\ln(S_v)}}$$

To obtain a prediction interval for $S_v$, you need to take the antilogs of the endpoints of the prediction interval for $S_v^*$.† Find a 95% prediction interval for $S_v$ when the sintering time is 150 minutes.

*To see the stabilizing effect of the log transform, use your calculator to take the logs of the standard deviations for $S_v$ shown in the table. Note that the transformed values appear to be much less variable.

†Unfortunately, you cannot take antilogs to find the confidence interval for the mean response $E(y)$. This is because the mean value of $\ln(y)$ is not equal to the natural logarithm of the mean of $y$.

CHAPTER 12

# Model Building

## OBJECTIVE

*To show you why the choice of the deterministic portion of a linear model is crucial to the acquisition of a good prediction equation; to present some basic concepts and procedures for constructing good linear models*

## CONTENTS



*STATISTICS IN ACTION*
Deregulation of the Intrastate Trucking Industry

## *STATISTICS IN ACTION*
### Deregulation of the Intrastate Trucking Industry

We illustrate the modeling techniques outlined in this chapter with an actual study from engineering economics. Consider the problem of modeling the price charged for motor transport service (e.g., trucking) in Florida. In the early 1980s, several states removed regulatory constraints on the rate charged for intrastate trucking services. (Florida was the first state to embark on a deregulation policy on July 1, 1980.) Prior to this time, the state determined price schedules for motor transport service with review and approval by the Public Service Commission. Once approved, individual carriers were not allowed to deviate from these official rates. The objective of the analysis is twofold: (1) assess the impact of deregulation on the prices charged for motor transport service in the state of Florida, and (2) estimate a regression model of the supply price for predicting future prices.

The data employed for this purpose ($n = 134$ observations) were obtained from a population of over 27,000 individual shipments in Florida made by major intrastate carriers before and after deregulation. The shipments of interest were made by one particular carrier whose trucks originated from either the city of Jacksonville or Miami. The dependent variable of interest is $y$, the natural logarithm of the price (measured in 1980 dollars) charged per ton-mile. The independent variables available for predicting $y$ are listed and described in Table SIA12.1. These data are saved in the **TRUCKING** file.

**TRUCKING**

**TABLE SIA12.1 Independent Variables for Predicting Trucking Prices**

| Variable Name | Description |
|---|---|
| DISTANCE | Miles traveled (in hundreds) |
| WEIGHT | Weight of product shipped (in 1,000 pounds) |
| PCTLOAD | Percent of truck load capacity |
| ORIGIN | City of origin (JAX or MIA) |
| MARKET | Size of market destination (LARGE or SMALL) |
| DEREG | Deregulation in effect (YES or NO) |
| PRODUCT | Product classification (100, 150, or 200)—Value roughly corresponds to the value-to-weight ratios of the goods being shipped (more valuable goods are categorized in the higher classification) |

In the *Statistics in Action Revisited* example at the end of this chapter, we apply the model building techniques presented in this chapter to estimate a model for trucking prices and use the model to examine the impact of deregulation.

## 12.1 Introduction: Why Model Building Is Important

We have emphasized in Chapters 10 and 11 that one of the first steps in the construction of a regression model is to hypothesize the form of the deterministic portion of the probabilistic model. This **model building**, or model construction, stage is the key to the success (or failure) of the regression analysis. If the hypothesized model does not reflect, at least approximately, the true nature of the relationship between the mean response $E(y)$ and the independent variables $x_1, x_2, \ldots, x_k$, the modeling effort will usually be unrewarded.

By *model building*, we mean writing a model that will provide a good fit to a set of data and that will give good estimates of the mean value of $y$ and good predictions of future values of $y$ for given values of the independent variables. To illustrate, suppose you want to relate the breaking strength $y$ for a certain type of plastic to the amount of pressure $x$ used to produce the plastic. Unknown to you, the second-order model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

would permit you to predict $y$ with a very small error of prediction (see Figure 12.1a). Unfortunately, you have erroneously chosen the first-order model

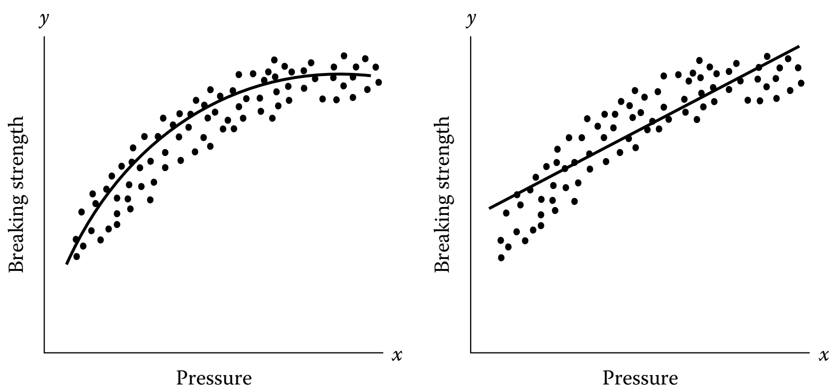
$$E(y) = \beta_0 + \beta_1 x$$

to explain the relationship between $y$ and $x$ (see Figure 12.1b).

The consequence of choosing the wrong model is clearly demonstrated by comparing Figures 12.1a and 12.1b. The errors of prediction for the second-order model are relatively small in comparison to those for the first-order model. The lesson to be learned from this simple example is clear. Choosing a good set of independent (predictor) variables $x_1, x_2, \ldots, x_k$ will not guarantee a good prediction equation. In addition to selecting independent variables that contain information about $y$, you must specify an equation relating $y$ to $x_1, x_2, \ldots, x_k$ that will provide a good fit to your data.

In this chapter, we discuss the most difficult part of a multiple regression analysis—the formulation of a good model for $E(y)$. Although several of the models presented in this chapter have already been introduced in Chapter 11, we assume the reader has little or no background in model building. This chapter serves as a basic reference guide to model building for multiple regression users.

**FIGURE 12.1**
Two models for relating breaking strength $y$ to amount of pressure $x$




a. Second-order model    b. First-order model

## 12.2 The Two Types of Independent Variables: Quantitative and Qualitative

Recall from Chapter 1 the two types of data that arise in experimental situations: **quantitative** and **qualitative**. For the types of regression analyses considered in this text, the dependent variable will always be quantitative, but the independent variables may be either quantitative or qualitative. As you will see, the way an independent variable enters the model depends on its type. We repeat the definitions of quantitative and qualitative variables from Chapter 1 here.

**Definition 12.1**

A **quantitative** independent variable is one that assumes numerical values corresponding to the points on a line. An independent variable that is not quantitative but categorical in nature is called **qualitative**.

The waiting time before a computer begins to process data, the number of defects in a product, and the kilowatt-hours of electricity used per day are all examples of quantitative independent variables. On the other hand, recall that three species of fish—channel catfish, largemouth bass, and smallmouth buffalofish—were found in the contaminated Tennessee River. The variable species is qualitative, since it is not measured on a numerical scale. Since it is likely that the different species have different mean levels of DDT contamination, we would want to include it as an independent variable in a model predicting the level of DDT contamination, $y$, in fish found in the Tennessee River.

**Definition 12.2**

The different intensity settings (i.e., values) of an independent variable are called its **levels**.

For a quantitative independent variable, the levels correspond to the numerical values it assumes. For example, if the number of defects in a product ranges from 0 to 3, the independent variable has four levels: 0, 1, 2, and 3.

The levels of a qualitative variable are not numerical. They can be defined only by describing them. For example, the independent variable for the species of fish was observed at three levels: channel catfish, largemouth bass, and smallmouth buffalofish.

**Example 12.1**

Identifying the Type of Variable

Suppose our task is to predict the salary of a corporate executive at an engineering firm as a function of the following four independent variables:

a. Experience of an employee (years)
b. Gender of the employee
c. Net asset value of the firm
d. Rank of the employee

For each of these independent variables, give its type and describe the levels you would expect to observe.

Solution

a. The independent variable, experience, is quantitative, since its values are numerical. We would expect to observe levels ranging from 0 to 40 (approximately) years.
b. The independent variable for gender is qualitative, since its levels can be described only by the nonnumerical labels "female" and "male."
c. The independent variable, net asset value of the firm, is quantitative, with a large number of possible levels corresponding to the range of dollar values representing the net asset values of the various firms.
d. Suppose the independent variable for the rank of the employee is observed at three levels: supervisor, assistant vice president, and vice president. Since we cannot assign a realistic numerical measure of relative importance to each position, rank is a qualitative independent variable.

Quantitative independent variables are treated differently from qualitative variables in regression modeling. In the next section, we will begin our discussion of how quantitative variables are used in the modeling effort.

## Applied Exercises

12.1 *Chemical composition of rainwater.* Researchers at the University of Aberdeen (Scotland) developed a statistical model for estimating the chemical composition of water (*Journal of Agricultural, Biological, and Environmental Statistics*, March 2005). For one application, the nitrate concentration (milligrams per liter) in a water sample collected after a heavy rainfall was modeled as a function of water source (groundwater, subsurface flow, or overground flow).

a. Identify the dependent variable, $y$, for the study.
b. Identify the independent variable and give its type (quantitative or qualitative).

12.2 *Properties of biodiesel fuels.* The performance of a diesel engine with blends of biodiesel fuels was the topic of research in the *International Journal of Energy and Environmental Engineering* (Dec. 2013). Several of the many variables measured in the study are listed below. Identify the type (quantitative or qualitative) of each variable.

a. Diesel fuel type (HSD, MO, MB100, SRO, or B20)
b. Water content (parts per million)
c. Flash point temperature (degrees Centigrade)
d. Fuel density (kilograms per cubic meter)
e. Location of soot deposits (cylinder head, piston crown, or fuel injector)

12.3 *Design of cold-formed steel walls.* The behavior and design of cold-formed steel buildings and walls was investigated in the *Journal of Structural Engineering* (May 2013). Several of the many variables measured in the study are listed below. Identify the type (quantitative or qualitative) of each variable. (Note: The dependent variable in the study was peak load of a single stud.)

a. Type of sheathing (Bare, Gypsum, or OSB)
b. Limit state observed at peak strength (local buckling, weak-axis flexural, or flexural-torsional)
c. Peak load of single stud (kilo-Newtons)
d. Linear position transducer displacement (millimeters)

12.4 *Properties of cement mortar.* In the *International Journal of Engineering Research & Applications* (May-June 2013), structural engineers examined the properties of Portland cement mortar made from rice ash husk. The variables measured in the study included the following. Identify the type (quantitative or qualitative) of each variable. (Note: The dependent variable in the study was compressive strength.)

a. Proportion of cement mix that contains rice ash husk
b. Quantity of sand in the cement mix (grams)
c. Quantity of water in the cement mix (grams)
d. Compressive strength of cement (Newtons per millimeters-squared)
e. Setting time of cement (number of minutes)
f. Type of Portland cement used (Type I, Type II, Type III, Type IV, Type V, White)

12.5 *Emotional distress of firefighters.* The *Journal of Human Stress* (Summer 1987) reported on a study of "psychological response of firefighters to a chemical fire." The researchers used multiple regression to predict emotional distress as a function of the following independent variables. Identify each independent variable as quantitative or qualitative. For qualitative variables, suggest several levels that might be observed. For quantitative variables, give a range of values (levels) for which the variable might be observed.

a. Number of preincident psychological symptoms
b. Years experience
c. Cigarette smoking behavior
d. Level of social support
e. Marital status
f. Age
g. Ethnic status
h. Exposure to a chemical fire
i. Educational level
j. Distance lived from site of incident
k. Gender

12.6 *Flow rate of land waste.* An experiment was conducted to investigate the sheet flow rate of a land waste treatment plant. Classify each of the following independent variables as quantitative or qualitative and describe the levels the variables might assume.

a. Amount of rainfall
b. Method of treatment
c. Irrigation rate
d. Slope of grass mat
e. Type of sod

12.7 *Sorption rate of organic vapors.* *Environmental Science & Technology* (Oct. 1993) published an article that investigated the variables that affect the sorption of organic vapors on clay minerals. The independent variables and levels considered in the study are listed here. Identify the type (quantitative or qualitative) of each.

a. Temperature (50°F, 60°F, 75°F, 90°F)
b. Relative humidity (30%, 50%, 70%)
c. Organic compound (benzene, toluene, chloroform, methanol, anisole)

## 12.3 Models with a Single Quantitative Independent Variable

The most common linear models relating $y$ to a single quantitative independent variable $x$ are those derived from a **polynomial** expression of the type shown in the box. Specific models, obtained by assigning particular values to $p$, are listed subsequently.

### Formula for a $p$th-Order Polynomial with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_p x^p$$

where $p$ is a positive integer and $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters that must be estimated.

### First-Order (Straight-Line) Model with One Quantitative Independent Variable

$$E(y) = \beta_0 + \beta_1 x$$

Interpretation of model parameters

$\beta_0$: $y$-intercept; the value of $E(y)$ when $x = 0$

$\beta_1$: Slope of the line; the change in $E(y)$ for a 1-unit increase in $x$

The first-order model is used when you expect the rate of change in $y$ per unit change in $x$ to remain fairly stable over the range of values of $x$ for which you wish to predict $y$ (see Figure 12.2). Most relationships between $E(y)$ and $x$ are curvilinear, but the curvature over the range of values of $x$ for which you wish to predict $y$ may be very slight. When this occurs, a first-order (straight-line) model should provide a good fit to your data.

### Second-Order (Quadratic) Model with One Quantitative Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Interpretation of model parameters

$\beta_0$: $y$-intercept; the value of $E(y)$ when $x = 0$

$\beta_1$: Shift parameter; changing the value of $\beta_1$ shifts the parabola to the right or left (increasing the value of $\beta_1$ causes the parabola to shift to the right)

$\beta_2$: Rate of curvature

A second-order model traces a parabola, one that opens either downward $(\beta_2 < 0)$ or upward $(\beta_2 > 0)$, as shown in Figure 12.3. Since most relationships will possess some curvature, a second-order model will often be a good choice to relate $y$ to $x$.

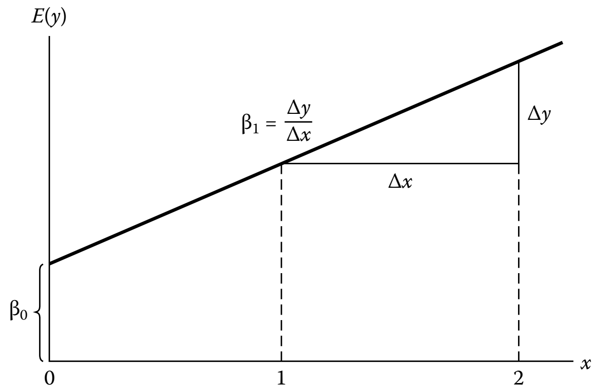



**FIGURE 12.2**
Graph of a first-order model

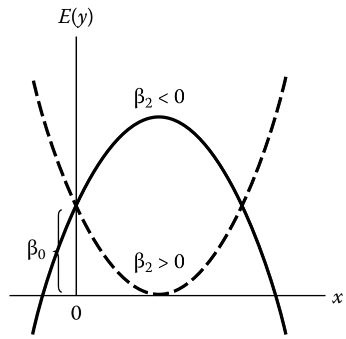



**FIGURE 12.3**
The graphs of two second-order models

### Third-Order Model with One Quantitative Independent Variable

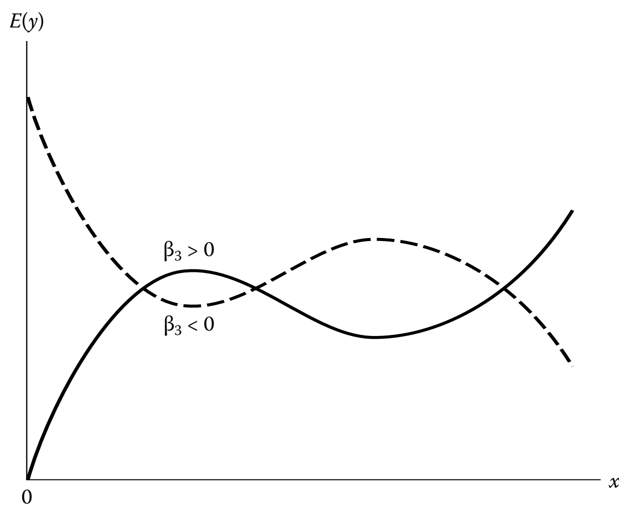
$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Interpretation of model parameters

$\beta_0$: $y$-intercept; the value of $E(y)$ when $x = 0$
$\beta_1$: Shift parameter (shifts the polynomial right or left on the $x$-axis)
$\beta_2$: Rate of curvature
$\beta_3$: The magnitude of $\beta_3$ controls the rate of reversal of curvature for the curve

Reversals in curvature are not common, but such relationships can be modeled by third- and higher-order polynomials. As can be seen in Figure 12.3, a second-order model contains no reversals in curvature. The slope continues to either increase or decrease as $x$ increases and produces either a trough (minimum) or a peak (maximum). A third-order model (see Figure 12.4) contains one reversal in curvature and produces one peak and one trough. In general, the graph of a $p$th-order polynomial will contain at most $(p - 1)$ peaks and troughs.

**FIGURE 12.4**
The graphs of two third-order models



Most functional relationships in nature seem to be smooth (except for random error)—that is, they are not subject to rapid and irregular reversals in direction. Consequently, the second-order polynomial model is perhaps the most useful of those previously described. To develop a better understanding of how this model is used, consider the following example.

## Example 12.2

**Higher-order Polynomial Models for Power Loads**

To operate efficiently, power companies must be able to predict the peak power load at their various stations. Peak power load is the maximum amount of power that must be generated each day to meet demand. A power company wants to use daily high temperature, $x$, to model daily peak power load, $y$, during the summer months when demand is greatest. Although the company expects peak load to increase as the temperature increases, the *rate* of increase in $E(y)$ might not remain constant as $x$ increases. For example, a 1-unit increase in high temperature from 100°F to 101°F might result in a larger increase in power demand than would a 1-unit increase from 80°F to 81°F. Therefore, the company postulates that the model for $E(y)$ will include a second-order (quadratic) term and, possibly, a third-order (cubic) term.

A random sample of 25 summer days is selected and both the peak load (measured in megawatts) and high temperature (in degrees) recorded for each day. The data are listed in Table 12.1.

a. Construct a scatterplot for the data. What type of model is suggested by the plot?

b. Fit the third-order model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, to the data. Is there evidence that the cubic term, $\beta_3 x^3$, contributes information for the prediction of peak power load? Test at $\alpha = .05$.

c. Fit the second-order model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, to the data. Test the hypothesis that the power load increases at an increasing rate with temperature. Use $\alpha = .05$.

d. Give the prediction equation for the second-order model, part **c**. Are you satisfied with using this model to predict peak power loads?

**Solution**

a. The scatterplot of the data, produced using MINITAB, is shown in Figure 12.5. The nonlinear, upward-curving trend indicates that a second-order model would likely fit the data well.

b. The third-order model is fit to the data using MINITAB and the resulting printout is shown in Figure 12.6. The $p$-value for testing

$$H_0: \quad \beta_3 = 0$$
$$H_a: \quad \beta_3 \neq 0$$

highlighted on the printout, is .911. Since this value exceeds $\alpha = .05$, there is insufficient evidence of a third-order relationship between peak load and high temperature. Consequently, we will drop the cubic term, $\beta_3 x^3$, from the model.

**POWERLOADS**

**TABLE 12.1 Power Load Data**

| Temperature °F | Peak Load megawatts | Temperature °F | Peak Load megawatts | Temperature °F | Peak Load megawatts |
|---|---|---|---|---|---|
| 94 | 136.0 | 106 | 178.2 | 76 | 100.9 |
| 96 | 131.7 | 67 | 101.6 | 68 | 96.3 |
| 95 | 140.7 | 71 | 92.5 | 92 | 135.1 |
| 108 | 189.3 | 100 | 151.9 | 100 | 143.6 |
| 67 | 96.5 | 79 | 106.2 | 85 | 111.4 |
| 88 | 116.4 | 97 | 153.2 | 89 | 116.5 |
| 89 | 118.5 | 98 | 150.1 | 74 | 103.9 |
| 84 | 113.4 | 87 | 114.7 | 86 | 105.1 |
| 90 | 132.0 | | | | |

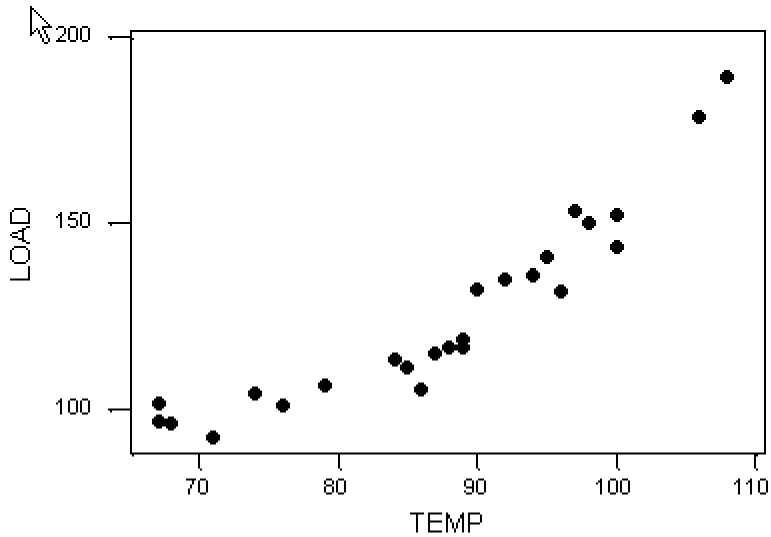



**FIGURE 12.5**
MINITAB scatterplot for power load data

```
The regression equation is
LOAD = 331 - 6.4 TEMP + 0.038 TEMP2 +0.000084 TEMP3

Predictor         Coef     SE Coef         T        P
Constant         331.3       477.1      0.69    0.495
TEMP             -6.39       16.79     -0.38    0.707
TEMP2           0.0378      0.1945      0.19    0.848
TEMP3        0.0000843   0.0007426      0.11    0.911

S = 5.501       R-Sq = 95.9%     R-Sq(adj) = 95.4%

Analysis of Variance

Source           DF        SS        MS        F        P
Regression        3   15012.2    5004.1   165.36    0.000
Residual Error   21     635.5      30.3
Total            24   15647.7
```

**FIGURE 12.6**
MINITAB output for third-order model of power load

c. The second-order model is fit to the data using MINITAB and the resulting printout is shown in Figure 12.7. For this quadratic model, if $\beta_2$ is positive, then the peak power load $y$ increases at an increasing rate with temperature $x$. Consequently, we test

$$H_0: \quad \beta_2 = 0$$
$$H_a: \quad \beta_2 > 0$$

The test statistic, $T = 7.93$, and two-tailed $p$-value, are both highlighted on Figure 12.7. Since the one-tailed $p$-value, $p = 0/2 = 0$, is less than $\alpha = .05$, we reject $H_0$ and conclude that peak power load increases at an increasing rate with temperature.

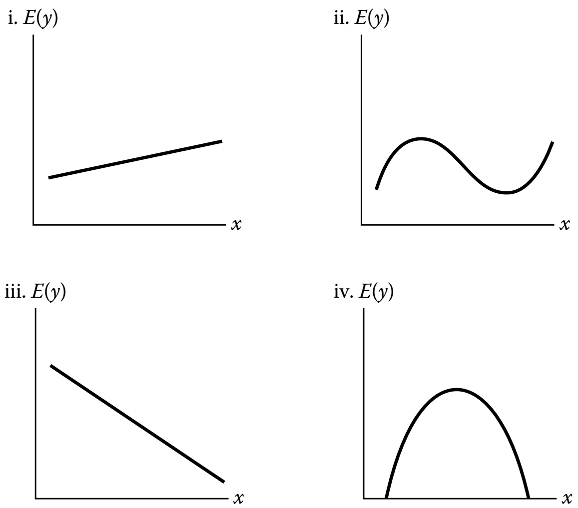
**FIGURE 12.7**
MINITAB output for second-order model of power load

```
The regression equation is
LOAD = 385 - 8.29 TEMP + 0.0598 TEMP2

Predictor        Coef      SE Coef        T       P
Constant       385.05        55.17     6.98   0.000
TEMP           -8.293        1.299    -6.38   0.000
TEMP2        0.059823     0.007549     7.93   0.000

S = 5.376       R-Sq = 95.9%     R-Sq(adj) = 95.6%

Analysis of Variance

Source            DF          SS        MS        F        P
Regression         2     15011.8    7505.9   259.69    0.000
Residual Error    22       635.9      28.9
Total             24     15647.7
```

## Applied Exercises

12.8 *Identifying polynomials.* The accompanying graphs depict $p$th-order polynomials for one quantitative independent variable.

a. For each graph, identify the order of the polynomial.
b. Using the parameters $\beta_0$, $\beta_1$, $\beta_2$, etc., write an appropriate model relating $E(y)$ to $x$ for each graph.
c. The signs (+ or −) of many of the parameters in the models of part **b** can be determined by examining the graphs. Give the signs of those parameters that can be determined.




12.9 *Graphing polynomials.* Graph the following polynomials and identify the order of each on your graph:

a. $E(y) = 2 + 3x$
b. $E(y) = 2 + 3x^2$
c. $E(y) = 1 + 2x + 2x^2 + x^3$
d. $E(y) = 2x + 2x^2 + x^3$
e. $E(y) = 2 - 3x^2$
f. $E(y) = -2 + 3x$

12.10 *Chlorophyll in Florida Everglades water.* The Organic Geochemistry Group at Florida Atlantic University studied the photosynthetic pigments in the waters of the Florida Everglades (*Florida Scientist*, Fall 2004). The researchers measured the amount of chlorophyll in a liter of water collected from the Florida Bay using each of two methods: spectrophotometry ($y$) and high-performance liquid chromatography ($x$).

a. Write a first-order (straight-line) model for $E(y)$. Interpret the betas in the model.
b. Theoretically, if there is no chlorophyll in the water specimen, then both $x = 0$ and $y = 0$. Rewrite the model, part **a**, assuming that the line will go through the origin, (0, 0).
c. Write a second-order (quadratic) model for $E(y)$.
d. What is the expected sign of $\beta_2$ in the model, part **c**, if theory indicates that as the high-performance liquid chromatography measurement ($x$) increases, the spectrophotometry measurement ($y$) will increase at a decreasing rate?

12.11 *Sorption of organic vapors.* Refer to the *Environmental Science & Technology* study of sorption of organic vapors, Exercise 12.7 (p. 646). Consider modeling the vapor retention coefficient $y$ as a function of one of the two quantitative variables, temperature ($x_1$) and relative humidity ($x_2$).

a. Propose a model that hypothesizes a curvilinear relationship between mean retention $E(y)$ and relative humidity $x_2$. Draw a sketch of the model.
b. Propose a model that hypothesizes a third-order relationship between mean retention $E(y)$ and temperature $x_1$. Draw a sketch of the model.

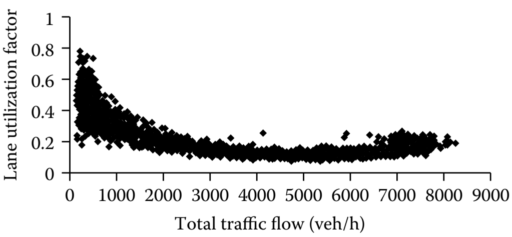
12.12 *Strength of plastic.* The amount of pressure used to produce a certain plastic is thought to be related to the strength of the plastic. Researchers believe that, as pressure is increased, the strength of the plastic increases until, at some point, increases in pressure will have a detrimental effect on strength. Write a model to relate the strength, $y$, of the plastic to pressure, $x$, that would reflect these beliefs. Sketch the model.

12.13 *Highway lane utilization.* The lane utilization of a highway is measured by how the traffic flow in one direction is distributed among the available lanes. A lane utilization model for highways in the United Kingdom was developed in the

*Journal of Transportation Engineering* (May 2013). The dependent variable in the analysis was lane utilization $y$, measured as the percentage of vehicles in the lane. Previous studies used a quadratic model for lane utilization as a function of $x$ = total traffic flow (total number of vehicles per hour). An analysis of traffic data collected over several weeks for different sections of Lane 2 of the 4-lane M25 highway yielded the following results:

$$\hat{y} = .46 - .0000764x + .00000000627x^2, R^2 = .73$$

a. Interpret the value of $R^2$ for this model.
b. Assuming $n = 2{,}000$ observations, conduct a test of overall model utility. Use $\alpha = .01$.
c. Use the $\beta$ estimates to draw a sketch of the estimated relationship between lane utilization and total traffic flow.
d. A graph of the data for another lane of the M25 highway is shown below. Hypothesize a polynomial model that you believe will fit the data.




12.14 *Level of overcrowding at a train station.* The level of overcrowding for passengers waiting at a train station was investigated in the *Journal of Transportation Engineering* (June 2013). The researchers measured "crowdedness" as the average distance between the forefront edge of the passenger waiting on a platform and the rear edge of the one in front of this passenger. This variable reflects the degree of closeness of passengers waiting in line for the next train. The shorter the distance, the more crowded the platform. The average distance ($y$, in meters) was hypothesized to be related to the volume of passengers waiting ($x$, number of persons) on the train platform. Data (simulated from information provided in the journal article) on these two variables for a sample of 20 train stops are listed in the accompanying table. Consider the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. A Minitab printout of the regression analysis follows. In theory, the rate of decrease of distance with increasing passenger volume should level off for more crowded platforms. Use the model to test this theory at $\beta = .01$.

**TRAINWAIT**

| Stop | Volume | Distance |
|---|---|---|
| 1 | 21 | 0.08 |
| 2 | 2 | 0.78 |
| 3 | 4 | 0.62 |
| 4 | 25 | 0.06 |
| 5 | 16 | 0.11 |
| 6 | 26 | 0.06 |
| 7 | 6 | 0.35 |
| 8 | 22 | 0.07 |
| 9 | 23 | 0.06 |
| 10 | 9 | 0.19 |
| 11 | 11 | 0.15 |
| 12 | 20 | 0.09 |
| 13 | 17 | 0.10 |
| 14 | 18 | 0.10 |
| 15 | 19 | 0.11 |
| 16 | 7 | 0.27 |
| 17 | 29 | 0.05 |
| 18 | 5 | 0.50 |
| 19 | 14 | 0.12 |
| 20 | 8 | 0.20 |

### Minitab Output for Exercise 12.14

**Regression Analysis: DISTANCE versus WAITVOL, WAITVOL-SQ**

```
The regression equation is
DISTANCE = 0.784 - 0.0707 WAITVOL + 0.00167 WAITVOL-SQ


Predictor        Coef    SE Coef      T      P
Constant      0.78415    0.05959  13.16  0.000
WAITVOL     -0.070680   0.009175  -7.70  0.000
WAITVOL-SQ  0.0016728  0.0002991   5.59  0.000


S = 0.0735672   R-Sq = 88.5%   R-Sq(adj) = 87.1%


Analysis of Variance

Source          DF       SS       MS      F      P
Regression       2  0.70785  0.35392  65.39  0.000
Residual Error  17  0.09201  0.00541
Total           19  0.79985
```

12.15 *Immunity and exercise.* Does exercise improve the human immune system? An experiment was conducted by a physiologist at the University of Florida to determine whether such a relationship exists. Thirty subjects volunteered to participate in the study. The amount of immunoglobulin known as IgG (an indicator of long-term immunity) and the maximal oxygen uptake (a measure of aerobic fitness level) were recorded for each subject. The resulting data are given in the accompanying table.

**IMMUNE**

| Subject | IgG $y$ | Maximal Oxygen Uptake $x$ | Subject | IgG $y$ | Maximal Oxygen Uptake $x$ |
|---|---|---|---|---|---|
| 1 | 881 | 34.6 | 16 | 1,660 | 52.5 |
| 2 | 1,290 | 45.0 | 17 | 2,121 | 69.9 |
| 3 | 2,147 | 62.3 | 18 | 1,382 | 38.8 |
| 4 | 1,909 | 58.9 | 19 | 1,714 | 50.6 |
| 5 | 1,282 | 42.5 | 20 | 1,959 | 69.4 |
| 6 | 1,530 | 44.3 | 21 | 1,158 | 37.4 |
| 7 | 2,067 | 67.9 | 22 | 965 | 35.1 |
| 8 | 1,982 | 58.5 | 23 | 1,456 | 43.0 |
| 9 | 1,019 | 35.6 | 24 | 1,273 | 44.1 |
| 10 | 1,651 | 49.6 | 25 | 1,418 | 49.8 |
| 11 | 752 | 33.0 | 26 | 1,743 | 54.4 |
| 12 | 1,687 | 52.0 | 27 | 1,997 | 68.5 |
| 13 | 1,782 | 61.4 | 28 | 2,177 | 69.5 |
| 14 | 1,529 | 50.2 | 29 | 1,965 | 63.0 |
| 15 | 969 | 34.1 | 30 | 1,264 | 43.2 |

a. Construct a scattergram for the IgG–maximal oxygen uptake data.
b. Hypothesize a probabilistic model relating IgG to maximal oxygen uptake.
c. Fit the model to the data and give the least-squares prediction equation.
d. Assess the adequacy of the model.

12.16 *Glacier snow pit temperatures.* The National Snow and Ice Data Center at the University of New Hampshire collected data on the temperature of ice core samples from a glacier snow pit. The ice core samples were collected from depths ranging from 2 meters to 175.5 meters. The data for depths up to 20 meters are listed in the next table.

a. Plot the data in a scattergram, with temperature on the $y$-axis and depth on the $x$-axis. What trend do you observe?
b. Propose a polynomial model for $E(y)$ that you think will fit the data. What is the order of the model?
c. Fit the model, part **b**, to the data using the method of least squares. Assess the adequacy of the model.

**SNOWTEMP20**

| Depth (meters) | Temperature (degrees, Celsius) | Depth (meters) | Temperature (degrees, Celsius) |
|---|---|---|---|
| 19.60 | −28.77 | 7.90 | −29.41 |
| 15.00 | −28.84 | 7.00 | −29.56 |
| 14.00 | −28.88 | 6.00 | −29.68 |
| 13.80 | −28.89 | 5.00 | −29.68 |
| 13.00 | −28.90 | 4.00 | −29.39 |
| 12.35 | −28.93 | 3.00 | −28.33 |
| 12.00 | −28.94 | 2.00 | −25.24 |
| 11.00 | −29.02 | 2.00 | −25.19 |
| 10.00 | −29.11 | 2.00 | −25.25 |
| 9.00 | −29.25 | | |

*Source*: Mayewski, P., and Whitlow, S. "Newall glacier snow pit and ice core." National Snow and Ice Data Center, Boulder, CO., 2000.

12.17 *Creases in deployable space membranes.* The *Journal of Space Engineering* (Vol. 4, 2011) investigated the properties of large, deployable space membranes used to store structures such as solar sails, antennas, sunshields, and solar power satellites. These membranes are susceptible to creases during folding and packaging, which can have a detrimental effect on deployment. The study examined the relationship between the size of the mesh (in millimeters) around the crease and the contact force (Newtons per millimeter) exerted on the membrane. Data for $n = 4$ experiments on folded space membranes are provided in the table.

**CREASE**

| Experiment | FORCE | MESH |
|---|---|---|
| 1 | 0.14 | 0.125 |
| 2 | 0.15 | 0.250 |
| 3 | 0.16 | 0.500 |
| 4 | 0.41 | 1.000 |

*Source:* Satou, Y. & Furuya, H. "Mechanical Properties of Z-Fold Membrane under Elasto-Plastic Deformation", *Journal of Space Engineering,* Vol. 4, No. 1, 2011 (adapted from Figure 9).

a. Fit the straight-line model, $E(y) = \beta_0 + \beta_1 x$, to the data, where $y$ = contact force and $x$ = mesh size.
b. Obtain influence diagnostics for the model, part a. Identify any influential observations.
c. Examine a scatterplot of the data. Locate the influential observation on the graph. What does this point suggest about the relationship between $y$ and $x$?
d. Now fit the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, to the data. Comment on the fit of this model as compared to the straight-line model.

## 12.4 Models with Two or More Quantitative Independent Variables

Like models with a single quantitative independent variable, models with two or more quantitative independent variables are classified as first-order, second-order, and so forth. Since we rarely encounter third- or higher-order relationships in practice, we focus our discussion on first- and second-order models.

### First-Order Model with $k$ Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Interpretation of model parameters

$\beta_0$: $y$-intercept of a $(k + 1)$ dimensional surface (see Figure 12.8 for $k = 2$); the value of $E(y)$ when $x_1 = x_2 = 0 \cdots = x_k = 0$

$\beta_i$: Change in $E(y)$ for a 1-unit increase in $x_i$, when all other $x$'s are held fixed, $i = 1, 2, ..., k$

The graph in Figure 12.8 traces a **response surface** [in contrast to the **response curve** that is used to relate $E(y)$ to a *single* quantitative variable]. In particular, a first-order model relating $E(y)$ to two independent quantitative variables, $x_1$ and $x_2$, graphs as a plane in a three-dimensional space. The plane traces the value of $E(y)$ for every combination of values $(x_1, x_2)$ that correspond to points in the $(x_1, x_2)$-plane. Most response surfaces in the real world are well behaved (smooth) and they have curvature. Consequently, a first-order model is appropriate only if the response surface is fairly flat over the $(x_1, x_2)$-region that is of interest to you.

The assumption that a first-order model will adequately characterize the relationship between $E(y)$ and the variables $x_1$ and $x_2$ is equivalent to assuming that $x_1$ and $x_2$ do not "interact"; that is, you assume that the effect on $E(y)$ of a change in $x_1$ (for a fixed value of $x_2$) is the same regardless of the value of $x_2$ (and vice versa). Thus, "no interaction" is equivalent to saying that the effect of changes in one variable (say, $x_1$) on $E(y)$ is *independent* of the value of the second variable (say, $x_2$). For example, if we assign values to $x_2$ in a first-order model, the graph of $E(y)$ as a function of $x_1$ would produce parallel lines as shown in Figure 12.9. These lines, called **contour lines**, show the contours of the surface when it is sliced by three planes, each of which is parallel to the $[E(y), x_1]$-plane, at distances $x_2 = 1, 2,$ and 3 from the origin.

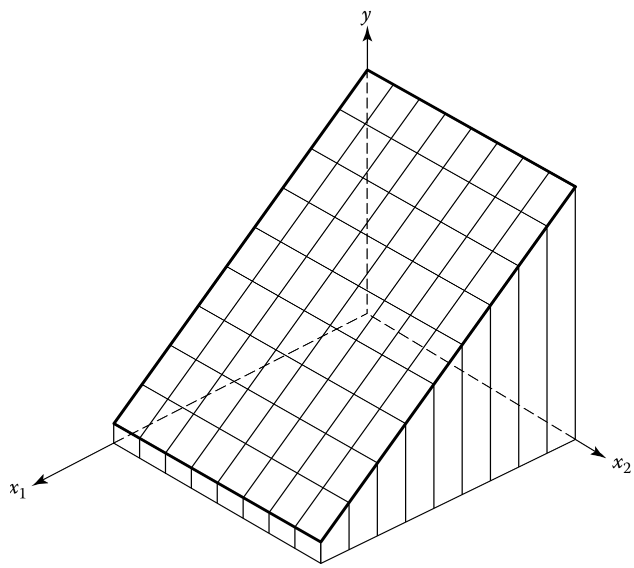



**FIGURE 12.8**
Response surface for first-order model with two independent variables

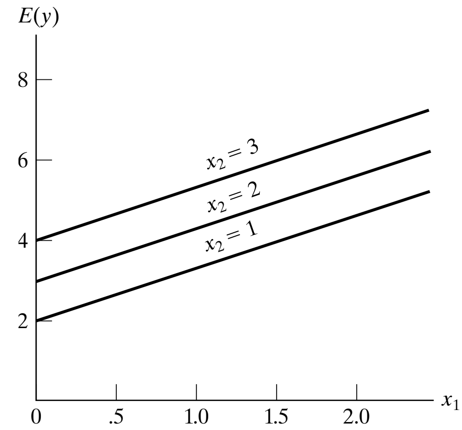



**FIGURE 12.9**
A graph indicating no interaction between $x_1$ and $x_2$

**Definition 12.3**

Two variables $x_1$ and $x_2$ are said to **interact** if the change in $E(y)$ for a 1-unit change in $x_1$ (when $x_2$ is held fixed) is dependent on the value of $x_2$.

### Interaction Model (Second-Order) with Two Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Interpretation of model parameters

$\beta_0$: $y$-intercept; the value of $E(y)$ when $x_1 = x_2 = 0$

$\beta_1$ and $\beta_2$: Changing $\beta_1$ and $\beta_2$ causes the surface to shift along the $x_1$- and $x_2$-axes

$\beta_3$: Controls the rate of twist in the ruled surface (see Figure 12.10)

When one independent variable is held fixed, the model produces straight lines with the following slopes:

$\beta_1 + \beta_3 x_2$: Change in $E(y)$ for a 1-unit increase in $x_1$, when $x_2$ is held fixed

$\beta_2 + \beta_3 x_1$: Change in $E(y)$ for a 1-unit increase in $x_2$, when $x_1$ is held fixed

The interaction model is said to be second-order because the order of the highest-order ($x_1 x_2$) term in $x_1$ and $x_2$ is 2; i.e., the sum of the exponents of $x_1$ and $x_2$ equals 2. This model traces a ruled surface in a three-dimensional space (see Figure 12.10). You could produce such a surface by placing a pencil perpendicular to a line and moving it along the line, while rotating it around the line. The resulting surface would appear as a twisted plane. A graph of $E(y)$ as a function of $x_1$ for given values of $x_2$ (say, $x_2 = 1, 2,$ and 3) produces nonparallel contour lines (see Figure 12.11), thus indicating that the change in $E(y)$ for a given change in $x_1$ is dependent on the value of $x_2$ and, therefore, that $x_1$ and $x_2$ interact. Interaction is an extremely important concept because it is easy to get in the habit of fitting first-order models and individually examining the relationships between $E(y)$ and each of a set of independent variables, $x_1, x_2, \ldots, x_k$. Such a procedure is meaningless when interaction exists (which is, at least to some extent, almost always the case), and it can lead to gross errors in interpretation. For example, suppose that the relationship between $E(y)$ and $x_1$ and $x_2$ is as shown in Figure 12.11 and that you have observed $y$ for each of the $n = 9$ combinations of values of $x_1$ and $x_2$, ($x_1 = 0, 1, 2,$ and $x_2 = 1, 2, 3$). If you fit a first-order model in $x_1$ and $x_2$ to the data, the fitted plane would be (except for random error) approximately parallel to the ($x_1$, $x_2$)-plane, thus suggesting that $x_1$ and $x_2$ contribute very little information about $E(y)$. That this is not the case is clearly indicated by Figure 12.10. Fitting a first-order model to the data would not allow for the twist in the true surface and would therefore give a false impression of the relationship between

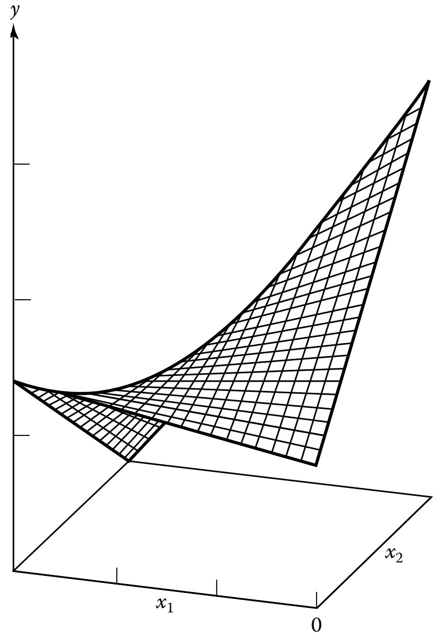



**FIGURE 12.10**
Response surface for an interaction model (second-order)

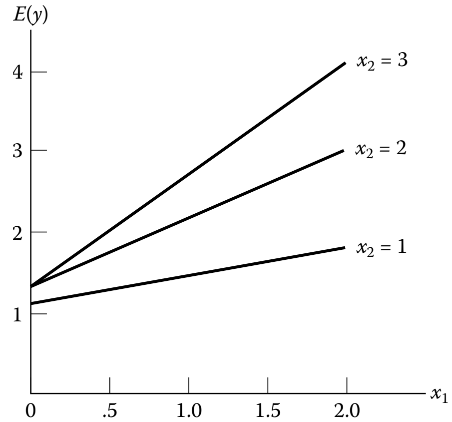



**FIGURE 12.11**
A graph indicating interaction between $x_1$ and $x_2$

$E(y)$ and $x_1$ and $x_2$. The procedure for detecting interaction between two independent variables can be seen by examining the model. The interaction model differs from the noninteraction first-order model only in the inclusion of the $\beta_3 x_1 x_2$ term:

*Interaction model*: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

*First-order model*: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Therefore, to test for the presence of interaction, we test

$H_0$: $\beta_3 = 0$ (no interaction)

against the alternative hypothesis

$H_a$: $\beta_3 \neq 0$ (interaction)

using the familiar Student's $T$ test of Section 11.4.

### Complete Second-Order Model with Two Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Interpretation of model parameters

$\beta_0$: $y$-intercept; the value of $E(y)$ when $x_1 = x_2 = 0$

$\beta_1$ and $\beta_2$: Changing $\beta_1$ and $\beta_2$ causes the surface to shift along the $x_1$- and $x_2$-axes

$\beta_3$: The value of $\beta_3$ controls the rotation of the surface

$\beta_4$ and $\beta_5$: Signs and values of these parameters control the type of surface and the rates of curvature

The following three types of surfaces may be produced by a second-order model:*

A paraboloid that opens upward (Figure 12.12a)
A paraboloid that opens downward (Figure 12.12b)
A saddle-shaped surface (Figure 12.12c)

*The saddle-shaped surface (Figure 12.12c) is produced when $\beta_3^2 > 4\beta_4\beta_5$. For $\beta_3^2 < 4\beta_4\beta_5$, the paraboloid opens upward (Figure 12.12a) when $\beta_4 + \beta_5 > 0$ and opens downward (Figure 12.12b) when $\beta_4 + \beta_5 < 0$.

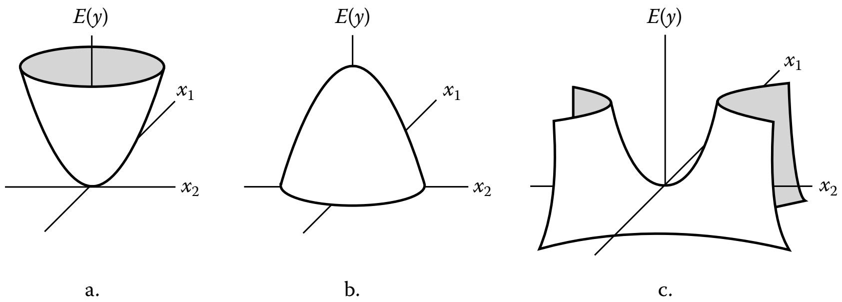



**FIGURE 12.12**
Graphs of three second-order surfaces

A complete second-order model is the three-dimensional equivalent of a second-order model in a single quantitative variable. Instead of tracing parabolas, it traces **paraboloids** and **saddle surfaces**. Since you fit only a portion of the complete surface to your data, a complete second-order model provides a very large variety of gently curving surfaces. It is a good choice for a model if you expect curvature in the response surface relating $E(y)$ to $x_1$ and $x_2$.

## Example 12.3

**Complete 2nd-order Model for Product Quality**

Many companies manufacture products (e.g., steel, paint, gasoline) that are at least partially produced using chemicals. In many instances, the quality of the finished product is a function of the temperature and pressure at which the chemical reactions take place. Suppose you want to model the quality, $y$, of a product as a function of the temperature, $x_1$, and the pressure, $x_2$, at which it is produced. Four inspectors independently assign a quality score between 0 and 100 to each product, and then the quality, $y$, is calculated by averaging the four scores. An experiment is conducted by varying temperature between 80°F and 100°F and pressure between 50 and 60 pounds per square inch. The resulting data are given in Table 12.2.

a. Fit the complete second-order model to the data.
b. Sketch the response surface.
c. Test the overall utility of the model.

**Solution**

a. The complete second-order model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

**PRODQUAL**

**TABLE 12.2 Temperature, Pressure, and Quality of the Finished Product**

| $x_1$, °F | $x_2$, psi | $y$ | $x_1$, °F | $x_2$, psi | $y$ | $x_1$, °F | $x_2$, psi | $y$ |
|---|---|---|---|---|---|---|---|---|
| 80 | 50 | 50.8 | 90 | 50 | 63.4 | 100 | 50 | 46.6 |
| 80 | 50 | 50.7 | 90 | 50 | 61.6 | 100 | 50 | 49.1 |
| 80 | 50 | 49.4 | 90 | 50 | 63.4 | 100 | 50 | 46.4 |
| 80 | 55 | 93.7 | 90 | 55 | 93.8 | 100 | 55 | 69.8 |
| 80 | 55 | 90.9 | 90 | 55 | 92.1 | 100 | 55 | 72.5 |
| 80 | 55 | 90.9 | 90 | 55 | 97.4 | 100 | 55 | 73.2 |
| 80 | 60 | 74.5 | 90 | 60 | 70.9 | 100 | 60 | 38.7 |
| 80 | 60 | 73.0 | 90 | 60 | 68.8 | 100 | 60 | 42.5 |
| 80 | 60 | 71.2 | 90 | 60 | 71.3 | 100 | 60 | 41.4 |

The data in Table 12.2 were used to fit this model, and a portion of the SAS output is shown in Figure 12.13.

The least-squares prediction equation is

$$\hat{y} = -5{,}127.90 + 31.10x_1 + 139.75x_2 - .146x_1x_2 - .133x_1^2 - 1.14x_2^2$$

b. A three-dimensional graph of this prediction model is shown in Figure 12.14. Note that the mean quality seems to be greatest for temperatures of about 85–90°F and for pressures of about 55–57 pounds per square inch.* Further experimentation in these ranges might lead to a more precise determination of the optimal temperature–pressure combination.

c. A look at the coefficient of determination, $R_a^2 = .991$, the $F$ value for testing the entire model, $F = 596.32$, and the $p$-value for the test, $p = .0001$ (in Figure 12.13), leaves little doubt that the complete second-order model is useful for explaining mean quality as a function of temperature and pressure. This, of course, will not always be the case. The additional complexity of second-order models is worthwhile only if a better model results. Consequently, it is important to determine whether the higher-order terms in the model (e.g., the curvilinear terms, $\beta_4$ and $\beta_5$) are statistically useful. A test for $H_0$: $\beta_4 = \beta_5 = 0$ is presented in Section 12.8.

Dependent Variable: QUALITY

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 8402.26454 | 1680.45291 | 596.32 | <.0001 |
| Error | 21 | 59.17843 | 2.81802 | | |
| Corrected Total | 26 | 8461.44296 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.67870 | R-Square | 0.9930 |
| Dependent Mean | 66.96296 | Adj R-Sq | 0.9913 |
| Coeff Var | 2.50690 | | |

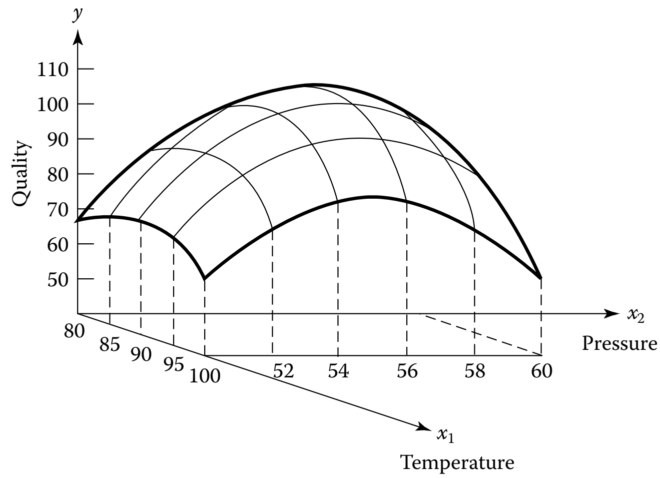
Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -5127.89907 | 110.29601 | -46.49 | <.0001 |
| TEMP | 1 | 31.09639 | 1.34441 | 23.13 | <.0001 |
| PRESSURE | 1 | 139.74722 | 3.14005 | 44.50 | <.0001 |
| TEMPRESS | 1 | -0.14550 | 0.00969 | -15.01 | <.0001 |
| TEMPSQ | 1 | -0.13339 | 0.00685 | -19.46 | <.0001 |
| PRESSQ | 1 | -1.14422 | 0.02741 | -41.74 | <.0001 |

**FIGURE 12.13**
SAS printout for complete second-order model

*We can estimate the values of temperature and pressure that maximize quality in the least-squares model by solving $\partial\hat{y}/\partial x_1 = 0$ and $\partial\hat{y}/\partial x_2 = 0$ for $x_1$ and $x_2$. These estimated optimal values are $x_1 = 86.25$°F and $x_2 = 55.58$ pounds per square inch.

**FIGURE 12.14**
Plot of second-order least-squares model for Example 12.3




## Applied Exercises

12.18 *Graphing a first-order model.* Suppose the true relationship between $E(y)$ and the quantitative independent variables $x_1$ and $x_2$ is described by the following first-order model:

$$E(y) = 4 - x_1 + 2x_2$$

a. Describe the corresponding response surface.
b. Plot the contour lines of the response surface for $x_1 = 2, 3, 4$, where $0 \leq x_2 \leq 5$.
c. Plot the contour lines of the response surface for $x_2 = 2, 3, 4$, where $0 \leq x_1 \leq 5$.
d. Use the contour lines you plotted in parts **b** and **c** to explain how changes in the settings of $x_1$ and $x_2$ affect $E(y)$.
e. Use your graph from part **b** to determine how much $E(y)$ changes when $x_1$ is changed from 4 to 2 and $x_2$ is simultaneously changed from 1 to 2.

12.19 *Graphing a second-order model.* Suppose the true relationship between $E(y)$ and the quantitative independent variables $x_1$ and $x_2$ is

$$E(y) = 4 - x_1 + 2x_2 + x_1x_2$$

Answer the questions posed in Exercise 12.16. Explain the effect of the interaction term on the mean response $E(y)$.

12.20 *Sorption of organic vapors.* Refer to the *Environmental Science & Technology* study of sorption of organic vapors, Exercise 12.7 (p. 646). Consider modeling the retention coefficient $y$ as a function of both

$$x_1 = \text{Temperature (degrees)}$$
$$x_2 = \text{Relative humidity (percent)}$$

a. Write a first-order model for $E(y)$.
b. Write a complete second-order model for $E(y)$.
c. Write a model for $E(y)$ that hypothesizes: (i) linear relationships, and (ii) that the relationship between retention ($y$) and temperature ($x_1$) depends on relative humidity ($x_2$).

12.21 *Modeling eel speed.* A Harvard University biologist used multiple regression to model the swimming speed of an American eel. (*Proceedings of the Royal Society, B*, Dec. 2004.) Steady swimming speed, $y$ (body lengths per second), was modeled as a function of the quantitative variables body wave speed, $x_1$ (body lengths per second), tail amplitude deviation, $x_2$ (body lengths), and tail velocity deviation, $x_3$ (body lengths per second).

a. Write a first-order model for $E(y)$ as a function of the three independent variables.
b. Give an interpretation of the value of $\beta_1$ in the model, part **a**.
c. Write a model for $E(y)$ as a function of the three independent variables that hypothesizes an interaction between body wave speed, $x_1$, and tail amplitude deviation, $x_2$.
d. In terms of the $\beta$'s in the model, part **c**, what is the change in $E(y)$ for every 1-unit increase in tail velocity deviation, $x_3$, for fixed values of $x_1$ and $x_2$?
e. In terms of the $\beta$'s in the model, part **c**, what is the change in $E(y)$ for every 1-unit increase in tail amplitude deviation, $x_2$, for fixed values of $x_1$ and $x_3$?

12.22 *Highway lane utilization.* Refer to the *Journal of Transportation Engineering* (May 2013) study of the lane utilization of a highway, Exercise 12.13 (p. 651). Recall that the dependent variable in the analysis was lane utilization $y$, measured as the percentage of vehicles in the lane. The researchers used two independent variables used to model $y$: $x_1$ = total traffic flow (total number of vehicles per hour) and $x_2$ = HGV flow (number of heavy-goods vehicles per hour). An analysis of traffic data collected over several weeks for different sections of Lane 1 of the M42 highway yielded the following results:

$$\hat{y} = .976 - .0000285x_1 - .002004x_2, \; R^2 = .70$$

a. Interpret the value of $R^2$ for this model.
b. Assuming $n = 2{,}000$ observations, conduct a test of overall model utility. Use $\alpha = .01$.
c. Use the $\beta$ estimates to draw a sketch of the estimated relationship between lane utilization and total traffic flow.
d. Use the $\beta$ estimates to draw a sketch of the estimated relationship between lane utilization and HGV flow.

12.23 *Muscle activity of harvesting foresters.* Research in the *International Journal of Foresting Engineering* (Vol. 19, 2008) investigated the muscle activity patterns in the neck and upper extremities exhibited during a work day among forestry vehicle operators. For one portion of the study, the researchers compared the muscle activity of operators of two types of harvesting vehicles—Timberjack and Valmet. (See Exercise 7.38.) In another part of the study, the researchers identified the key explanatory variables of $y$ = the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. A list of the potential predictors is provided below:

$x_1$ = Age of operator (years)
$x_2$ = Duration of lunch break (minutes)
$x_3$ = Dominant hand power level (percentage)
$x_4$ = Perceived stress at work (5-point scale)
$x_5$ = {1 if married, 0 if not}
$x_6$ = {1 if day shift, 0 if night shift}
$x_7$ = {1 if operating a Timberjack vehicle, 0 if operating a Valmet vehicle}

a. Write the equation of a first-order model for $E(y)$ as a function of the four quantitative independent variables.
b. Which of the $\beta$'s in the model, part **a** represents the change in $E(y)$ for every 1 percent increase in dominant hand power when the operator is 50 years old, takes a 30 minute lunch break, and has a perceived work stress of 2 points?
c. Add terms to the model, part **a** that allow for interactions between all possible pairs of the quantitative variables.
d. What function of the $\beta$'s in the model, part **c** represents the change in $E(y)$ for every 1 percent increase in dominant hand power, when the operator is 50 years old, takes a 30 minute lunch break, and has a perceived work stress of 2 points?

12.24 *Semiconductor wafer thickness.* Data on the polysilicon thickness of semiconductor wafers processed using rapid thermal chemical vapor deposition was analyzed in the *Journal of the American Statistical Association* (March 1998). The polysilicon thickness measurements (in angstroms) as well as the thickness of oxide applied to the wafer (in angstroms) and the deposition time (in seconds) for 22 wafers processed at a particular location are listed in the accompanying table.

a. Write a complete second-order model for polysilicon thickness ($y$) as a function of oxide thickness ($x_1$) and deposition time ($x_2$).
b. Fit the model to the data. Give the least-squares prediction equation.
c. Conduct a test to determine if the quadratic terms in the model are necessary. Test using $\alpha = .05$.
d. Conduct a test to determine if oxide thickness ($x_1$) and deposition time ($x_2$) interact. Test using $\alpha = .05$.
e. Based on the results, parts **c** and **d**, what model modifications do you recommend? Explain.

**WAFER2**

| Oxide Thickness ($x_1$, angstroms) | Time ($x_2$, seconds) | Polysilicon Thickness ($y$, angstroms) |
|---|---|---|
| 1059 | 18 | 494 |
| 1049 | 35 | 853 |
| 1039 | 52 | 1090 |
| 1026 | 52 | 1058 |
| 1001 | 18 | 517 |
| 986 | 35 | 882 |
| 1447 | 35 | 732 |
| 458 | 35 | 1143 |
| 1263 | 23 | 608 |
| 1283 | 23 | 590 |
| 1301 | 47 | 940 |
| 1287 | 47 | 920 |
| 1300 | 47 | 917 |
| 1307 | 23 | 581 |
| 632 | 23 | 738 |
| 621 | 23 | 732 |
| 623 | 23 | 750 |
| 620 | 47 | 1205 |
| 613 | 47 | 1194 |
| 615 | 47 | 1221 |
| 478 | 35 | 1209 |
| 1498 | 35 | 708 |

*Source*: Hughes-Oliver, J., Lu, J., and Gyurcsik, R. "Achieving uniformity in a semiconductor fabrication process using spatial modeling." *Journal of the American Statistical Association*, Vol. 93, March 1998.

12.25 *Seismic wave study.* An exploration seismologist wants to develop a model that will allow him to estimate the average signal-to-noise ratio of an earthquake's seismic wave, $y$, as a function of two independent variables:

$x_1$ = Frequency (cycles per second)
$x_2$ = Amplitude of the wavelet

a. Identify the independent variables as quantitative or qualitative.
b. Write the first-order model for $E(y)$.
c. Write a model for $E(y)$ that contains all first-order and interaction terms. Sketch typical response curves showing $E(y)$, the mean signal-to-noise ratio, versus $x_2$, the amplitude of the wavelet, for different values of $x_1$ (assume that $x_1$ and $x_2$ interact).
d. Write the complete second-order model for $E(y)$.

12.26 *Speech recognition device.* A study reported in *Human Factors* (Apr. 1990) investigated the effects of recognizer accuracy and vocabulary size on the performance of a computerized speech recognition device. Accuracy ($x_1$) of the device, measured as the percentage of correctly recognized spoken utterances, was set at three levels: 90%, 95%, and 99%. Vocabulary size ($x_2$), measured as the percentage of words needed for the task, was also set at three levels: 75%, 87.5%, and 100%. The dependent variable of primary interest was task completion time ($y$, in minutes), measured from when a user of the recognition device spoke the first input until the recognizer displayed the last spoken word of the task. Data collected for $n = 162$ trials were used to fit a complete second-order model for task completion time ($y$), as a function of the quantitative independent variables accuracy ($x_1$) and vocabulary ($x_2$). The coefficient of determination for the model was $R^2 = .75$.

a. Write the complete second-order model for $E(y)$.
b. Interpret the value of $R^2$.
c. Conduct a test of overall model adequacy. Use $\alpha = .05$.

12.27 *Tablet formulation study.* Researchers at the Upjohn Company utilized multiple regression analysis in the development of a sustained-release tablet.* One of the objectives of the research was to develop a model relating the dissolution $y$ of a tablet (i.e., the percentage of the tablet dissolved over a specified period of time) to the following independent variables:

$x_1$ = Excipient level (i.e., amount of nondrug ingredient in the tablet)
$x_2$ = Process variable (e.g., machine setting under which tablet is processed)

a. Write the complete second-order model for $E(y)$.
b. Write a model that hypothesizes straight-line relationships between $E(y)$, $x_1$, and $x_2$. Assume that $x_1$ and $x_2$ do not interact.
c. Repeat part **b**, but add interaction to the model.
d. For the model in part **c**, what is the slope of the linear relationship between $E(y)$ and $x_1$ for fixed $x_2$?
e. For the model in part **c**, what is the slope of the linear relationship between $E(y)$ and $x_2$ for fixed $x_1$?

**Source*: Klassen, R. A. "The Application of Response Surface Methods to a Tablet Formulation Problem." Paper presented at Joint Statistical Meetings, American Statistical Association and Biometric Society, Aug. 1986, Chicago, IL.

12.28 *Emotional intelligence and team performance.* Refer to the *Engineering Project Organizational Journal* (Vol. 3., 2013) study of how the emotional intelligence of individual team members relates directly to the performance of their team, Exercise 11.23 (p. 589). Recall that students enrolled in the course, *Introduction to the Building Industry*, completed an emotional intelligence test and received an interpersonal score, stress management score, and mood score. Students were then grouped into $n = 23$ teams and each team received an average project score. Three independent variables—range of interpersonal scores ($x_1$), range of stress management scores ($x_2$), and range of mood scores ($x_3$)—were used to model mean project score ($y$). The data for the analysis are reproduced below.

**TEAMPERF**

| Team | Intrapersonal (Range) | Stress (Range) | Mood (Range) | Project (Average) |
|---|---|---|---|---|
| 1 | 14 | 12 | 17 | 88.0 |
| 2 | 21 | 13 | 45 | 86.0 |
| 3 | 26 | 18 | 6 | 83.5 |
| 4 | 30 | 20 | 36 | 85.5 |
| 5 | 28 | 23 | 22 | 90.0 |
| 6 | 27 | 24 | 28 | 90.5 |
| 7 | 21 | 24 | 38 | 94.0 |
| 8 | 20 | 30 | 30 | 85.5 |
| 9 | 14 | 32 | 16 | 88.0 |
| 10 | 18 | 32 | 17 | 91.0 |
| 11 | 10 | 33 | 13 | 91.5 |
| 12 | 28 | 43 | 28 | 91.5 |
| 13 | 19 | 19 | 21 | 86.0 |
| 14 | 26 | 31 | 26 | 83.0 |
| 15 | 25 | 31 | 11 | 85.0 |
| 16 | 40 | 35 | 24 | 84.0 |
| 17 | 27 | 12 | 14 | 85.5 |
| 18 | 30 | 13 | 29 | 85.0 |
| 19 | 31 | 24 | 28 | 84.5 |
| 20 | 25 | 26 | 16 | 83.5 |
| 21 | 23 | 28 | 12 | 85.0 |
| 22 | 20 | 32 | 10 | 92.5 |
| 23 | 35 | 35 | 17 | 89.0 |

a. Hypothesize a complete 2nd-order model for project score ($y$) as a function of $x_1$, $x_2$, and $x_3$.
b. Fit the model, part **a** to the data using statistical software.

c. Evaluate the overall adequacy of the model using both a test of hypothesis and a numerical measure of model adequacy.
d. Is there sufficient evidence to indicate that the range of interpersonal scores ($x_1$) is curvilinearly related to average project score, $y$? Test at $\alpha = .01$.
e. Repeat part **d** for range of stress management scores ($x_2$).
f. Repeat part **d** for range of mood scores ($x_3$).

## 12.5 Coding Quantitative Independent Variables (*Optional*)

In fitting higher-order polynomial regression models (e.g., second- or third-order models), it is often a good practice to code the quantitative independent variables. For example, suppose one of the independent variables in a regression analysis is temperature, $T$, and $T$ is observed at three levels: 50°F, 100°F, and 150°F. We can code (or transform) the temperature measurements using the formula

$$x = \frac{T - 100}{50}$$

Then the coded levels $x = -1, 0$, and 1 correspond to the original levels 50°, 100°, and 150°.

In a general sense, **coding** means transforming a set of independent variables (qualitative or quantitative) into a new set of independent variables. For example, if we observe two independent variables,

$$T = \text{Temperature}$$
$$P = \text{Pressure}$$

then we can transform the two independent variables, $T$ and $P$, into two new coded variables, $x_1$ and $x_2$, where $x_1$ and $x_2$ are related to $T$ and $P$ by two functional equations,

$$x_1 = f_1(T, P) \qquad x_2 = f_2(T, P)$$

The functions $f_1$ and $f_2$ are algebraic relations that establish a one-to-one correspondence between combinations of levels of $T$ and $P$ with combinations of the coded values of $x_1$ and $x_2$.

Since qualitative independent variables are not numerical, it is necessary to code their values to fit the regression model. However, you might ask why we would bother to code the quantitative independent variables. There are two related reasons for coding quantitative variables. At first glance, it would appear that a computer would be oblivious to the values assumed by the independent variables in a regression analysis, but this is not the case. Recall from Section 11.3 that the computer must invert the $(\boldsymbol{X}'\boldsymbol{X})$ matrix to obtain the least-squares estimates of the model parameters. Considerable rounding error may occur during the inversion process if the numbers in the $(\boldsymbol{X}'\boldsymbol{X})$ matrix vary greatly in absolute value. This can produce sizable errors in the computed values of the least-squares estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots$. **Coding makes it computationally easier for the computer to invert the $(X'X)$ matrix, thus leading to more accurate estimates**.

A second reason for coding quantitative variables pertains to the problem of multicollinearity discussed in Section 11.11. When polynomial regression models (e.g., second-order models) are fit, the problem of multicollinearity is unavoidable, especially when higher-order terms are fit. For example, consider the quadratic model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

If the range of the values of $x$ is narrow, then the two variables, $x_1 = x$ and $x_2 = x^2$, will generally be highly correlated. As we pointed out in Section 11.11, the likelihood of rounding errors in the regression coefficients is increased in the presence of multicollinearity.

The best way to cope with the rounding error problem is to

1. Code the quantitative variable so that the new coded origin is in the center of the coded values. For example, by coding temperature, $T$, as

$$x = \frac{T - 100}{50}$$

we obtain coded values $-1, 0, 1$. This places the coded origin, 0, in the middle of the range of coded values ($-1$ to $1$).

2. Code the quantitative variable so that the range of the coded values is approximately the same for all coded variables. You need not hold exactly to this requirement. The range of values for one independent variable could be double or triple the range of another without causing any difficulty, but it would not be desirable to have a sizable disparity in the ranges, say, a ratio of 100 to 1.

When the data are observational (the values assumed by the independent variables are uncontrolled), the coding procedure described in the next box satisfies, reasonably well, these two requirements. The coded variable $u$ is similar to the standardized normal $z$ statistic of Section 5.5. Thus, the $u$ value is the deviation (the distance) between an $x$ value and the mean of the $x$ values, $\bar{x}$, expressed in units of $s_x$.* Since we know that most (approximately 95%) measurements in a set will lie within 2 standard deviations of their mean, it follows that most of the coded $u$ values will lie in the interval $-2$ to $+2$.

## Coding Procedure for Observational Data

Let

$x$ = Uncoded quantitative independent variable
$u$ = Coded quantitative independent variable

Then if $x$ takes values $x_1, x_2, \ldots, x_n$ for the $n$ data points in the regression analysis, let

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

where $s_x$ is the standard deviation of the $x$ values, i.e.,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

If you apply this coding to each quantitative variable, the range of values for each will be approximately $-2$ to $+2$. The variation in the absolute values of the elements of the coefficient matrix will be moderate, and rounding errors generated in finding the inverse of the matrix will be reduced. Additionally, the correlation between $x$ and $x^2$ will be reduced.†

**Example 12.4**

**Coding $x$ to Reduce Multicollinearity**

Carbon dioxide–baited traps are typically used by entomologists to monitor mosquito populations. An article in the *Journal of the American Mosquito Control Association* (Mar. 1995) investigated whether temperature influences the number of mosquitoes caught in a trap. Six mosquito samples were collected on each of nine consecutive days. For each day, two variables were measured: $x$ = average temperature (in

*The divisor of the deviation, $x - \bar{x}$, need not equal $s_x$ exactly. Any number approximately equal to $s_x$ would suffice. Other candidate denominators are range/4 and the interquartile range (IQR).

†Another by-product of coding is that the $\beta$ coefficients of the model have slightly different interpretations. For example, in the model $E(y) = \beta_0 + \beta_1 u$, where $u = (x - 10)/5$, the change in $y$ for every 1-unit increase in $x$ is not $\beta_1$, but $\beta_1/5$. In general, for first-order models with coded independent quantitative variables, the slope associated with $x_i$ is represented by $\beta_i/s_{x_i}$ where $s_{x_i}$ is the divisor of the coded $x_i$.

degrees Centigrade) and $y$ = mosquito catch ratio (the number of mosquitoes caught in each sample divided by the largest sample caught). The data are reported in Table 12.3.

The researchers are interested in relating catch ratio $y$ to average temperature $x$. Suppose we consider using a quadratic model.

a. Give the equation relating the coded variable $u$ to the temperature $x$ using the coding system for observational data.
b. Calculate the coded values, $u$, for the nine $x$ values.
c. Find the sum of the $n = 9$ values for $u$.

MOSQUITO

**TABLE 12.3 Data for Example 12.4**

| Date | Average Temperature, $x$ | Catch Ratio, $y$ |
|---|---|---|
| July 24 | 16.8 | .66 |
| 25 | 15.0 | .30 |
| 26 | 16.5 | .46 |
| 27 | 17.7 | .44 |
| 28 | 20.6 | .67 |
| 29 | 22.6 | .99 |
| 30 | 23.3 | .75 |
| 31 | 18.2 | .24 |
| Aug. 1 | 18.6 | .51 |

*Source*: Petric, D., et. al., "Dependence of $CO_2$-baited suction trap captures on temperature variations." *Journal of the American Mosquito Control Association*, Vol. 11, No. 1. Mar. 1995, p. 8.

Solution

a. We first find $\bar{x}$ and $s_x$. From the MINITAB printout, Figure 12.15, which provides summary statistics for temperature, $x$, we obtain

$$\bar{x} = 18.811 \quad \text{and} \quad s_x = 2.812$$

Then the equation relating $u$ and $x$ is

$$u = \frac{x - 18.8}{2.8}$$

b. When temperature $x = 16.8$

$$u = \frac{x - 18.8}{2.8} = \frac{16.8 - 18.8}{2.8} = -.71$$

Similarly, when $x = 15.0$

$$u = \frac{x - 18.8}{2.8} = \frac{15.0 - 18.8}{2.8} = -1.36$$

Table 12.4 gives the coded values for all $n = 9$ observations. (*Note*: You can see that all the $n = 9$ values for $u$ lie in the interval from $-2$ to $+2$.)

c. If you ignore rounding error, the sum of the $n = 9$ values for $u$ will equal 0. This is because the sum of the deviations of a set of measurements about their mean is always equal to 0.

```
Variable          N      Mean    Median    TrMean     StDev   SE Mean
X                 9    18.811    18.200    18.811     2.812     0.937

Variable    Minimum   Maximum        Q1        Q3
X            15.000    23.300    16.650    21.600
```

**FIGURE 12.15**
MINITAB descriptive statistics for temperature, $x$

**TABLE 12.4 Coded Values of $x$, Example 12.4**

| Temperature, $x$ | Coded Values, $u$ |
|---|---|
| 16.8 | −.71 |
| 15.0 | −1.36 |
| 16.5 | −.82 |
| 17.7 | −.39 |
| 20.6 | .64 |
| 22.6 | 1.36 |
| 23.3 | 1.61 |
| 18.2 | −.21 |
| 18.6 | −.07 |

To illustrate the advantage of coding, consider fitting the second-order model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

to the data of Example 12.4. The coefficient of correlation between the two variables, $x$ and $x^2$, shown at the top of the MINITAB printout displayed in Figure 12.16, is $r = .998$. However, the coefficient of correlation between the corresponding coded values, $u$ and $u^2$, shown at the bottom of Figure 12.16, is only $r = .448$. Thus, we can avoid potential rounding error caused by highly correlated $x$ values by fitting, instead, the model

$$E(y) = \beta_0^* + \beta_1^* u + \beta_2^* u^2$$

Other methods of coding have been developed to reduce rounding errors and multicollinearity. One of the more complex coding systems involves fitting **orthogonal**

**FIGURE 12.16**
MINITAB correlations for temperature, $x$, and coded temperature, $u$

```
Correlations: X, XSQ

Pearson correlation of X and XSQ = 0.998
P-Value = 0.000

Correlations: U, USQ

Pearson correlation of U and USQ = 0.448
P-Value = 0.227
```

**polynomials**. An orthogonal system of coding guarantees that the coded independent variables will be uncorrelated. For a discussion of orthogonal polynomials, consult the references given at the end of this chapter.

## Applied Exercises

12.29 *Processed straw as thermal insulation.* Refer to the *Engineering Structures and Technologies* (Sep. 2012) study on the use of processed straw as thermal insulation for homes, Exercise 11.4 (p. 579). You used data on $n = 25$ straw specimens (see right column) to fit a quadratic model relating $y =$ thermal conductivity (watts per meter-Kelvin) and $x =$ density (kilograms per cubic meter).

a. Demonstrate that the correlation between $x$ and $x^2$ is high. What are potential consequences of estimating the quadratic model using $x$ and $x^2$?
b. Give the equation relating the coded variable $u$ for density ($x$), using the coding system for observational data.
c. Demonstrate that the correlation between $u$ and $u^2$ is near 0.
d. Fit the model, $E(y) = \beta_0 + \beta_1 u + \beta_2 u^2$, using available statistical software. Interpret the results.

12.30 *Tire pressure and mileage.* Suppose you want to use the coding system for observational data to fit a second-order model to the tire pressure–automobile mileage data given in the next table below.

**TIRES**

| Pressure $x$, pounds per square inch | Mileage $y$, thousands | Pressure $x$, pounds per square inch | Mileage $y$, thousands |
|---|---|---|---|
| 30 | 29.5 | 33 | 37.6 |
| 30 | 30.2 | 34 | 37.7 |
| 31 | 32.1 | 34 | 36.1 |
| 31 | 34.5 | 35 | 33.6 |
| 32 | 36.3 | 35 | 34.2 |
| 32 | 35.0 | 36 | 26.8 |
| 33 | 38.2 | 36 | 27.4 |

**STRAW**

| Specimen | Thermal Conductivity ($y$) | Density ($x$) |
|---|---|---|
| 1 | 0.052 | 49 |
| 2 | 0.045 | 50 |
| 3 | 0.055 | 51 |
| 4 | 0.042 | 56 |
| 5 | 0.048 | 57 |
| 6 | 0.049 | 62 |
| 7 | 0.046 | 64 |
| 8 | 0.047 | 65 |
| 9 | 0.051 | 66 |
| 10 | 0.047 | 68 |
| 11 | 0.049 | 78 |
| 12 | 0.048 | 79 |
| 13 | 0.048 | 82 |
| 14 | 0.052 | 83 |
| 15 | 0.051 | 84 |
| 16 | 0.053 | 98 |
| 17 | 0.054 | 100 |
| 18 | 0.055 | 100 |
| 19 | 0.057 | 101 |
| 20 | 0.055 | 103 |
| 21 | 0.074 | 115 |
| 22 | 0.075 | 116 |
| 23 | 0.077 | 118 |
| 24 | 0.076 | 119 |
| 25 | 0.074 | 120 |

a. Give the equation relating the coded variable $u$ to pressure, $x$, using the coding system for observational data.
b. Calculate the coded values, $u$.
c. Calculate the coefficient of correlation $r$ between the variables $x$ and $x^2$.
d. Calculate the coefficient of correlation $r$ between the variables $u$ and $u^2$. Compare this value to the value computed in part **c**.
e. Fit the model

$$E(y) = \beta_0 + \beta_1 u + \beta_2 u^2$$

using available statistical software. Interpret the results.

12.31 *Level of overcrowding at a train station.* Refer to the *Journal of Transportation Engineering* (June 2013) study of overcrowding at a train station, Exercise 12.14 (p. 652). Using data collected for a sample of 20 train stops (see below), you analyzed the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $y$ = average distance (meters) between passengers and $x$ = volume of passengers waiting (number of persons) on the train platform.

**TRAINWAIT**

| Stop | Volume | Distance |
|---|---|---|
| 1 | 21 | 0.08 |
| 2 | 2 | 0.78 |
| 3 | 4 | 0.62 |
| 4 | 25 | 0.06 |
| 5 | 16 | 0.11 |
| 6 | 26 | 0.06 |
| 7 | 6 | 0.35 |
| 8 | 22 | 0.07 |
| 9 | 23 | 0.06 |
| 10 | 9 | 0.19 |
| 11 | 11 | 0.15 |
| 12 | 20 | 0.09 |
| 13 | 17 | 0.10 |
| 14 | 18 | 0.10 |
| 15 | 19 | 0.11 |
| 16 | 7 | 0.27 |
| 17 | 29 | 0.05 |
| 18 | 5 | 0.50 |
| 19 | 14 | 0.12 |
| 20 | 8 | 0.20 |

a. Find the correlation between $x$ and $x^2$. What potential problems may occur due to this correlation? Do you recommend coding the independent variable, $x$?
b. Give the equation relating the coded variable $u$ for volume($x$), using the coding system for observational data.
c. Find the correlation between $u$ and $u^2$. Has the multicollinearity problem been diminished?
d. Fit the model, $E(y) = \beta_0 + \beta_1 u + \beta_2 u^2$, using available statistical software. Interpret the results.

12.32 *Estimating repair and replacement costs of water pipes.* Refer to the *IHS Journal of Hydraulic Engineering* (September 2012) study of the repair and replacement of water pipes, Exercise 11.37 (p. 602). Recall that a team of civil engineers used regression analysis to model $y$ = the ratio of repair to replacement cost of commercial pipe as a function of $x$ = the diameter (in millimeters) of the pipe. Using data for a sample of 13 different pipe sizes (see below) you fit the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Is there a high level of multicollinearity in the independent variables? If so, propose an alternative model that does not suffer from the same high level of multicollinearity. Fit the model to the data and interpret the results.

**WATERPIPE**

| DIAMETER | RATIO |
|---|---|
| 80 | 6.58 |
| 100 | 6.97 |
| 125 | 7.39 |
| 150 | 7.61 |
| 200 | 7.78 |
| 250 | 7.92 |
| 300 | 8.20 |
| 350 | 8.42 |
| 400 | 8.60 |
| 450 | 8.97 |
| 500 | 9.31 |
| 600 | 9.47 |
| 700 | 9.72 |

*Source:* Suribabu, C.R. & Neelakantan, T.R. “Sizing of water distribution pipes based on performance measure and breakage-repair replacement economics”, *IHS Journal of Hydraulic Engineering,* Vol. 18, No. 3, September 2012 (Table 1).

## 12.6 Models with One Qualitative Independent Variable

Suppose we want to write a model for the mean performance, $E(y)$, of a diesel engine as a function of type of fuel. (For the purpose of explanation, we will ignore other independent variables that might affect the response.) Further, suppose there are three fuel types available: a petroleum-based fuel (P), a coal-based fuel (C), and a blended fuel (B). The fuel type is a single qualitative variable with three levels corresponding to fuels P, C, and B. Note that with a qualitative independent variable, we cannot attach a quantitative meaning to a given level. All we can do is describe it.

To simplify our notation, let $\mu_P$ be the mean performance for fuel P, and let $\mu_C$ and $\mu_B$ be the corresponding mean performances for fuels C and B. Our objective is to write a single prediction equation that will give the mean value of $y$ for the three fuel types. This can be done as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$$x_1 = \begin{cases} 1 & \text{if fuel P is used} \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if fuel C is used} \\ 0 & \text{if not} \end{cases}$$

The values of $x_1$ and $x_2$ for each of the three fuel types are shown in Table 12.5.

The variables $x_1$ and $x_2$ are not meaningful independent variables as for the case of the models with quantitative independent variables. Instead, they are **dummy (indicator) variables** that make the model function. To see how they work, let $x_1 = 0$ and $x_2 = 0$. This condition will apply when we are seeking the mean response for fuel B (neither fuel P nor C is used; hence, it must be B). Then the mean value of $y$ when fuel B is used is

$$\mu_B = E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

This tells us that the mean performance level for fuel B is $\beta_0$. Or, it means that $\beta_0 = \mu_B$.

Now suppose we want to represent the mean response, $E(y)$, when fuel P is used. Checking the dummy variable definitions, we see that we should let $x_1 = 1$ and $x_2 = 0$:

$$\mu_P = E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

or, since $\beta_0 = \mu_B$,

$$\mu_P = \mu_B + \beta_1$$

Then it follows that the interpretation of $\beta_1$ is

$$\beta_1 = \mu_P - \mu_B$$

which is the difference in the mean performance levels for fuels P and B.

**TABLE 12.5 Mean Response for the Model with Three Diesel Fuel Types**

| Fuel Type | $x_1$ | $x_2$ | Mean Response, $E(y)$ |
|---|---|---|---|
| Blended (B) | 0 | 0 | $\beta_0 = \mu_B$ |
| Petroleum (P) | 1 | 0 | $\beta_0 + \beta_1 = \mu_P$ |
| Coal (C) | 0 | 1 | $\beta_0 + \beta_2 = \mu_C$ |

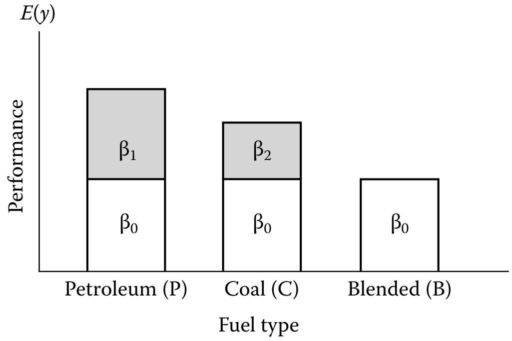



**FIGURE 12.17**
Bar chart comparing $E(y)$ for three diesel fuel types

Finally, if we want the mean value of $y$ when fuel C is used, we set $x_1 = 0$ and $x_2 = 1$:

$$\mu_C = E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

or, since $\beta_0 = \mu_B$,

$$\mu_C = \mu_B + \beta_2$$

Then it follows that the interpretation of $\beta_2$ is

$$\beta_2 = \mu_C - \mu_B$$

Note that we were able to describe *three levels* of the qualitative variable with only *two dummy variables*, because the mean of the base level (fuel B, in this case) is accounted for by the intercept $\beta_0$.

Since fuel type is a qualitative variable, we will use a bar graph to show the value of mean performance, $E(y)$, for the three levels of fuel type (see Figure 12.17). In particular, note that the height of the bar, $E(y)$, for each level of fuel type is equal to the sum of the model parameters shown in the preceding equations. You can see that the height of the bar corresponding to fuel B is $\beta_0$; i.e., $E(y) = \beta_0$. Similarly, the heights of the bars corresponding to P and C are $E(y) = \beta_0 + \beta_1$ and $E(y) = \beta_0 + \beta_2$, respectively.*

Now, carefully examine the model for a single qualitative independent variable with three levels, because we will use exactly the same pattern for any number of levels. Arbitrarily select one level to be the **base level**, (i.e., the level assigned all 0-values for dummy variables), then set up 1–0 dummy variables for the remaining levels.† This setup always leads to the interpretation of the parameters given in the box.

### Procedure for Writing a Model with One Qualitative Independent Variable at $k$ Levels $(A, B, C, D, \ldots)$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$$

where

$$x_i = \begin{cases} 1 & \text{if qualitative variable at level } i + 1 \\ 0 & \text{otherwise} \end{cases}$$

*Either $\beta_1$ or $\beta_2$, or both, could be negative. If, for example, $\beta_1$ were negative, the height of the bar corresponding to fuel P would be *reduced* (rather than increased) from the height of the bar for fuel B by the amount $\beta_1$. Figure 12.17 is constructed assuming that $\beta_1$ and $\beta_2$ are positive quantities.

†We do not have to use a 1–0 system of coding for the dummy variables. Any two-value system will work, but the interpretation given to the model parameters will depend on the code. Using the 1–0 system makes the model parameters easy to interpret.

The number of dummy variables for a single qualitative variable is always 1 less than the number of levels for the variable. Then, assuming the base level is A, the mean for each level is

$$\begin{aligned}\mu_A &= \beta_0\\ \mu_B &= \beta_0 + \beta_1\\ \mu_C &= \beta_0 + \beta_2\\ \mu_D &= \beta_0 + \beta_3\\ &\vdots\end{aligned}$$

$\beta$ Interpretations:

$$\begin{aligned}\beta_0 &= \mu_A\\ \beta_1 &= \mu_B - \mu_A\\ \beta_2 &= \mu_C - \mu_A\\ \beta_3 &= \mu_D - \mu_A\\ &\vdots\end{aligned}$$

## Example 12.5

### Cost Model with a Qualitative Independent Variable

A large consulting firm markets a computerized system for monitoring road construction bids to various state departments of transportation. Since the high cost of maintaining the system is partially absorbed by the firm, the firm wants to compare the mean annual maintenance costs accrued by system users in three different states: Kansas, Kentucky, and Texas. A sample of 10 users is selected from each state installation and the maintenance cost accrued by each is recorded, as shown in Table 12.6.

a. Do the data provide sufficient evidence (at $\alpha = .05$) to indicate that the mean annual maintenance costs accrued by system users differ for the three state installations?

b. Find and interpret a 95% confidence interval for the difference between the mean costs in Texas and Kansas.

**Solution**

a. The model relating $E(y)$ to the single qualitative variable, state installation, is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$$x_1 = \begin{cases} 1 & \text{if Kentucky} \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if Texas} \\ 0 & \text{if not} \end{cases}$$

**BIDMAINT**

**TABLE 12.6 Annual Maintenance Costs**

| | State Installation | | |
|---|---|---|---|
| | 1: Kansas | 2: Kentucky | 3: Texas |
| | $ 198 | $ 563 | $ 385 |
| | 126 | 314 | 693 |
| | 443 | 483 | 266 |
| | 570 | 144 | 586 |
| | 286 | 585 | 178 |
| | 184 | 377 | 773 |
| | 105 | 264 | 308 |
| | 216 | 185 | 430 |
| | 465 | 330 | 644 |
| | 203 | 354 | 515 |
| Totals | $2,796 | $3,599 | $4,778 |

and

$$\beta_1 = \mu_2 - \mu_1$$
$$\beta_2 = \mu_3 - \mu_1$$

where $\mu_1$, $\mu_2$, and $\mu_3$ are the mean responses for Kansas, Kentucky, and Texas, respectively. Testing the null hypothesis that the means for the three states are equal, i.e., $\mu_1 = \mu_2 = \mu_3$, is equivalent to testing

$$H_0\text{:} \quad \beta_1 = \beta_2 = 0$$

because if $\beta_1 = \mu_2 - \mu_1 = 0$ and $\beta_2 = \mu_3 - \mu_1 = 0$, then $\mu_1$, $\mu_2$, and $\mu_3$ must be equal. The alternative hypothesis is

$H_a$: At least one of the parameters, $\beta_1$ or $\beta_2$, differs from 0

We conduct the $F$ test for the complete model (Section 11.5), which tests the null hypothesis that all parameters in the model, with the exception of $\beta_0$, equal 0. The SPSS printout for fitting the complete model,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is shown in Figure 12.18. The value of the $F$ statistic for testing the complete model (shaded on Figure 12.18) is $F = 3.482$; the $p$-value for the test (also shaded) is $p = .045$. Since our choice of $\alpha$, $\alpha = .05$, exceeds the $p$-value, we reject $H_0$ and

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .453[a] | .205 | .146 | 168.948 |

a. Predictors: (Constant), X2, X1

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 198772.5 | 2 | 99386.233 | 3.482 | .045[a] |
| | Residual | 770670.9 | 27 | 28543.367 | | |
| | Total | 969443.4 | 29 | | | |

a. Predictors: (Constant), X2, X1

b. Dependent Variable: COST

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 279.600 | 53.426 | | 5.233 | .000 | 169.979 | 389.221 |
| | X1 | 80.300 | 75.556 | .211 | 1.063 | .297 | -74.728 | 235.328 |
| | X2 | 198.200 | 75.556 | .520 | 2.623 | .014 | 43.172 | 353.228 |

a. Dependent Variable: COST

**FIGURE 12.18**
SPSS printout for dummy variable model

conclude that at least one of the parameters, $\beta_1$ or $\beta_2$, differs from 0. Or, equivalently, we conclude that the data provide sufficient evidence to indicate that the mean user maintenance cost does vary among the three state installations.

b. Since $\beta_2 = \mu_3 - \mu_1 =$ the difference between the mean costs of Texas and Kansas, we want a 95% confidence interval for $\beta_2$. The interval, highlighted on Figure 12.18, is (43.172, 353.228). Consequently, we are 95% confident that the difference, $\mu_3 - \mu_1$, falls in our interval. This implies that the mean cost of users in Texas is anywhere from \$43.17 to \$353.23 higher than the mean cost of Kansas users.

A second method of analyzing the data of Table 12.6 is known as analysis of variance, or ANOVA. ANOVA is the topic of Chapters 13 and 14.

## Applied Exercises

12.33 *Chemical composition of rainwater.* Refer to the *Journal of Agricultural, Biological, and Environmental Statistics* (March 2005) study of the chemical composition of rainwater, Exercise 12.1 (p. 646). Recall that the nitrate concentration, $y$ (milligrams per liter), in a rainwater sample was modeled as a function of water source (groundwater, subsurface flow, or overground flow).

a. Write a model for $E(y)$ as a function of the qualitative independent variable.

b. Give an interpretation of each of the $\beta$ parameters in the model, part **c**.

12.34 *Emotional stress of firefighters.* Refer to the *Journal of Human Stress* study of firefighters, Exercise 12.5 (p. 646). Consider using the qualitative variable, level of social support, as a predictor of emotional stress $y$. Suppose that four social support levels were studied: none, low, moderate, and high.

a. Write a model for $E(y)$ as a function of social support at four levels.

b. Interpret the $\beta$ parameters in the model.

c. Explain how to test for differences among the emotional stress means for the four social support levels.

12.35 *Sorption rate of organic vapors.* Refer to the *Environmental Science & Technology* study of sorption of organic vapors, Exercise 12.7 (p. 646). Consider using the qualitative variable, organic compound, as a predictor of the retention coefficient $y$. Recall that five organic compounds were studied: benzene, toluene, chloroform, methanol, and anisole.

a. Write a model for $E(y)$ as a function of organic compound at two levels.

b. Interpret the $\beta$ parameters in the model.

c. Explain how to test for differences among the mean retention coefficients of the five organic compounds.

12.36 *Muscle activity of harvesting foresters.* Refer to the *International Journal of Foresting Engineering* (Vol. 19, 2008) study of neck muscle activity patterns among forestry vehicle operators, Exercise 12.23 (p. 660). Recall that the researchers identified the key explanatory variables of $y =$ the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. A list of the potential predictors is reproduced below:

$x_1 =$ Age of operator (years)
$x_2 =$ Duration of lunch break (minutes)
$x_3 =$ Dominant hand power level (percentage)
$x_4 =$ Perceived stress at work (5-point scale)
$x_5 = \{1$ if married, 0 if not$\}$
$x_6 = \{1$ if day shift, 0 if night shift$\}$
$x_7 = \{1$ if operating a Timberjack vehicle, 0 if operating a Valmet vehicle$\}$

a. Write the equation of a model for $E(y)$ as a function of the day/night shift qualitative independent variable.

b. Which of the $\beta$'s in the model, part **a**, represents the difference in $E(y)$ values between the day shift and night shift operators?

c. Write the equation of a model for $E(y)$ as a function of the vehicle type qualitative independent variable.

d. If the $\beta$ multiplied by $x_7$ in the model, part c, is negative, what does this imply practically?

12.37 *Whales entangled in fishing gear.* Refer to the *Marine Mammal Science* (April 2010) study of whales entangled in fishing gear, Exercise 11.18 (p. 588). These entanglements involved one of three types of fishing gear: set nets, pots, and gill nets. Consequently, the researchers used gear type as a predictor of the body length ($y$, in meters) of the entangled whale. Consider the regression model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $x_1 = \{1$ if set net, 0 if not$\}$ and $x_2 = \{1$ if pots, 0 if not$\}$. [Note: Gill nets is the "base" level of gear type.]

a. The researchers want to know the mean body length of whales entangled in gill nets. Give an expression for this value in terms of the $\beta$'s in the model.

b. Practically interpret the value of $\beta_1$ in the model.

c. In terms of the $\beta$'s in the model, how would you test to determine if the mean body lengths of entangled whales differ for the three types of fishing gear?

12.38 *Magnetron tube study.* An electrical engineer wants to compare the mean lifelengths (in hours) of five different brands of magnetron tubes. Data are gathered on 10 magnetron tubes selected at random from each of the five brands. Write a model that will give the mean lifelength for the five brands and interpret all the $\beta$ parameters used in the model.

12.39 *Improving milk production with shade.* Because of the hot, humid weather conditions in Florida, the growth rates of beef cattle and the milk production of dairy cows typically decline during the summer. However, agricultural and environmental engineers have found that a well-designed shade structure can significantly increase the milk production of dairy cows. In one experiment, 30 cows were selected and divided into three groups of 10 cows each. Group 1 cows were provided with an artificial shade structure, group 2 cows with tree shade, and group 3 cows with no shade. Of interest was the mean milk production (in gallons) of the cows in each group.

a. Identify the independent variables in the experiment.
b. Write a model relating the mean milk production, $E(y)$, to the independent variables. Identify and code all dummy variables.
c. Interpret the $\beta$ parameters of the model.

12.40 *Comparing insect repellents.* Which insect repellents protect best against mosquitos? *Consumer Reports* (June 2000) tested 14 products that all claim to be an effective mosquito repellent. Each product was classified as either lotion/cream or aerosol/spray. The cost of the product (in dollars) was divided by the amount of the repellent needed to cover exposed areas of the skin (about 1/3 ounce) to obtain a cost-per-use value. Effectiveness was measured as the maximum number of hours of protection (in half-hour increments) provided when human testers exposed their arms to 200 mosquitos. The data from the report are listed in the table on bottom of page.

a. Suppose you want to use repellent type to model the cost per use ($y$). Create the appropriate number of dummy variables for repellent type and write the model.
b. Fit the model, part **a**, to the data.
c. Give the null hypothesis for testing whether repellent type is a useful predictor of cost per use ($y$).
d. Conduct the test, part **c**, and give the appropriate conclusion. Use $\alpha = .10$.
e. Repeat parts **a–d** if the dependent variable is maximum number of hours of protection ($y$).

**NZBIRDS**

12.41 *Extinct New Zealand birds. Evolutionary Ecology Research* (July 2003) published a study of the patterns of extinction in the New Zealand bird population. The **NZBIRDS** file contains qualitative data on flight capability (volant or flightless), habitat (aquatic, ground terrestrial, or aerial terrestrial), nesting site (ground, cavity within ground, tree, cavity above ground), nest density (high or low), diet (fish, vertebrates, vegetables, or invertebrates), and extinct status (extinct, absent from island, present), and quantitative data on body mass (grams) and egg length (millimeters) for 132 bird species at the time of the Maori colonization of New Zealand.

a. Write a model for mean body mass as a function of flight capability.
b. Write a model for mean body mass as a function of diet.
c. Write a model for mean egg length as a function of nesting site.
d. Fit the model, part **a**, to the data and interpret the estimates of the $\beta$'s.

**REPELLENT**

| Insect Repellent | Type | Cost/Use | Maximum Protection |
|---|---|---|---|
| Amway HourGuard 12 | Lotion/Cream | $2.08 | 13.5 hours |
| Avon Skin-So-Soft | Aerosol/Spray | 0.67 | 0.5 |
| Avon BugGuard Plus | Lotion/Cream | 1.00 | 2.0 |
| Ben's Backyard Formula | Lotion/Cream | 0.75 | 7.0 |
| Bite Blocker | Lotion/Cream | 0.46 | 3.0 |
| BugOut | Aerosol/Spray | 0.11 | 6.0 |
| Cutter Skinsations | Aerosol/Spray | 0.22 | 3.0 |
| Cutter Unscented | Aerosol/Spray | 0.19 | 5.5 |
| Muskoll Ultra6Hours | Aerosol/Spray | 0.24 | 6.5 |
| Natrapel | Aerosol/Spray | 0.27 | 1.0 |
| Off! Deep Woods | Aerosol/Spray | 1.77 | 14.0 |
| Off! Skintastic | Lotion/Cream | 0.67 | 3.0 |
| Sawyer Deet Formula | Lotion/Cream | 0.36 | 7.0 |
| Repel Permanone | Aerosol/Spray | 2.75 | 24.0 |

*Source*: "Buzz off." *Consumer Reports*, June 2000.

e. Conduct a test to determine if the model, part **a**, is statistically useful (at $\alpha = .01$) for estimating mean body mass.
f. Fit the model, part **b**, to the data and interpret the estimates of the $\beta$'s.
g. Conduct a test to determine if the model, part **b**, is statistically useful (at $\alpha = .01$) for estimating mean body mass.
h. Fit the model, part **c**, to the data and interpret the estimates of the $\beta$'s.
i. Conduct a test to determine if the model, part **c**, is statistically useful (at $\alpha = .01$) for estimating mean egg length.

12.42 *Greenhouse gas emissions.* Wastewater treatment systems are designed to maintain the chemical, physical, and biological integrity of water. These systems, however, tend to generate various greenhouse gases, such as methane ($CH_4$). The *Journal of the Institute of Engineering* (Vol. 8, 2011) published a study of the amount of methane gas (milligrams per liter) emitted from wastewater treatment sludge. Two different types of treated sludge were investigated: (1) sludge without nutrients added and (2) sludge with nutrients added. The specific nutrient studied was volatile fatty acids (VFA). Data for $n = 27$ sludge specimens are listed in the table. Use regression to determine if the mean amount of methane gas emitted differs for the two types of sludge, and if so, provide a 95% confidence interval for the magnitude of the difference.

**SLUDGE**

| Specimen | Methane($CH_4$) | Time | Nutrients(VHA) |
|---|---|---|---|
| 1 | 5 | 20 | No |
| 2 | 9 | 21 | No |
| 3 | 18 | 24 | No |
| 4 | 35 | 26 | No |
| 5 | 61 | 29 | No |
| 6 | 65 | 32 | No |
| 7 | 105 | 35 | No |
| 8 | 120 | 37 | No |
| 9 | 117 | 42 | No |
| 10 | 154 | 44 | No |
| 11 | 200 | 47 | No |
| 12 | 198 | 49 | No |
| 13 | 203 | 51 | No |
| 14 | 21 | 20 | Yes |
| 15 | 25 | 21 | Yes |
| 16 | 61 | 24 | Yes |
| 17 | 75 | 26 | Yes |
| 18 | 102 | 29 | Yes |
| 19 | 150 | 32 | Yes |
| 20 | 183 | 34 | Yes |
| 21 | 194 | 36 | Yes |
| 22 | 245 | 37 | Yes |
| 23 | 308 | 42 | Yes |
| 24 | 295 | 44 | Yes |
| 25 | 272 | 47 | Yes |
| 26 | 280 | 49 | Yes |
| 27 | 287 | 51 | Yes |

*Source:* Devkota, R.P. " Greenhouse Gas Emissions from Wastewater Treatment System", *Journal of the Institute of Engineering,* Vol. 8, No. 1, 2011 (adapted from Figure 4).

12.43 *Corporate sustainability and firm characteristics. Corporate sustainability* refers to business practices designed around social and environmental considerations (e.g., "going green" and energy conservation). *Business and Society* (March 2011) published a paper on how firm size and firm type impacts sustainability behaviors. Nearly 1,000 senior managers were surveyed on their firms' likelihood of reporting sustainability policies (measured as a probability between 0 and 1). The managers were divided into four groups depending on firm size (large or small) and firm type (public or private): large/public, large/private, small/public, and small/private. One goal of the analysis is to determine whether the mean likelihood of reporting sustainability policies differs depending on firm size and firm type.

a. Consider a single qualitative variable representing the four size/type categories. Create the appropriate dummy variables for representing this qualitative variable as an independent variable in a regression model for predicting likelihood of reporting sustainability policies ($y$).
b. Give the equation of the model, part **a**, and interpret each of the model parameters.
c. The global $F$-test for the model resulted in $p$-value $<.001$. Give a practical interpretation of this result.
d. Now consider treating firm size and firm type as two different qualitative independent variables in a model for likelihood of reporting sustainability policies ($y$). Create the appropriate dummy variables for representing these qualitative variables in the model.
e. Refer to part **d**. Write a model for $E(y)$ as a function of firm size and firm type, but do not include interaction. (This model is called the *main effects* model.)
f. Refer to the model, part **e**. For each combination of firm size and firm type (e.g., large/public), write $E(y)$ as a function of the model parameters.
g. Use the results, part **f**, to show that for the main effects model, the difference between the mean likelihoods for large and small firms does not depend on firm type.
h. Write a model for $E(y)$ as a function of firm size, firm type, and size $\times$ type interaction. [Hint: For this model, include all possible interactions between pairs of dummy variables, where one dummy variable

represents firm size and the other dummy variable represents firm type.]

i. Refer to the model, part **h**. For each combination of firm size and firm type (e.g., large/public), write $E(y)$ as a function of the model parameters.

j. Use the results, part **i**, to show that for the interaction model, the difference between the mean likelihoods for large and small firms does depend on firm type.

## 12.7 Models with Both Quantitative and Qualitative Independent Variables

Perhaps the most interesting data analysis problems are those that involve both quantitative and qualitative independent variables. For example, suppose you want to relate the mean performance, $E(y)$, of a diesel engine to engine speed (rpm), $x$, for three different fuel types—petroleum, coal, and blended—and you wish to use first-order (straight-line) models to model the responses for all three fuels. Graphs of these three relationships might appear as shown in Figure 12.19.

Since the lines in Figure 12.19 are hypothetical, a number of practical questions arise. Does one fuel type perform as well as any other; that is, do the three mean performance lines differ for the three fuel types? Does the rate of increase in mean performance level with engine speed differ for the three fuel types; that is, do the slopes of the three lines differ? Note that each of the two practical questions has been rephrased into a question about the parameters that define the three lines of Figure 12.19. To answer them, we must write a single linear statistical model that will characterize the three lines of Figure 12.19. Then the practical questions can be answered by testing hypotheses about the model parameters.

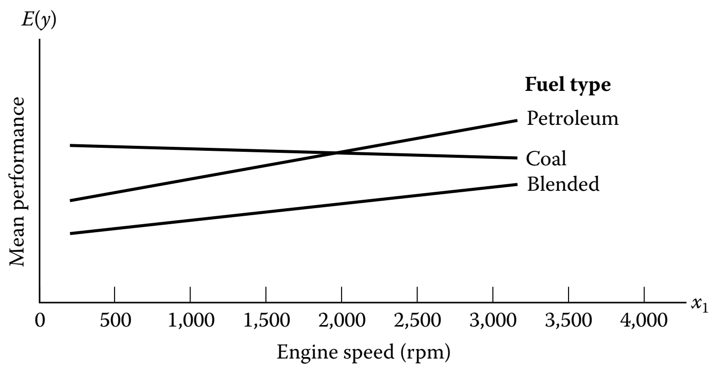
In the previous example, the response (engine performance) is a function of *two* independent variables: one quantitative (engine speed, $x$) and one qualitative (type of fuel). We will examine the different models that can be constructed relating $E(y)$ to these two independent variables.

1. The straight-line relationship between mean performance, $E(y)$, and engine speed is the same for all three fuels; that is, a single line will describe the relationship between $E(y)$ and speed, $x_1$, for all the fuel types (see Figure 12.20).

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 \\ x_1 &= \text{Engine speed} \end{aligned}$$

2. The straight lines relating mean performance, $E(y)$, to engine speed differ from one fuel to another, but the rate of increase in $E(y)$ per 1-rpm increase in speed, $x_1$,

**FIGURE 12.19**
Graphs of the relationship between mean performance, $E(y)$, and engine speed, $x_1$



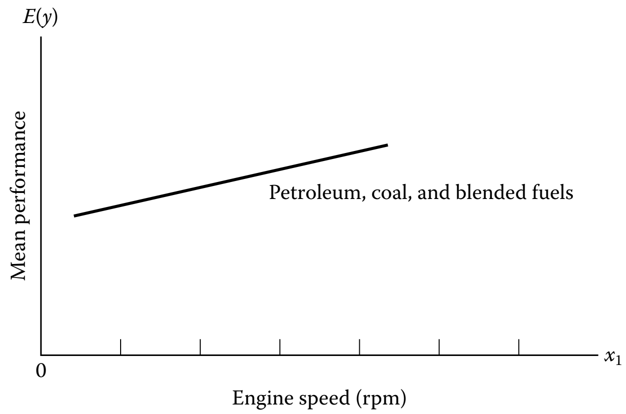



**FIGURE 12.20**
The relationship between $E(y)$ and $x_1$ is the same for all fuel types

is the same for all fuel types. That is, the lines are parallel but possess different $y$-intercepts (see Figure 12.21).

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$x_1 = \text{Engine speed}$$

$$x_2 = \begin{cases} 1 & \text{if petroleum fuel} \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if coal fuel} \\ 0 & \text{if not} \end{cases}$$

Notice that this model is essentially a combination of the bolded terms in the first-order model with a single quantitative variable and the model with a single qualitative variable:

| | |
|---|---|
| *First-order model with a single quantitative variable*: | $E(y) = \beta_0 + \boldsymbol{\beta_1 x_1}$ |
| *Model with a single qualitative variable at three levels*: | $E(y) = \beta_0 + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_3 x_3}$ |

where $x_1$, $x_2$, and $x_3$ are defined as above. The model described implies no interaction between the two independent variables, engine speed $x_1$ and the qualitative variable, type of fuel. The change in $E(y)$ for a 1-unit change in $x_1$ is identical (i.e., the slopes of the lines are equal) for all three fuel types. The terms

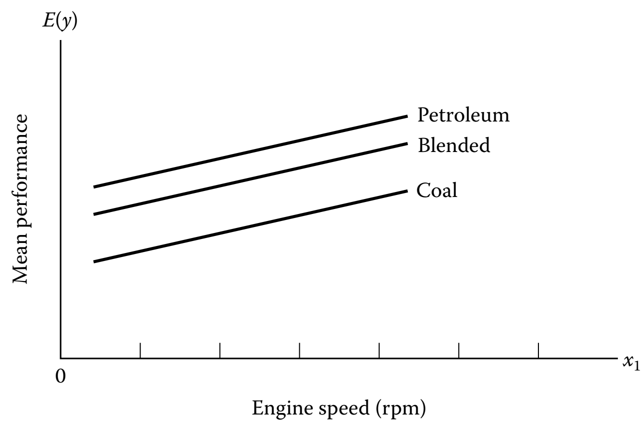



**FIGURE 12.21**
Parallel response lines for the three fuel types

corresponding to each of the independent variables are called **main effect terms** because they imply no interaction.

**Definition 12.4**
All noninteraction terms in a regression model involving a particular variable (quantitative or qualitative) represent the **main effect** of that independent variable on *y*.

3. The straight lines relating mean performance, $E(y)$, to engine speed, $x_1$, differ for the three fuel types; that is, the intercepts and slopes differ for the three lines (see Figure 12.22). As you will see, this interaction model is obtained by adding interaction terms (those involving the cross-product terms, one each from each of the two independent variables):

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1}^{\text{Main effect, engine speed}} + \overbrace{\beta_2 x_2 + \beta_3 x_3}^{\text{Main effect, type of fuel}} + \overbrace{\beta_4 x_1 x_2 + \beta_5 x_1 x_3}^{\text{Interaction}}$$

Note that each of the preceding models is obtained by adding terms to model 1, the single first-order model used to model the responses for all three fuels. Model 2 is obtained by adding the main effect terms for the qualitative variable, type of fuel; and model 3 is obtained by adding the interaction terms to model 2. Consequently, the models are *nested* (model 1 is nested within models 2 and 3; model 2 is nested within model 3). We learn how to compare nested models in Section 12.8.

Will a single line (Figure 12.20) characterize the responses for all three fuels, or do the three response lines differ as shown in Figure 12.22? A test of the null hypothesis that a single first-order model adequately describes the relationship between $E(y)$ and engine speed $x_1$ for all three fuels is a test of the null hypothesis that the parameters of model 3, $\beta_2$, $\beta_3$, $\beta_4$, and $\beta_5$, equal 0, i.e.,

$$H_0\colon \quad \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

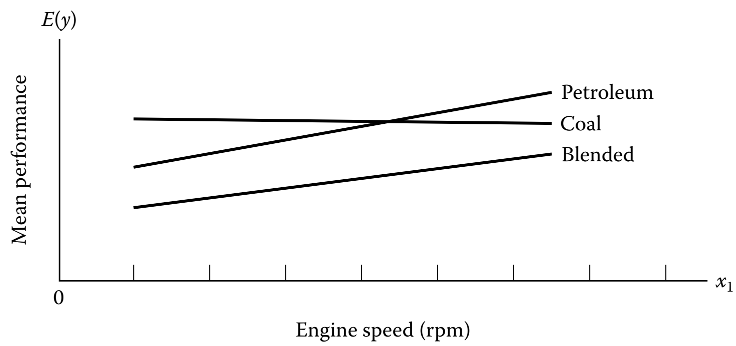
As we will see in the next section, this hypothesis is tested by comparing the complete model (model 3) to the reduced model (model 1).

Suppose we assume that the response lines for the three fuels will differ, but we wonder whether the data present sufficient evidence to indicate differences in the slopes of the lines. To test the null hypothesis that model 2 adequately describes the relationship between $E(y)$ and engine speed $x_1$, we want to test

$$H_0\colon \quad \beta_4 = \beta_5 = 0$$

that is, that the two independent variables, engine speed $x_1$ and the qualitative variable, type of fuel, do not interact. This test can be conducted by comparing the complete model (model 3) to the reduced model (model 2).

**FIGURE 12.22**
Different response lines for the three fuel types



**Example 12.6**

A Model with One Quantitative and One Qualitative Independent Variable

Substitute the appropriate values of the dummy variables in model 3 to obtain the equations of the three response lines in Figure 12.22.

Solution

The complete model that characterizes the three lines in Figure 12.22 is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

where

$$x_1 = \text{Engine speed}$$
$$x_2 = \begin{cases} 1 & \text{if petroleum fuel} \\ 0 & \text{if not} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{if coal fuel} \\ 0 & \text{if not} \end{cases}$$

Examining the coding, you can see that $x_2 = x_3 = 0$ when the fuel used is blended. Substituting these values into the expression for $E(y)$, we obtain the blended fuel line as:

*Blended fuel line*

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) \\ &= \beta_0 + \beta_1 x_1 \end{aligned}$$

Similarly, we substitute the appropriate values of $x_2$ and $x_3$ into the expression for $E(y)$ to obtain:

*Petroleum fuel line*

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0) \\ &= \overbrace{(\beta_0 + \beta_2)}^{y\text{-intercept}} + \overbrace{(\beta_1 + \beta_4)}^{\text{Slope}} x_1 \end{aligned}$$

*Coal fuel line*

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1) \\ &= \overbrace{(\beta_0 + \beta_3)}^{y\text{-intercept}} + \overbrace{(\beta_1 + \beta_5)}^{\text{Slope}} x_1 \end{aligned}$$

If you were to fit model 3, obtain estimates of $\beta_0, \beta_1, \ldots, \beta_5$, and substitute them into the equations for the three fuel-type lines shown above, you would obtain exactly the same prediction equations as you would obtain if you fit three separate straight lines, one to each of the three sets of fuel data. You may ask why we would not fit the three lines separately. Why fit a model (model 3) that combines all three lines into the same equation? The answer is that you need to use this procedure if you want to use statistical tests to compare the three fuel-type lines. We need to be able to express a practical question about the lines in terms of a hypothesis that each of a set of parameters in the model equals 0. You could not do this if you perform three separate regression analyses and fit a line to each set of fuel data.

CASTINGS

**TABLE 12.7 Productivity Data for Example 12.7**

| Type of Plant | Incentive 20¢/casting | | | 30¢/casting | | | 40¢/casting | | |
|---|---|---|---|---|---|---|---|---|---|
| Union | 1,435 | 1,512 | 1,491 | 1,583 | 1,529 | 1,610 | 1,601 | 1,574 | 1,636 |
| Nonunion | 1,575 | 1,512 | 1,488 | 1,635 | 1,589 | 1,661 | 1,645 | 1,616 | 1,689 |

## Example 12.7

Interaction in a Model for Productivity

An industrial engineer conducted an experiment to investigate the relationship between worker productivity and a measure of salary incentive for two manufacturing plants, one, A, with union representation and the other, B, with nonunion representation. The productivity, $y$, per worker was measured by recording the number of acceptable machined castings that a worker could produce in a 4-week, 40 hour-per-week period. The incentive was the amount, $x_1$, of bonus (in cents per casting) paid for all castings produced in excess of 1,000 per worker for the 4-week period. Nine workers were selected from each plant and three from each group of nine were assigned to receive a 20¢ bonus per casting, three a 30¢ bonus, and three a 40¢ bonus per casting. The productivity data for the 18 workers, three for each plant type and incentive combination, are shown in Table 12.7.

Assume that a first-order model* is adequate to detect a change in mean productivity, $E(y)$, as a function of incentive, $x_1$. The model that produces two productivity lines, one for each plant, is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where

$$x_1 = \text{Incentive}$$

$$x_2 = \begin{cases} 1 & \text{if nonunion plant} \\ 0 & \text{if union plant} \end{cases}$$

a. Fit the model to the data and graph the prediction equations for the two productivity lines.

b. Do the data provide sufficient evidence to indicate that the rate of increase of worker productivity with incentive is different for union and nonunion plants? Test at $\alpha = .10$.

Solution

a. The MINITAB printout for the regression analysis is shown in Figure 12.23. The prediction equation is obtained by reading the parameter estimates from the printout:

$$\hat{y} = 1{,}365.83 + 6.217x_1 + 47.78x_2 + .033x_1x_2$$

**FIGURE 12.23**
MINITAB printout of the complete model for the casting data

```
The regression equation is
CASTINGS = 1366 + 6.22 INCENTIVE + 47.8 PDUMMY + 0.03 INC_PDUM


Predictor     Coef  SE Coef      T      P
Constant   1365.83    51.84  26.35  0.000
INCENTIVE    6.217    1.667   3.73  0.002
PDUMMY       47.78    73.31   0.65  0.525
INC_PDUM     0.033    2.358   0.01  0.989


S = 40.8387   R-Sq = 71.1%   R-Sq(adj) = 64.9%


Analysis of Variance

Source          DF     SS     MS      F      P
Regression       3  57332  19111  11.46  0.000
Residual Error  14  23349   1668
Total           17  80682
```

*Although the model contains a term involving $x_1x_2$, it is first-order (graphs as a straight line) in the quantitative variable $x_1$. The variable $x_2$ is a dummy variable that introduces or deletes terms in the model. The order of a term is determined only by the quantitative variables that appear in the term.

The prediction equation for the union plant can be obtained by substituting $x_2 = 0$ into the general prediction equation. Then,

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(0) + \hat{\beta}_3 x_1(0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ &= 1{,}365.83 + 6.217x_1\end{aligned}$$

Similarly, the prediction equation for the nonunion plant is obtained by substituting $x_2 = 1$ into the general prediction equation. Then,

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(1) + \hat{\beta}_3 x_1(1) \\ &= \overbrace{(\hat{\beta}_0 + \hat{\beta}_2)}^{y\text{-intercept}} + \overbrace{(\hat{\beta}_1 + \hat{\beta}_3)}^{\text{Slope}} x_1 \\ &= (1{,}365.83 + 47.78) + (6.217 + .033)x_1 \\ &= 1{,}413.61 + 6.250x_1\end{aligned}$$

A MINITAB graph of these prediction equations is shown in Figure 12.24. Note that the slopes of the two lines are nearly identical (6.217 for union and 6.250 for nonunion).

b. If the rate of increase of productivity with incentive (i.e, the slope) for nonunion plants is different from the corresponding slope for union plants, then the interaction $\beta$ (i.e., $\beta_3$) will differ from 0. Consequently, we want to test

$H_0$: $\beta_3 = 0$ (no interaction)
$H_a$: $\beta_3 \neq 0$ (interaction)

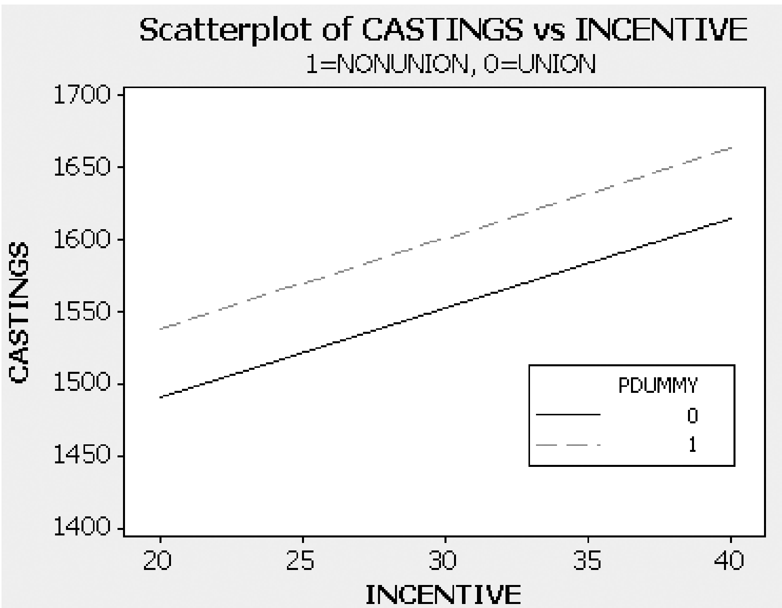



**FIGURE 12.24**
MINITAB plot of prediction equations

This test is conducted using the $T$ test. From the MINITAB printout, the test statistic and corresponding $p$-value (highlighted) are

$$T = .01 \qquad p\text{-value} = .989$$

Since $\alpha = .10$ is less than the $p$-value, we fail to reject $H_0$. There is insufficient evidence to conclude that the union and nonunion shapes differ. Thus, the test supports our observation of two nearly identical slopes in part b. Since interaction is not significant, we will drop the $x_1x_2$ term from the model and use the simpler model, $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2$, to predict productivity.

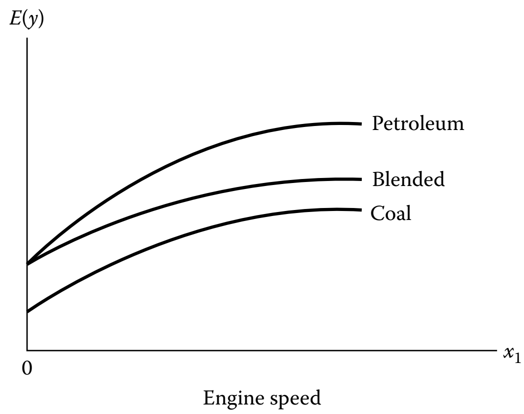
## Example 12.8

**A Second-order Model with One Quantitative and One Qualitative Independent Variable**

Refer to Example 12.6. Suppose we think that the relationship between mean diesel engine performance, $E(y)$, and engine speed, $x_1$, is second-order, as illustrated in Figure 12.25.

a. Write the equation of the model for $E(y)$ that yields the response curves shown in Figure 12.25.
b. Give the equation of the curve for petroleum fuel.
c. How would you determine whether or not the model, part **a**, gives a better prediction of $y$ than the first-order model of Example 12.6, $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3$?

**FIGURE 12.25**
The response curves for the three fuel types




Solution

a. Let

$$x_1 = \text{Engine speed}$$

$$x_2 = \begin{cases} 1 & \text{if petroleum fuel} \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if coal fuel} \\ 0 & \text{if not} \end{cases}$$

Since Figure 12.25 shows a curvilinear relationship between $E(y)$ and engine speed ($x_1$), and the response curves for the three fuel types are different (i.e., engine speed and type of fuel interact), the appropriate model is

$$E(y) = \beta_0 + \overbrace{\beta_1x_1 + \beta_2x_1^2}^{\substack{\text{Main effects,}\\ \text{engine}\\ \text{speed}}} + \overbrace{\beta_3x_2 + \beta_4x_3}^{\substack{\text{Main effects,}\\ \text{fuel type}}}$$

$$+ \underbrace{\beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1^2x_2 + \beta_8x_1^2x_3}_{\text{Interaction}}$$

Note that each of the main effect terms for engine speed ($x_1$ and $x_1^2$) is multiplied by each of the main effect terms for fuel type ($x_2$ and $x_3$) to obtain the four interaction terms in the model.

b. The model characterizes the relationship between $E(y)$ and $x_1$ for the petroleum fuel (see the coding) when $x_2 = 1$ and $x_3 = 0$. Substituting these values into the model, we obtain

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 \\ &\quad + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_3 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3(1) + \beta_4(0) + \beta_5 x_1(1) + \beta_6 x_1(0) \\ &\quad + \beta_7 x_1^2(1) + \beta_8 x_1^2(0) \\ &= \underbrace{(\beta_0 + \beta_3)}_{\text{y-intercept}} \quad + \quad \underbrace{(\beta_1 + \beta_5)x_1}_{\text{Shift}} \quad + \quad \underbrace{(\beta_2 + \beta_7)x_1^2}_{\text{Rate of curvature}} \end{aligned}$$

c. The only difference between the model, part **a**, and the first-order model of Example 12.6 are those terms involving $x_1^2$. Therefore, we want to test the null hypothesis, "the second-order terms contribute no information for the prediction of $y$," i.e.,

$$H_0: \quad \beta_2 = \beta_7 = \beta_8 = 0$$

Note that this test is neither a global $F$ test on all the $\beta$'s in the model nor a $T$ test on a single $\beta$. We learn how to conduct this test of a "partial" set of $\beta$'s in the next section.

The models described in the preceding sections provide only an introduction to statistical modeling. Models can be constructed to relate $E(y)$ to any number of quantitative and/or qualitative independent variables. You can compare response curves and surfaces for different levels of a qualitative variable or for different combinations of levels of two or more qualitative independent variables.

## Applied Exercises

12.44 *Chemical composition of rainwater.* Refer to the *Journal of Agricultural, Biological, and Environmental Statistics* (March 2005) study of the chemical composition of rainwater, Exercise 12.1 (p. 646). Recall that the nitrate concentration, $y$ (milligrams per liter), in a rainwater sample was modeled as a function of water source (groundwater, subsurface flow, or overground flow). Now consider adding a second independent variable, silica concentration (milligrams per liter), to the model.

a. Write a first-order model for $E(y)$ as a function of the independent variables. Assume that the rate of increase of nitrate concentration with silica concentration is the same for all three water sources. Sketch the relationships hypothesized by the model on a graph.

b. Write a first-order model for $E(y)$ as a function of the independent variables, but now assume that the rate of increase of nitrate concentration with silica concentration differs for the three water sources. Sketch the relationships hypothesized by the model on a graph.

12.45 *Sorption rates of organic vapors.* Refer to the *Environmental Science & Technology* study of sorption of organic vapors, Exercise 12.7 (p. 646). The independent variables used to model the retention coefficient $y$ are

$x_1$ = Temperature (degrees)

$x_2$ = Relative humidity (percent)

Organic compound = (benzene, toluene, chloroform, methanol, and anisole)

a. Write a first-order, main effects model for $E(y)$ as a function of temperature and organic compound. Draw a sketch of the model.

b. Interpret the $\beta$ parameters of the model, part **a**.

c. Write a model for $E(y)$ as a function of relative humidity and organic compound that hypothesizes different retention–relative humidity slopes for the five compounds. Draw a sketch of the model.

d. Give the slopes of the five compounds (in terms of the $\beta$'s) for the model, part **c**.

12.46 *Whales entangled in fishing gear.* Refer to the *Marine Mammal Science* (April 2010) study of whales entangled in fishing gear, Exercise 11.37 (p. 602). Now consider a model for the length ($y$) of an entangled whale (in meters) that is a function of water depth of the entanglement (in meters) and gear type (set nets, pots, or gill nets).

a. Write a main-effects only model for $E(y)$.
b. Sketch the relationships hypothesized by the model, part **a**. (Hint: Plot length on the vertical axis and water depth on the horizontal axis.)
c. Add terms to the model, part **a**, that includes interaction between water depth and gear type. (Hint: Be sure to interact each dummy variable for gear type with water depth.)
d. Sketch the relationships hypothesized by the model, part **c**.
e. In terms of the $\beta$'s in the model of part **c**, give the rate of change of whale length with water depth for set nets.
f. Repeat part **e** for pots.
g. Repeat part **e** for gill nets.
h. In terms of the $\beta$'s in the model of part **c**, how would you test to determine if the rate of change of whale length with water depth is the same for all three types of fishing gear?

12.47 *Muscle activity of harvesting foresters.* Refer to the *International Journal of Foresting Engineering* (Vol. 19, 2008) study of neck muscle activity patterns among forestry vehicle operators, Exercises 12.23 and 12.36 (p. 660, 671). Recall that the researchers identified the key explanatory variables of $y$ = the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. A list of the potential predictors is reproduced below:

$x_1$ = Age of operator (years)
$x_2$ = Duration of lunch break (minutes)
$x_3$ = Dominant hand power level (percentage)
$x_4$ = Perceived stress at work (5-point scale)
$x_5$ = {1 if married, 0 if not}
$x_6$ = {1 if day shift, 0 if night shift}
$x_7$ = {1 if operating a Timberjack vehicle, 0 if operating a Valmet vehicle}

a. Write the equation of a first-order, main effects model for $E(y)$ as a function of dominant hand power level ($x_3$) and vehicle type ($x_7$).
b. Interpret, in the words of the problem, the value of $\beta_2$ in the model, part **a**.
c. Add term(s) to the model, part **a**, that allow for interaction between dominant hand power level and vehicle type. Sketch the relationships hypothesized by this model.
d. What function of the $\beta$'s in the model, part **c**, represents the change in $E(y)$ for every 1 percent increase in dominant hand power when operating a Valmet vehicle?
e. What function of the $\beta$'s in the model, part **c**, represents the difference in $E(y)$ values between the Timberjack operators and Valmet operators who have a dominant hand power level of 75% ?
f. Write the equation of a complete second-order model for $E(y)$ as a function of dominant hand power level ($x_3$) and vehicle type ($x_7$). Sketch the relationships hypothesized by this model.
g. What function of the $\beta$'s in the model, part **f**, represents the rate of curvature between $E(y)$ and dominant hand power when operating a Valmet vehicle?

12.48 *RNA analysis of wheat genes.* Engineers from the Department of Crop and Soil Sciences at Washington State University used regression to estimate the number of copies of a gene transcript in an aliquot of RNA extracted from a wheat plant. (*Electronic Journal of Biotechnology*, April 15, 2004.) The proportion ($x_1$) of RNA extracted from a cold-exposed wheat plant was varied, and the transcript copy number ($y$, in thousands) was measured for each of two cloned genes: Mn Superoxide Dismutose (MnSOD) and Phospholipose D (PLD). The data are listed in the accompanying table.

a. Write a first-order model for number of copies ($y$) as a function of proportion ($x_1$) of RNA extracted and the gene type (MnSOD or PLD). Assume that proportion of RNA and gene type interact to effect $y$.
b. Fit the model, part **a**, to the data. Give the least-squares prediction equation for $y$.
c. Conduct a test to determine if, in fact, proportion of RNA and gene type interact. Test using $\alpha = .01$.

**WHEATRNA**

| RNA Proportion ($x_1$) | Number of copies ($y$, thousands) MnSOD | PLD |
|---|---|---|
| 0.00 | 401 | 80 |
| 0.00 | 336 | 83 |
| 0.00 | 337 | 75 |
| 0.33 | 711 | 132 |
| 0.33 | 637 | 148 |
| 0.33 | 602 | 115 |
| 0.50 | 985 | 147 |
| 0.50 | 650 | 142 |
| 0.50 | 747 | 146 |
| 0.67 | 904 | 146 |
| 0.67 | 1007 | 150 |
| 0.67 | 1047 | 184 |
| 0.80 | 1151 | 173 |
| 0.80 | 1098 | 201 |
| 0.80 | 1061 | 181 |
| 1.00 | 1261 | 193 |
| 1.00 | 1272 | 187 |
| 1.00 | 1256 | 199 |

*Source*: Baek, K. H., and Skinner, D. Z. "Quantitative real-time PCR method to detect changes in specific transcript and total RNA amounts." *Electronic Journal of Biotechnology*, Vol. 7, No. 1, April 15, 2004 (adapted from Figure 2).

d. Use the results, part **b**, to estimate the rate of increase of number of copies ($y$) with proportion ($x_1$) of RNA extracted for the MnSOD gene type.
e. Repeat part **d** for the PLD gene type.

12.49 *Shelf life of a pharmaceutical.* Eli Lilly and Company has developed three methods (G, $R_1$, and $R_2$) for estimating the shelf life of its drug products based on the potency of the drug.* One way to compare the three methods is to build a regression model for the dependent variable, estimated shelf life $y$ (as a percent of true shelf life), with potency of the drug ($x_1$) as a quantitative predictor and method as a qualitative predictor.

a. Write a first-order, main effects model for $E(y)$ as a function of potency ($x_1$) and method.
b. Interpret the $\beta$ coefficients of the model, part **a**.
c. Write a first-order model for $E(y)$ that will allow the slopes to differ for the three methods.
d. Refer to part **c**. For each method, write the slope of the $y$–$x_1$ line in terms of the $\beta$'s.

12.50 *Evaluating Web browser graphics.* An experiment was conducted to compare four Web browsers on their ability to display graphics. (*Journal of Graphic Engineering Design,* Vol. 3, 2012.) The four browsers were the latest versions of Google Chrome, Mozilla Firefox, Opera and Apple Safari. Each browser was tested on its ability to generate 50, 250, 500, and 750 simple objects (graphs). For each of these four tasks, the average completion time (in hundredths of a second) was determined. The data (simulated from information provided in the journal article) are provided in the accompanying table.

a. Hypothesize a first-order, interaction model for completion time ($y$) as a function of number of objects ($x_1$) and browser type.
b. Fit the model, part **a**, to the data. Conduct a test of overall model adequacy (at $\alpha = .05$).
c. Give an estimate of the rate of change of completion time with number of objects when using the Chrome browser.

**BROWSER**

| Task | Completion Time (hundredths of a second) | Number of Objects | Browser |
|---|---|---|---|
| 1 | 3.0 | 50 | Chrome |
| 2 | 6.0 | 50 | Firefox |
| 3 | 2.0 | 50 | Safari |
| 4 | 3.5 | 50 | Opera |
| 5 | 5.0 | 250 | Chrome |
| 6 | 7.0 | 250 | Firefox |
| 7 | 4.0 | 250 | Safari |
| 8 | 10.0 | 250 | Opera |
| 9 | 8.0 | 500 | Chrome |
| 10 | 7.5 | 500 | Firefox |
| 11 | 7.0 | 500 | Safari |
| 12 | 17.5 | 500 | Opera |
| 13 | 13.0 | 750 | Chrome |
| 14 | 8.0 | 750 | Firefox |
| 15 | 12.0 | 750 | Safari |
| 16 | 22.0 | 750 | Opera |

12.51 *Greenhouse gas emissions.* Refer to the *Journal of the Institute of Engineering* (Vol. 8, 2011) study of methane gas (milligrams per liter) emitted from wastewater treatment

**SLUDGE**

| Specimen | Methane($CH_4$) | Time | Nutrients(VHA) |
|---|---|---|---|
| 1 | 5 | 20 | No |
| 2 | 9 | 21 | No |
| 3 | 18 | 24 | No |
| 4 | 35 | 26 | No |
| 5 | 61 | 29 | No |
| 6 | 65 | 32 | No |
| 7 | 105 | 35 | No |
| 8 | 120 | 37 | No |
| 9 | 117 | 42 | No |
| 10 | 154 | 44 | No |
| 11 | 200 | 47 | No |
| 12 | 198 | 49 | No |
| 13 | 203 | 51 | No |
| 14 | 21 | 20 | Yes |
| 15 | 25 | 21 | Yes |
| 16 | 61 | 24 | Yes |
| 17 | 75 | 26 | Yes |
| 18 | 102 | 29 | Yes |
| 19 | 150 | 32 | Yes |
| 20 | 183 | 34 | Yes |
| 21 | 194 | 36 | Yes |
| 22 | 245 | 37 | Yes |
| 23 | 308 | 42 | Yes |
| 24 | 295 | 44 | Yes |
| 25 | 272 | 47 | Yes |
| 26 | 280 | 49 | Yes |
| 27 | 287 | 51 | Yes |

*Source:* Devkota, R.P. "Greenhouse Gas Emissions from Wastewater Treatment System", *Journal of the Institute of Engineering*, Vol. 8, No. 1, 2011 (adapted from Figure 4).

*Murphy, J. R., and Weisman, D. "Using Random Slopes for Estimating Shelf Life." Paper presented at Joint Statistical Meetings, Anaheim, Calif., Aug. 1990.

sludge, Exercise 12.42 (p. 673). Recall that two different types of treated sludge were investigated: (1) sludge without nutrients added and (2) sludge with nutrients (VHA) added. Data for $n = 27$ sludge specimens are reproduced in the table on p. 683. Note that the table includes the variable, length of time (in days) the sludge was processed The researcher estimated the *emission rate* (i.e., the rate of increase in emitted methane gas for each additional day of treatment) for both untreated and treated sludge. Use regression to determine if the emission rates differ for the two types of sludge, and if so, provide an estimate of each emission rate.

12.52 *Study of lead in moss.* A study of the atmospheric pollution on the slopes of the Blue Ridge Mountains (Tennessee) was conducted. The file **LEADMOSS** contains the levels of lead found in 70 fern moss specimens (in micrograms of lead per gram of moss tissue) collected from the mountain slopes, as well as the elevation of the moss specimen (in feet) and the direction (1 if east, 0 if west) of the slope face. The first five and last five observations of the data set are listed in the table.

**LEADMOSS**

| Specimen | Lead Level | Elevation | Slope Face |
|---|---|---|---|
| 1 | 3.475 | 2000 | 0 |
| 2 | 3.359 | 2000 | 0 |
| 3 | 3.877 | 2000 | 0 |
| 4 | 4.000 | 2500 | 0 |
| 5 | 3.618 | 2500 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 66 | 5.413 | 2500 | 1 |
| 67 | 7.181 | 2500 | 1 |
| 68 | 6.589 | 2500 | 1 |
| 69 | 6.182 | 2000 | 1 |
| 70 | 3.706 | 2000 | 1 |

*Source*: Schilling, J. "Bioindication of atmospheric heavy metal deposition in the Blue Ridge using the moss, *Thuidium delicatulum*." Master of Science Thesis, Spring 2000.

a. Write the equation of a first-order model relating mean lead level, $E(y)$, to elevation ($x_1$) and slope face ($x_2$). Include interaction between elevation and slope face in the model.
b. Graph the relationship between mean lead level and elevation for the different slope faces that is hypothesized by the model, part **a**.
c. In terms of the $\beta$'s of the model, part **a**, give the change in lead level for every 1 foot increase in elevation for moss specimens on the east slope.
d. Fit the model, part **a**, to the data using an available statistical software package. Is the overall model statistically useful for predicting lead level? Test using $\alpha = .10$.
e. Write the equation of the complete second-order model relating mean lead level, $E(y)$, to elevation ($x_1$) and slope face ($x_2$).

12.53 *Storing nuclear waste in glass.* Since glass is not prone to radiation damage, encapsulation of waste in glass is considered to be one of the most promising solutions to the problem of low-level nuclear waste in the environment. However, glass undergoes chemical changes when exposed to extreme environmental conditions and certain of its constituents leach into the surroundings. In addition, these chemical reactions may possibly weaken the glass. These concerns led to a study undertaken jointly by the Department of Materials Science and Engineering at the University of Florida and the U.S. Department of Energy to assess the utility of glass as a waste encapsulant material.* Corrosive chemical solutions (called corrosion baths) were prepared and applied directly to glass samples containing one of three types of waste (TDS-3A, FE, and AL); the chemical reactions were observed over time. A few of the key variables measured were

$y =$ Amount of silicon (in parts per million) found in solution at end of experiment. (This is both a measure of the degree of breakdown in the glass and a proxy for the amount of radioactive species released into the environment.)

$x_1 =$ Temperature (°C) of the corrosion bath

$$x_2 = \begin{cases} 1 & \text{if waste type TDS-3A} \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if waste type FE} \\ 0 & \text{if not} \end{cases}$$

Waste type AL is the base level. Suppose we want to model amount $y$ of silicon as a function of temperature ($x_1$) and type of waste ($x_2, x_3$).

a. Write a model that proposes parallel straight-line relationships between amount of silicon and temperature, one line for each of the three waste types.
b. Add terms for the interaction between temperature and waste type to the model of part **a**.
c. Refer to the model of part **b**. For each waste type, give the slope of the line relating amount of silicon to temperature.
d. Give the null hypothesis in a test for the presence of temperature–waste type interaction.

*The background information for this exercise was provided by Dr. David Clark, Department of Materials Science and Engineering, University of Florida.

## 12.8 Tests for Comparing Nested Models

In regression analysis, we often want to determine (with a high degree of confidence) which one among a set of candidate models best fits the data. In this section, we present such a method for **nested models**.

**Definition 12.5**

Two models are **nested** if one model contains all the terms of the second model and at least one additional term.

To illustrate, suppose you have collected data on a response, $y$, and two quantitative independent variables, $x_1$ and $x_2$, and you are considering the use of either a first-order or a second-order model to relate $E(y)$ to $x_1$ and $x_2$. Will the second-order model provide better predictions of $y$ than the first-order model? To answer this question, examine the two models, and note that the second-order model contains all terms contained in the first-order model plus three additional terms—those involving $\beta_3$, $\beta_4$, and $\beta_5$:

*First-order model*: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

*Second-order model*:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \overbrace{\beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2}^{\text{Second-order terms}}$$

Consequently, these are nested models. Since the first-order model is the simpler of the two, we say that the *first-order model is nested within the more complex second-order model*.

In general, the more complex of two nested models is called the **complete** (or **full**) **model**, and the simpler of the two is called the **reduced model**. Asking whether the second-order (or *complete*) model contributes more information for the prediction of $y$ than the first-order (or *reduced*) model is equivalent to asking whether at least one of the parameters, $\beta_3$, $\beta_4$, or $\beta_5$, differs from 0—i.e., whether the terms involving $\beta_3$, $\beta_4$, and $\beta_5$ should be retained in the model. Therefore, to test whether the second-order terms should be included in the model, we test the null hypothesis

$$H_0: \quad \beta_3 = \beta_4 = \beta_5 = 0$$

(i.e., the second-order terms do not contribute information for the prediction of $y$) against the alternative hypothesis

$H_a$: At least one of the parameters, $\beta_3$, $\beta_4$, or $\beta_5$, differs from 0

(i.e., at least one of the second-order terms contributes information for the prediction of $y$).

The procedure for conducting this test is intuitive: First, we use the method of least squares to fit the reduced model and calculate the corresponding sum of squares for error, $\text{SSE}_\text{R}$ (the sum of squares of the deviations between observed and predicted $y$ values). Next, we fit the complete model and calculate its sum of squares for error, $\text{SSE}_\text{C}$. Then, we compare $\text{SSE}_\text{R}$ to $\text{SSE}_\text{C}$ by calculating the difference $\text{SSE}_\text{R} - \text{SSE}_\text{C}$. If the second-order terms contribute to the model, then $\text{SSE}_\text{C}$ should be much smaller than $\text{SSE}_\text{R}$, and the difference $\text{SSE}_\text{R} - \text{SSE}_\text{C}$ will be large. The larger the difference, the greater the weight of evidence that the complete model provides better predictions of $y$ than does the reduced model.

The sum of squares for error will always decrease when new terms are added to the model. The question is whether this decrease is large enough to conclude that it is due to more than just an increase in the number of model terms and to chance. To test the null hypothesis that the parameters of the second-order terms, $\beta_3$, $\beta_4$, and $\beta_5$, simultaneously equal 0, we use an $F$ statistic calculated as follows:

$$F = \frac{\text{Drop in SSE/Number of } \beta \text{ parameters being tested}}{s^2 \text{ for the second-order model}}$$
$$= \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/3}{\text{SSE}_\text{C}/[n - (5 + 1)]}$$

When the standard regression assumptions about the error term $\varepsilon$ are satisfied and the $\beta$ parameters for the second-order terms are all 0 (i.e., $H_0$ is true), this $F$ statistic has an $F$ distribution with $\nu_1 = 3$ and $\nu_2 = n - 6$ degrees of freedom. Note that $\nu_1$ is the number of $\beta$ parameters being tested and $\nu_2$ is the number of degrees of freedom associated with $s^2$ in the second-order model.

If the second-order terms *do* contribute to the model (i.e., $H_\text{a}$ is true), we expect the $F$ statistic to be large. Thus, we use a one-tailed test and reject $H_0$ if $F$ exceeds some critical value, $F_\alpha$.

### *F* Test for Comparing Nested Models

*Reduced model*: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$

*Complete model*: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k$

$H_0$: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

$H_\text{a}$: At least one of the $\beta$ parameters under test is nonzero

*Test statistic*: $$F_c = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(k - g)}{\text{SSE}_\text{C}/[n - (k + 1)]} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/\text{\# of } \beta\text{'s tested in } H_0}{\text{MSE}_\text{C}}$$

where

$\text{SSE}_\text{R}$ = Sum of squared errors for the reduced model
$\text{SSE}_\text{C}$ = Sum of squared errors for the complete model
$\text{MSE}_\text{C}$ = Mean square error ($s^2$) for the complete model
$k - g$ = Number of $\beta$ parameters specified in $H_0$ (i.e., number of $\beta$ parameters tested)
$k + 1$ = Number of $\beta$ parameters in the complete model (including $\beta_0$)
$n$ = Total sample size

*Rejection region*: $F_c > F_\alpha$
*p*-value: $P(F > F_c)$
where $F$ is based on $\nu_1 = k - g$ numerator degrees of freedom and $\nu_2 = n - (k + 1)$ denominator degrees of freedom

**Example 12.9**
Comparing Nested Models

Refer to the second-order model relating temperature ($x_1$) and pressure ($x_2$) to the quality ($y$) of a finished product, Example 12.3. (p. 657). Do the data provide sufficient evidence to indicate that the quadratic terms, $x_1^2$ and $x_2^2$, contribute information for the prediction of $y$? Test at $\alpha = .05$.

Solution

To determine whether the quadratic (i.e., curvilinear) terms contribute information for the prediction of $y$, we test

$H_0$: $\beta_4 = \beta_5 = 0$

against the alternative hypothesis

$H_\text{a}$: At least one of the parameters, $\beta_4$ or $\beta_5$, differs from 0

```
                              The REG Procedure
                                Model: MODEL1
                        Dependent Variable: QUALITY

                  Number of Observations Read          27
                  Number of Observations Used          27

                            Analysis of Variance

                                      Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       3     2425.04194      808.34731       3.08    0.0475
Error                      23     6036.40102      262.45222
Corrected Total            26     8461.44296

             Root MSE               16.20038    R-Square     0.2866
             Dependent Mean         66.96296    Adj R-Sq     0.1935
             Coeff Var              24.19304

                            Parameter Estimates

                               Parameter       Standard
     Variable        DF         Estimate          Error    t Value    Pr > |t|

     Intercept        1       -614.13981      466.16828      -1.32      0.2007
     TEMP             1          7.08639        5.15846       1.37      0.1828
     PRESSURE         1         13.88278        8.45253       1.64      0.1141
     TEMP_PRESS       1         -0.14550        0.09353      -1.56      0.1335
```

**FIGURE 12.26**
SAS printout for reduced model

The nested models to be compared are:

$$\textit{Completed model:}\quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$
$$\textit{Reduced model:}\quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

The SAS printout for the reduced model is shown in Figure 12.26.

The sum of squares for error for the reduced model (shaded in Figure 12.26) is

$$\text{SSE}_\text{R} = 6{,}036.40102$$

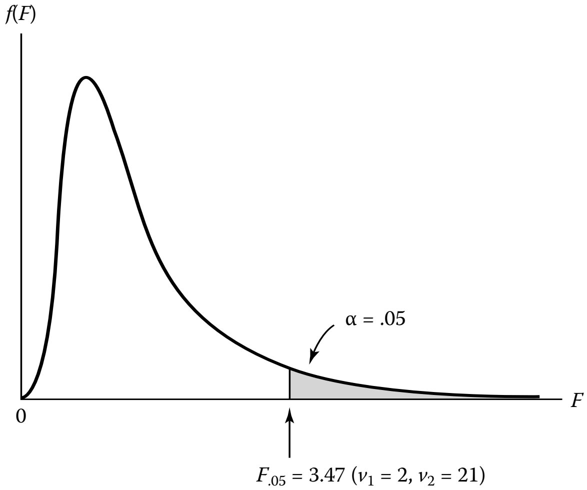
To conduct the test, we also need the SSE and MSE for the complete model. These values, shown in Figure 12.13 (p. 658), are

$$\text{SSE}_\text{C} = 59.17843$$
$$\text{MSE}_\text{C} = 2.81802$$

For this test, $n = 27$, $k = 5$, $g = 3$, and the number of $\beta$'s tested is $(k - g) = 2$. Therefore, the calculated value of the $F$ statistic, based on $\nu_1 = k - g = 2$ and $\nu_2 = n - (k + 1) = 21$ degrees of freedom is

$$\textit{Test statistic:}\quad F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/\#\beta\text{'s tested in model}}{\text{MSE}_\text{C}}$$
$$= \frac{(6{,}036.40102 - 59.17843)/2}{2.81802} = 1{,}060.5$$

**FIGURE 12.27**
Rejection region for the $F$ test $H_0$: $\beta_4 = \beta_5 = 0$




The final step in the test is to compare this computed value of $F$ with the tabulated value based on $\nu_1 = 2$ and $\nu_2 = 21$ degrees of freedom. For $\alpha = .05$, $F_{.05} = 3.47$. Then the rejection region is

*Rejection region*: $F > 3.47$ (see Figure 12.27)

Since the computed value of $F$ falls in the rejection region—i.e., it exceeds $F_{.05} = 3.47$—we reject $H_0$ and conclude (at $\alpha = .05$) that at least one of the quadratic terms contributes information for the prediction of $y$. In other words, the data support the contention that the curvature we see in the response surface is not due simply to random variation in the data. The complete second-order model appears to provide better predictions of $y$ than does the reduced model.

(*Note*: The test statistic and $p$-value for this nested model $F$ test can be obtained using statistical software. These values are highlighted on the SAS printout, Figure 12.28. Since the $p$-value is less than $\alpha = .05$, we arrive at the same conclusion: Reject $H_0$.)

Test CURV Results for Dependent Variable QUALITY

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 2 | 2988.61130 | 1060.54 | <.0001 |
| Denominator | 21 | 2.81802 | | |

**FIGURE 12.28**
SAS printout for nested-model $F$ test

Suppose the $F$ test in Example 12.9 yielded a test statistic that did not fall in the rejection region. That is, suppose there was insufficient evidence (at $\alpha = .05$) to say that the curvature terms contribute information for the prediction of product quality. As with any statistical test of hypothesis, we must be cautious about accepting $H_0$ since the probability of a Type II error is unknown. Nevertheless, most practitioners of regression analysis adopt the principle of **parsimony**. That is, in situations where two competing models are found to have essentially the same predictive power, the model with the fewest number of $\beta$'s (i.e., the more **parsimonious model**) is selected. The principle of parsimony would lead us to choose the simpler (reduced) model over the more complex complete model when we fail to reject $H_0$ in the $F$ test for nested models.

**Definition 12.6**

A **parsimonious model** is a general linear model with a small number of $\beta$ parameters. In situations where two competing models have essentially the same predictive power (as determined by an $F$ test), choose the more parsimonious of the two.

When the candidate models in model building are nested models, the $F$ test developed in this section is the appropriate procedure to apply to compare the models. However, if the models are not nested, this $F$ test is not applicable. In this situation, the analyst must base the choice of the best model on statistics such as $R_a^2$ and $s$. It is important to remember that decisions based on these and other numerical descriptive measures of model adequacy cannot be supported with a measure of reliability and are often very subjective in nature.

## Applied Exercises

12.54 *Muscle activity of harvesting foresters.* Refer to the *International Journal of Foresting Engineering* (Vol. 19, 2008) study of neck muscle activity patterns among forestry vehicle operators, Exercise 12.23 (p. 660). Recall that the researchers identified the key explanatory variables of $y$ = the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. A list of the potential predictors is reproduced below:

$x_1$ = Age of operator (years)

$x_2$ = Duration of lunch break (minutes)

$x_3$ = Dominant hand power level (percentage)

$x_4$ = Perceived stress at work (5-point scale)

$x_5$ = {1 if married, 0 if not}

$x_6$ = {1 if day shift, 0 if night shift}

$x_7$ = {1 if operating a Timberjack vehicle, 0 if operating a Valmet vehicle}

a. Write a model for $E(y)$ as a function of the seven independent variables that (1) hypothesizes straight-line relationships between each quantitative $x$ and $y$, and (2) allows for interactions among all possible pairs of independent variables.

b. Refer to the model, part **a**. Specify the null hypothesis to test to determine if the impact of any of the four qualitative independent variables on $E(y)$ depends on the levels of the three qualitative variables.

c. How would you conduct the test, part **b**. Identify the reduced model as part of your answer.

d. The researchers collected data for $n = 13$ forestry vehicle operators. Is this sufficient to fit the model, part **a**, and carry out the test, part **c**? Explain.

12.55 *Whales entangled in fishing gear.* Refer to the *Marine Mammal Science* (April 2010) study of whales entangled in fishing gear, Exercise 12.37 (p. 671). A first-order model for the length ($y$) of an entangled whale that is a function of water depth of the entanglement ($x_1$) and gear type (set nets, pots, or gill nets) is written as follows: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$, where $x_2$ = {1 if set net, 0 if not} and $x_2$ = {1 if pot, 0 if not}. Consider this model the complete model in a nested model $F$ test.

a. Suppose you want to determine if there are any differences in the mean lengths of entangled whales for the three gear types. Give the appropriate null hypothesis to test.

b. Refer to part **a**. Give the reduced model for the test.

c. Refer to parts **a** and **b**. If you reject the null hypothesis, what would you conclude?

d. Suppose you want to determine if the rate of change of whale length ($y$) with water depth ($x_1$) is the same for all three types of fishing gear. Give the appropriate null hypothesis to test.

e. Refer to part **d**. Give the reduced model for the test.

f. Refer to parts **d** and **e**. If you fail to reject the null hypothesis, what would you conclude?

**SLUDGE**

12.56 *Greenhouse gas emissions.* Refer to the *Journal of the Institute of Engineering* (Vol. 8, 2011) study of methane gas (milligrams per liter) emitted from wastewater treatment sludge, Exercises 12.42 and 12.51 (p. 673, 683). Recall that two different types of treated sludge were investigated: (1) sludge without nutrients added and (2) sludge with nutrients (VHA) added. Data for $n = 27$ sludge specimens were collected on the variables methane gas emitted ($y$), treatment time ($x_1$), and sludge type ($x_2 = 1$ if VHA added, 0 if not).

a. Write a complete 2nd-order model for $E(y)$ as a function of $x_1$ and $x_2$.

b. Specify the null hypothesis to test in order to determine whether curvature exists in the relationship between methane gas emission and treatment time.

c. Give the equation of the reduced model to be compared to the complete model, part **a**, in order to conduct the test, part **b**.

d. A SAS printout of the analysis is shown on the next page. Locate the $p$-value of the test for curvature on the printout. Interpret the results.

## SAS Output for Exercise 12.56

```
                         The REG Procedure
                           Model: MODEL1
                     Dependent Variable: CH4

          Number of Observations Read          27
          Number of Observations Used          27

                       Analysis of Variance

                                  Sum of         Mean
Source                DF         Squares       Square    F Value    Pr > F

Model                  5          244222        48844     123.76    <.0001
Error                 21      8287.79860    394.65708
Corrected Total       26          252510

        Root MSE             19.86598    R-Square     0.9672
        Dependent Mean      140.29630    Adj R-Sq     0.9594
        Coeff Var            14.16002

                       Parameter Estimates

                     Parameter       Standard
    Variable    DF    Estimate          Error    t Value    Pr > |t|

    Intercept    1   -95.38759       76.49017      -1.25      0.2261
    TIME         1     4.14559        4.60805       0.90      0.3785
    VHA          1  -331.58792      105.95577      -3.13      0.0051
    TIME_VHA     1    21.74633        6.35646       3.42      0.0026
    TIMESQ       1     0.03639        0.06472       0.56      0.5799
    TIMESQ_VHA   1    -0.26449        0.08924      -2.96      0.0074


     Test CURVE Results for Dependent Variable CH4

                                 Mean
Source             DF          Square    F Value    Pr > F

Numerator           2      2781.83832       7.05    0.0045
Denominator        21       394.65708
```

12.57 *Students' ability in science.* The *American Educational Research Journal* (Fall 1998) published a study of students' perceptions of their science ability in hands-on classrooms. A first-order, main effects model that was used to predict ability perception ($y$) included the following independent variables:

*Control Variables*

$x_1$ = Prior science attitude score

$x_2$ = Science ability test score

$x_3$ = 1 if boy, 0 if girl

$x_4$ = 1 if classroom 1 student, 0 if not

$x_5$ = 1 if classroom 3 student, 0 if not

$x_6$ = 1 if classroom 4 student, 0 if not

$x_7$ = 1 if classroom 5 student, 0 if not

$x_8$ = 1 if classroom 6 student, 0 if not

*Performance Behaviors*

$x_9$ = Active-leading behavior score

$x_{10}$ = Passive-assisting behavior score

$x_{11}$ = Active-manipulating behavior score

a. Hypothesize the equation of the first-order, main effects model for $E(y)$.

b. The researchers also considered a model that included all possible interactions between the control variables and the performance behavior variables. Write the equation for this model for $E(y)$.

c. The researchers determined that the interaction terms in the model, part **b**, were not significant, and therefore used the model, part **a**, to make inferences. Explain the best way to conduct this test for interaction. Give the null hypothesis of the test.

**WATEROIL**

12.58 *Extracting water from oil.* Refer to the *Journal of Colloid and Interface Science* study of water/oil mixtures, Exercise 11.27 (p. 591). Recall that three of the seven variables used to predict voltage ($y$) were volume ($x_1$), salinity ($x_2$), and surfactant concentration ($x_5$). The model the researchers fit is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_5 + \beta_4 x_1 x_2 + \beta_5 x_1 x_5$$

a. Note that the model includes interaction between disperse phase volume ($x_1$) and salinity ($x_2$) as well as interaction between disperse phase volume ($x_1$) and surfactant concentration ($x_5$). Discuss how these interaction terms affect the hypothetical relationship between $y$ and $x_1$. Draw a sketch to support your answer.
b. Fit the interaction model to the data. Does this model appear to fit the data better than the first-order model in Exercise 11.27? Explain.
c. Interpret the $\beta$ estimates of the interaction model.
d. Conduct a test to determine whether the interaction terms contribute significantly to the prediction of voltage ($y$). Use $\alpha = .05$.

12.59 *Seismic wave study.* Refer to Exercise 12.25 (p. 660), in which an exploration seismologist wants to develop a regression model for estimating the mean signal-to-noise ratio of seismic waves from earthquakes. The model under consideration is a complete second-order model:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

where

$y$ = Signal-to-noise ratio
$x_1$ = Frequency of wavelet
$x_2$ = Amplitude of wavelet

Both the complete model and the reduced model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, were fit to $n = 12$ data points, with the following results: $SSE_C = 159.94$, $MSE_C = 26.66$, $SSE_R = 2094.4$, $MSE_R = 232.7$. Compare the two models using a nested model $F$ test at $\alpha = .05$. What do you conclude?

12.60 *Speech recognition device.* Refer to the *Human Factors* study of the performance of a computerized speech recognizer, Exercise 12.26 (p. 661). Recall that the researchers built a complete second-order model for task completion time ($y$) as a function of accuracy ($x_1$) and vocabulary ($x_2$).

a. Give the null hypothesis for testing whether the quadratic terms in the model are useful predictors of $y$.
b. The test, part **a**, resulted in a $p$-value of less than .01. Interpret this result.

12.61 *Emotional stress of firefighters.* Refer to the *Journal of Human Stress* study of firefighters, Exercise 12.5 (p. 646). It is thought that a complete second-order model, shown here, will be adequate to describe the relationship between emotional distress and years of experience for two groups of firefighters—those exposed to a chemical fire and those unexposed.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2$$

where

$y$ = Emotional distress
$x_1$ = Experience (years)

$$x_2 = \begin{cases} 1 & \text{if exposed to chemical fire} \\ 0 & \text{if not} \end{cases}$$

a. What hypothesis would you test to determine whether the *rate* of increase of emotional distress with experience is different for the two groups of firefighters?
b. What hypothesis would you test to determine whether there are differences in mean emotional distress levels that are attributable to exposure group?
c. The second-order model was fit to data collected for a sample of 200 firefighters, resulting in SSE = 783.90. The reduced model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$, is fit to the same data, resulting in SSE = 795.23. Is there sufficient evidence to support the claim that the mean emotional distress levels differ for the two groups of firefighters? Use $\alpha = .05$.

12.62 *Sorption rate of organic vapors.* Refer to the *Environmental Science & Technology* study of sorption of organic vapors, Exercise 12.45 (p. 681). Consider using the quantitative variable, relative humidity, and the qualitative variable, organic compound (at five levels), to model the retention coefficient $y$.

a. Write a complete second-order model that relates $E(y)$ to relative humidity and organic compound.
b. Under what circumstances will the response curves of the model of part **a** possess the same shape but have different $y$-intercepts?
c. Under what circumstances will the response curves of the model of part **a** be parallel lines?
d. Under what circumstances will the response curves of the model of part **a** be identical?

12.63 *Concrete strength experiment.* A building materials engineer is experimenting with three different cement mixes—dry, damp, and wet—for laying concrete. Since the compressive strength of a concrete slab varies as a function of hardening time and cement mix, the following main effects model is proposed:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where

$y$ = Compressive strength (thousands of pounds per square inch)
$x_1$ = Hardening time of cement mix (days)

$$x_2 = \begin{cases} 1 & \text{if damp cement} \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if wet cement} \\ 0 & \text{if not} \end{cases}$$

Dry cement is the base level.

a. What hypothesis would you test to determine whether mean compressive strength differs for the three cement mixes?
b. Using data collected for a sample of 50 batches of concrete, the main effects model is fit, with the result SSE = 140.5. Then the reduced model $E(y) = \beta_0 + \beta_1 x_1$ is fit to the same data, with the result SSE = 183.2. Test the hypothesis you formulated in part **a**. Use $\alpha = .05$.
c. Explain how you would test the hypothesis that the slope of the linear relationship between mean compressive strength $E(y)$ and hardening time $x_1$ varies according to type of cement mix.
d. Write a second-order model that allows different response curves for the three types of cement mixes.
e. Explain how you would test the hypothesis that the three response curves have the same shape, but different $y$-intercepts.

12.64 *Modeling machine downtime.* An operations manager is interested in modeling $E(y)$, the expected length of time per month (in hours) that a machine will be shut down for repairs, as a function of the type of machine (001 or 002) and the age of the machine (in years).

a. Write a complete second-order model for machine downtime ($y$) as a function of age and machine type.
b. Give the reduced model required to determine whether the second-order terms in the model are necessary.
c. Give the reduced model required to determine whether the machine-type terms are necessary.

## 12.9 External Model Validation (*Optional*)

Regression analysis is one of the most widely used statistical tools for estimation and prediction. All too frequently, however, a regression model deemed to be an adequate predictor of some response $y$ performs poorly when applied in practice. For example, a model developed for predicting the heat rate of a gas turbine engine, although found to be statistically useful based on a test for overall model adequacy, may fail to take into account any extreme changes in temperature that did not occur when the experimental data were collected. This points out an important problem. *Models that fit the sample data well may not be successful predictors of y when applied to new data*. For this reason, it is important to assess the **validity** of the regression model in addition to its **adequacy** before using it in practice.

In Chapter 11, we presented several techniques for checking *model adequacy* (for example, tests of overall model adequacy, partial $F$ tests, $R_a^2$, and $s$). In short, checking model adequacy involves determining whether the regression model adequately fits the *sample data*. **Model validation**, however, involves an assessment of how the fitted regression model will perform in practice—that is, how successful it will be when applied to new or future data. A number of different model validation techniques have been proposed, several of which are briefly discussed in this section. You will need to consult the references for more details on how to apply these techniques.

1. *Examining the predicted values:* Sometimes, the predicted values $\hat{y}$ of the fitted regression model can help to identify an invalid model. Nonsensical or unreasonable predicted values may indicate that the form of the model is incorrect or that the $\beta$ coefficients are poorly estimated. For example, a model for a binary response $y$, where $y$ is 0 or 1, may yield predicted probabilities that are negative or greater than 1. In this case, the user may want to consider a model that produces predicted values between 0 and 1 in practice.* On the other hand, if the predicted values of the fitted model all seem reasonable, the user should refrain from using the model in practice until further checks of model validity are carried out.
2. *Examining the estimated model parameters:* Typically, the user of a regression model has some knowledge of the relative size and sign (positive or negative) of the model parameters. This information should be used as a check on the estimated $\beta$ coefficients. Coefficients with signs opposite to what is expected or with unusually small or large values, or unstable coefficients (i.e., coefficients with large

*A model developed for a binary response $y$ is a **logistic regression model**.

standard errors), forewarn that the final model may perform poorly when applied to new or different data.

3. *Collecting new data for prediction:* One of the most effective ways of validating a regression model is to use the model to predict $y$ for a new sample. By directly comparing the predicted values to the observed values of the new data, we can determine the accuracy of the predictions and use this information to assess how well the model performs in practice.

   Several measures of model validity have been proposed for this purpose. One simple technique is to calculate the percentage of variability in the new data explained by the model, ***R*²- prediction** (denoted, $R^2_{\text{pred}}$), and compare it to the coefficient of determination $R^2$ for the least-squares fit of the final model. Let $y_1, y_2, \ldots, y_n$ represent the $n$ observations used to build and fit the final regression model and $y_{n+1}, y_{n+2}, \ldots, y_{n+m}$ represent the $m$ observations in the new data set. Then

$$R^2_{\text{pred}} = 1 - \left\{ \frac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2} \right\}$$

   where $\hat{y}_i$ is the predicted value for the $i$th observation using the $\beta$ estimates from the fitted model and $\bar{y}$ is the sample mean of the original data.* If $R^2_{\text{pred}}$ compares favorably to $R^2$ from the least-squares fit, we will have increased confidence in the usefulness of the model. However, if a significant drop in $R^2$ is observed, we should be cautious about using the model for prediction in practice.

   A similar type of comparison can be made between the mean square error, MSE, for the least-squares fit and the **mean squared prediction error**

$$\text{MSE}_{\text{pred}} = \frac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{m - (k + 1)}$$

   where $k$ is the number of $\beta$ coefficients (excluding $\beta_0$) in the model. Whichever measure of model validity you decide to use, the number of observations in the new data set should be large enough to reliably assess the model's prediction performance. Montgomery, Peck, and Vining (2001), for example, recommend 15–20 new observations, *at minimum*.

4. *Data-splitting (cross-validation):* For those applications where it is impossible or impractical to collect new data, the original sample data can be split into two parts, with one part used to estimate the model parameters and the other part used to assess the fitted model's predictive ability. **Data-splitting** (or **cross-validation**, as it is sometimes known) can be accomplished in a variety of ways. A common technique is to randomly assign half the observations to the estimation data set and the other half to the prediction data set.† Measures of model validity, such as $R^2_{\text{p}}$ or $\text{MSE}_{\text{p}}$ can then be calculated. Of course, a sufficient number of observations must be available for data-splitting to be effective. For the estimation and prediction data sets of equal size, it has been recommended that the entire sample consist of *at least* $n = 2k + 25$ observations, where $k$ is the number of $\beta$ parameters in the model [see Snee (1977)].

5. *Jackknifing:* In situations where the sample data set is too small to apply data-splitting, a method called the **jackknife** can be applied. Let $y_{(i)}$ denote the predicted

*Alternatively, the sample mean of the new data may be used.

†Random splits are usually applied in cases where there is no logical basis for dividing the data. Consult the references for other, more formal, data-splitting techniques.

value for the $i$th observation obtained when the regression model is fit with the data point for $y_i$ omitted (or deleted) from the sample. The jackknife method involves leaving each observation out of the data set, one at a time, and calculating the difference, $y_i - \hat{y}_{(i)}$, for all $n$ observations in the data set. Measures of model validity, such as $R^2$ and MSE, are then calculated:

$$R^2_{\text{jackknife}} = 1 - \frac{\sum (y_i - \hat{y}_{(i)})^2}{\sum (y_i - \bar{y})^2}$$

$$\text{MSE}_{\text{jackknife}} = \frac{\sum (y_i - \hat{y}_{(i)})^2}{n - (k + 1)}$$

The numerator of both $R^2_{\text{jackknife}}$ and $\text{MSE}_{\text{jackknife}}$ is called the **prediction sum of squares**, or **PRESS**. In general, PRESS will be larger than the SSE of the fitted model. Consequently, $R^2_{\text{jackknife}}$ will be smaller than the $R^2$ of the fitted model and $\text{MSE}_{\text{jackknife}}$ will be larger than the MSE of the fitted model. These jackknife measures, then, give a more conservative (and more realistic) assessment of the ability of the model to predict future observations than the usual measures of model adequacy.

The appropriate model validation technique(s) will vary from application to application. Keep in mind that a favorable result is still no guarantee that the model will always perform successfully in practice. However we have much greater confidence in a validated model than in one that simply fits the sample data well.

## 12.10 Stepwise Regression

In building a model to describe a response variable $y$, we must choose the important terms to be included in the model. The list of potentially important independent variables, with their associated main effect and interaction terms, may be extremely large. Therefore, we need some objective method of screening out those that are not important. The screening procedure that we present in this chapter is known as a **stepwise regression analysis**.

The most commonly used stepwise regression procedure, available in most popular statistical software packages, works as follows: The user first identifies the response, $y$, and the set of potentially important independent variables, $x_1, x_2, \ldots, x_k$, where $k$ will generally be large. (Note that this set of variables could represent both first- and higher-order terms, as well as any interaction terms that might be important information contributors.) The response and independent variables are then entered into the computer, and the stepwise procedure begins.

**Step 1** The computer fits all possible one-variable models of the form

$$E(y) = \beta_0 + \beta_1 x_i$$

to the data. For each model, the test of the null hypothesis

$$H_0: \quad \beta_1 = 0$$

against the alternative hypothesis

$$H_a: \quad \beta_1 \neq 0$$

is conducted using the $T$ (or the equivalent $F$) test for a single $\beta$ parameter. The independent variable that produces the largest (absolute) $T$ value is declared the best one-variable predictor of $y$.*

*Note that the variable with the largest $T$ value will also be the one with the largest Pearson product moment correlation $r$ with $y$.

**Step 2** The stepwise program now begins to search through the remaining $(k - 1)$ independent variables for the best two-variable model of the form

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_i$$

This is done by fitting all two-variable models containing $x_1$ and each of the other $(k - 1)$ options for the second variable $x_i$. The $T$ values for the test $H_0$: $\beta_2 = 0$ are computed for each of the $(k - 1)$ models (corresponding to the remaining independent variables $x_i$, $i = 2, 3, \ldots, k$), and the variable having the largest $T$ is retained. Call this variable $x_2$.

At this point, some software packages diverge in methodology. The better packages now go back and check the $T$ value of $\hat{\beta}_1$ *after* $\hat{\beta}_2 x_2$ *has been added to the model*. If the $T$ value has become nonsignificant at some specified $\alpha$ level (say, $\alpha = .10$), the variable $x_1$ is removed and a search is made for the independent variable with a $\beta$ parameter that will yield the most significant $T$ value in the presence of $\hat{\beta}_2 x_2$. Other packages do not recheck $\hat{\beta}_1$, but proceed directly to step 3.

The best-fitting model may yield a different value for $\hat{\beta}_1$ than that obtained in step 1, because $x_1$ and $x_2$ may be correlated. Thus, both the value of $\hat{\beta}_1$ and its significance usually change from step 1 to step 2. For this reason, the software packages that recheck the $T$ values at each step are preferred.

**Step 3** The stepwise procedure now checks for a third independent variable to include in the model with $x_1$ and $x_2$. That is, we seek the best model of the form

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_i$$

To do this, we fit all the $(k - 2)$ models using $x_1$, $x_2$, and each of the $(k - 2)$ remaining variables, $x_i$, as a possible $x_3$. The criterion is again to include the independent variable with the largest $T$ value. Call this best third variable $x_3$.

The better programs now recheck the $T$ values corresponding to the $x_1$ and $x_2$ coefficients, replacing the variables that have $T$ values that have become nonsignificant. This procedure is continued until no further independent variables can be found that yield significant $T$ values (at the specified $\alpha$ level) in the presence of the variables already in the model.

The result of the stepwise procedure is a model containing only those terms with $T$ values that are significant at the specified $\alpha$ level. Thus, in most practical situations, only several of the large number of independent variables will remain. However, it is very important *not* to jump to the conclusion that all the independent variables important for predicting $y$ have been identified or that the unimportant independent variables have been eliminated. Remember, the stepwise procedure is using only *sample estimates* of the true model coefficients ($\beta$'s) to select the important variables. An extremely large number of single $\beta$ parameter $t$ tests have been conducted, and the probability is very high that one or more errors have been made in including or excluding variables. That is, we have very probably included some unimportant independent variables in the model (Type I errors) and eliminated some important ones (Type II errors).

There is a second reason why we might not have arrived at a good model. When we choose the variables to be included in the stepwise regression, we may often omit high-order terms (to keep the number of variables manageable). Consequently, we may have initially omitted several important terms from the model. Thus, we should recognize stepwise regression for what it is: an objective screening procedure.

Now, we will consider interactions and quadratic terms (for quantitative variables) among variables screened by the stepwise procedure. It would be best to develop this response surface model with a second set of data independent of that used for the screening, so the results of the stepwise procedure can be partially verified with new data. However, this is not always possible, because in many practical modeling situations only a small amount of data is available.

### Warning

Be cautious when using the results of stepwise regression to make inferences about the relationship between $E(y)$ and the independent variables in the resulting first-order model. First, an extremely large number of $T$ tests have been conducted, leading to a high probability of making either one or more Type I or Type II errors. Second, the stepwise model does not include any higher-order or interaction terms. Stepwise regression should be used only when necessary, i.e., when you want to determine which of a large number of potentially important independent variables should be used in the model-building process.

Remember, do not be deceived by the impressive looking $t$ values that result from the stepwise procedure—it has retained only the independent variables with the largest $T$ values. Also, if you have used a main effects model for your stepwise procedure, remember that it may be greatly improved by the addition of interaction and quadratic terms.

## Example 12.10

Variable Screening with Stepwise Regression

Suppose a large civil engineering firm wants to use multiple regression to model an executive's salary as a function of experience, education, gender, and other factors. A preliminary step in the construction of the model is to determine the most important independent variables. Ten independent variables to be considered are listed in Table 12.8. Since it would be very difficult to perform a regression analysis on a complete second-order model using 10 independent variables, we need to eliminate those variables (or terms) that do not contribute much information for the prediction of salary. Salary data for a sample of 100 executives is saved in the **CIVILENG** file. Use stepwise regression to decide which of the 10 variables should be included in the construction of the final model.

**Solution** We ran a stepwise regression with the main effects of the 10 independent variables to identify the most important variables. The dependent variable $y$ is the natural logarithm of the executive salaries. The MINITAB stepwise regression printout is shown in Figure 12.29. (*Note*: MINITAB automatically enters the constant term ($\beta_0$) into the model in the first step. The remaining steps follow the procedure outlined above. The $T$ values and associated $p$-values for the tests at each step are highlighted on the printout.)

**CIVILENG**

**TABLE 12.8 Independent Variables in Example 12.10**

| Independent Variable | Description |
|---|---|
| $x_1$ | Experience (years)—quantitative |
| $x_2$ | Education (years)—quantitative |
| $x_3$ | Engineering degree (1 if yes, 0 if no)—qualitative |
| $x_4$ | Number of employees supervised—quantitative |
| $x_5$ | Corporate assets (millions of dollars)—quantitative |
| $x_6$ | Board member (1 if yes, 0 if no)—qualitative |
| $x_7$ | Age (years)—quantitative |
| $x_8$ | Company profits (past 12 months, millions of dollars)—quantitative |
| $x_9$ | Has international responsibility (1 if yes, 0 if no)—qualitative |
| $x_{10}$ | Company's total sales (past 12 months, millions of dollars)—quantitative |

**Stepwise Regression: Y versus X1, X2, X3, X4, X5, X6, X7, X8, X9, X10**

```
 Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is    Y     on 10 predictors, with N =  100


   Step          1         2         3         4         5
Constant    11.091    10.968    10.783    10.278     9.962

X1          0.0278    0.0273    0.0273    0.0273    0.0273
T-Value      12.62     15.13     18.80     24.68     26.50
P-Value      0.000     0.000     0.000     0.000     0.000

X3                     0.197     0.233     0.232     0.225
T-Value                 7.10     10.17     13.30     13.74
P-Value                0.000     0.000     0.000     0.000

  X4                              0.00048   0.00055   0.00052
  T-Value                            7.32     10.92     11.06
  P-Value                           0.000     0.000     0.000

  X2                                         0.0300    0.0291
  T-Value                                      8.38      8.72
  P-Value                                     0.000     0.000

  X5                                                  0.00196
  T-Value                                                3.95
  P-Value                                               0.000

  S              0.161     0.131     0.106    0.0807    0.0751
  R-Sq           61.90     74.92     83.91     90.75     92.06
  R-Sq(adj)      61.51     74.40     83.41     90.36     91.64
  C-p            343.9     195.5      93.8      16.8       3.6
  PRESS        2.66387   1.78796   1.17124  0.695637  0.610197
  R-Sq(pred)     60.14     73.24     82.47     89.59     90.87
```

**FIGURE 12.29**
MINITAB stepwise regression results for salaries of civil engineers

Note that the first variable included in the model is $x_1$, years of experience. At the second step, $x_3$, engineering degree, enters the model. In succeeding steps, $x_4$, $x_2$, and $x_5$ are entered. None of the other $x$'s can meet the $\alpha = .15$ (default value in MINITAB) level of significance for entry into the model. Thus, stepwise regression suggests that we concentrate on the five variables $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$ in our final modeling effort.

In addition to stepwise regression, other more subjective methods are designed to aid in the selection of the "most important" independent variables. For example, the **all-possible regressions selection procedure** considers regression models involving all possible subsets of the potentially important predictors. The "best" subset of variables is selected (by the researcher) based on model statistics, such as the familiar $R^2$ and MSE, and other statistics, such as PRESS (prediction sum of squares) and Mallows $C_p$. (Consult the references for details on the $C_p$ statistic.) These methods, however, lack the objectivity of stepwise regression, and (as with stepwise regression) analysts typically omit interactions and higher-order terms in the list of potential variables when using them.

## Applied Exercises

12.65 *Stepwise regression.* There are six independent variables, $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, and $x_6$, that might be useful in predicting a response $y$. A total of $n = 50$ observations are available, and it is decided to employ stepwise regression to help in selecting the independent variables that appear to be useful.

a. How many models are fit to the data in step 1? Give the general form of these models.
b. The table gives the estimate of $\beta_1$ and standard error for each independent variable fit in the model in step 1. Use the results to determine which independent variable is declared the best one-variable predictor of $y$.

| Independent Variable | $\hat{\beta}_1$ | $s_{\hat{\beta}_1}$ |
|---|---|---|
| $x_1$ | 1.6 | .42 |
| $x_2$ | $-.9$ | .01 |
| $x_3$ | 3.4 | 1.14 |
| $x_4$ | 2.5 | 2.06 |
| $x_5$ | $-4.4$ | .73 |
| $x_6$ | .3 | .35 |

c. How many models are fit to the data in step 2? Give the general form of these models.
d. Explain how the procedure determines when to stop adding independent variables to the model.
e. Give two major drawbacks to using the final stepwise model as the "best" model for $E(y)$.

12.66 *An analysis of footprints in sand.* Fossilized human footprints provide a direct source of information on the gait dynamics of extinct species. How paleontologists and anthropologists interpret these prints, however, may vary. To gain insight into this phenomenon, a group of scientists used human subjects (16 young adults) to generate footprints in sand. (*American Journal of Physical Anthropology,* April 2010.) One dependent variable of interest was Heel depth ($y$) of the footprint (in millimeters). The scientists wanted to find the best predictors of depth from among six possible independent variables. Three variables were related to the human subject (Foot mass, Leg length, and Foot type) and three variables were related to walking in sand (Velocity, Pressure, and Impulse). A stepwise regression run on these six variables yielded the following results:

Selected independent variables: Pressure and Leg length

$R^2 = 771$,
Global $F$ test $p$-value $< .001$

a. Write the hypothesized equation of the final stepwise regression model.
b. Interpret the value of $R^2$ for the model.
c. Conduct a test of the overall utility of the final stepwise model.
d. At minimum, how many $T$-tests on individual $\beta$'s were conducted to arrive at the final stepwise model?
e. Based on your answer to part **d**, comment on the probability of making at least one Type I error during the stepwise analysis.

12.67 *Using corn in a duck diet.* Corn is high in starch content; consequently, it is considered excellent feed for domestic chickens. Does corn possess the same potential in feeding ducks bred for broiling? This was the subject of research published in *Animal Feed Science and Technology* (April 2010). The objective of the study was to establish a prediction model for the true metabolizable energy (TME) of corn regurgitated from ducks. The researchers considered 11 potential predictors of TME: dry matter (DM), crude protein (CP), ether extract (EE), ash (ASH), crude fiber (CF), neutral detergent fiber (NDF), acid detergent fiber (ADF), gross energy (GE), amylose (AM), amylopectin (AP), and amylopectin/amylose (AMAP). Stepwise regression was used to find the best subset of predictors. The final stepwise model yielded the following results:

$$\widehat{TME} = 7.70 + 2.14(\text{AMAP}) + .16(\text{NDF}),$$
$$R^2 = .988,\ s = .07,\ \text{Global } F\ p\text{-value} = .001$$

a. Determine the number of $T$-tests performed in Step 1 of the stepwise regression.
b. Determine the number of $T$-tests performed in Step 2 of the stepwise regression.
c. Give a full interpretation of the final stepwise model regression results.
d. Explain why it is dangerous to use the final stepwise model as the "best" model for predicting TME?
e. Using the independent variables selected by the stepwise routine, write a complete 2nd-order model for TME.
f. Refer to part **e**. How would you determine if the terms in the model that allow for curvature are statistically useful for predicting TME?

12.68 *Modeling species abundance.* A marine biologist was hired by the EPA to determine whether the hot-water runoff from a particular power plant located near a large gulf is having an adverse effect on the marine life in the area. The biologist's goal is to acquire a prediction equation for the number of marine animals located at certain designated areas, or stations, in the gulf. Based on past experience, the EPA considered the following environmental factors as predictors for the number of animals at a particular station:

$x_1$ = Temperature of water (TEMP)
$x_2$ = Salinity of water (SAL)
$x_3$ = Dissolved oxygen content of water (DO)
$x_4$ = Turbidity index, a measure of the turbidity of the water (TI)
$x_5$ = Depth of the water at the station (ST_DEPTH)
$x_6$ = Total weight of sea grasses in sampled area (TGRSWT)

**SPSS Output for Exercise 12.68**

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | ST_DEPTH | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |
| 2 | TGRSWT | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |
| 3 | TI | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |

a. Dependent Variable: LOGNUM

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .329[a] | .122 | .121 | .7615773 |
| 2 | .427[b] | .182 | .180 | .7348470 |
| 3 | .432[c] | .187 | .184 | .7348469 |

a. Predictors: (Constant), ST_DEPTH

b. Predictors: (Constant), ST_DEPTH, TGRSWT

c. Predictors: (Constant), ST_DEPTH, TGRSWT, TI

As a preliminary step in the construction of this model, the biologist used a stepwise regression procedure to identify the most important of these six variables. A total of 716 samples were taken at different stations in the gulf, producing the SPSS printout shown above. (The response measured was $y$, the logarithm of the number of marine animals found in the sampled area.)

a. According to the SPSS printout, which of the six independent variables should be used in the model?
b. Are we able to assume that the marine biologist has identified all the important independent variables for the prediction of $y$? Why?
c. Using the variables identified in part **a**, write the first-order model with interaction that may be used to predict $y$.
d. How would the marine biologist determine whether the model specified in part **c** is better than the first-order model?
e. Note the small value of $R^2$. What action might the biologist take to improve the model?

12.69 *Bus Rapid Transit study.* Bus Rapid Transit (BRT) is a rapidly growing trend in the provision of public transportation in America. The Center for Urban Transportation Research (CUTR) at the University of South Florida conducted a survey of BRT customers in Miami. (*Transportation Research Board* Annual Meeting, Jan. 2003.) Data on the following variables (all measured on a 5-point scale, where 1 = "very unsatisfied" and 5 = "very satisfied") were collected for a sample of over 500 bus riders: overall satisfaction with BRT ($y$), safety on bus ($x_1$), seat availability ($x_2$), dependability ($x_3$), travel time ($x_4$), cost ($x_5$), information/maps ($x_6$), convenience of routes ($x_7$), traffic signals ($x_8$), safety at bus stops ($x_9$), hours of service ($x_{10}$), and frequency of service ($x_{11}$). CUTR analysts used stepwise regression to model overall satisfaction ($y$).

a. How many models are fit at step 1 of the stepwise regression?
b. How many models are fit at step 2 of the stepwise regression?
c. How many models are fit at step 11 of the stepwise regression?
d. The stepwise regression selected the following eight variables to include in the model (in order of selection): $x_{11}$, $x_4$, $x_2$, $x_7$, $x_{10}$, $x_1$, $x_9$, and $x_3$. Write the equation for $E(y)$ that results from stepwise regression.
e. The model, part **d**, resulted in $R^2 = .677$. Interpret this value.
f. Explain why the CUTR analysts should be cautious in concluding that the "best" model for $E(y)$ has been found.

12.70 *Muscle activity of harvesting foresters.* Refer to the *International Journal of Foresting Engineering* (Vol. 19, 2008) study of neck muscle activity patterns among forestry vehicle operators, Exercise 12.23 (p. 660). Recall that the researchers identified the key explanatory variables of $y$ = the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. A list of the potential predictors is reproduced in the table on p. 700. The researchers collected data for $n = 13$ forestry vehicle operators and applied the stepwise regression method (using $\alpha = .10$) in order to determine the best subset of predictor variables.

| | |
|---|---|
| $x_1$ = | Age of operator (years) |
| $x_2$ = | Duration of lunch break (minutes) |
| $x_3$ = | Dominant hand power level (percentage) |
| $x_4$ = | Perceived stress at work (5-point scale) |
| $x_5$ = | {1 if married, 0 if not} |
| $x_6$ = | {1 if day shift, 0 if night shift} |
| $x_7$ = | {1 if operating a Timberjack vehicle, 0 if operating a Valmet vehicle} |

a. The stepwise regression identified $x_7$ as the best one-variable predictor of $y$. How many one-variable models were fit and tested to arrive at this result?
b. The stepwise regression method did not select any other variables as "significant" predictors of $y$. How many more models were fit and tested to arrive at this result?
c. The model identified by stepwise regression took the form, $E(y) = \beta_0 + \beta_1 x_7$. The $p$-value for testing $H_0$: $\beta_1 = 0$ was $p = .067$, with $R^2 = .30$. Would you advise the researchers to use this model for predicting $y$ = the number of sustained low-level muscle activity (SULMA) periods exhibited by an operator that exceed 8 minutes. Fully explain.

12.71 *Accuracy of software effort estimates.* Periodically, software engineers must provide estimates of their effort in developing new software. In the *Journal of Empirical Software Engineering* (Vol. 9, 2004), multiple regression was used to predict the accuracy of these effort estimates. The dependent variable, defined as the relative error in estimating effort,

$$y = (\text{Actual effort} - \text{Estimated effort})/(\text{Actual effort})$$

was determined for each in a sample of $n = 49$ software development tasks. Eight independent variables were evaluated as potential predictors of relative error using stepwise regression. Each of these was formulated as a dummy variable, as shown in the table.

*Company role of estimator:* $x_1$ = 1 if developer, 0 if project leader

*Task complexity:* $x_2$ = 1 if low, 0 if medium/high

*Contract type:* $x_3$ = 1 if fixed price, 0 if hourly rate

*Customer importance:* $x_4$ = 1 if high, 0 if low/medium

*Customer priority:* $x_5$ = 1 if time-of-delivery, 0 if cost or quality

*Level of knowledge:* $x_6$ = 1 if high, 0 if low/medium

*Participation:* $x_7$ = 1 if estimator participates in work, 0 if not

*Previous accuracy:* $x_8$ = 1 if more than 20% accurate, 0 if less than 20% accurate

a. In Step 1 of the stepwise regression, how many different one-variable models are fit to the data?
b. In Step 1, the variable $x_1$ is selected as the "best" one-variable predictor. How is this determined?
c. In Step 2 of the stepwise regression, how many different two-variable models (where $x_1$ is one of the variables) are fit to the data?
d. The only two variables selected for entry into the stepwise regression model were $x_1$ and $x_8$. The stepwise regression yielded the following prediction equation:

$$\hat{y} = .12 - .28x_1 + .27x_8$$

Give a practical interpretation of the $\beta$-estimates multiplied by $x_1$ and $x_8$.
e. Why should a researcher be wary of using the model, part **d**, as the final model for predicting effort ($y$)?

12.72 *Yield strength of steel alloy.* Industrial engineers at the University of Florida used regression modeling as a tool to reduce the time and cost associated with developing new metallic alloys. (*Modelling and Simulation in Materials Science and Engineering*, Vol. 13, 2005.) To illustrate, the engineers built a regression model for the tensile yield strength ($y$) of a new steel alloy. The potential important predictors of yield strength are listed in the accompanying table.

| | |
|---|---|
| $x_1$ = | Carbon amount (% weight) |
| $x_2$ = | Manganese amount (% weight) |
| $x_3$ = | Chromium amount (% weight) |
| $x_4$ = | Nickel amount (% weight) |
| $x_5$ = | Molybdenum amount (% weight) |
| $x_6$ = | Copper amount (% weight) |
| $x_7$ = | Nitrogen amount (% weight) |
| $x_8$ = | Vanadium amount (% weight) |
| $x_9$ = | Plate thickness (millimeters) |
| $x_{10}$ = | Solution treating (milliliters) |
| $x_{11}$ = | Aging temperature (degrees, Celsius) |

a. The engineers discovered that the variable Nickel ($x_4$) was highly correlated with the other potential independent variables. Consequently, Nickel was dropped from the model. Do you agree with this decision? Explain.
b. The engineers used stepwise regression on the remaining 10 potential independent variables in order to search for a parsimonious set of predictor variables. Do you agree with this decision? Explain.
c. The stepwise regression selected the following independent variables: $x_1$ = Carbon, $x_2$ = Manganese, $x_3$ = Chromium, $x_5$ = Molybdenum, $x_6$ = Copper, $x_8$ = Vanadium, $x_9$ = Plate thickness, $x_{10}$ = Solution treating, and $x_{11}$ = Aging temperature. All these variables were statistically significant in the stepwise model, with $R^2 = .94$. Consequently, the engineers used the estimated stepwise model to predict yield strength. Do you agree with this decision? Explain.

## *STATISTICS IN ACTION REVISITED*

### Deregulation in the Intrastate Trucking Industry

We now return to the problem of assessing the impact of deregulation in the intrastate trucking industry using data collected in the state of Florida. Recall that our objective is to build a good regression model for the natural logarithm of the supply price charged per ton-mile (y). The potential independent variables are listed in Table SIA12.1 (p. 643). The data for 134 shipments are saved in the **TRUCKING** file. Note that the first three variables—distance shipped, weight of product, and percent of truckload capacity—are quantitative in nature, while the last four variables—city of origin, market size, deregulation status, and product classification—are all qualitative in nature. Of course, these qualitative independent variables will require the creation of the appropriate number of dummy variables—1 dummy variable for city of origin, 1 for market size, 1 for deregulation, and 2 for product classification. The variable assignments are given in Table SIA12.2.

**TRUCKING**

**TABLE SIA12.2 Independent Variable Assignments for Trucking Price Model**

| | |
|---|---|
| $x_1$ = | Distance shipped (hundreds of miles) |
| $x_2$ = | Weight of product shipped (thousands of pounds) |
| $x_3$ = | {1 if Deregulation in effect, 0 if not} |
| $x_4$ = | {1 if trip originates in Miami, 0 if in Jacksonville} |
| $x_5$ = | Truck load (percentage of capacity) |
| $x_6$ = | {1 if destination is Large market, 0 if Small market} |
| $x_7$ = | {1 if Product classification of 100, 0 if not} |
| $x_8$ = | {1 if Product classification of 150, 0 if not} |

*Variable Screening:* One strategy to finding the best model for *y* is to use a "build-down" approach, i.e., start with a complete 2nd-order model and conduct tests to eliminate terms in the model that are not statistically useful. However, a complete 2nd-order model with these 3 quantitative predictors and 4 qualitative predictors will involve 240 terms. (Check this as an exercise.) Since the sample size is $n = 134$, it will be impossible to fit this model. Hence, we require a screening procedure to find a subset of the independent variables that best predict *y*.

We employed stepwise regression to obtain the "best" subset of predictors of the natural logarithm of supply price. The SAS stepwise regression printout is shown in Figure SIA12.1. This analysis leads us to select the following variables to begin the model building process: Distance ($x_1$), Weight ($x_2$), Deregulation ($x_3$), and Origin ($x_4$).

Summary of Stepwise Selection

| Step | Variable Entered | Variable Removed | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
|---|---|---|---|---|---|---|---|---|
| 1 | DISTANCE | | 1 | 0.2969 | 0.2969 | 417.090 | 55.74 | <.0001 |
| 2 | DEREG | | 2 | 0.3127 | 0.6096 | 175.795 | 104.91 | <.0001 |
| 3 | WEIGHT | | 3 | 0.1897 | 0.7993 | 30.1997 | 122.84 | <.0001 |
| 4 | ORIGIN | | 4 | 0.0362 | 0.8355 | 4.0122 | 28.40 | <.0001 |

**FIGURE SIA12.1**
Portion of SAS stepwise regression output

*Model Building:* We begin the model-building process by specifying four models. These models, named Model 1–4, are shown in Table SIA12.3. Notice that Model 1 is the complete second-order model. Recall from Section 12.7 that the complete second-order model contains quadratic (curvature) terms for quantitative variables and interactions among the quantitative and qualitative terms. For the trucking data, Model 1 traces a parabolic surface for mean natural log of price, $E(y)$, as a function of distance ($x_1$) and weight ($x_2$), and the response surfaces differ for the $2 \times 2 = 4$ combinations of the levels of deregulation ($x_3$) and origin ($x_4$). Generally, the complete second-order model is a good place to start the model-building process

**TABLE SIA12.3 Hypothesized Models for Natural Log of Trucking Price**

| | |
|---|---|
| *Model 1:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 + \beta_8 x_3 x_4 + \beta_9 x_1 x_3 + \beta_{10} x_1 x_4 + \beta_{11} x_1 x_3 x_4 + \beta_{12} x_2 x_3 + \beta_{13} x_2 x_4 + \beta_{14} x_2 x_3 x_4 + \beta_{15} x_1 x_2 x_3 + \beta_{16} x_1 x_2 x_4 + \beta_{17} x_1 x_2 x_3 x_4 + \beta_{18} x_1^2 x_3 + \beta_{19} x_1^2 x_4 + \beta_{20} x_1^2 x_3 x_4 + \beta_{21} x_2^2 x_3 + \beta_{22} x_2^2 x_4 + \beta_{23} x_2^2 x_3 x_4$ |
| *Model 2:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_6 x_3 + \beta_7 x_4 + \beta_8 x_3 x_4 + \beta_9 x_1 x_3 + \beta_{10} x_1 x_4 + \beta_{11} x_1 x_3 x_4 + \beta_{12} x_2 x_3 + \beta_{13} x_2 x_4 + \beta_{14} x_2 x_3 x_4 + \beta_{15} x_1 x_2 x_3 + \beta_{16} x_1 x_2 x_4 + \beta_{17} x_1 x_2 x_3 x_4$ |
| *Model 3:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 + \beta_8 x_3 x_4$ |
| *Model 4:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 + \beta_8 x_3 x_4 + \beta_9 x_1 x_3 + \beta_{10} x_1 x_4 + \beta_{11} x_1 x_3 x_4 + \beta_{12} x_2 x_3 + \beta_{13} x_2 x_4 + \beta_{14} x_2 x_3 x_4 + \beta_{15} x_1 x_2 x_3 + \beta_{16} x_1 x_2 x_4 + \beta_{17} x_1 x_2 x_3 x_4$ |
| *Model 5:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_9 x_1 x_3 + \beta_{12} x_2 x_3 + \beta_{15} x_1 x_2 x_3$ |
| *Model 6:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_7 x_4 + \beta_{10} x_1 x_4 + \beta_{13} x_2 x_4 + \beta_{16} x_1 x_2 x_4$ |
| *Model 7:* | $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 + \beta_9 x_1 x_3 + \beta_{10} x_1 x_4 + \beta_{12} x_2 x_3 + \beta_{13} x_2 x_4 + \beta_{15} x_1 x_2 x_3 + \beta_{16} x_1 x_2 x_4$ |

Dependent Variable: LNPRICE

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 23 | 83.90934 | 3.64823 | 65.59 | <.0001 |
| Error | 110 | 6.11860 | 0.05562 | | |
| Corrected Total | 133 | 90.02794 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.23585 | R-Square | 0.9320 |
| Dependent Mean | 10.57621 | Adj R-Sq | 0.9178 |
| Coeff Var | 2.22997 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 12.51593 | 0.95441 | 13.11 | <.0001 |
| X1 | 1 | -0.89923 | 0.73410 | -1.22 | 0.2232 |
| X2 | 1 | 0.02421 | 0.02886 | 0.84 | 0.4034 |
| X1X2 | 1 | -0.02071 | 0.00673 | -3.08 | 0.0026 |
| X1SQ | 1 | 0.15145 | 0.13455 | 1.13 | 0.2628 |
| X2SQ | 1 | -0.00010196 | 0.00076963 | -0.13 | 0.8948 |
| X3 | 1 | -1.12650 | 1.49104 | -0.76 | 0.4516 |
| X4 | 1 | 0.27615 | 0.96332 | 0.29 | 0.7749 |
| X3X4 | 1 | 0.49697 | 1.50290 | 0.33 | 0.7415 |
| X1X3 | 1 | 0.48205 | 1.15882 | 0.42 | 0.6782 |
| X1X4 | 1 | 0.06958 | 0.73882 | 0.09 | 0.9251 |
| X1X3X4 | 1 | -0.54037 | 1.16440 | -0.46 | 0.6435 |
| X2X3 | 1 | -0.09486 | 0.04477 | -2.12 | 0.0363 |
| X2X4 | 1 | -0.05261 | 0.03528 | -1.49 | 0.1387 |
| X2X3X4 | 1 | 0.06826 | 0.05220 | 1.31 | 0.1937 |
| X1X2X3 | 1 | 0.02207 | 0.01078 | 2.05 | 0.0429 |
| X1X2X4 | 1 | 0.02355 | 0.00709 | 3.32 | 0.0012 |
| X1X2X3X4 | 1 | -0.02694 | 0.01127 | -2.39 | 0.0185 |
| X1SQX3 | 1 | -0.11674 | 0.21918 | -0.53 | 0.5954 |
| X1SQX4 | 1 | -0.07276 | 0.13510 | -0.54 | 0.5913 |
| X1SQX3X4 | 1 | 0.13424 | 0.21984 | 0.61 | 0.5427 |
| X2SQX3 | 1 | 0.00043756 | 0.00119 | 0.37 | 0.7127 |
| X2SQX4 | 1 | 0.00011095 | 0.00107 | 0.10 | 0.9174 |
| X2SQX3X4 | 1 | -0.00027597 | 0.00157 | -0.18 | 0.8609 |

**FIGURE SIA12.2**
SAS regression printout for Model 1

since most real-world relationships are curvilinear. (Keep in mind, however, that you must have a sufficient number of data points to find estimates of all the parameters in the model.) Model 1 is fit to the data for the 134 shipments using SAS. The results are shown in Figure SIA12.2. Note that the *p*-value for the global model *F* test is less than .0001, indicating that the complete second-order model is statistically useful for predicting trucking price.

Model 2 contains all the terms of Model 1, except that the quadratic terms (i.e., terms involving $x_1^2$ and $x_2^2$) are dropped. This model also proposes four different response surfaces for the combinations of levels of deregulation and origin, but the surfaces are twisted planes (see Figure 12.10) rather than paraboloids. A direct comparison of Models 1 and 2 will allow us to test for the importance of the curvature terms.

Model 3 contains all the terms of Model 1, except that the quantitative-qualitative interaction terms are omitted. This model proposes four curvilinear paraboloids corresponding to the four deregulation-origin combinations, that differ only with respect to the *y*-intercept. By directly comparing Models 1 and 3, we can test for the importance of all the quantitative-qualitative interaction terms.

**FIGURE SIA12.3**
SAS nested model $F$ tests for terms in Model 1

Test QUADRATIC Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 8 | 0.75727 | 13.61 | <.0001 |
| Denominator | 110 | 0.05562 | | |

Test QN_QL_INTERACT Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 15 | 0.25574 | 4.60 | <.0001 |
| Denominator | 110 | 0.05562 | | |

Test QL_QUAD_INTERACT Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 6 | 0.01407 | 0.25 | 0.9572 |
| Denominator | 110 | 0.05562 | | |

Model 4 is identical to Model 1, except that it does not include any interactions between the quadratic terms and the two qualitative variables, deregulation ($x_3$) and origin ($x_4$). Although curvature is included in this model, the rates of curvature for both distance ($x_1$) and weight ($x_2$) are the same for all levels of deregulation and origin.

Figure SIA12.3 shows the results of the nested model $F$ tests described in the preceding paragraphs. Each of these tests is summarized as follows:

Test for significance of all quadratic terms (Model 1 vs. Model 2)

$H_0$: $\beta_4 = \beta_5 = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$

$H_a$: At least one of the quadratic $\beta$'s in Model 1 differs from 0

$F = 13.61$, $p$-value $< .0001$ (shaded at the top of Figure SIA12.3)

*Conclusion:* There is sufficient evidence (at $\alpha = .01$) of curvature in the relationships between $E(y)$ and distance ($x_1$) and weight ($x_2$). Model 1 is a statistically better predictor of trucking price than Model 2.

Test for significance of all quantitative-qualitative interaction terms (Model 1 vs. Model 3)

$H_0$: $\beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19}$
$= \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$

$H_a$: At least one of the QN $\times$ QL interaction $\beta$'s in Model 1 differs from 0

$F = 4.60$, $p$-value $< .0001$ (shaded at the middle of Figure SIA12.3)

*Conclusion:* There is sufficient evidence (at $\alpha = .01$) of interaction between the quantitative variables, distance ($x_1$) and weight ($x_2$), and the qualitative variables, deregulation ($x_3$) and origin ($x_4$). Model 1 is a statistically better predictor of trucking price than Model 3.

Test for significance of qualitative-quadratic interaction (Model 1 vs. Model 4)

$H_0$: $\beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$

$H_a$: At least one of the qualitative-quadratic interaction $\beta$'s in Model 1 differs from 0

$F = .25$, $p$-value $= .9572$ (shaded at the bottom of Figure SIA12.3)

Dependent Variable: LNPRICE

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 17 | 83.82495 | 4.93088 | 92.21 | <.0001 |
| Error | 116 | 6.20299 | 0.05347 | | |
| Corrected Total | 133 | 90.02794 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.23124 | R-Square | 0.9311 |
| Dependent Mean | 10.57621 | Adj R-Sq | 0.9210 |
| Coeff Var | 2.18646 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 12.08482 | 0.25871 | 46.71 | <.0001 |
| X1 | 1 | -0.55296 | 0.09648 | -5.73 | <.0001 |
| X2 | 1 | 0.01889 | 0.02120 | 0.89 | 0.3748 |
| X1X2 | 1 | -0.02041 | 0.00649 | -3.15 | 0.0021 |
| X1SQ | 1 | 0.08738 | 0.00827 | 10.56 | <.0001 |
| X2SQ | 1 | 0.00008202 | 0.00037354 | 0.22 | 0.8266 |
| X3 | 1 | -0.38504 | 0.40009 | -0.96 | 0.3379 |
| X4 | 1 | 0.76041 | 0.27135 | 2.80 | 0.0059 |
| X3X4 | 1 | -0.35311 | 0.42661 | -0.83 | 0.4095 |
| X1X3 | 1 | -0.13163 | 0.14172 | -0.93 | 0.3549 |
| X1X4 | 1 | -0.33374 | 0.08995 | -3.71 | 0.0003 |
| X1X3X4 | 1 | 0.18215 | 0.14746 | 1.24 | 0.2192 |
| X2X3 | 1 | -0.08259 | 0.02956 | -2.79 | 0.0061 |
| X2X4 | 1 | -0.04830 | 0.02049 | -2.36 | 0.0201 |
| X2X3X4 | 1 | 0.05937 | 0.03217 | 1.85 | 0.0675 |
| X1X2X3 | 1 | 0.02136 | 0.01043 | 2.05 | 0.0428 |
| X1X2X4 | 1 | 0.02320 | 0.00685 | 3.39 | 0.0010 |
| X1X2X3X4 | 1 | -0.02601 | 0.01090 | -2.39 | 0.0186 |

**FIGURE SIA12.4**
SAS regression printout for Model 4

*Conclusion:* There is insufficient evidence (at $\alpha = .01$) of interaction between the quadratic terms for distance ($x_1$) and weight ($x_2$), and the qualitative variables, deregulation ($x_3$) and origin ($x_4$). Since these terms are not statistically useful, we will drop these terms from Model 1 and conclude that Model 4 is a statistically better predictor of trucking price.*

Based on the three nested-model *F* tests, we found Model 4 to be the "best" of the first four models. The SAS printout for Model 4 is shown in Figure SIA12.4. Looking at the results of the global *F* test (*p*-value less than .0001), you can see that the overall model is statistically useful for predicting trucking price. Also, $R_a^2 = .9210$ implies that about 92% of the sample variation in the natural log of trucking price can be explained by the model. Although these model statistics are impressive, we may be able to find a simpler model that fits the data just as well.

*There is always danger in dropping terms from the model. Essentially, we are accepting $H_0$: $\beta_{18} = \beta_{19} = \beta_{20} = \cdots = \beta_{23} = 0$ when $P(\text{Type II error}) = P(\text{Accepting } H_0 \text{ when } H_0 \text{ is false}) = \beta$ is unknown. In practice, however, many researchers are willing to risk making a Type II error rather than use a more complex model for $E(y)$ when simpler models that are nearly as good as predictors (and easier to apply and interpret) are available. Note that we used a relatively large amount of data ($n = 134$) in fitting our models and that $R_a^2$ for Model 4 is actually larger than $R_a^2$ for Model 1. If the quadratic interaction terms are, in fact, important (i.e., we have made a Type II error), there is little lost in terms of explained variability in using Model 4.

Table SIA12.3 gives three additional models. Model 5 is identical to Model 4, but all terms for the qualitative variable origin ($x_4$) have been dropped. A comparison of Model 4 to Model 5 will allow us to determine whether origin really has an impact on trucking price. Similarly, Model 6 is identical to Model 4, but now all terms for the qualitative variable deregulation ($x_3$) have been dropped. By comparing Model 4 to Model 6, we can determine whether deregulation has an impact on trucking price. Finally, we propose Model 7, which is obtained by dropping all the qualitative-qualitative interaction terms. A comparison of Model 4 to Model 7 will allow us to see whether deregulation and origin interact to affect the natural log of trucking price.

Figure SIA12.5 is a SAS printout showing the results of the nested model $F$ tests described above. A summary of each of these tests follows:

Test for significance of all origin terms (Model 4 vs. Model 5)

$H_0$: $\beta_7 = \beta_8 = \beta_{10} = \beta_{11} = \beta_{13} = \beta_{14} = \beta_{16} = \beta_{17} = 0$

$H_a$: At least one of the origin $\beta$'s in Model 4 differs from 0

$F = 3.55$, $p$-value $= .001$ (shaded at the top of Figure SIA12.5)

*Conclusion:* There is sufficient evidence (at $\alpha = .01$) to indicate that origin ($x_4$) has an impact on trucking price. Model 4 is a statistically better predictor of trucking price than Model 5.

Test for significance of all deregulation terms (Model 4 vs. Model 6)

$H_0$: $\beta_6 = \beta_8 = \beta_9 = \beta_{11} = \beta_{12} = \beta_{14} = \beta_{15} = \beta_{17} = 0$

$H_a$: At least one of the deregulation $\beta$'s in Model 4 differs from 0

$F = 75.44$, $p$-value $< .0001$ (shaded at the middle of Figure SIA12.5)

*Conclusion*: There is sufficient evidence (at $\alpha = .01$) to indicate that deregulation ($x_3$) has an impact on trucking price. Model 4 is a statistically better predictor of trucking price than Model 6.

Test for significance of all deregulation-origin interaction terms (Model 4 vs. Model 7)

$H_0$: $\beta_8 = \beta_{11} = \beta_{14} = \beta_{17} = 0$

$H_a$: At least one of the QL $\times$ QL interaction $\beta$'s in Model 4 differs from 0

$F = 2.13$, $p$-value $= .0820$ (shaded at the bottom of Figure SIA12.5)

*Conclusion:* There is insufficient evidence (at $\alpha = .01$) to indicate that deregulation ($x_3$) and origin ($x_4$) interact. Thus, we will drop these interaction terms from Model 4 and conclude that Model 7 is a statistically better predictor of trucking price.

In summary, the nested-model $F$ tests suggest that Model 7 is the best for modeling the natural log of trucking price. The SAS printout for Model 7 is shown in Figure SIA 12.6. The $\beta$ estimates used for making predictions of trucking price are highlighted on the printout.

*A note of caution*: Just as with $T$ tests on individual $\beta$ parameters, you should avoid conducting too many partial $F$ tests. Regardless of the type of test ($t$ test or $F$ test), the more tests that are performed, the higher the overall Type I error rate will be. In practice, you should limit the number of models that you propose for $E(y)$ so that the overall Type I error rate $\alpha$ for conducting partial $F$ tests remains reasonably small.*

*Impact of Deregulation:* In addition to estimating a model of the supply price for prediction purposes, a goal of the regression analysis was to assess the impact of deregulation on the trucking prices. To do this, we examine the $\beta$-estimates in Model 7, specifically the $\beta$'s associated with the deregulation dummy variable, $x_3$. From Figure SIA12.6, the prediction equation is:

$$\begin{aligned}\hat{y} = {} & 12.192 - .598x_1 - .00598x_2 - .01078x_1x_2 + .086x_1^2 + .00014x_2^2 \\ & + .677x_4 - .275x_1x_4 - .026x_2x_4 + .013x_1x_2x_4 \\ & - .782x_3 + .0399x_1x_3 - .021x_2x_3 - .0033x_1x_2x_3\end{aligned}$$

*A technique suggested by Bonferroni is often applied to maintain control of the overall Type I error rate $\alpha$. If $c$ tests are to be performed, then conduct each individual test at significance level $\alpha/c$. This will guarantee an overall Type I error rate less than or equal to $\alpha$. For example, conducting each of $c = 5$ tests at the $.05/5 = .01$ level of significance guarantees an overall $\alpha \leq .05$.

**FIGURE SIA12.5**
SAS nested model *F*-tests for terms in Model 4

Test ORIGIN Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 8 | 0.18987 | 3.55 | 0.0010 |
| Denominator | 116 | 0.05347 | | |

Test DEREG Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 8 | 4.03417 | 75.44 | <.0001 |
| Denominator | 116 | 0.05347 | | |

Test ORG_DEREG_INTERACTION Results for Dependent Variable LNPRICE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 4 | 0.11367 | 2.13 | 0.0820 |
| Denominator | 116 | 0.05347 | | |

Dependent Variable: LNPRICE

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 13 | 83.37026 | 6.41310 | 115.59 | <.0001 |
| Error | 120 | 6.65767 | 0.05548 | | |
| Corrected Total | 133 | 90.02794 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.23554 | R-Square | 0.9260 |
| Dependent Mean | 10.57621 | Adj R-Sq | 0.9180 |
| Coeff Var | 2.22710 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 12.19150 | 0.21583 | 56.49 | <.0001 |
| X1 | 1 | -0.59800 | 0.08425 | -7.10 | <.0001 |
| X2 | 1 | -0.00598 | 0.01857 | -0.32 | 0.7480 |
| X1X2 | 1 | -0.01078 | 0.00530 | -2.03 | 0.0442 |
| X1SQ | 1 | 0.08575 | 0.00834 | 10.28 | <.0001 |
| X2SQ | 1 | 0.00014207 | 0.00037728 | 0.38 | 0.7072 |
| X3 | 1 | -0.78192 | 0.12900 | -6.06 | <.0001 |
| X4 | 1 | 0.67679 | 0.21035 | 3.22 | 0.0017 |
| X1X3 | 1 | 0.03991 | 0.03999 | 1.00 | 0.3203 |
| X1X4 | 1 | -0.27464 | 0.07267 | -3.78 | 0.0002 |
| X2X3 | 1 | -0.02094 | 0.01045 | -2.00 | 0.0473 |
| X2X4 | 1 | -0.02619 | 0.01610 | -1.63 | 0.1063 |
| X1X2X3 | 1 | -0.00332 | 0.00303 | -1.10 | 0.2757 |
| X1X2X4 | 1 | 0.01298 | 0.00544 | 2.39 | 0.0186 |

**FIGURE SIA12.6**
SAS regression printout for Model 7

Note that the terms in the equation were rearranged so that the $\beta$'s associated with the deregulation variable are shown together at the end of the equation. Because of some interactions, simply examining the signs of the $\beta$-estimates can be confusing and lead to erroneous conclusions.

A good way to assess the impact of deregulation is to hold fixed all but one of the other independent variables in the model. For example, suppose we fix the weight of the shipment at 15 thousand pounds and consider only shipments originating from Jacksonville, i.e., set $x_2 = 15$ and $x_4 = 0$. Substituting these values into the prediction equation and combining like terms yields:

$$\begin{aligned}\hat{y} = {} & 12.192 - .598x_1 - .00598(15) - .01078x_1(15) + .086x_1^2 + .00014(15)^2 \\ & + .677(0) - .275x_1(0) - .026(15)(0) + .013x_1(15)(0) \\ & - .782x_3 + .0399x_1x_3 - .021(15)x_3 - .0033x_1(15)x_3 \\ = {} & 12.134 - .760x_1 + .086x_1^2 - 1.097x_3 - .0096x_1x_3\end{aligned}$$

To see the impact of deregulation on the estimated curve relating log price ($y$) to distance shipped ($x_1$), compare the prediction equations for the two conditions, $x_3 = 0$ (regulated prices) and $x_3 = 1$ (deregulation):

$$\begin{aligned}\text{Regulated } (x_3 = 0)\text{:}\quad \hat{y} &= 12.134 - .760x_1 + .086x_1^2 - 1.097(0) - .0096x_1(0) \\ &= 12.134 - .760x_1 + .086x_1^2 \\ \text{Deregulation } (x_3 = 1)\text{:}\quad \hat{y} &= 12.134 - .760x_1 + .086x_1^2 - 1.097(1) - .0096x_1(1) \\ &= 11.037 - .7696x_1 + .086x_1^2\end{aligned}$$

Notice that the $y$-intercept for the regulated prices (12.134) is larger than the y-intercept for the deregulated prices (11.037). Also, although the equations have the same rate of curvature, the estimated shift parameters differ.

These prediction equations are portrayed graphically in the SAS printout, Figure SIA12.7. The graph clearly shows the impact of deregulation on the prices charged when the carrier leaves from Jacksonville with a cargo of 15,000 pounds. As expected from economic theory, the curve for the regulated prices lies above the curve for deregulated prices.

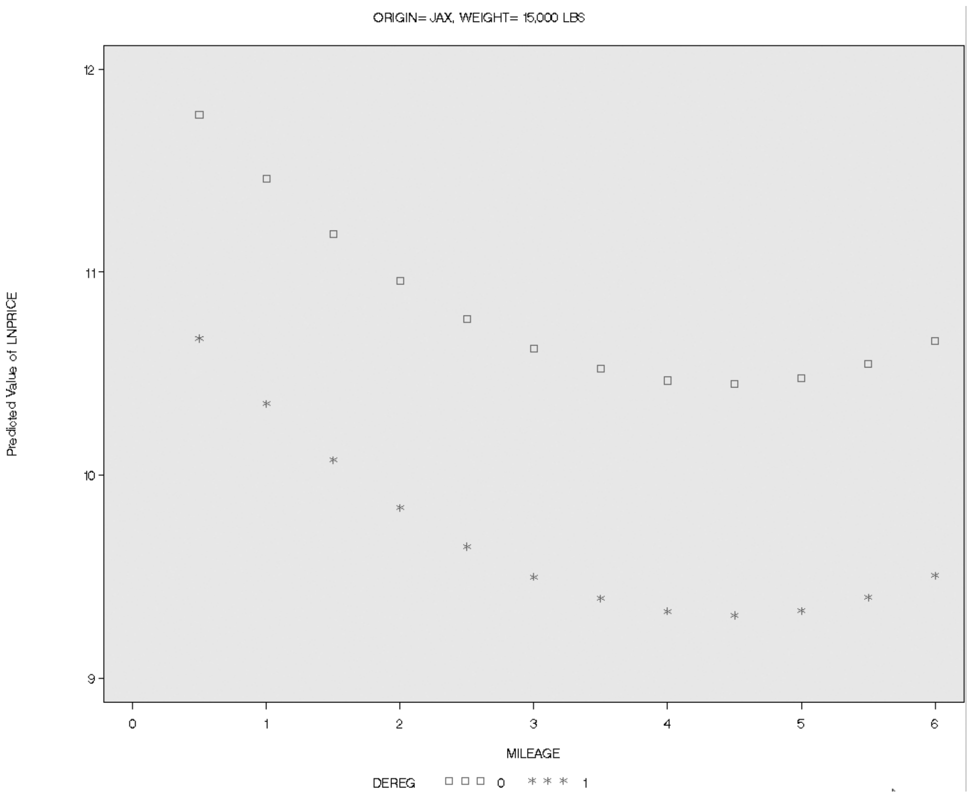



**FIGURE SIA12.7**
SAS Graph of the Prediction Equation for Log Price

# Quick Review

## Key Terms

*[Note: Starred (*) items are from the optional sections in this chapter.]*

All-possible-regressions selection procedure, 697
Base level, 668
*Coding variables, 662
Complete (nested) model, 685
Complete second-order model, 656
Contour lines, 654
*Cross-validation, 693
*Data-splitting, 693
Dummy variable, 667
Indicator variable, 667
Interaction of independent variables, 667
*Jackknife procedure, 693
Levels of a variable, 000
Main effect, 676
*Mean squared prediction error, 693
Model building, 644
*Model validation, 692
Nested models, 685
Nested-model F test, 685
*Orthogonal polynomials, 664
Paraboloid, 657
Parsimonious model, 688
Parsimony, 688
Polynomial model, 647
*Prediction sum of squares (PRESS), 694
Qualitative independent variable, 667
Quantitative independent variable, 662
Reduced (nested) model, 685
Response surface, 654
*$R^2$-prediction, 693
Saddle-shaped surface, 657
Stepwise regression, 694

## Key Formulas

| | |
|---|---|
| $E(y) = \beta_0 + \beta_1 x$ | First-order model with one quantitative $x$ 647 |
| $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ | Second-order model with one quantitative $x$ 647 |
| $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$ | $p$th-order polynomial model with one quantitative $x$ 647 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ | First-order model with $k$ quantitative $x$'s 654 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ | Interaction model with two quantitative $x$'s 655 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 (x_1)^2 + \beta_5 (x_2)^2$ | Complete second-order model with two quantitative $x$'s 656 |
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$, where $x_i = 1$ if level $i$, 0 if not | Dummy variable model with one qualitative $x$ at $k$ levels 668 |
| $F = \dfrac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(\#\beta\text{'s tested})}{\text{MSE}_\text{C}}$ | Test statistic for comparing complete and reduced nested models 686 |
| $u = (x - \bar{x})/s_x$ | Coding a quantitative variable $x$ 663 |
| $R^2_{\text{pred}} = 1 - \dfrac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2}$ | $R^2$ for cross-validation 693 |
| $\text{MSE}_{\text{pred}} = \dfrac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{m - (k+1)}$ | MSE for cross-validation 693 |
| $R^2_{\text{jackknife}} = 1 - \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$ | $R^2$ for model validation using jackknife 694 |
| $\text{MSE}_{\text{jackknife}} = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2}{n - (k+1)}$ | MSE for model validation using jackknife 694 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $SSE_R$ | S-S-E-sub-R | Sum of squared errors for reduced (nested) model |
| $SSE_C$ | S-S-E-sub-C | Sum of squared errors for complete (nested) model |
| $MSE_C$ | M-S-E-sub-C | Mean squared error for complete (nested) model |
| $R^2_{\text{pred}}$ | R-squared-predict | $R^2$ for cross-validation ($R^2$-prediction) |
| $MSE_{\text{pred}}$ | M-S-E-predict | MSE for cross-validation (Mean square prediction error) |
| $R^2_{\text{jackknife}}$ | R-squared-jackknife | $R^2$ for model validation using the jackknife method |
| $MSE_{\text{jackknife}}$ | M-S-E-jackknife | MSE for model validation using the jackknife method |
| PRESS | | Prediction sum of squares using the jackknife method |
| $\hat{y}_{(i)}$ | y-hat-sub-i | Predicted value of $y$ when $i$th observation is deleted |

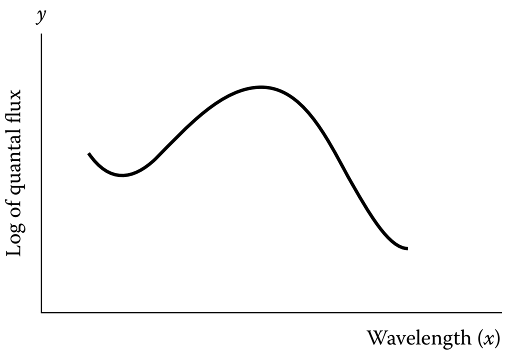
## Chapter Summary Notes

- **Model building** is a process that involves fitting and evaluating a series of models for $E(y)$, culminating in the selection of a single, "best" model.
- The different intensity settings (values) of an independent variable are called **levels**.
- Two quantitative independent variables **interact** if the change in $E(y)$ for every 1-unit change in $x_1$ depends on the value of $x_2$ held fixed.
- **Coding** a quantitative independent variable $x$ will aid in reducing the inherent multicollinearity that occurs when both $x$ and $x^2$ are in the model.
- Two models are **nested** if one model (called the **complete** model) contains all the terms of the other (**reduced**) model and at least one additional term.
- A **parsimonious model** is a model with a small number of $\beta$ parameters.
- **Dummy (indicator) variables** are used to represent qualitative independent variables in a model; for a qualitative variable with $k$ levels, there will be $(k - 1)$ dummy variables.
- **Model validation** involves an assessment of how the estimated regression model will perform when applied to new or future data.
- **Data-splitting** (or **cross-validation**) is a model validation technique that requires you to split the sample data set. One subset of data is used to estimate the model, while the other subset is used to validate the model.
- **Jackknifing** is a model validation method used when the sample size is small. The jackknife predicted value, $\hat{y}_{(i)}$, is the predicted value of $y$ obtained from a regression model fit to $(n - 1)$ data points, with the $i$th observation deleted.
- **Stepwise regression** is an objective method for screening out the least important independent variables from a lengthy list of potential independent variables.
- Two problems with using the stepwise regression model as the "final" model for predicting $y$: (1) *extremely large number of t tests are performed*, inflating the probabilities of at least one Type I error and at least one Type II error; (2) *no higher-order terms* (e.g., interaction or squared terms) are included in the final stepwise regression model.

## Supplementary Applied Exercises

12.73 *Air pollution model.* Air pollution regulations for power plants are often written so that the maximum amount of pollution that can be emitted increases as the plant's output increases. Assuming this is true, write a model relating the maximum amount of pollution permitted (in parts per million) to a plant's output (in megawatts).

12.74 *Optomotor response of frogs.* The optomotor responses of tree frogs were studied in the *Journal of Experimental Zoology* (Sept. 1993). Microspectrophotometry was used to measure the threshold quantal flux (the light intensity at which the optomotor response was first observed) of tree frogs tested at different spectral wavelengths. The data revealed the relationship between the log of quantal flux $(y)$




and wavelength ($x$) shown in the above graph. Hypothesize a model for $E(y)$ that corresponds to the graph.

12.75 *Work output study.* An engineer has proposed the following model to describe the relationship between the number of acceptable items produced per day (output) and the number of work-hours expended per day (input) in a particular production process:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

where

$y$ = Number of acceptable items produced per day

$x$ = Number of work-hours per day

A portion of the MINITAB computer printout that results from fitting this model to a sample of 25 weeks of production data is shown above (right column). Test the hypothesis that as amount of input increases, the amount of output also increases but at a decreasing rate. Do the data provide sufficient evidence to indicate that the *rate* of increase in output per unit increase of input decreases as the input increases? Test using $\alpha = .05$.

**MINITAB Output for Exercise 12.75**

```
The regression equation is
Y = -6.17 + 2.04 X1 - .0323 X2

Predictor       Coef   SE Coef       T      P
Constant      -6.173     1.666   -3.71  0.002
X1             2.036      .185   11.02  0.000
X2           -.03231    .00489   -6.60  0.000


S = 1.243         R-Sq = 95.5%   R-Sq(adj) = 95.1%


Analysis of Variance

Source          DF       SS      MS       F      P
Regression       2  718.168  359.08  232.41  0.000
Residual Error  22   33.992   1.545
Total           24  752.160
```

12.76 *Modeling gasoline sales.* Many service stations offer self-service gasoline at reduced prices when customers pay with cash rather than credit. Suppose an oil company wants to model the mean monthly gasoline sales, $E(y)$, of its affiliated stations as a function of the type of service they offer: cash only, cash or credit (same price), cash or credit (cash at reduced price).

a. How many dummy variables will be needed to describe the qualitative independent variable type of service?
b. Write the main effects model relating $E(y)$ to the type of service. Describe the coding of the dummy variables.

12.77 *A second-order polynomial.* To model the relationship between $y$, a dependent variable, and $x$, an independent variable, a researcher has taken one measurement on $y$ at each of three different $x$ values. Drawing on his mathematical expertise, the researcher realizes that he can fit the second-order polynomial model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

and it will pass exactly through all three points, yielding SSE = 0. The researcher, delighted with the "excellent" fit of the model, eagerly sets out to use it to make inferences. What problems will he encounter in attempting to make inferences?

12.78 *Deep space survey of quasars.* Refer to *The Astronomical Journal* (July 1995) study of quasars detected by a deep space survey, Exercise 11.72 (p. 638). Recall that several

**QUASAR**

(*First five quasars shown*)

| Quasar | Redshift ($x_1$) | Lineflux ($x_2$) | Line Luminosity ($x_3$) | $AB_{1450}$ ($x_4$) | Absolute Magnitude ($x_5$) | Rest Frame Equivalent Width ($y$) |
|---|---|---|---|---|---|---|
| 1 | 2.81 | −13.48 | 45.29 | 19.50 | −26.27 | 117 |
| 2 | 3.07 | −13.73 | 45.13 | 19.65 | −26.26 | 82 |
| 3 | 3.45 | −13.87 | 45.11 | 18.93 | −27.17 | 33 |
| 4 | 3.19 | −13.27 | 45.63 | 18.59 | −27.39 | 92 |
| 5 | 3.07 | −13.56 | 45.30 | 19.59 | −26.32 | 114 |

*Source*: Schmidt, M., Schneider, D.P., and Gunn, J. E. "Spectroscopic CCD surveys for quasars at large redshift." *The Astronomical Journal*, Vol. 110, No. 1, July 1995, p. 70 (Table 1).

quantitative independent variables were used to model the quasar characteristic, rest frame equivalent width ($y$). The data for 25 quasars are saved in the **QUASAR** file. (Data for the first five quasars are listed in the table on p. 711.)

a. Write a complete second-order model for $y$ as a function of redshift ($x_1$), lineflux ($x_2$), and $AB_{1450}$ ($x_4$).
b. Fit the model, part **a**, to the data using a statistical software package. Is the overall model statistically useful for predicting $y$?
c. Conduct a test to determine if any of the curvilinear terms in the model, part **a**, are statistically useful predictors of $y$.

**TRUCKING**

12.79 *Deregulation of intrastate trucking prices.* Refer to the *Statistics in Action* problem for this chapter (p. 643). Recall that we built a model for $y$ = natural logarithm of supply price for an intrastate trucking shipment as a function of the following independent variables: $x_1$ = Distance shipped (hundreds of miles), $x_2$ = Weight of product shipped (thousands of pounds), $x_3 = \{1$ if Deregulation in effect, 0 if not$\}$, and $x_4 = \{1$ if trip originates in Miami, 0 if in Jacksonville$\}$. The estimated equation for the best model of $y$ (from the SAS printout for Model 7) was:

$$
\begin{aligned}
\hat{y} = {} & 12.192 - .598x_1 - .00598x_2 \\
& -.01078x_1x_2 + .086x_1^2 + .00014x_2^2 \\
& +.677x_4 - .275x_1x_4 - .026x_2x_4 + .013x_1x_2x_4 \\
& -.782x_3 + .0399x_1x_3 - .021x_2x_3 - .0033x_1x_2x_3
\end{aligned}
$$

a. Based on the equation, give an estimate of the difference between the predicted regulated price and predicted deregulated price for any fixed value of mileage, weight and origin.
b. Demonstrate the impact of deregulation on price charged using the estimated $\beta$'s, but now hold distance fixed at 100 miles, origin fixed at Miami, and weight fixed at 10,000 pounds.
c. The data file **TRUCKING** contains data on trucking prices for four Florida carriers (A, B, C, and D). These carriers are identified by the variable CARRIER. (Note: Carrier B is the carrier analyzed in the SIA.) Using Model 7 as a base model, add terms that allow for different response curves for the four carriers. Conduct the appropriate test to determine if the curves differ.

12.80 *Performance of a diesel engine.* An experiment was conducted to evaluate the performances of a diesel engine run on synthetic (coal-derived) and petroleum-derived fuel oil (*Journal of Energy Resources Technology*, Mar. 1990). The petroleum-derived fuel used was a number 2 diesel fuel (DF-2) obtained from Phillips Chemical Company. Two synthetic fuels were used: a blended fuel (50% coal-derived and 50% DF-2) and a blended fuel with advanced timing. The brake power (kilowatts) and fuel type were varied in test runs, and engine performance was measured. The following table gives the experimental results for the performance measure, mass burning rate per degree of crank angle.

**SYNFUELS**

| Brake Power, $x_1$ | Fuel Type | Mass Burning Rate, $y$ |
|---|---|---|
| 4 | DF-2 | 13.2 |
| 4 | Blended | 17.5 |
| 4 | Advanced Timing | 17.5 |
| 6 | DF-2 | 26.1 |
| 6 | Blended | 32.7 |
| 6 | Advanced Timing | 43.5 |
| 8 | DF-2 | 25.9 |
| 8 | Blended | 46.3 |
| 8 | Advanced Timing | 45.6 |
| 10 | DF-2 | 30.7 |
| 10 | Blended | 50.8 |
| 10 | Advanced Timing | 68.9 |
| 12 | DF-2 | 32.3 |
| 12 | Blended | 57.1 |

*Source*: Litzinger, T. A., and Buzza, T. G. "Performance and emissions of a diesel engine using a coal-derived fuel." *Journal of Energy Resources Technology*, Vol. 112, Mar. 1990, p. 32, Table 3.

The researchers fit the interaction model

$$
\begin{aligned}
E(y) = {} & \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 \\
& + \beta_4x_1x_2 + \beta_5x_1x_3
\end{aligned}
$$

where

$y$ = Mass burning rate

$x_1$ = Brake power (kW)

$$x_2 = \begin{cases} 1 & \text{DF-2 fuel} \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if blended fuel} \\ 0 & \text{if not} \end{cases}$$

The results are shown in the SAS printout on p. 713.

a. Conduct a test to determine whether brake power and fuel type interact. Test using $\alpha = .01$.
b. Refer to the model, part **a**. Give the estimates of the slope of the $y$–$x_1$ line for each of the three fuel types.

12.81 *Passive solar retrofit.* A 40-year-old masonry duplex structure has recently undergone a passive solar retrofit with features including insulated exterior walls, heat distribution systems, storm sashes, and air-lock entries. To gauge the effectiveness of the improvements, architectural engineers monitored the winter energy usage of the structure for 2 years prior to the retrofit and for 2 years after the retrofit. The engineers want to use the data to fit a regression model relating monthly energy usage $y$ (therms per billing day) to weather intensity $x_1$ (ddh/bd) and $x_2$, where

$$x_2 = \begin{cases} 1 & \text{if prior to retrofit} \\ 0 & \text{if after retrofit} \end{cases}$$

## SAS Output for Exercise 12.80

```
                          Dependent Variable: BURNRATE

                               Analysis of Variance

                                     Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       5    3253.97929      650.79586      25.65    <.0001
Error                       8     203.01000       25.37625
Corrected Total            13    3456.98929


              Root MSE              5.03748    R-Square     0.9413
              Dependent Mean       36.29286    Adj R-Sq     0.9046
              Coeff Var            13.88010

                               Parameter Estimates

                           Parameter       Standard
     Variable      DF       Estimate          Error    t Value    Pr > |t|

     Intercept      1      -10.83000        8.27743      -1.31      0.2271
     POWER          1        7.81500        1.12642       6.94      0.0001
     X2             1       19.35000       10.68612       1.81      0.1078
     X3             1       12.79000       10.68612       1.20      0.2656
     POWERX2        1       -5.67500        1.37957      -4.11      0.0034
     POWERX3        1       -2.95000        1.37957      -2.14      0.0649


          Test INTERACT Results for Dependent Variable BURNRATE

                                          Mean
        Source              DF          Square    F Value    Pr > F

        Numerator            2       223.03750       8.79    0.0096
        Denominator          8        25.37625
```

a. Write the complete second-order model for $E(y)$.
b. Graph the contour lines for the model of part **a**.
c. Hypothesize a first-order model that allows for a constant difference between the mean monthly usage prior to and after the retrofit at different levels of weather intensity.
d. Graph the contour lines for the model of part **c**.

**DDT**

12.82 *Study of contaminated fish.* Refer to Exercise 11.26 (p. 591) and the model relating the mean DDT level $E(y)$ of contaminated fish to $x_1$ = miles captured upstream, $x_2$ = length, and $x_3$ = weight. Now consider a model for $E(y)$ as a function of both weight and species (channel catfish, large-mouth bass, and small-mouth buffalofish).

a. Set up the appropriate dummy variables for species.
b. Write the equation of a model that proposes parallel straight-line relationships between mean DDT level $E(y)$ and weight, one line for each species.
c. Write the equation of a model that proposes nonparallel straight-line relationships between mean DDT level $E(y)$ and weight, one line for each species.
d. Fit the model, part **b**, to the data saved in the **DDT** file. Give the least-squares prediction equation.
e. Refer to part **d**. Interpret the value of the least-squares estimate of the beta coefficient multiplied by weight.
f. Fit the model, part **c**, to the data saved in the **DDT** file. Give the least-squares prediction equation.
g. Refer to part **f**. Find the estimated slope of the line relating DDT level $(y)$ to weight for the channel catfish species.

12.83 *Work safety study.* A company is studying three different safety programs, A, B, and C, in an attempt to reduce the number of work-hours lost because of accidents. Each program is to be tried at three of the company's nine factories. The plan is to monitor the lost work-hours, $y$, for a 1-year period beginning 6 months after the new safety program is instituted.

a. Write a main effects model relating $E(y)$ to the lost work-hours, $x_1$, the year before the plan is instituted and to the type of program that is instituted.
b. In terms of the model parameters from part **a**, what hypothesis would you test to determine whether the mean work-hours lost differ for the three safety programs?
c. After the three safety programs have been in effect for 18 months, the complete main effects model is fit to the

$n = 9$ data points. Using safety program A as the base level, the following results are obtained:

$$\hat{y} = -2.1 + .88x_1 - 150x_2 + 35x_3$$
$$\text{SSE} = 1{,}527.27$$

Then the reduced model $E(y) = \beta_0 + \beta_1 x_1$ is fit, with the result

$$\hat{y} = 15.3 + .84x_1 \qquad \text{SSE} = 3{,}113.14$$

Test to determine whether the mean work-hours lost differ for the three programs. Use $\alpha = .05$.

12.84 *Density of mosquito larvae.* A field experiment was conducted to assess the effect of organic enrichment on the mean density of mosquito larvae. (*Journal of the American Mosquito Control Association*, June 1995.) Larval specimens were collected from a pond 3 days after the pond was flooded with canal water. A second sample of specimens was collected 3 weeks after flooding and enriching the pond with rabbit pellets. All specimens were returned to the laboratory and the number $y$ of mosquito larvae counted in each specimen.

a. Write a model that will allow you to compare the mean number of mosquito larvae found in the enriched pond to the corresponding mean for the natural pond.
b. Interpret the $\beta$ coefficients in the model, part **a**.
c. Set up the null and alternative hypotheses for testing whether the mean larval density for the enriched pond exceeds the mean for the natural pond.
d. The $p$-value associated with the global $F$ test for the model, part **a**, was determined to be .004. Interpret this result.

12.85 *Walking study.* Refer to the *American Scientist* (Jul–Aug. 1998) study of the relationship between number of self-avoiding and unrooted walks, Exercise 11.71 (p. 638). The data for the analysis are repeated in the accompanying table.

**WALK**

| Walk Length | Number of Walks | |
|---|---|---|
| (number of steps) | Unrooted | Self-Avoiding |
| 1 | 1 | 4 |
| 2 | 2 | 12 |
| 3 | 4 | 36 |
| 4 | 9 | 100 |
| 5 | 22 | 284 |
| 6 | 56 | 780 |
| 7 | 147 | 2172 |
| 8 | 388 | 5916 |

*Source*: Hayes, B. "How to avoid yourself." *American Scientist*, Vol. 86, No. 4, Jul–Aug. 1998 (Figure 5).

a. Give the equation relating the coded variable $u$ to walk length, $x$, using the coding system for observational data.
b. Calculate the coded values, $u$.
c. Calculate the coefficient of correlation $r$ between the variables $x$ and $x^2$.
d. Calculate the coefficient of correlation $r$ between the variables $u$ and $u^2$. Compare this value to the value computed in part **c**.
e. Let $y$ = number of unrooted walks. Fit the model

$$E(y) = \beta_0 + \beta_1 u + \beta_2 u^2$$

using available statistical software. Interpret the result.

12.86 *Yields of orange juice extractors.* The Florida Citrus Commission is interested in evaluating the performance of two orange juice extractors, brand A and brand B. It is believed that the size of the fruit used in the test may influence the juice yield (amount of juice per pound of oranges) obtained by the extractors. The commission wants to find a regression model relating the mean yield, $E(y)$, to the type of orange juice extractor (brand A or brand B) and the size of orange (diameter), $x_1$.

a. Identify the independent variables as qualitative or quantitative.
b. Write a model that describes the relationship between $E(y)$ and size of orange as two parallel lines, one for each brand of extractor.
c. Modify the model of part **b** to permit the slopes of the two lines to differ.
d. Sketch typical response lines for the model of part **b**. Do the same for the model of part **c**. Label your graphs carefully.
e. Specify the null and alternative hypotheses you would employ to determine whether the model of part **c** provides more information for predicting yield than does the model of part **b**.

12.87 *Modeling dissolved oxygen in water.* The dissolved oxygen content, $y$, in rivers and streams is related to the amount, $x_1$, of nitrogen compounds per liter of water and the temperature, $x_2$, of the water. Write the complete second-order model relating $E(y)$ to $x_1$ and $x_2$.

12.88 *Profitability of airlines.* When the U.S. airline industry was deregulated, researchers have questioned whether the deregulation has ensured a truly competitive environment. If so, the profitability of any major airline would be related only to overall industry conditions (e.g, disposable income and market share) but not to any unchanging feature of that airline. This profitability hypothesis was tested using multiple regression (*Transportation Journal*, Winter 1990). Data for $n = 234$ carrier-years were used to fit the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{30} x_{30}$$

where

$y$ = Profit rate

$x_1$ = Real personal disposable income

$x_2$ = Industry market share

$x_3 - x_{30}$ = Dummy variables (coded 0–1) for the 29 air carriers investigated in the study

The results of the regression are summarized in the table. Interpret the results. Is the profitability hypothesis supported?

| Variable | $\beta$ estimate | $t$ value | $p$-value |
|---|---|---|---|
| Intercept | 1.2642 | .09 | .9266 |
| $x_1$ | $-.0022$ | $-.99$ | .8392 |
| $x_2$ | 4.8405 | 3.57 | .0003 |
| $x_3 - x_{30}$ | (not given) | — | — |
| $R^2 = .3402$ | $F$(full model) = 3.49 | $p$-value = .0001 | |
| $F$ (for testing carrier dummies) = 3.59 | | $p$-value = .0001 | |

*Source*: Leigh, L. E. "Contestability in deregulated airline markets: Some empirical tests." *Transportation Journal*, Winter 1990, p. 55 (Table 4). Reprinted from the Winter 1990 issue of *Transportation Journal* with the express permission of the publisher, the American Society of Transportation and Logistics, Inc., for educational purposes only.

12.89 *Halogen lightbulb operation.* A firm has developed a new type of halogen lightbulb and is interested in evaluating its performance to decide whether to market the bulb. It is known that the level of light output of the bulb depends on the cleanliness of its surface area and the length of time the bulb has been in operation. The data are presented in the accompanying table. Use these data and the procedures you learned in this chapter to build a regression model that relates drop in light output to bulb surface cleanliness and length of operation.

**HALOGEN**

| Drop in Light Output percent original output | Bulb Surface C = Clean, D = Dirty | Length of Operation hours |
|---|---|---|
| 0 | C | 0 |
| 16 | C | 400 |
| 22 | C | 800 |
| 27 | C | 1,200 |
| 32 | C | 1,600 |
| 36 | C | 2,000 |
| 38 | C | 2,400 |
| 0 | D | 0 |
| 4 | D | 400 |
| 6 | D | 800 |
| 8 | D | 1,200 |
| 9 | D | 1,600 |
| 11 | D | 2,000 |
| 12 | D | 2,400 |

CHAPTER 13

# Principles of Experimental Design

## OBJECTIVE

*To present an overview of experiments designed to compare two or more population means; to explain the statistical principles of experimental design*

## CONTENTS

## *STATISTICS IN ACTION*

### Anti-corrosive Behavior of Epoxy Coatings Augmented with Zinc

Organic coatings that use epoxy resins are widely used for protecting steel and metal against weathering and corrosion. The anti-corrosion performance of a coating depends on several factors, including the characteristics of the coating system. The recent trend has been to use epoxy coatings that contain zinc dust or zinc phosphate. These zinc-augmented epoxy coatings are believed to offer the best corrosion inhibition available.

Researchers at the Department of Materials Science and Engineering, National Technical University (Athens, Greece) examined the steel anticorrosive behavior of different epoxy coatings formulated with zinc pigments in an attempt to find the epoxy coating with the best corrosion inhibition. (*Pigment & Resin Technology,* Vol. 32, 2003.) The experimental materials were flat, rectangular panels cut from steel sheets with the following composition: iron, 99.7%; carbon, .063%; manganese, .022%, phosphorous, .009%; and, sulfur, .007%. Each panel was coated with one of four different coating systems, S1, S2, S3, and S4. Three panels were prepared for each coating system. (These panels are labeled, S1-A, S1-B, S1-C, S2-A, S2-B, …, S4-C.) The characteristics of the four coating systems are listed in Table SIA13.1.

**TABLE SIA13.1 Characteristics of Four Epoxy Coating Systems**

| Coating System | 1st Layer | 2nd Layer |
|---|---|---|
| S1 | Zinc dust | Epoxy paint, 100 micro-meters thick |
| S2 | Zinc phosphate | Epoxy paint, 100 micro-meters thick |
| S3 | Zinc phosphate with mica | Finish layer, 100 micro-meters thick |
| S4 | Zinc phosphate with mica | Finish layer, 200 micro-meters thick |

Each coated panel was immersed in de-ionized and de-aerated water and then tested for corrosion. Since exposure time is likely to have a strong influence on anti-corrosive behavior, the researchers attempted to remove this extraneous source of variation through experimental design. Exposure times were fixed at 24 hours, 60 days, and 120 days. For each of the coating systems, one panel was exposed to water for 24 hours, one exposed to water for 60 days, and one exposed to water for 120 days in random order. The design is illustrated in Figure SIA13.1.

**FIGURE SIA13.1**
Diagram of the Experimental Design

| *Exposure Time* | *Coating system/panel exposed* |
|---|---|
| 24 Hours | S1-A, S2-C, S3-C, S4-B |
| 60 Days | S1-C, S2-A, S3-B, S4-A |
| 120 Days | S1-B, S2-B, S3-A, S4-C |

Following exposing, the corrosion rate (nanoamperes per square centimeter) was determined for each panel. The lower the corrosion rate, the greater the anti-corrosion performance of the coating system. The objective was to compare the mean corrosion rates of the four epoxy coating systems (S1, S2, S3, and S4).

In the *Statistics in Action Revisited* example at the end of this chapter, we apply the methodology of chapter to determine if the mean corrosion rates differ for the four coating systems.

## 13.1 Introduction

In Chapters 11–12, we learned how to analyze multivariable sample data using a multiple regression analysis. The data for regression can be collected **observationally** (where the values of the independent variables are observed in their natural setting) or **experimentally** (where the values of the $x$'s are controlled, i.e., set in advance). With observational data, however, there is a caveat: *A statistically significant relationship between a response y and a predictor x does not necessarily imply a cause-and-effect relationship!* Since the values of other relevant independent variables—both those in the model and those omitted from the model—are uncontrolled, we are unsure whether it is these other variables or $x$ that is causing the increase (or decrease) in $y$.

To illustrate, a Department of Transportation engineer is interested in modeling the price, $y$, of a road contract, where the price is determined by the lowest bidder in a sealed bid process. Suppose that the engineer finds that number of bidders, $x$, is negatively related to $y$ in the first-order model, $E(y) = \beta_0 + \beta_1 x$ and the relationship is statistically significant. Does this imply that when fewer contractors bid on the road contract, the contract price will always be higher? Not necessarily so. The engineer's knowledge of the bidding process reveals that more contractors tend to bid on longer roads. And the low-bid prices for these longer-road contracts will obviously be larger than the low-bid prices for short-road contracts. In other words, an unmeasured variable, length of road, is causing both $y$ and $x$ to change.

This caveat can be overcome by controlling the values of all the relevant $x$'s via a planned experiment. With experimental data, we usually select the $x$'s so that we can compare the mean responses, $E(y)$, for several different combinations of the $x$ values.

The procedure for selecting sample data with the $x$'s set in advance is called the **design of the experiment**. The statistical procedure for comparing the population means is called an **analysis of variance**. The objective of this chapter is to introduce some key aspects of experimental design. The analysis of the data from such experiments using an analysis of variance is the topic of Chapter 14.

## 13.2 Experimental Design Terminology

The study of experimental design originated with R. A. Fisher in the early 1900s in England. During these early years, it was associated solely with agricultural experimentation. The need for experimental design in agriculture was very clear: It takes a full year to obtain a single observation on the yield of a new variety of most crops. Consequently, the need to save time and money led to a study of ways to obtain more information using smaller samples. Similar motivations led to its subsequent acceptance and wide use in all fields of scientific experimentation. Despite this fact, the terminology associated with experimental design clearly indicates its early association with the biological sciences.

We will call the process of collecting sample data an **experiment** and the (*dependent*) variable to be measured, the **response** $y$. The planning of the sampling procedure is called the **design** of the experiment. The object upon which the response measurement $y$ is taken is called an **experimental unit**.

**Definition 13.1**
The process of collecting sample data is called an **experiment**.

**Definition 13.2**
The plan for collecting the sample is called the **design of the experiment**.

**Definition 13.3**
The variable measured in the experiment is called the **response variable**.

**Definition 13.4**
The object upon which the response $y$ is measured is called an **experimental unit**.

*Independent* variables that may be related to a response variable $y$ are called **factors**. The value—that is, the intensity setting—assumed by a factor in an experiment is called a **level**. The combinations of levels of the factors for which the response will be observed are called **treatments**.

**Definition 13.5**
The independent variables, quantitative or qualitative, that are related to a response variable $y$ are called **factors**.

**Definition 13.6**
The intensity setting of a factor (i.e., the value assumed by a factor in an experiment) is called a **level**.

**Definition 13.7**
A **treatment** is a particular combination of levels of the factors involved in an experiment.

## Example 13.1

Designed Experiment
Plastic Hardness Study

An experiment is conducted to determine the effects of pressure and temperature on the hardness of a new type of plastic, where hardness is rated on a scale of 1 (very soft) to 10 (very hard). At the time of molding, the pressure will be set at 200, 300, or 400 pounds per square inch (psi), while the temperature will be set at either 200 or 300 degrees Fahrenheit (°F). Three plastic molds were randomly assigned to each of the $3 \times 2 = 6$ combinations of pressure and temperature, and the hardness rating of each mold was measured. A layout of the design is illustrated in Figure 13.1. For this experiment, identify

a. The experimental unit
b. The response, $y$
c. The factors
d. The factor levels
e. The treatments.

Solution

a. Since measurements are made on the 18 plastic molds, the experimental unit is a plastic mold.
b. The response variable of interest, i.e., the variable measured after the molds are randomly assigned, is $y$ = plastic hardness level. Note that $y$ is a quantitative variable. For the types of designs studied in this text, the response will always be a quantitative variable.
c. Since the objective of the experiment is to investigate the effect of both pressure and temperature on plastic hardness, pressure and temperature are the factors.

| Temperature | Pressure: 200 psi | 300 psi | 400 psi |
|---|---|---|---|
| 200°F | Mold 1<br>Mold 9<br>Mold 14 | Mold 2<br>Mold 7<br>Mold 16 | Mold 4<br>Mold 12<br>Mold 17 |
| 300°F | Mold 5<br>Mold 10<br>Mold 13 | Mold 3<br>Mold 8<br>Mold 18 | Mold 6<br>Mold 11<br>Mold 15 |

**FIGURE 13.1**
Layout for designed experiment of Example 13.1

d. For this experiment, pressure is set at three levels (200, 300, and 400 psi) and temperature is set at two levels (200°F and 300°F).

e. A treatment is a combination of factor levels. For this experiment, there are $3 \times 2 = 6$ treatments, or pressure-temperature combinations, as shown in Figure 13.1: (200 psi, 200°F), (200 psi, 300°F), (300 psi, 200°F), (300 psi, 300°F), (400 psi, 200°F), and (400 psi, 300°F).

Now that you understand some of the terminology, it is helpful to think of the design of an experiment in four steps.

**Designing an Experiment**

*Step 1* Select the factors to be included in the experiment, and identify the parameters that are the object of the study. Usually, the target parameters are the population means associated with the factor level combinations (i.e., treatment).

*Step 2* Choose the treatments (the factor level combinations to be included in the experiment).

*Step 3* Determine the number of observations (sample size) to be made for each treatment. [This will usually depend on the standard error(s) that you desire.]

*Step 4* Plan how the treatments will be assigned to the experimental units. That is, decide on which design to use.

By following these steps, you can control the quantity of information in an experiment. We shall explain how this is done in Section 13.3.

## 13.3 Controlling the Information in an Experiment

The problem of acquiring good experimental data is analogous to the problem faced by a communications engineer. The receipt of any signal, verbal or otherwise, depends on the volume of the signal and the amount of background noise. The greater the volume of the signal, the greater will be the amount of information transmitted to the receiver. Conversely, the amount of information transmitted is reduced when the background noise is great. These intuitive thoughts about the factors that affect the information in an experiment are supported by the following fact: The standard errors of most estimators of the target parameters are proportional to $\sigma$ (a measure of data variation or noise) and inversely proportional to the sample size (a measure of the volume of the signal). To illustrate, take the simple case where we wish to estimate a population mean $\mu$ by the sample mean $\bar{y}$. The standard error of the sampling distribution of $\bar{y}$ is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \qquad \text{(see Section 6.9)}$$

For a fixed sample size $n$, the smaller the value of $\sigma$, which measures the **variability** (**noise**) in the population of measurements, the smaller will be the standard error $\sigma_{\bar{y}}$. Similarly, by increasing the sample size $n$ (**volume of the signal**) in a given experiment, you decrease $\sigma_{\bar{y}}$.

The first three steps in the design of an experiment—selecting the factors and treatments to be included in an experiment and specifying the sample sizes—determine the volume of the signal. You must select the treatments so that the observed values of $y$ provide information on the parameters of interest. Then the larger

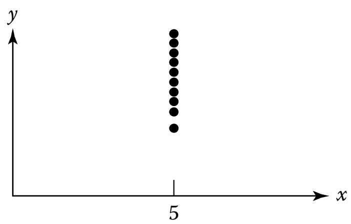



**FIGURE 13.2**
Data set with $n = 10$ responses, all at $x = 5$

the treatment sample sizes, the greater will be the quantity of information in the experiment. We present an example of a volume-increasing experiment in Section 13.5.

Is it possible to observe $y$ and obtain no information on a parameter of interest? The answer is yes. To illustrate, suppose that you attempt to fit a first-order model

$$E(y) = \beta_0 + \beta_1 x$$

to a set of $n = 10$ data points, all of which were observed for a single value of $x$, say, $x = 5$. The data points might appear as shown in Figure 13.2. Clearly, there is no possibility of fitting a line to these data points. The only way to obtain information on $\beta_0$ and $\beta_1$ is to observe $y$ for *different* values of $x$. Consequently, the $n = 10$ data points in this example contain absolutely no information on the parameters $\beta_0$ and $\beta_1$.

Step 4 in the design of an experiment provides an opportunity to reduce the noise (or experimental error) in an experiment. As we illustrate in Section 13.4, known sources of data variation can be reduced or eliminated by **blocking**—that is, observing all treatments within relatively homogeneous **blocks** of experimental material. When the treatments are compared within each block, any background noise produced by the block is canceled, or eliminated, allowing us to obtain better estimates of treatment differences.

**Summary of Steps in Experimental Design**

| | |
|---|---|
| *Volume-increasing:* | 1. Select the factors. |
| | 2. Choose the treatments (factor level combinations). |
| | 3. Determine the sample size for each treatment. |
| *Noise-reducing:* | 4. Assign the treatments to the experimental units. |

In summary, it is useful to think of experimental designs as being either "noise reducers" or "volume increasers." We will learn, however, that most designs are multifunctional. That is, they tend to both reduce the noise and increase the volume of the signal at the same time. Nevertheless, we will find that specific designs lean heavily toward one or the other objective.

## 13.4 Noise-Reducing Designs

Noise reduction in an experimental design, i.e., the removal of extraneous experimental variation, can be accomplished by an appropriate assignment of treatments to the experimental units. The idea is to compare treatments within blocks of relatively homogeneous experimental units. The most common of this type is called a **randomized block design**.

To illustrate, suppose we want to compare the mean length of time required to assemble a digital watch using three different methods of assembly, A, B, and C. Thus, we want to compare the three means $\mu_A$, $\mu_B$, and $\mu_C$, where $\mu_i$ is the mean assembly time for method $i$. One way to design the experiment is to select 15 workers (where the workers are the experimental units) and randomly assign one of the three assembly methods (treatments) to each worker. A diagram of this design, called a **completely randomized design** (since the treatments are randomly assigned to the experimental units) is shown in Table 13.1.

**Definition 13.8**

A **completely randomized design** to compare $p$ treatments is one in which the treatments are randomly assigned to the experimental units.

This design has the obvious disadvantage that the assembly times would vary greatly from worker to worker depending on manual dexterity, experience, etc. A better

**TABLE 13.1 Completely Randomized Design with $p = 3$ Treatments**

| Worker | Treatment (Method) Assigned |
|---|---|
| 1 | B |
| 2 | A |
| 3 | B |
| 4 | C |
| 5 | C |
| 6 | A |
| 7 | B |
| 8 | C |
| 9 | A |
| 10 | A |
| 11 | C |
| 12 | A |
| 13 | B |
| 14 | C |
| 15 | B |

| Blocks (Workers) | Treatments (Methods) | | |
|---|---|---|---|
| 1 | B | A | C |
| 2 | A | C | B |
| 3 | B | C | A |
| 4 | A | B | C |
| 5 | A | C | B |

**FIGURE 13.3**
Diagram for a randomized block design containing $b = 5$ blocks and $p = 3$ treatments

design—one that contains more information on the mean assembly times—would be to use only five workers and have each worker assemble three digital watches using each of the three methods. This *randomized block* procedure acknowledges the fact that the length of time required to assemble a watch varies substantially from worker to worker. By comparing the three assembly times for each worker, we eliminate worker-to-worker variation from the comparison.

The randomized block design that we have just described is shown diagrammatically in Figure 13.3. The figure shows that there are five workers. Each worker can be viewed as a **block** of three experimental units—watches assembled—one corresponding to the use of each of the assembly methods, A, B, and C. The blocks are said to be **randomized** because the treatments (assembly methods) are randomly assigned to the experimental units within a block. For our example, the watches would be assembled in a random order to avoid bias introduced by other unknown and unmeasured variables that may affect the assembly time.

In general, a randomized block design to compare $p$ treatments will contain $b$ relatively homogeneous blocks, with each block containing $p$ experimental units. Each treatment appears once in every block with the $p$ treatments randomly assigned to the experimental units within each block.

**Definition 13.9**

A **randomized block design** to compare $p$ treatments involves $b$ blocks, each containing $p$ relatively homogeneous experimental units. The $p$ treatments are randomly assigned to the experimental units within each block, with one experimental unit assigned per treatment.

**Example 13.2**

Noise-Reducing Design: Engineer Cost Estimation

Suppose you want to compare the abilities of four Department of Transportation (DOT) engineers, A, B, C, D, to estimate the cost of road construction contracts. One way to make the comparison would be to randomly allocate a number of road contracts—say, 40—10 to each of the four DOT engineers. Each engineer would then estimate the cost $y$ of each contract. The treatment allocation to experimental units that we have described is a completely randomized design.

a. Discuss the problems that result when the completely randomized design is used for this experiment.
b. Explain how you would employ a randomized block design.

Solution

a. The problem with using a completely randomized design for the DOT experiment is that comparison of mean construction costs will be influenced by the nature of the road contracts. Some contracts will be easier to estimate than others, and the variation in costs that can be attributed to this fact will make it more difficult to compare treatment means.

| Contracts (Blocks) | Engineers (Treatments) | | | |
|---|---|---|---|---|
| 1 | A | C | D | B |
| 2 | B | C | A | D |
| 3 | D | A | B | C |
| ⋮ | | ⋮ | | |
| 10 | B | D | C | A |

**FIGURE 13.4**
Diagram for a randomized block design: Example 13.2

b. To eliminate contract-to-contract variability in comparing mean engineers' estimates, you would select only 10 road contracts and require each DOT engineer to estimate the cost of each of the 10 contracts. Although in this case there is probably no need for randomization, it might be desirable to randomly assign the order (in time) of the estimates. This randomized block design, consisting of $p = 4$ treatments and $b = 10$ blocks would appear as shown in Figure 13.4.

Each experimental design can be represented by a multiple regression model relating the response $y$ to the factors (treatments, blocks, etc.) in the experiment. When the factors are qualitative in nature (as is often the case), the model includes dummy variables. For example, consider the completely randomized design portrayed in Table 13.1. Since the experiment involves three treatments (methods), we require two dummy variables. The model for this completely randomized design would appear as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

$$x_1 = \begin{cases} 1 & \text{if method A} \\ 0 & \text{if not} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if method B} \\ 0 & \text{if not} \end{cases}$$

Note that we have arbitrarily selected method C as the base level. From our discussion of dummy-variable models in Chapter 12 we know that the mean responses for the three methods are

$$\mu_A = \beta_0 + \beta_1$$
$$\mu_B = \beta_0 + \beta_2$$
$$\mu_C = \beta_0$$

Recall that $\beta_1 = \mu_A - \mu_C$ and $\beta_2 = \mu_B - \mu_C$. Thus, to estimate the differences between the treatment means, we require estimates of $\beta_1$ and $\beta_2$.

Similarly, we can write the model for the randomized block design in Figure 13.3 as follows:

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Treatments effects}} + \underbrace{\beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6}_{\text{Block effects}} + \varepsilon$$

**TABLE 13.2 The Response for the Randomized Block Design Shown in Figure 13.3**

| | *Treatments (Methods)* | | |
|---|---|---|---|
| *Blocks (Workers)* | $A(x_1 = 1, x_2 = 0)$ | $B(x_1 = 0, x_2 = 1)$ | $C(x_1 = 0, x_2 = 0)$ |
| 1 $(x_3 = 1, x_4 = x_5 = x_6 = 0)$ | $y_{A1} = \beta_0 + \beta_1 + \beta_3 + \varepsilon_{A1}$ | $y_{B1} = \beta_0 + \beta_2 + \beta_3 + \varepsilon_{B1}$ | $y_{C1} = \beta_0 + \beta_3 + \varepsilon_{C1}$ |
| 2 $(x_4 = 1, x_3 = x_5 = x_6 = 0)$ | $y_{A2} = \beta_0 + \beta_1 + \beta_4 + \varepsilon_{A2}$ | $y_{B2} = \beta_0 + \beta_2 + \beta_4 + \varepsilon_{B2}$ | $y_{C2} = \beta_0 + \beta_4 + \varepsilon_{C2}$ |
| 3 $(x_5 = 1, x_3 = x_4 = x_6 = 0)$ | $y_{A3} = \beta_0 + \beta_1 + \beta_5 + \varepsilon_{A3}$ | $y_{B3} = \beta_0 + \beta_2 + \beta_5 + \varepsilon_{B3}$ | $y_{C3} = \beta_0 + \beta_5 + \varepsilon_{C3}$ |
| 4 $(x_6 = 1, x_3 = x_4 = x_5 = 0)$ | $y_{A4} = \beta_0 + \beta_1 + \beta_6 + \varepsilon_{A4}$ | $y_{B4} = \beta_0 + \beta_2 + \beta_6 + \varepsilon_{B4}$ | $y_{C4} = \beta_0 + \beta_6 + \varepsilon_{C4}$ |
| 5 $(x_3 = x_4 = x_5 = x_6 = 0)$ | $y_{A5} = \beta_0 + \beta_1 + \varepsilon_{A5}$ | $y_{B5} = \beta_0 + \beta_2 + \varepsilon_{B5}$ | $y_{C5} = \beta_0 + \varepsilon_{C5}$ |

where

$$x_1 = \begin{cases} 1 & \text{if method A} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if method B} \\ 0 & \text{if not} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if worker 1} \\ 0 & \text{if not} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if worker 2} \\ 0 & \text{if not} \end{cases} \quad x_5 = \begin{cases} 1 & \text{if worker 3} \\ 0 & \text{if not} \end{cases} \quad x_6 = \begin{cases} 1 & \text{if worker 4} \\ 0 & \text{if not} \end{cases}$$

In addition to the treatment terms, the model includes four dummy variables representing the five blocks (workers). Note that we have arbitrarily selected worker 5 as the base level. Using this model, we can write each response $y$ in the experiment of Figure 13.3 as a function of $\beta$'s, as shown in Table 13.2.

For example, to obtain the model for the response $y$ for method A and worker 1 (denoted $y_{A1}$), we substitute $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 0$, $x_5 = 0$ and $x_6 = 0$ into the equation. The resulting model is

$$y_{A1} = \beta_0 + \beta_1 + \beta_3 + \varepsilon_{A1}$$

Now we will use Table 13.2 to illustrate how a randomized block design reduces experimental noise.

Since each treatment appears in each of the five blocks, there are five measured responses per treatment. Averaging the five responses for treatment A shown in Table 13.2, we obtain

$$\begin{aligned}
\bar{y}_A &= \frac{y_{A1} + y_{A2} + y_{A3} + y_{A4} + y_{A5}}{5} \\
&= [(\beta_0 + \beta_1 + \beta_3 + \varepsilon_{A1}) + (\beta_0 + \beta_1 + \beta_4 + \varepsilon_{A2}) \\
&\quad + (\beta_0 + \beta_1 + \beta_5 + \varepsilon_{A3}) + (\beta_0 + \beta_1 + \beta_6 + \varepsilon_{A4}) \\
&\quad + (\beta_0 + \beta_1 + \varepsilon_{A5})]/5 \\
&= \frac{5\beta_0 + 5\beta_1 + (\beta_3 + \beta_4 + \beta_5 + \beta_6) + (\varepsilon_{A1} + \varepsilon_{A2} + \varepsilon_{A3} + \varepsilon_{A4} + \varepsilon_{A5})}{5} \\
&= \beta_0 + \beta_1 + \frac{(\beta_3 + \beta_4 + \beta_5 + \beta_6)}{5} + \bar{\varepsilon}_A
\end{aligned}$$

Similarly, the mean responses for treatments B and C are obtained:

$$\begin{aligned}
\bar{y}_B &= \frac{y_{B1} + y_{B2} + y_{B3} + y_{B4} + y_{B5}}{5} \\
&= \beta_0 + \beta_2 + \frac{(\beta_3 + \beta_4 + \beta_5 + \beta_6)}{5} + \bar{\varepsilon}_B
\end{aligned}$$

$$\bar{y}_C = \frac{y_{C1} + y_{C2} + y_{C3} + y_{C4} + y_{C5}}{5}$$
$$= \beta_0 + \frac{(\beta_3 + \beta_4 + \beta_5 + \beta_6)}{5} + \bar{\varepsilon}_C$$

Since the objective is to compare treatment means, we are interested in the differences $\bar{y}_A - \bar{y}_B$, $\bar{y}_A - \bar{y}_C$, and $\bar{y}_B - \bar{y}_C$. These differences are calculated as follows:

$$\begin{aligned}
\bar{y}_A - \bar{y}_B &= [\beta_0 + \beta_1 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_A] \\
&\quad - [\beta_0 + \beta_2 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_B] \\
&= (\beta_1 - \beta_2) + (\bar{\varepsilon}_A - \bar{\varepsilon}_B) \\
\bar{y}_A - \bar{y}_C &= [\beta_0 + \beta_1 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_A] \\
&\quad - [\beta_0 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_C] \\
&= \beta_1 + (\bar{\varepsilon}_A - \bar{\varepsilon}_C) \\
\bar{y}_B - \bar{y}_C &= [\beta_0 + \beta_2 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_B] \\
&\quad - [\beta_0 + (\beta_3 + \beta_4 + \beta_5 + \beta_6)/5 + \bar{\varepsilon}_C] \\
&= \beta_2 + (\bar{\varepsilon}_B - \bar{\varepsilon}_C)
\end{aligned}$$

Note that for each pairwise comparison, the block $\beta$'s ($\beta_3$, $\beta_4$, $\beta_5$, and $\beta_6$) cancel out, leaving only the treatment $\beta$'s ($\beta_1$ and $\beta_2$). That is, the experimental noise resulting from differences between blocks is eliminated when treatment means are compared. The quantities $\bar{\varepsilon}_A - \bar{\varepsilon}_B$, $\bar{\varepsilon}_A - \bar{\varepsilon}_C$, and $\bar{\varepsilon}_B - \bar{\varepsilon}_C$ are the errors of estimation and represent the noise that tends to obscure the true differences between the treatment means.

What would occur if we employed the completely randomized design of Table 13.1 rather than the randomized block design? Since each worker assembles a watch using only one of the methods, each treatment does not appear in each block. Consequently, when we compare the treatment means, the worker-to-worker variation (i.e., the block effects) will not cancel. For example, the difference between $\bar{y}_A$ and $\bar{y}_C$ would be

$$\bar{y}_A - \bar{y}_C = \beta_1 + \underbrace{(\text{Block } \beta\text{'s that do not cancel}) + (\bar{\varepsilon}_A - \bar{\varepsilon}_C)}_{\text{Error of estimation}}$$

Thus, for the completely randomized design, the error of estimation will be increased by an amount involving the block effects ($\beta_3$, $\beta_4$, $\beta_5$, and $\beta_6$) that do not cancel. These effects, which inflate the error of estimation, cancel out for the randomized block design, thereby reducing the noise in the experiment.

### Example 13.3

Randomized Block Design Model

Refer to Example 13.2 and the randomized block design used to compare the mean construction cost estimates of the four DOT engineers. The design is illustrated in Figure 13.4.

a. Write the model for the randomized block design.
b. Interpret the $\beta$ parameters of the model, part **a**.
c. How can we use the model, part **a**, to test for differences among the mean estimates of the four engineers?

Solution

a. The experiment involves a qualitative factor, engineer, at four levels, which represents the treatments. The blocks for the experiment are the 10 road contracts. Therefore, the model is

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\substack{\text{Treatments} \\ \text{(Engineers)}}} + \underbrace{\beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_{12} x_{12}}_{\substack{\text{Blocks} \\ \text{(Contracts)}}}$$

where

$$x_1 = \begin{cases} 1 & \text{if engineer A} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if engineer B} \\ 0 & \text{if not} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if engineer C} \\ 0 & \text{if not} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if contract 1} \\ 0 & \text{if not} \end{cases} \quad x_5 = \begin{cases} 1 & \text{if contract 2} \\ 0 & \text{if not} \end{cases} \cdots x_{12} = \begin{cases} 1 & \text{if contract 9} \\ 0 & \text{if not} \end{cases}$$

b. Note that we have arbitrarily selected engineer D and contract 10 as the base levels. The interpretations of the $\beta$'s follow:

$$\begin{aligned} \beta_1 &= \mu_A - \mu_D && \text{for a given contract} \\ \beta_2 &= \mu_B - \mu_D && \text{for a given contract} \\ \beta_3 &= \mu_C - \mu_D && \text{for a given contract} \\ \beta_4 &= \mu_1 - \mu_{10} && \text{for a given engineer} \\ \beta_5 &= \mu_2 - \mu_{10} && \text{for a given engineer} \\ &\vdots \\ \beta_{12} &= \mu_9 - \mu_{10} && \text{for a given engineer} \end{aligned}$$

c. One way to determine whether the means for the four engineers differ is to test the null hypothesis

$$H_0\text{:} \quad \mu_A = \mu_B = \mu_C = \mu_D$$

From our $\beta$ interpretations in part **b**, this hypothesis is equivalent to testing

$$H_0\text{:} \quad \beta_1 = \beta_2 = \beta_3 = 0$$

To test this hypothesis, we drop the treatment $\beta$'s ($\beta_1$, $\beta_2$, and $\beta_3$) from the complete model and fit the reduced model

$$E(y) = \beta_0 + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_{12} x_{12}$$

Then we conduct the nested model $F$ test (see Section 12.8), where

$$F = \frac{(\text{SSE}_{\text{Reduced}} - \text{SSE}_{\text{Complete}})/3}{\text{MSE}_{\text{Complete}}}$$

The randomized block design represents one of the simplest types of noise-reducing designs. Other, more complex, designs that use the principle of blocking are available to remove trends or variation in two or more directions. The **Latin square design** is useful when you want to eliminate two sources of variation, i.e., when you want to block in two directions. **Latin cube designs** allow you to block in three directions. A further variation in blocking occurs when the block contains fewer experimental units than the number of treatments. By properly assigning the treatments to a specified number of blocks, one can still obtain an estimate of the difference between a pair of treatments free of block effects. These are known as **incomplete block designs**. Consult the references for details on how to set up these more complex block designs.

## Applied Exercises

13.1 *Designed experiment.* In a designed experiment,

a. What two factors affect the quantity of information?
b. How does blocking increase the quantity of information?

13.2 *Health risks to beachgoers.* According to a University of Florida researcher, the longer a beachgoers sits in wet sand or stays in the water, the higher the risk of gastroenteritis (*University of Florida News,* Jan. 29, 2008). The result is based on a study of over 1,000 adults conducted at three popular Florida beaches. The adults were divided into three groups: (1) beachgoers who were recently exposed to wet sand and water for at least two consecutive hours, (2) beachgoers who were not recently exposed to wet sand and water, and (3) people who had not recently visited a beach. Suppose the researcher wants to compare the mean levels of intestinal bacteria for the three groups. For this study, identify each of the following:

a. experimental unit
b. response variable
c. factor
d. factor levels

13.3 *Corrosion prevention of buried steel structures.* Refer to the *Materials Performance* (March 2013) study to compare two tests for steel corrosion of underground piping, Exercise 1.2 (p. 5). The two tests, which engineers call "instant-off" and "instant-on" potential, were applied to buried piping at a petrochemical plant in Turkey. Recall that both the "instant-off" and "instant-on" tests were used to make predictions of corrosion at each of 19 different randomly selected pipe locations. The researchers want to compare the mean accuracy of the corrosion predictions for the two tests in order to determine if one test is more desirable than the other when applied to buried steel piping.

a. What are the experimental units for this study?
b. What type of design is employed? Identify the features (e.g., treatments, blocks) of the design.
c. What is the dependent (response) variable of interest?
d. Write the model for this design.

13.4 *Visual attention skills test.* Refer to the *Journal of Articles in Support of the Null Hypothesis* (Vol. 6, 2009) study to determine whether video game players have superior visual attention skills over non-video game players, Exercise 1.3 (p. 5). Recall that each in a sample of 65 male students was classified as a video game player or a non-player. The two groups were then subjected to a series of visual attention tasks that included the "field of view" test. The researchers compared the mean test scores of the two groups.

a. What are the experimental units for this study?
b. What type of design is employed? Identify the features (e.g., treatments, blocks) of the design.
c. What is the dependent (response) variable of interest?
d. Write the model for this design.

13.5 *Taste preferences of cockatiels. Applied Animal Behaviour Science* (October 2000) published a study of the taste preferences of caged cockatiels. A sample of birds bred at the University of California, Davis, was randomly divided into three experimental groups. Group 1 was fed purified water in bottles on both sides of the cage. Group 2 was fed water on one side and a liquid sucrose (sweet) mixture on the opposite side of the cage. Group 3 was fed water on one side and a liquid sodium chloride (sour) mixture on the opposite side of the cage. One variable of interest to the researchers was total consumption of liquid by each cockatiel.

a. What is the experimental unit for this study?
b. Is the study a designed experiment? What type of design is employed?
c. What are the factors in the study?
d. Give the levels of each factor.
e. How many treatments are in the study? Identify them.
f. What is the response variable?
g. Write the regression model for the designed experiment.

13.6 *CT scanning for lung cancer.* A University of South Florida clinical trial of 50,000 smokers was carried out to compare the effectiveness of CT scans with X-rays for detecting lung cancer. (*Today's Tomorrows*, Fall 2002.) Each participating smoker was randomly assigned to one of two screening methods, CT or chest X-ray, and the age (in years) at which the scanning method first detects a tumor was to be determined. One goal of the study is to compare the mean ages when cancer is first detected by the two screening methods.

a. Identify the response variable of the study.
b. Identify the experimental units of the study.
c. Identify the factor(s) in the study.
d. Identify the treatments in the study.
e. What type of design, completely randomized or randomized block, was employed?

13.7 *Rotary oil rigs.* A petroleum engineer wants to compare the average monthly number of rotary oil rigs running in three states—California, Utah, and Alaska. In order to account for month-to-month variation, three months were randomly selected over a 2-year period and the number of oil rigs running in each state in each month was obtained from data provided from *World Oil* (Jan. 2002) magazine. The data are reproduced in the accompanying table.

**OILRIGS**

| Month | California | Utah | Alaska |
|---|---|---|---|
| 1 | 27 | 17 | 11 |
| 2 | 34 | 20 | 14 |
| 3 | 36 | 15 | 14 |

a. Why is a randomized block design preferred over a completely randomized design for comparing the mean number of oil rigs running monthly in California, Utah, and Alaska?
b. Identify the treatments for the experiment.
c. Identify the blocks for the experiment.

13.8 *Properties of cemented soils.* Refer to the *Bulletin of Engineering Geology and the Environment* (Vol. 69, 2010) study of cemented sandy soils, Exercise 1.9 (p. 7). The researchers applied one of three different sampling methods (rotary core, metal tube, or plastic tube) to randomly selected soil specimens, then measured the effective stress level (Newtons per meters-squared) of each specimen. Suppose that each method was applied to 10 soil specimens—a total of 30 measurements in all. Consider an analysis to compare the mean effective stress levels of the three sampling methods.

a. What are the experimental units for this study?
b. What type of design is employed? Identify the features (e.g., treatments, blocks) of the design.
c. What is the dependent (response) variable of interest?
d. Write the model for this design.

13.9 *DOT road construction cost estimate.* Refer to the randomized block design setup to compare the mean costs estimated by four DOT engineers, Examples 13.2 and 13.3.

a. Write the model for each observation of estimated cost $y$ for engineer B. Sum the observations to obtain the average for engineer B.
b. Repeat part **a** for engineer D.
c. Show that $(\bar{y}_B - \bar{y}_D) = \beta_2 + (\bar{\varepsilon}_B - \bar{\varepsilon}_D)$ Note that the $\beta$'s for blocks cancel when computing this difference.

## 13.5 Volume-Increasing Designs

In this section, we focus on how the proper choice of the treatments associated with two or more factors can increase the "volume" of information extracted from the experiment. The volume-increasing designs we will discuss are commonly known as **factorial designs** because they involve careful selection of the combinations of **factor levels** (i.e., treatments) in the experiment.

Consider a utility company that charges its customers a less expensive rate for using electricity during off-peak (less-demanded) hours. The company is experimenting with several time-of-day pricing schedules. Two factors (i.e., independent variables) that the company can manipulate to form the schedule are price ratio, $x_1$, measured as the ratio of peak to off-peak prices, and peak period length, $x_2$, measured in hours. Suppose the utility company wants to investigate pricing ratio at two levels, 200% and 400%, and peak period length at two levels, 6 and 9 hours. The company will measure customer satisfaction, $y$, for several different schedules (i.e., combinations of $x_1$ and $x_2$) with the goal of comparing the mean satisfaction levels of the schedules. How should the company select the treatments for the experiment?

One method of selecting the price ratio–peak period length levels to be assigned to the experimental units (customers) would be to use the "one-at-a-time" approach. According to this procedure, one independent variable is varied while the remaining independent variables are held constant. This process is repeated for each of the independent variables in the experiment. This plan would *appear* to be extremely logical and consistent with the concept of blocking introduced in Section 13.4—that is, making comparisons within relatively homogeneous conditions—but this is not the case, as we will demonstrate.

The one-at-a-time approach applied to price ratio ($x_1$) and peak period length ($x_2$) is illustrated in Figure 13.5. When length is held constant at $x_2 = 6$ hours, we will observe the response $y$ at a ratio of $x_1 = 200\%$ and $x_1 = 400\%$, thus yielding one pair of $y$ values to estimate the average change in customer satisfaction as a result of changing the pricing ratio ($x_1$). Also, when pricing ratio is held constant at $x_1 = 200\%$, we observe the response $y$ at a peak period length of $x_2 = 9$ hours. This observation, along with the one at (200%, 6 hours), allows us to estimate the average change in customer satisfaction as result of a change in peak period length ($x_2$). The three treatments just described, (200%, 6 hours), (400%, 6 hours), and (200%, 9 hours), are indicated as points in Figure 13.5. Note that the figure shows two measurements (points) for each treatment. This is necessary to obtain an estimate of the standard deviation of the differences of interest.

A second method of selecting the factor–level combinations would be to choose the same three treatments as implied by the one-at-a-time approach and then to choose

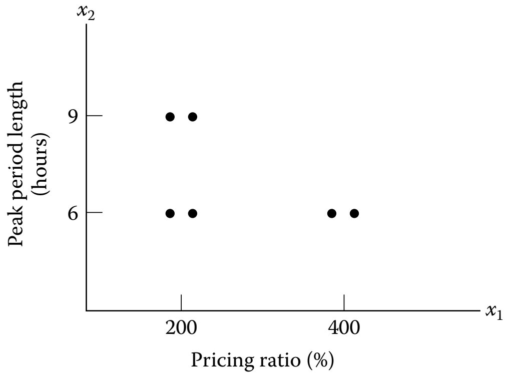



**FIGURE 13.5**
"One-at-a-time" approach to selecting treatments

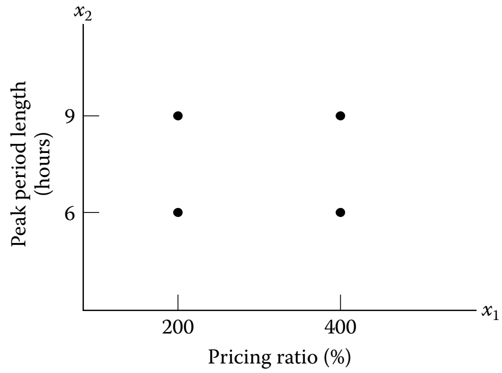



**FIGURE 13.6**
Selecting all possible treatments

the fourth treatment at (400%, 9 hours) as shown in Figure 13.6. In other words, we have varied both variables $x_1$ and $x_2$, at the same time.

Which of the two designs yields the most information about the treatment differences? Surprisingly, the design of Figure 13.6, with only four observations, yields more accurate information than the one-at-a-time approach with its six observations. First, note that both designs yield two estimates of the difference between the mean response $y$ at $x_1 = 200\%$ and $x_1 = 400\%$ when peak period length ($x_2$) is held constant, and both yield two estimates of the difference between the mean response $y$ at $x_2 = 6$ hours and $x_2 = 9$ hours when pricing ratio ($x_1$) is held constant. But what if the difference between the mean response $y$ at $x_1 = 200\%$ and at $x_1 = 400\%$ depends on which level of $x_2$ is held fixed, i.e., what if pricing ratio ($x_1$) and peak period length ($x_2$) *interact*? Then, we require estimates of the mean difference $(\mu_{200} - \mu_{400})$ when $x_2 = 6$ and the mean difference $(\mu_{200} - \mu_{400})$ when $x_2 = 9$. Estimates of both these differences are obtainable from the second design, Figure 13.6. However, since no estimate of the mean response for $x_1 = 400$ and $x_2 = 9$ is available from the one-at-a-time method, the interaction will go undetected for this design!

The importance of interaction between independent variables was emphasized in Chapters 11 and 12. If interaction is present, we cannot study the effect of one variable (or factor) on the response $y$ independent of the other variable. Consequently, we require experimental designs that provide information on factor interaction.

Designs that accomplish this objective are called **factorial experiments**. A **complete factorial experiment** is one that includes all possible combinations of the levels of the factors as treatments. For the experiment on time-of-day pricing, we have two levels of pricing ratio (200% and 400%) and two levels of peak period length (6 and 9 hours). Hence, a complete factorial experiment will include $(2 \times 2 = 4)$ treatments, as shown in Figure 13.6, and is called a $2 \times 2$ factorial design.

**Definition 13.10**

A **factorial design** is a method for selecting the treatments (that is, the factor–level combinations) to be included in an experiment. A complete factorial experiment is one in which the treatments consist of all factor–level combinations.

If we were to include a third factor, say, season, at four levels, then a complete factorial experiment would include all $2 \times 2 \times 4 = 16$ combinations of pricing ratio, peak period length, and season. The resulting collection of data would be called a $2 \times 2 \times 4$ factorial design.

**Example 13.4**

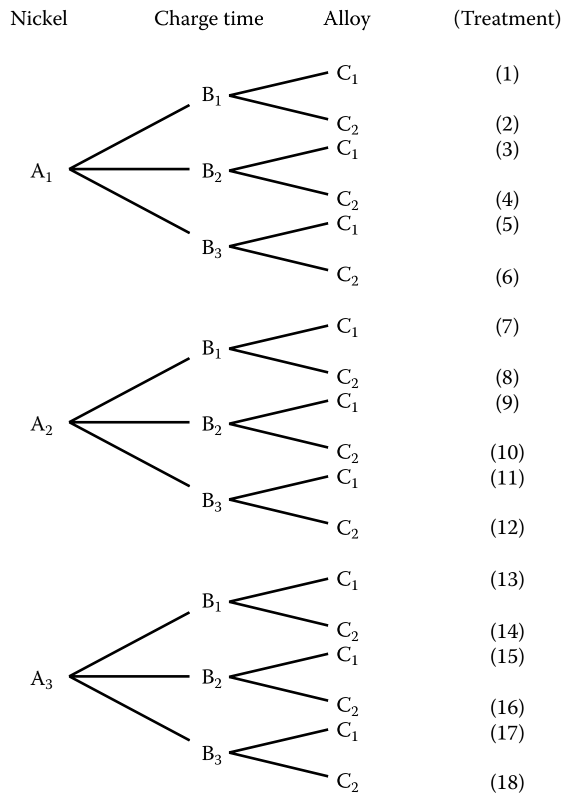
Factorial Design: Yield Strength of Nickel Alloy

Suppose you want to conduct an experiment to compare the yield strengths of nickel alloy tensile specimens charged in a sulfuric acid solution. In particular, you want to investigate the effect on mean strength of three factors: nickel composition at three levels ($A_1$, $A_2$, and $A_3$), charging time at three levels ($B_1$, $B_2$, and $B_3$), and alloy type at two levels ($C_1$ and $C_2$). Consider a complete factorial experiment. Identify the treatments for this $3 \times 3 \times 2$ factorial design.

Solution

The complete factorial experiment includes all possible combinations of nickel composition, charging time, and alloy type. We therefore would include the following treatments: $A_1B_1C_1$, $A_1B_1C_2$, $A_1B_2C_1$, $A_1B_2C_2$, $A_1B_3C_1$, $A_1B_3C_2$, $A_2B_1C_1$, $A_2B_1C_2$, $A_2B_2C_1$, $A_2B_2C_2$, $A_2B_3C_1$, $A_2B_3C_2$, $A_3B_1C_1$, $A_3B_1C_2$, $A_3B_2C_1$, $A_3B_2C_2$, $A_3B_3C_1$, $A_3B_3C_2$. These 18 treatments are shown diagrammatically in Figure 13.7

**FIGURE 13.7**
The 18 treatments for the $3 \times 3 \times 2$ factorial of Example 13.4




The multiple regression model for a factorial design includes terms for each of the factors in the experiment—called **main effects**—and terms for factor interactions. For example, the model for the $2 \times 2$ factorial for the time-of-day pricing experiment includes a first-order term for the quantitative factor pricing ratio ($x_1$), a first-order term for the quantitative factor peak period length ($x_2$), and an interaction term:

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Main effects}} + \underbrace{\beta_3 x_1 x_2}_{\text{Interaction}} + \varepsilon$$

In general, the model for a complete factorial design for $k$ factors contains terms for the following:

The main effects for each of the $k$ factors
Two-way interaction terms for all pairs of factors
Three-way interaction terms for all combinations of three factors
⋮
$k$-way interaction terms of all combinations of $k$ factors.

If the factors are qualitative, then we set up dummy variables and proceed as in the next example.

**Example 13.5**

Factorial Design Model

Write the model for the $3 \times 3 \times 2$ factorial experiment of Example 13.4.

Solution

Since the factors are qualitative, we set up dummy variables as follows:

$$x_1 = \begin{cases} 1 & \text{if nickel } A_1 \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if nickel } A_2 \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if charge } B_1 \\ 0 & \text{if not} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{if charge } B_2 \\ 0 & \text{if not} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if alloy } C_1 \\ 0 & \text{if alloy } C_2 \end{cases}$$

Then the appropriate model is

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\substack{\text{Nickel main} \\ \text{effects}}} + \underbrace{\beta_3 x_3 + \beta_4 x_4}_{\substack{\text{Charge main} \\ \text{effects}}} + \underbrace{\beta_5 x_5}_{\substack{\text{Alloy main} \\ \text{effect}}}$$

$$+ \underbrace{\beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \beta_9 x_2 x_4}_{\text{Nickel} \times \text{Charge}} + \underbrace{\beta_{10} x_1 x_5 + \beta_{11} x_2 x_5}_{\text{Nickel} \times \text{Alloy}}$$

$$+ \underbrace{\beta_{12} x_3 x_5 + \beta_{13} x_4 x_5}_{\text{Charge} \times \text{Alloy}}$$

$$+ \underbrace{\beta_{14} x_1 x_3 x_5 + \beta_{15} x_1 x_4 x_5 + \beta_{16} x_2 x_3 x_5 + \beta_{17} x_2 x_4 x_5}_{\text{Nickel} \times \text{Charge} \times \text{Alloy}}$$

Note that the number of parameters in the model for the $3 \times 3 \times 2$ factorial design of Example 13.5 is 18, which is equal to the number of treatments contained in the experiment. This is always the case for a complete factorial experiment. Consequently, if we fit the complete model to a single replication of the factorial treatments (i.e., one $y$ observation measured per treatment), we will have no degrees of freedom available for estimating the error variance, $\sigma^2$. One way to solve this problem is to add additional data points to the sample. Researchers usually accomplish this by **replicating** the complete set of factorial treatments. That is, we collect two or more observed $y$ values for each treatment in the experiment. This provides sufficient degrees of freedom for estimating $\sigma^2$.

One potential disadvantage of a complete factorial experiment is that it may require a large number of treatments. For example, an experiment involving 10 factors each at two levels would require $2^{10} = 1{,}024$ treatments! This might occur in an exploratory study where we are attempting to determine which of a large set of factors affect the response $y$. Several volume-increasing designs are available that employ only a fraction of the total number of treatments in a complete factorial experiment. For this reason, they are called **fractional factorial experiments**. Fractional factorials permit the estimation of the $\beta$ parameters of lower-order terms (e.g., main effects and two-way interactions); however, $\beta$ estimates of certain higher-order terms (e.g., three-way and four-way interactions) will be the same as some lower-order terms, thus *confounding* the results of the experiment. Consequently, a great deal of expertise is required to run and interpret fractional factorial experiments. Consult the references for details on fractional factorials and other more complex, volume-increasing designs.

## Applied Exercises

13.10 *Baker's versus brewer's yeast.* *The Electronic Journal of Biotechnology* (Dec. 15, 2003) published an article on a comparison of two yeast extracts, baker's yeast and brewer's yeast. Brewer's yeast is a surplus by-product obtained from a brewery, hence it is less expensive than primary-grown baker's yeast. Samples of both yeast extracts were prepared at four different temperatures (45, 48, 51, and 54°C), and the autolysis yield (recorded as a percentage) was measured for each of the yeast–temperature combinations. The goal of the analysis is to investigate the impact of yeast extract and temperature on mean autolysis yield.

a. Identify the factors (and factor levels) in the experiment.
b. Identify the response variable.
c. How many treatments are included in the experiment?
d. What type of experimental design is employed?

13.11 *Removing bacteria from water.* A coagulation–microfiltration process for removing bacteria from water was investigated in *Environmental Science & Engineering* (Sept. 1, 2000). Chemical engineers at Seoul National University performed a designed experiment to estimate the effect of both the level of the coagulant and acidity (pH) level on the coagulation efficiency of the process. Six levels of coagulant (5, 10, 20, 50, 100, and 200 milligrams per liter) and six pH levels (4.0, 5.0, 6.0, 7.0, 8.0, and 9.0) were employed. Water specimens collected from the Han River in Seoul, Korea, were placed in jars, and each jar randomly assigned to receive one of the $6 \times 6 = 36$ combinations of coagulant level and pH level.

a. What type of experimental design was applied in this study?
b. Give the factors, factor levels, and treatments for the study.

13.12 *Back/knee strength, gender, and lifting strategy.* *Human Factors* (December 2009) investigated whether back and knee strength dictates the load lifting strategies of males and females. A sample of 32 healthy adults (16 men and 16 women) participated in a series of strength tests on the back and the knees. Following the tests, the participants were randomly divided into two groups, where each group consisted of 8 men and 8 women. One group was provided with knowledge of their strength test results, while the other group was not provided with this knowledge. The final phase of the study required the participants to lift heavy cast iron plates out of a bin. Based on the different angles used to lift the plates, a quantitative measure of posture—called a postural index—was measured for each participant. The goal of the research was to determine the effect of gender and strength knowledge (provided or not provided) on the mean postural index. For this study, identify each of the following:

a. experimental unit
b. response variable
c. factors
d. levels of each factor
e. treatments.
f. Write the model appropriate for analyzing the data.

13.13 *Steel ingot experiment.* A quality control supervisor measures the quality of a steel ingot on a scale from 0 to 10. He designs an experiment in which three different temperatures (ranging from 1,100 to 1,200°F) and five different pressures (ranging from 500 to 600 psi) are utilized, with 20 ingots produced at each temperature–pressure combination. Identify the following elements of the experiment:

a. response
b. factor(s) and factor type(s)
c. treatments
d. experimental units.

13.14 *Factorial models.*

a. Write the complete factorial model for a $2 \times 3$ factorial experiment where both factors are qualitative.
b. Write the complete factorial model for a $2 \times 3 \times 3$ factorial experiment where the factor at two levels is quantitative and the other two factors are qualitative.

13.15 *Factorial interaction.* Consider a factorial design with two factors, $A$ and $B$, each at three levels. Suppose we select the following treatment (factor–level) combinations to be included in the experiment: $A_1B_1$, $A_2B_1$, $A_3B_1$, $A_1B_2$, and $A_1B_3$.

a. Is this a complete factorial experiment? Explain.
b. Explain why it is impossible to investigate $AB$ interaction in this experiment.

13.16 *Factorial design issues.* Suppose you wish to investigate the effect of three qualitative factors on a response $y$.

a. Explain why a factorial selection of treatments is better than varying each factor, one at a time, while holding the remaining two factors constant.
b. Why is the randomized block design a poor design to use?

## 13.6 Selecting the Sample Size

We demonstrated how to select the sample size for estimating a single population mean or comparing two population means in Chapter 7. We now show you how this problem can be solved for designed experiments.

As mentioned in Section 13.3, a measure of the quantity of information in an experiment that is pertinent to a particular population parameter is the standard error of the estimator of the parameter. A more practical measure is the half-width of the parameter's confidence interval, which will, of course, be a function of the standard error. For example, the half-width of a confidence interval for a population mean (given in Section 7.6) is

$$t_{\alpha/2}s_{\bar{y}} = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Similarly, the half-width of a confidence interval for the slope $\beta_1$ of a straight-line model relating $y$ to $x$ (given in Section 10.6) is

$$(t_{\alpha/2})s_{\hat{\beta}_1} = t_{\alpha/2}\left(\frac{s}{\sqrt{\text{SS}_{xx}}}\right) = t_{\alpha/2}\left(\sqrt{\frac{\text{SSE}}{n-2}}\right)\left(\frac{1}{\sqrt{\text{SS}_{xx}}}\right)$$

In both cases, the half-width is a function of the total number of data points in the experiment; each interval half-width gets smaller as the total number of data points $n$ increases. The same is true for a confidence interval for a parameter $\beta_i$ of a general linear model, for a confidence interval for $E(y)$, and for a prediction interval for $y$. Since each designed experiment can be represented by a linear model, this result can be used to select, approximately, the number of **replications** r (i.e., the number of observations measured for each treatment) in the experiment.

We illustrate this procedure with the following examples.

### Example 13.6

**Determining the Sample Size: Completely Randomized Design**

Refer to Example 13.1 (p. 719). Consider a simpler experiment, where the researcher wants to determine the effect of Temperature only on the Hardness (rated on a scale of 1 to 10 points) of plastic. Three levels of temperature (200º, 250º, and 300º) are selected for investigation. A completely randomized design will be employed, with r plastic molds formed for each of the 3 levels of temperature. How many replicates, r, of plastic molds are required at each temperature level in order to estimate the difference between the mean hardness value for any two temperature levels to within .5 point? Assume a 95% confidence interval will be used to estimate the difference.

**Solution**

For this experiment, we have a single factor—temperature—at 3 levels (200º, 250º, and 300º). In this completely randomized design, r plastic molds will be randomly assigned to each level of temperature. The model for this design follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where $y$ = hardness level, $x_1 = \{1 \text{ if } 200º, 0 \text{ if not}\}$, and $x_2 = \{1 \text{ if } 250º, 0 \text{ if not}\}$. [Note: 300º is the base level for temperature.]

Recall that for this dummy variable model, $\beta_1 = \mu_{200} - \mu_{300}$, i.e., the difference in hardness means for temperatures 200 and 300 degrees. Similarly, $\beta_2 = \mu_{250} - \mu_{300}$. Consequently, a confidence interval on one of these $\beta$'s will yield a confidence interval for the difference between mean hardness values set at two different temperatures.

Consider $\beta_1 = \mu_{200} - \mu_{300}$. We know (from Chapter 7) that the estimate of the difference between two population means is the difference between the sample means, i.e., $\hat{\beta}_1 = \bar{y}_{200} - \bar{y}_{300}$. Then

$$V(\hat{\beta}_1) = V(\bar{y}_{200}) + V(\bar{y}_{300}) = V(y)/r + V(y)/r = 2V(y)/r,$$

since $r$ is the sample size associated with each mean. The variance of hardness, $V(y)$, is equal to the pooled variance obtained from the completely randomized design model, $\sigma^2$. Assume that $V(y) = \sigma^2 = 1$. Then the estimated variance of $\hat{\beta}_1$ is $(s_{\beta_1})^2 = 2(1)/r = 2/r$.

Now, a confidence interval for $\beta_1$ is given by the formula: $\hat{\beta}_1 \pm t_{\alpha/2}(s_{\beta_1})$. Hence, the bound on the error of estimation is $B = t_{\alpha/2}(s_{\beta_1})$; and since $(s_{\beta_1})^2 = 2/r$ for this completely randomized design, we have

$$B = t_{\alpha/2}\sqrt{(2/r)},$$

$$\text{or } r = (t_{\alpha/2})^2(2)/B^2$$

Here, we want to estimate the difference in means to within .5 point; consequently, the bound on the error of estimation is $B = .5$. For a 95% confidence interval, $\alpha = .05$, $\alpha/2 = .025$, and $t_{.025} \approx 2$. Substituting $t_{.025} \approx 2$ and $B = .5$ into the equation, we obtain

$$r = (2)^2(2)/(.5^2) = 32$$

Therefore, our completely randomized design requires $r = 32$ replicates of plastic molds at each temperature level. A layout of the design is shown in Figure 13.8.

**FIGURE 13.8**
Layout of Completely Randomized Design, Example 13.6

| | Temperature | | |
|---|---|---|---|
| | 200° | 250° | 300° |
| Replicate 1 | ___ | ___ | ___ |
| 2 | ___ | ___ | ___ |
| 3 | ___ | ___ | ___ |
| ⋮ | | | |
| 32 | ___ | ___ | ___ |

## Example 13.7

**Determining the Sample Size: Factorial Design**

Consider a $2 \times 2$ factorial experiment to investigate the effect of two factors on the light output $y$ of flashbulbs (measured as a percentage) used in cameras. The two factors (and their levels) are $x_1$ = Amount of foil contained in the bulb (100 and 200 milligrams) and $x_2$ = Speed of sealing machine (1.2 and 1.3 revolutions per minute). The complete model for this $2 \times 2$ factorial design is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where $x_1 = \{1 \text{ if } 100 \text{ mg}, 0 \text{ if } 200 \text{ mg}\}$, and $x_2 = \{1 \text{ if } 1.2 \text{ rpm}, 0 \text{ if } 1.3 \text{ rpm}\}$. [Note: 200 is the base level for amount of foil and 1.3 is the base level for speed.]

How many replicates, $r$, of flashbulbs produced at each of the $2 \times 2 = 4$ treatments are required to estimate the interaction $\beta$ to within 2.2% of its true value using a 95% confidence interval? Assume that an estimate for the standard deviation of light output, $y$, is 1%.

**Solution** For this designed experiment, let $\mu_{ij}$ represent the mean light output, $E(y)$, for amount of foil at level $i$ and speed at level $j$. Then the means for each of the $2 \times 2 = 4$ treatments can be expressed as follows:

$$\mu_{200,1.3} = \beta_0 \qquad (\text{since } x_1 = x_2 = 0)$$

$$\mu_{200,1.2} = \beta_0 + \beta_2 \qquad (\text{since } x_1 = 0, x_2 = 1)$$

$$\mu_{300,1.3} = \beta_0 + \beta_1 \qquad (\text{since } x_1 = 1, x_2 = 0)$$

$$\mu_{300,1.2} = \beta_0 + \beta_1 + \beta_2 + \beta_3 \qquad (\text{since } x_1 = 1 = x_2 = 1)$$

With some algebra, you can show that

$$\beta_3 = (\mu_{300,1.2} - \mu_{300,1.3}) - (\mu_{200,1.2} - \mu_{200,1.3})$$

That is, the interaction, is a linear combination of the four treatment means. Then $V(\hat{\beta}_3) = V(\bar{y}_{300,1.2}) + V(\bar{y}_{300,1.3}) + V(\bar{y}_{200,1.2}) + V(\bar{y}_{200,1.3}) = 4V(y)/r$, where $r$ is the sample size associated with each mean and $V(y)$ is the variance of the dependent variable, light output. Since the standard deviation of $y$ is assumed to be 1%, $V(y) \approx (1)^2 = 1$. Thus,

$$V(\hat{\beta}_3) = 4/r.$$

Like in the previous example, a confidence interval for $\beta_3$ is given by the formula, $\hat{\beta}_3 \pm t_{\alpha/2}(s_{\beta_3})$, and the bound on the error of estimation is $\mathrm{B} = t_{\alpha/2}(\mathrm{s}_{\beta_3})$ where $(\mathrm{s}_{\beta_3})^2 = V(\hat{\beta}_3) = 4/r$ for this factorial design. Consequently,

$$B = t_{\alpha/2}\sqrt{(4/r)},$$

$$\text{or } r = (t_{\alpha/2})^2(4)/B^2$$

Here, we desire a bound on the error of estimation of 2.2%; thus $\mathrm{B} = 2.2$. For a 95% confidence interval, $\alpha = .05$, $\alpha/2 = .025$, and $t_{.025} \approx 2$. Substituting $t_{.025} \approx 2$ and $B = 2.2$ into the equation, we obtain

$$r = (2)^2(4)/(2.2^2) = 3.31$$

Since we can run either three or four replications (but not 3.31), we should choose four replications to be reasonably certain that we will be able to estimate the interaction parameter, $\beta_3$, to within 2.2% of its true value. The $2 \times 2$ factorial with four replicates would be laid out as shown in Table 13.3.

**TABLE 13.3 $2 \times 2$ Factorial with Four Replicates**

| | | *Amount of Foil, $x_1$* | |
|---|---|---|---|
| | | **100** | **200** |
| **Machine Speed, $x_2$** | **1.2** | 4 observations on $y$ | 4 observations on $y$ |
| | **1.3** | 4 observations on $y$ | 4 observations on $y$ |

### Determining the Number of Replicates, $r$

*Completely Randomized Design*

To estimate the difference between two treatment means to within B units with $(1 - \alpha)100\%$ confidence:

$$r = 2(t_{\alpha/2})^2(s)^2/\mathrm{B}^2,$$

where $s$ is the estimated standard deviation of the response, $y$

*Factorial Design*

To estimate the interaction effect to within B units with $(1 - \alpha)100\%$ confidence:

$$r = 4(t_{\alpha/2})^2(s)^2/\mathrm{B}^2,$$

where $s$ is the estimated standard deviation of the response, $y$

## Applied Exercises

13.17 *Replication.* Why is replication important in a complete factorial experiment?

13.18 *Determining the number of replicates.* Consider a $2 \times 2$ factorial. How many replications are required to estimate the interaction $\beta$ to within two units with a 95% confidence interval? Assume that the standard deviation of the response variable, $y$, is approximately 3.

13.19 *Determining the number of blocks.* For a randomized block design with $b$ blocks, the estimated standard error of the estimated difference between any two treatment means is $\sqrt{2/b}$. Use this formula to determine the number of blocks required to estimate $(\mu_A - \mu_B)$, the difference between two treatment means, to within 10 units using a 95% confidence interval. Assume $s \approx 15$.

## 13.7 The Importance of Randomization

All the basic designs presented in this chapter involve randomization of some sort. In a completely randomized design and a basic factorial experiment, the treatments are randomly assigned to the experimental units. In a randomized block design, the blocks are randomly selected and the treatments within each block are assigned in random order. Why randomize? The answer is related to the assumptions we make about the random error $\varepsilon$ in the linear model. Recall (Section 11.2) our assumption that $\varepsilon$ follows a normal distribution with mean 0 and constant variance $\sigma^2$ for fixed settings of the independent variables (i.e., for each of the treatments). Further, we assume that the random errors associated with repeated observations are independent of each other in a probabilistic sense.

Experimenters rarely know all of the important variables in a process, nor do they know the true functional form of the model. Hence, the functional form chosen to fit the true relation is only an approximation, and the variables included in the experiment form only a subset of the total. The random error, $\varepsilon$, is thus a composite error caused by the failure to include all of the important factors as well as the error in approximating the function.

Although many unmeasured and important independent variables affecting the response $y$ do not vary in a completely random manner during the conduct of a designed experiment, we hope their behavior is such that their cumulative effect varies in a random manner and satisfies the assumptions upon which our inferential procedures are based. *The randomization in a designed experiment has the effect of randomly assigning these error effects to the treatments and assists in satisfying the assumptions on* $\varepsilon$.

## *STATISTICS IN ACTION REVISITED*

### Anti-Corrosive Behavior of Epoxy Coatings Augmented with Zinc

We now return to the *Statistics in Action* study of the steel anticorrosive behavior of different epoxy coatings formulated with zinc pigments. Recall that flat, rectangular panels cut from steel sheets represent the experimental units for the study. Each panel was coated with one of four different coating systems, labeled S1, S2, S3, and S4. These four systems represent the treatments in the study, with the goal to compare and rank the mean corrosion rates of the four coating systems.

Also recall that three panels were prepared for each coating system. One panel was exposed for 24 hours, another for 60 days, and the third for 120 days. This design was employed in an effort to remove the extraneous source of variation attributed to exposure time. Consequently, the researchers are using a randomized block design, with the three exposure times representing the blocks.

Following exposing, the corrosion rate (nanoamperes per square centimeter) was determined for each panel. (The lower the corrosion rate, the greater the anti-corrosion performance of the coating system.) The data for this randomized block design are shown in Table SIA13.2.

**EPOXY**

**TABLE SIA13.2 Corrosion Rates for Epoxy Coating Experiment**

| Exposure Time | System S1 | System S2 | System S3 | System S4 |
|---|---|---|---|---|
| 24 Hours | 6.7 | 7.5 | 8.2 | 6.1 |
| 60 Days | 8.7 | 9.1 | 10.5 | 8.3 |
| 120 Days | 11.8 | 12.6 | 14.5 | 11.8 |

*Source:* Kouloumbi, N., et al. "Anticorrosion performance of epoxy coatings on steel surface exposed to de-ionized water." *Pigment & Resin Technology*, Vol. 32, No. 2, 2003 (Table II).

The multiple regression model appropriate for analyzing the data follows:

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\substack{\text{Epoxy system terms} \\ \text{(Treatments)}}} + \underbrace{\beta_4 x_4 + \beta_5 x_5}_{\substack{\text{Exposure time terms} \\ \text{(Blocks)}}}$$

where

$$x_1 = \begin{cases} 1 & \text{if System S1} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if System S2} \\ 0 & \text{if not} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if System S3} \\ 0 & \text{if not} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if 24 hours exposure} \\ 0 & \text{if not} \end{cases} \quad x_5 = \begin{cases} 1 & \text{if 60 days exposure} \\ 0 & \text{if not} \end{cases}$$

If we let $\mu_{S1}$, $\mu_{S2}$, $\mu_{S3}$, and $\mu_{S4}$ represent the treatment (population) mean corrosion rates for epoxy systems $S_1$, $S_2$, $S_3$, and $S_4$, respectively; the $\beta$-parameters associated with the treatments (epoxy coating systems) are interpreted as follows:

$$\beta_1 = (\mu_{S1} - \mu_{S4}), \qquad \beta_2 = (\mu_{S2} - \mu_{S4}), \qquad \beta_3 = (\mu_{S3} - \mu_{S4})$$

Dependent Variable: CORRATE

| | |
|---|---|
| Number of Observations Read | 12 |
| Number of Observations Used | 12 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 72.68833 | 14.53767 | 155.30 | <.0001 |
| Error | 6 | 0.56167 | 0.09361 | | |
| Corrected Total | 11 | 73.25000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.30596 | R-Square | 0.9923 |
| Dependent Mean | 9.65000 | Adj R-Sq | 0.9859 |
| Coeff Var | 3.17056 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 11.75833 | 0.21635 | 54.35 | <.0001 |
| S1 | 1 | 0.33333 | 0.24981 | 1.33 | 0.2305 |
| S2 | 1 | 1.00000 | 0.24981 | 4.00 | 0.0071 |
| S3 | 1 | 2.33333 | 0.24981 | 9.34 | <.0001 |
| T1 | 1 | -5.55000 | 0.21635 | -25.65 | <.0001 |
| T2 | 1 | -3.52500 | 0.21635 | -16.29 | <.0001 |

Test TRTMNTS Results for Dependent Variable CORRATE

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 3 | 3.19444 | 34.12 | 0.0004 |
| Denominator | 6 | 0.09361 | | |

**FIGURE SIA13.2**
SAS output for randomized block design model

Now, if there are no differences among the four treatment means, then $\beta_1 = \beta_2 = \beta_3 = 0$. Consequently, a test of $H_0$: $\beta_1 = \beta_2 = \beta_3 = 0$ is appropriate for determining whether treatment differences exist. We can test this hypothesis using the nested model $F$ test of Chapter 12. The complete model above is compared to the reduced model:

$$E(y) = \beta_0 + \underbrace{\beta_4 x_4 + \beta_5 x_5}_{\substack{\text{Exposure time} \\ \text{terms (Blocks)}}}$$

A SAS printout of the analysis is shown in Figure SIA13.2. The $p$-value of the test, highlighted on the printout, is $p$-value $= .0004$. At $\alpha = .05$, there is sufficient evidence to reject $H_0$ and conclude that differences among the epoxy treatment means exist. Further analysis is required to determine which of the epoxy coating systems yields the lowest corrosion rate. We present the methodology for ranking treatment means in Chapter 14.

(*Note:* The importance of using the proper experimental design and regression model can be illustrated as follows. Suppose the block (exposure time) terms are omitted from the model. The SAS printout for the model with only treatment (epoxy system terms), $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, is shown in Figure SIA13.3. Note that the $p$-value for testing $H_0$: $\beta_1 = \beta_2 = \beta_3 = 0$ (i.e., the $p$-value for the global $F$ test) is $p$-value $= .7560$. Thus, if we used this inappropriate model to conduct the test, we would incorrectly conclude that there is no evidence of differences among the treatment means.)

Dependent Variable: CORRATE

| | |
|---|---|
| Number of Observations Read | 12 |
| Number of Observations Used | 12 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 9.58333 | 3.19444 | 0.40 | 0.7560 |
| Error | 8 | 63.66667 | 7.95833 | | |
| Corrected Total | 11 | 73.25000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 2.82105 | R-Square | 0.1308 |
| Dependent Mean | 9.65000 | Adj R-Sq | -0.1951 |
| Coeff Var | 29.23370 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 8.73333 | 1.62874 | 5.36 | 0.0007 |
| S1 | 1 | 0.33333 | 2.30338 | 0.14 | 0.8885 |
| S2 | 1 | 1.00000 | 2.30338 | 0.43 | 0.6757 |
| S3 | 1 | 2.33333 | 2.30338 | 1.01 | 0.3407 |

**FIGURE SIA13.3**
SAS output for model with only treatment effects

# Quick Review

## Key Terms

Analysis of variance 718
Blocking 721
Completely randomized design 721
Design of the experiment 718
Experiment 718
Experimental data 718
Experimental design 718, 721
Experimental unit 718
Factor 719
Factorial design 728
Fractional factorial experiments 731
Incomplete block design 726
Latin cube design 726
Latin square design 726
Level of a factor 728
Main effects 730
Noise-reducing designs 721
Observational data 718
Randomized block design 721
Replicates 735
Replication 733
Response variable 719
Treatment 719
Variability 720
Volume-increasing designs 728
Volume of the signal 720

## Key Formulas

| | | |
|---|---|---|
| $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$ where $x_i = 1$ if level $i$, 0 if not | Completely randomized design model for one factor at $p$ levels | 723 |
| $E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}}_{\text{Treatment dummy terms}} + \underbrace{\beta_p x_p + \beta_{p+1} x_{p+1} + \cdots + \beta_{p+b-2} x_{p+b-2}}_{\text{Block dummy terms}}$ where $x_i = 1$ if treatment/block $i$, 0 if not | Randomized block design model for $p$ treatments and $b$ blocks | 723 |
| $E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{a-2} x_{a-1}}_{\text{Factor A main effects}} \underbrace{+ \beta_a x_a + \beta_{a+1} x_{a+1} + \cdots + \beta_{a+b-2} x_{a+b-2}}_{\text{Factor B main effects}} \underbrace{+ \beta_{a+b-1} x_1 x_a + \beta_{a+b+1} x_2 x_a + \cdots + \beta_{ab-1} x_{a-1} x_{a+b-2}}_{\text{A} \times \text{B interaction terms}}$ | Factorial design model with factor A at $a$ levels and factor B at $b$ levels | 730 |
| $r = \dfrac{2(t_{\alpha/2})^2 (s)^2}{B^2}$ | Number of replicates required to estimate the difference between two treatment means in a completely randomized design to within $B$ units with $(1 - \alpha) \times 100\%$ confidence | 735 |
| $r = \dfrac{4(t_{\alpha/2})^2 (s)^2}{B^2}$ | Number of replicates required to estimate the interaction effect in a factorial design to within B units with $(1 - \alpha) \times 100\%$ confidence | 735 |

## LANGUAGE LAB

| Symbol | Pronunciation | Description |
|---|---|---|
| $p$ | | Number of treatments |
| $b$ | | Number of blocks |
| $a \times b$ | A-by-B | Two-factor factorial design with one factor at $a$ levels and one factor at $b$ levels |
| $a \times b \times c$ | A-by-B-by-C | Three-factor factorial design with first factor at $a$ levels, second factor at $b$ levels, and third factor at $c$ levels |
| $r$ | | Number of replicates |

### Chapter Summary Notes

- Independent variables (**factors**) in an experiment can be measured **observationally** (values are observed in their natural setting) or **experimentally** (values are controlled by the experimenter).
- **Treatments** are combinations of factor levels.
- **Experimental design** is a plan (strategy) for collecting the experimental data that involves four steps: (1) *select the factors*, (2) *select the treatments*, (3) *determine the sample size for each treatment*, (4) *assign the treatments to the experimental units*.
- **Volume-increasing designs** extract maximum information through careful selection of the factors, treatments, and sample size.
- **Noise-reducing designs** remove extraneous sources of information (noise) from the experiment through careful assignment of the treatments to the experimental units.
- Three basic experimental designs: *completely randomized, randomized block*, and *factorial designs*.
- A **completely randomized design** involves a single factor with a random assignment of the treatments to the experimental units.
- A **randomized block design** is a noise-reducing design involving a treatment factor and a blocking factor. The treatments are randomly assigned to the experimental units within each block.
- A **factorial design** is a volume-increasing design involving multiple factors. All possible treatments (factor–level combinations) are selected and randomly assigned to the experimental units.
- The data from a designed experiment is analyzed using a method called **analysis of variance**.

## Supplementary Applied Exercises

13.20 *Information in an experiment.* How do you measure the quantity of information in a sample that is pertinent to a particular population parameter?

13.21 *Increasing the volume.* What steps in the design of an experiment affect the volume of the signal pertinent to a particular population parameter?

13.22 *Reducing the noise.* In what step in the design of an experiment can you possibly reduce the variation produced by extraneous and uncontrolled variables?

13.23 *Choosing a design.* Explain the difference between a completely randomized design and a randomized block design. When is a randomized block design more advantageous?

13.24 *Factorial experiment.* Consider a two-factor factorial experiment where one factor is set at two levels and the other factor is set at four levels. How many treatments are included in the experiment? List them.

13.25 *Complete factorial model.* Write the complete factorial model for a $2 \times 2 \times 4$ factorial experiment where both factors at two levels are quantitative and the third factor at four levels is qualitative. If you conduct one replication of this experiment, how many degrees of freedom will be available for estimating $\sigma^2$?

13.26 *No factor interaction.* Refer to Exercise 13.25. Write the model for $y$ assuming that you wish to enter main effect terms for the factors, but no terms for factor interactions. How many degrees of freedom will be available for estimating $\sigma^2$?

13.27 *Flexible work schedules.* Researchers conducted an experiment to compare the mean job satisfaction rating $E(y)$ of workers using three types of work scheduling: flextime (which allows workers to set their individual work schedules), staggered starting hours, and fixed hours.

a. Identify the treatments in the experiment.
b. Suppose 60 workers are available for the study. Explain how you would employ a completely randomized design for this experiment.
c. Write the model for the completely randomized design.

13.28 *Drift ratio of a building.* A commonly used index to estimate the reliability of a building subjected to lateral loads is the drift ratio. Sophisticated computer programs such as STAAD-III have been developed to estimate the drift ratio based on variables such as beam stiffness, column stiffness, story height, moment of inertia, etc. Civil engineers at SUNY, Buffalo, and the University of Central Florida performed an experiment to compare drift ratio estimates using STAAD-III with the estimates produced by a new, simpler microcomputer program called DRIFT (*Microcomputers in Civil Engineering*, 1993). Data for a 21-story building were used as input to the programs. Two runs were made with STAAD-III: Run 1 considered axial deformation of the building columns, and run 2 neglected this information. The goal of the analysis is to compare the mean drift ratios (where drift is measured as lateral displacement) estimated by the three computer runs.

a. Identify the treatments in the experiment.
b. Because lateral displacement will vary greatly across building levels (floors), a randomized block design will be used to reduce the level-to-level variation in drift. Explain, diagrammatically, the setup of the design if all 21 levels are to be included in the study.
c. Write the linear model for the randomized block design.

13.29 *Worker productivity study.* Suppose you plan to investigate the effect of hourly pay rate and length of workday on some measure $y$ of worker productivity. Both pay rate and length of workday will be set at three levels, and $y$ will be observed for all combinations of these factors.

a. What type of experiment is this?
b. Identify the factors and state whether they are quantitative or qualitative.
c. Identify the treatments to be employed in the experiment.

13.30 *Diesel engine market share.* A study was conducted to compare market shares of diesel engine brands estimated by two different auditing methods.

a. Identify the treatments in the experiment.
b. Because of brand-to-brand variation in estimated market share, a randomized block design will be used. Explain how the treatments might be assigned to the experimental units if 10 diesel engine brands are to be included in the study.
c. Write the linear model for the randomized block design.

13.31 *Firefighting tasks.* Researchers investigated the effect of gender (male or female) and weight (light or heavy) on the length of time required by firefighters to perform a particular firefighting task (*Human Factors*). Eight firefighters were selected in each of the four gender–weight categories. Each firefighter was required to perform a certain task. The time (in minutes) needed to perform the task was recorded for each.

a. List the factors involved in the experiment.
b. For each factor, state whether it is quantitative or qualitative.
c. How many treatments are involved in this experiment? List them.

13.32 *Visual search task.* Many cognitively demanding jobs (e.g., air traffic controller, radar/sonar operator) require efficient processing of visual information. Researchers at Georgia Tech investigated the variables that affect the reaction time of subjects performing a visual search task (*Human Factors*, June 1993). College students were trained on microcomputers with one of two methods: continuously consistent or adjusted consistent. Each student was then assigned to one of six different practice sessions. Finally, the consistency of the search task was manipulated at four degrees: 100% consistency, 67%, 50%, or 33%. The goal of the researcher was to compare the mean reaction times of students assigned to each of the $2 \times 6 \times 4 = 48$ (training method) $\times$ practice session $\times$ (task consistency) experimental conditions.

a. List the factors involved in the experiment.
b. For each factor, state whether it is quantitative or qualitative.
c. How many treatments are involved in the experiment? List them.

CHAPTER 14

# The Analysis of Variance for Designed Experiments

## OBJECTIVE

*To present a method for analyzing data collected from designed experiments for comparing two or more population means; to define the relationship of the analysis of variance to regression analysis and to identify their common features*

## CONTENTS



*STATISTICS IN ACTION*
Pollutants at a Housing Development—A Case of Mishandling Small Samples

## STATISTICS IN ACTION

### Pollutants at a Housing Development—A Case of Mishandling Small Samples

According to the Environmental Protection Agency (EPA), "polycyclic aromatic hydrocarbons (PAHs) are a group of over 100 different chemicals that are formed during the incomplete burning of oil, gas, coal, garbage, or other organic substances like tobacco or charbroiled meat and from motor vehicle exhaust." (*www.epa.gov*). The EPA considers PAHs to be potential dangerous pollutants; consequently, industries are monitored regularly for the production of PAHs.

In this "Statistics in Action" we consider a legal case involving a developer who purchased a large parcel of Florida land that he planned to turn into a residential community. Unfortunately, the parcel turned out to have significant deposits of PAHs. Environmental regulatory agencies required the developer to remove the PAHs from the site prior to commencing development. The clean-up was finally completed, but the housing bubble burst and the development was a bust. The developer blamed the failure of his plan on the discovery of the pollutants, and filed suit against two industries that were within 25 miles of the site, both of which produced some PAH waste materials as part of their industrial processes. Not only did the developer want the industries to pay the costs of the clean-up, but he also wanted recompense for more than $100 million in lost profits he claimed would have been earned had the development been built out on schedule.

Both industries denied responsibility, and each hired experts to investigate the degree of similarity between pollutants at their industrial sites and those at the development site. Unfortunately, only limited PAH data had been collected at the proximate time that the pollution had been discovered at the development site. Nonetheless, one biochemical expert undertook a statistical analysis comparing two different types of PAH measurements for the three sites. The biochemical expert concluded that the data showed that Industry B was more likely to be responsible for the pollution at the development site than was his client, Industry A. Subsequently, an expert statistician hired by Industry B analyzed the same data and testified that "the data and statistical tests shed essentially no light on the matter."

Given the two contradictory expert opinions, how should the trial judge rule? To answer this question, we will analyze the data (saved in the **PAH** file) using the methods developed in this chapter and present the results in the *Statistics in Action Revisited* at the end of this chapter. Specifically, we want to (1) compare the mean PAH measurements at the different sites and (2) if the means differ, determine which industry is more likely to be responsible for the pollution at the housing development site.

## 14.1 Introduction

Once the data for a designed experiment have been collected, we will want to use the sample information to make inferences about the population means associated with the various treatments. The method used to compare the treatment means is traditionally known as **analysis of variance**, or **ANOVA**. The analysis of variance procedure provides a set of formulas that enable us to compute test statistics and confidence intervals required to make these inferences.

The formulas—one set for each experimental design—were developed in the early 1900s, well before the invention of computers. The formulas are easy to use, although the calculations can become quite tedious. However, you will recall from Chapter 13 that a linear model is associated with each experimental design. Consequently, the same inferences derived from the ANOVA calculation formulas can be obtained by properly analyzing the model using a regression analysis and the computer.

In this chapter, the main focus is on the regression approach to analyzing data from a designed experiment. Several common experimental designs—some of which were presented in Chapter 13—are analyzed. We also provide the ANOVA calculation formulas for each design and show their relationship to regression. First, we provide the logic behind an analysis of variance and these formulas in Section 14.2.

## 14.2 The Logic Behind an Analysis of Variance

The concept behind an analysis of variance can be explained using the following simple example.

Consider an experiment with a single factor at two levels (that is, two treatments). Suppose we want to decide whether the two treatment means differ based on the means of two independent random samples, each containing $n_1 = n_2 = 5$ measurements, and that the $y$ values appear as in Figure 14.1. Note that the five circles on the left are plots of the $y$ values for sample 1 and the five solid dots on the right are plots of the $y$ values for sample 2. Also, observe the horizontal lines that pass through the means for the two samples, $\bar{y}_1$ and $\bar{y}_2$. Do you think the plots provide sufficient evidence to indicate a difference between the corresponding population means?

If you are uncertain whether the population means differ for the data in Figure 14.1, examine the situation for two different samples in Figure 14.2a. We think that you will agree that for these data, it appears that the population means differ. Examine a third case in Figure 14.2b. For these data, it appears that there is little or no difference between the population means.

**FIGURE 14.1**
Plots of data for two samples

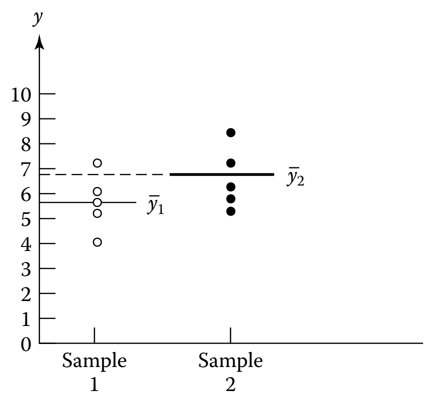
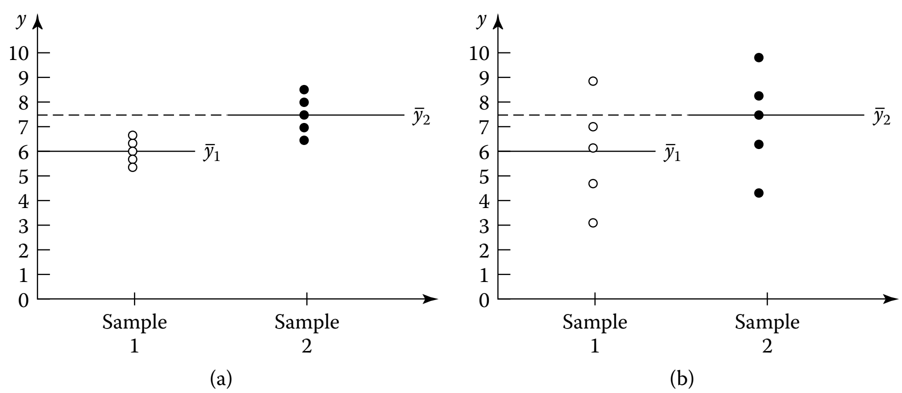



**FIGURE 14.2**
Plots of data for two cases



What elements of Figures 14.1 and 14.2 did we intuitively use to decide whether the data indicate a difference between the population means? The answer to the question is that we visually compared the distance (the variation) *between* the sample means to the variation *within* the $y$ values for each of the two samples. Since the difference between the sample means in Figure 14.2a is large relative to the within-sample variation, we inferred that the population means differ. Conversely, in Figure 14.2b, the variation between the sample means is small relative to the within-sample variation, and therefore there is little evidence to imply that the means are significantly different.

The variation within samples is measured by the pooled $s^2$ that we computed for the independent random samples $T$ test of Section 9.7, namely,

$$\textit{Within-sample variation:}\quad s^2 = \frac{\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{\text{SSE}}{n_1 + n_2 - 2}$$

where $y_{i1}$ is the $i$th observation in sample 1 and $y_{i2}$ is the $i$th observation in sample 2. The quantity in the numerator of $s^2$ is often denoted **SSE**, the **sum of squared errors**. As with regression analysis, SSE measures unexplained variability. But in this case, it measures variability *unexplained* by the differences between the sample means.

A measure of the between-sample variation is given by the weighted sum of squares of deviations of the individual sample means about the mean for all 10 observations, $\bar{y}$, divided by the number of samples minus 1, i.e.,

$$\textit{Between-sample variation:}\quad \frac{n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2}{2 - 1} = \frac{\text{SST}}{1}$$

The quantity in the numerator is often denoted **SST**, the **sum of squares for treatments**, since it measures the variability *explained* by the differences between the sample means of the two treatments.

For this experimental design, SSE and SST sum to a known total, namely,

$$\text{SS(Total)} = \sum (y_i - \bar{y})^2$$

[*Note*: SS(Total) is equivalent to $\text{SS}_{yy}$ in regression.] Also, the ratio

$$F = \frac{\text{Between-sample variation}}{\text{Within-sample variation}}$$

$$= \frac{\text{SST}/1}{\text{SSE}/(n_1 + n_2 - 2)}$$

has an $F$ distribution with $\nu_1 = 1$ and $\nu_2 = n_1 + n_2 - 2$ degrees of freedom (df) and therefore can be used to test the null hypothesis of no difference between the treatment means. The additivity property of the sums of squares led early researchers to view this analysis as a **partitioning** of $\text{SS(Total)} = \Sigma(y_i - \bar{y})^2$ into sources corresponding to the factors included in the experiment and to SSE. The simple formulas for computing the sums of squares, the additivity property, and the form of the test statistic made it natural for this procedure to be called **analysis of variance**. We demonstrate the analysis of variance procedure and its relation to regression for several common experimental designs in Sections 14.3–14.7.

## 14.3 One-Factor Completely Randomized Designs

Recall (Section 13.2) the first two steps in designing an experiment: (1) decide on the factors to be investigated and (2) select the factor level combinations (treatments) to be included in the experiment. For example, suppose you wish to compare the length of time to assemble a device in a manufacturing operation for workers who have completed one of three training programs, A, B, and C. Then this experiment involves a single factor, training program, at three levels, A, B, and C. Since training program is the only factor, these levels (A, B, and C) represent the treatments. Now we must decide the sample size for each treatment (step 3) and figure out how to assign the treatments to the experimental units, namely, the specific workers (step 4).

As we learned in Chapter 13, the most common assignment of treatments to experimental units is called a **completely randomized design**. To illustrate, suppose we wish to obtain equal amounts of information on the mean assembly times for the three training procedures; i.e., we decide to assign equal numbers of workers to each of the three training programs. Also, suppose we determine the number of workers in each of the three samples to be $n_1 = n_2 = n_3 = 10$. Then a completely randomized design is one in which the $n_1 = n_2 = n_3 = 30$ workers are **randomly assigned**, 10 to each of the three treatments. *A random assignment is one in which any one assignment is as probable as any other*. This eliminates the possibility of bias that might occur if the workers were assigned in some systematic manner. For example, a systematic assignment might accidentally assign most of the manually dexterous workers to training program A, thus underestimating the true mean assembly time corresponding to A.

Example 14.1 illustrates how a **random number generator** can be used to assign the 30 workers to the three treatments.

### Example 14.1

**Assigning Treatments in a Completely Randomized Design**

Use a random number generator to assign $n = 30$ experimental units (workers) to three treatment groups (training programs).

**Solution**

The first step is to number the 30 workers from 1 to 30. We used MINITAB's "Random Data" function to randomly reorder the 30 workers. That is, the integers between 1 and 30 are arranged in random order. The workers who have been assigned the first 10 numbers in the sequence are assigned to training program A, the second group of 10 workers are assigned to B, and the remaining workers are assigned to C.

Figure 14.3 is a MINITAB worksheet showing the random assignments. You can see that workers numbered 21, 5, 14, 7, 13, 18, 20, 22, 29 and 26 are assigned to program A; workers numbered 25, 27, 9, 2, 28, 3, 17, 16, 15 and 23 are assigned to Program B; and, the remaining workers to Program C.

In some experimental situations, we are unable to assign the treatments to the experimental units randomly because of the nature of the experimental units themselves. For example, the *Journal of Testing and Evaluation* described a study to compare the mean compression strengths (in pounds) of five different sizes of corrugated fiberboard shipping containers. The box sizes—labeled A, B, C, D, and E—are the treatments for this experiment. However, these treatments cannot be "assigned" to the corrugated fiberboard shipping containers (experimental units). A container is of size A, or size B, etc.; in other words, a container already has a size and cannot be randomly assigned one of the treatments. Rather, we view the treatments (box sizes) as populations from which we will select independent random samples of experimental units (containers).

**FIGURE 14.3**
MINITAB Random Assignment of Workers to Training Programs

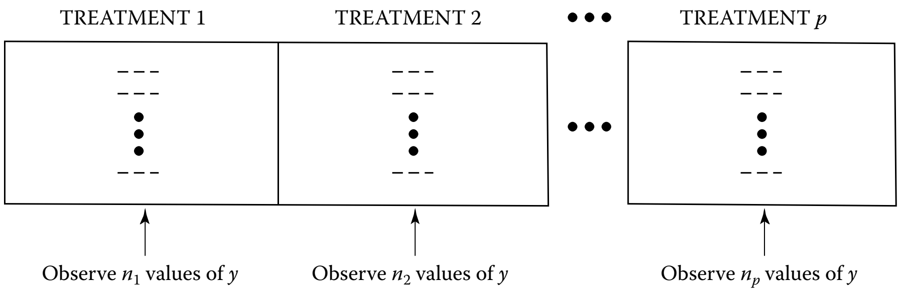
| ↓ | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| | Worker | ProgramA | ProgramB | ProgramC | |
| 1 | 1 | 21 | 25 | 6 | |
| 2 | 2 | 5 | 27 | 24 | |
| 3 | 3 | 14 | 9 | 1 | |
| 4 | 4 | 7 | 2 | 12 | |
| 5 | 5 | 13 | 28 | 10 | |
| 6 | 6 | 18 | 3 | 8 | |
| 7 | 7 | 20 | 17 | 11 | |
| 8 | 8 | 22 | 16 | 4 | |
| 9 | 9 | 29 | 15 | 30 | |
| 10 | 10 | 26 | 23 | 19 | |
| 11 | 11 | | | | |
| 12 | 12 | | | | |
| 13 | 13 | | | | |
| 14 | 14 | | | | |
| 15 | 15 | | | | |
| 16 | 16 | | | | |
| 17 | 17 | | | | |
| 18 | 18 | | | | |
| 19 | 19 | | | | |
| 20 | 20 | | | | |
| 21 | 21 | | | | |
| 22 | 22 | | | | |
| 23 | 23 | | | | |
| 24 | 24 | | | | |
| 25 | 25 | | | | |
| 26 | 26 | | | | |
| 27 | 27 | | | | |
| 28 | 28 | | | | |
| 29 | 29 | | | | |
| 30 | 30 | | | | |

A completely randomized design involves a comparison of the means for a number, say, $p$, of treatments, based on independent random samples of $n_1, n_2, \ldots, n_p$ observations, drawn from populations associated with treatments $1, 2, \ldots, p$, respectively. We repeat our definition of a completely randomized design (given in Section 13.4) with this modification. The general layout for a completely randomized design is shown in Figure 14.4.

**Definition 14.1**

A **completely randomized design** to compare $p$ treatment means is one in which the treatments are randomly assigned to the experimental units, or in which independent random samples are drawn from each of the $p$ populations.

**FIGURE 14.4**
Layout for a completely randomized design




After collecting the data from a completely randomized design, we want to make inferences about $p$ population means where $\mu_i$ is the mean of the population of measurements associated with treatment $i$, for $i = 1, 2, \ldots, p$. The null hypothesis to be tested is that the $p$ treatment means are equal, i.e., $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_p$, and the alternative hypothesis we wish to detect is that at least two of the treatment means differ. The appropriate linear model for the response $y$ is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

where

$$x_1 = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if treatment 3} \\ 0 & \text{if not} \end{cases} \quad \cdots x_{p-1} = \begin{cases} 1 & \text{if treatment } p \\ 0 & \text{if not} \end{cases}$$

and (arbitrarily) treatment 1 is the base level. Recall that this 0–1 system of coding implies that

$$\begin{aligned} \beta_0 &= \mu_1 \\ \beta_1 &= \mu_2 - \mu_1 \\ \beta_2 &= \mu_3 - \mu_1 \\ &\vdots \\ \beta_{p-1} &= \mu_p - \mu_1 \end{aligned}$$

The null hypothesis that the $p$ population means are equal is equivalent to the null hypothesis that all the treatment differences equal 0, i.e.,

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

To test this hypothesis using regression, we use the technique of Section 12.8; that is, we compare the sum of squares for error, $\text{SSE}_\text{R}$, for the nested *reduced* model

$$E(y) = \beta_0$$

to the sum of squares for error, $\text{SSE}_\text{C}$, for the *complete* model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

using the $F$ statistic

$$\begin{aligned} F &= \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/\text{Number of } \beta \text{ parameters in } H_0}{\text{SSE}_\text{C}/[n - (\text{Number of } \beta \text{ parameters in the complete model})]} \\ &= \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{SSE}_\text{C}/(n-p)} \\ &= \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{MSE}_\text{C}} \end{aligned}$$

where $F$ is based on $\nu_1 = (p - 1)$ and $\nu_2 = (n - p)$ df. If $F$ exceeds the upper critical value, $F_\alpha$, we reject $H_0$ and conclude that at least one of the treatment differences, $\beta_1, \beta_2, \ldots, \beta_{p-1}$, differs from zero; i.e., we conclude that at least two treatment means differ.

**Example 14.2**

ANOVA *F* Statistic for Completely Randomized Design

Show that the *F* statistic for testing the equality of treatment means in a completely randomized design is equivalent to a global *F* test of the complete model.

Solution

Since the reduced model contains only the $\beta_0$ term, the least-squares estimate of $\beta_0$ is $\bar{y}$, and it follows that

$$\text{SSE}_\text{R} = \sum (y - \bar{y})^2 = \text{SS}_{yy}$$

We called this quantity the sum of squares for total in Chapter 12. The difference $(\text{SSE}_\text{R} - \text{SSE}_\text{C})$ is simply $(\text{SS}_{yy} - \text{SSE})$ for the complete model. Since in regression $(\text{SS}_{yy} - \text{SSE}) = \text{SS (Model)}$, and the complete model has $(p - 1)$ terms (excluding $\beta_0$),

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{MSE}_\text{C}} = \frac{\text{SS (Model)}/(p-1)}{\text{MSE}} = \frac{\text{MS (Model)}}{\text{MSE}}$$

Thus, it follows that the test statistic for testing the null hypothesis,

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_p$$

in a completely randomized design is the same as the $F$ statistic for testing the global utility of the complete model for this design.

The regression approach to analyzing data from a completely randomized design is summarized in the next box. Note that the test requires several assumptions about the distributions of the response $y$ for the $p$ treatments and that these *assumptions are necessary regardless of the sizes of the samples*. (We have more to say about these assumptions in Section 14.9.)

### Model and *F* Test for a Completely Randomized Design with *p* Treatments

*Complete model*: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$

$$\text{where} \quad x_1 = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if treatment 3} \\ 0 & \text{if not} \end{cases}, \ldots,$$

$$x_{p-1} = \begin{cases} 1 & \text{if treatment } p \\ 0 & \text{if not} \end{cases}$$

$H_0$: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ (i.e., $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_p$)

$H_\text{a}$: At least one of the $\beta$ parameters listed in $H_0$ differs from 0 (i.e., $H_\text{a}$: At least two means differ)

*Test statistic:* $F = \dfrac{\text{MS(Model)}}{\text{MSE}}$

*Rejection region:* $F > F_\alpha$,
$p$-value: $P(F > F_c)$
where the distribution of $F$ is based on $\nu_1 = p - 1$ and $\nu_2 = (n - p)$ degrees of freedom, and $F_c$ is the computed value of the test statistic.

*Assumptions:*
1. All $p$ population probability distributions corresponding to the $p$ treatments are normal.
2. The population variances of the $p$ treatments are equal.

## Example 14.3

**Comparing Mean Wear Data for Three Paint Types: Regression**

An experiment was conducted to compare the wearing qualities of three types of paint when subjected to the abrasive action of a slowly rotating cloth-surfaced wheel. Ten paint specimens were tested for each paint type, and the number of hours until visible abrasion was apparent was recorded for each specimen. The data (with totals) are shown in Table 14.1. Is there sufficient evidence to indicate a difference in the mean time until abrasion is visibly evident for the three paint types? Test using $\alpha = .05$.

PAINTWEAR

**TABLE 14.1 Wear Data for Three Types of Paint**

| | Paint Type | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| | 148 | 513 | 335 |
| | 76 | 264 | 643 |
| | 393 | 433 | 216 |
| | 520 | 94 | 536 |
| | 236 | 535 | 128 |
| | 134 | 327 | 723 |
| | 55 | 214 | 258 |
| | 166 | 135 | 380 |
| | 415 | 280 | 594 |
| | 153 | 304 | 465 |
| Sample means: | $\bar{y}_1 = 229.6$ | $\bar{y}_2 = 309.9$ | $\bar{y}_3 = 427.8$ |

**Solution**

The experiment involves a single factor, paint type, at three levels. Thus, we have a completely randomized design with $p = 3$ treatments. Let $\mu_1$, $\mu_2$, and $\mu_3$ represent the mean abrasion times for paint types 1, 2, and 3, respectively. Then we want to test

$$H_0: \quad \mu_1 = \mu_2 = \mu_3$$

against

$H_a$: At least two of the three means differ

The appropriate linear model for $p = 3$ treatments is

$$\textit{Complete model}: \quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$$x_1 = \begin{cases} 1 & \text{if paint type 1} \\ 0 & \text{if not} \end{cases} \quad \text{and} \quad x_2 = \begin{cases} 1 & \text{if paint type 2} \\ 0 & \text{if not} \end{cases}$$

Thus, we want to test $H_0$: $\beta_1 = \beta_2 = 0$.

The MINITAB regression analysis for the complete model is shown in Figure 14.5. The $F$ statistic for testing the overall adequacy of the model (shaded on the printout) is $F = 3.48$, where the distribution of $F$ is based on $\nu_1 = (p - 1) = 3 - 1 = 2$ and $\nu_2 = (n - p) = 30 - 3 = 27$ df. For $\alpha = .05$, the critical value (obtained from Table 10 of Appendix B) is $F_{.05} = 3.35$ (see Figure 14.6).

Since the computed value of $F$, 3.48, exceeds the critical value, $F_{.05} = 3.35$, we reject $H_0$ and conclude (at the $\alpha = .05$ level of significance) that the mean time to visible abrasion differs for at least two of the three paint types. We can arrive at the same conclusion by noting that $\alpha = .05$ is greater than the $p$-value (.045) shaded on the printout.

```
Regression Analysis: HOURS versus X1, X2

The regression equation is
HOURS = 428 - 198 X1 - 118 X2


Predictor      Coef  SE Coef      T      P
Constant     427.80    53.43   8.01  0.000
X1          -198.20    75.56  -2.62  0.014
X2          -117.90    75.56  -1.56  0.130


S = 168.948   R-Sq = 20.5%   R-Sq(adj) = 14.6%


Analysis of Variance

Source          DF      SS     MS     F      P
Regression       2  198772  99386  3.48  0.045
Residual Error  27  770671  28543
Total           29  969443
```

**FIGURE 14.5**
MINITAB regression printout for the completely randomized design, Example 14.3

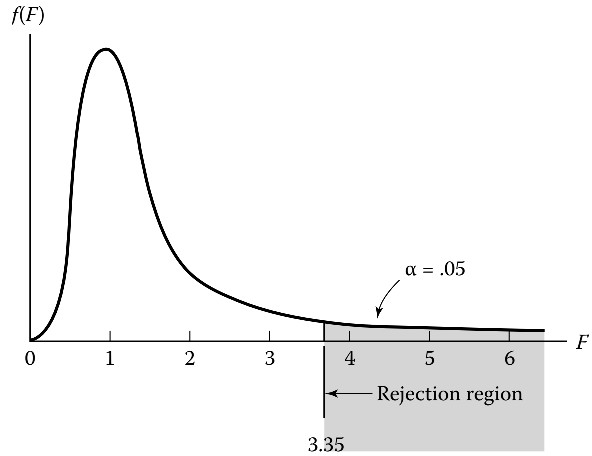
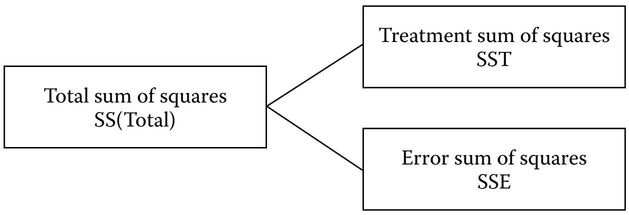



**FIGURE 14.6**
Rejection region for Example 14.3; numerator df $= 2$, denominator df $= 27$, $\alpha = .05$

**FIGURE 14.7**
The partitioning of SS(Total) for a completely randomized design




The analysis of the data in Example 14.3 can also be accomplished using ANOVA computing formulas. In Section 14.2, we learned that an analysis of variance partitions SS(Total) $= \Sigma(y - \bar{y})^2$ into two components, SSE and SST (see Figure 14.6).

Recall that the quantity SST denotes the sum of squares for treatments and measures the variation explained by the differences between the treatment means. The sum of squares for error, SSE, is a measure of the unexplained variability, obtained by calculating a pooled measure of the variability *within* the $p$ samples. If the treatment means truly differ, then SSE should be substantially smaller than SST. We compare the two sources of variability by forming an $F$ statistic:

$$F = \frac{\text{SST}/(p-1)}{\text{SSE}/(n-p)} = \frac{\text{MST}}{\text{MSE}}$$

where $n$ is the total number of measurements. The numerator of the $F$ statistic, MST $= \text{SST}/(p - 1)$, denotes **mean square for treatments** and is based on $(p - 1)$ degrees of freedom—one for each of the $p$ treatments minus one for the estimation of the overall mean. The denominator of the $F$ statistic, MSE $= \text{SSE}/(n - p)$, denotes **mean square for error** and is based on $(n - p)$ degrees of freedom—one for each of the $n$ measurements minus one for each of the $p$ treatment means being estimated. We have already demonstrated that this $F$ statistic is identical to the global $F$ value for the regression model specified earlier.

For completeness, we provide the computing formulas for an analysis of variance in the next box.

### ANOVA Computing Formulas for a Completely Randomized Design

$$\text{Sum of all } n \text{ measurements} = \sum_{i=1}^{n} y_i$$

$$\text{Mean of all } n \text{ measurements} = \bar{y}$$

$$\text{Sum of squares of all } n \text{ measurements} = \sum_{i=1}^{n} y_i^2$$

$$\begin{aligned}
\text{CM} &= \text{Correction for mean} \\
&= \frac{(\text{Total of all observations})^2}{\text{Total number of observations}} = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\
\text{SS(Total)} &= \text{Total sum of squares} \\
&= (\text{Sum of squares of all observations}) - \text{CM} \\
&= \sum_{i=1}^{n} y_i^2 - \text{CM} \\
\text{SST} &= \text{Sum of squares for treatments} \\
&= \left(\begin{array}{c}\text{Sum of squares of treatment totals with} \\ \text{each square divided by the number of} \\ \text{observations for that treatment}\end{array}\right) - \text{CM} \\
&= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \cdots + \frac{T_p^2}{n_p} - \text{CM} \\
\text{SSE} &= \text{Sum of squares for error} = \text{SS(Total)} - \text{SST} \\
\text{MST} &= \text{Mean square for treatments} = \frac{\text{SST}}{p-1} \\
\text{MSE} &= \text{Mean square for error} = \frac{\text{SSE}}{n-p} \\
F &= \frac{\text{MST}}{\text{MSE}}
\end{aligned}$$

**Example 14.4**

**Comparing Mean Wear Data for Three Paint Types: ANOVA**

Refer to Example 14.3. Analyze the data of Table 14.2 using the ANOVA approach. Use $\alpha = .05$.

**Solution**

Rather than perform the tedious calculations by hand (we leave this for the student as an exercise), we resort to a statistical software package.

The SAS ANOVA printout is shown in Figure 14.8. The value of the test statistic (shaded on the printout) is $F = 3.48$. Note that this is identical to the $F$ value obtained using the regression approach in Example 14.3. The $p$-value of the test (also shaded) is $p = .0452$. (Likewise, this quantity is identical to that in Example 14.3.) Since $\alpha = .05$ exceeds this $p$-value, we have sufficient evidence to conclude that the treatments differ.

```
                          The ANOVA Procedure

Dependent Variable: WEAR

                                         Sum of
    Source                  DF          Squares     Mean Square    F Value    Pr > F

    Model                    2      198772.4667     99386.2333       3.48    0.0452

    Error                   27      770670.9000     28543.3667

    Corrected Total         29      969443.3667

                   R-Square     Coeff Var      Root MSE     WEAR Mean

                   0.205038      52.39775      168.9478      322.4333

    Source                  DF         Anova SS     Mean Square    F Value    Pr > F

    TYPE                     2      198772.4667     99386.2333       3.48    0.0452
```

**FIGURE 14.8**
SAS ANOVA Output for Example 14.4

The results of an analysis of variance are often summarized in tabular form. The general form of an ANOVA table for a completely randomized design is shown in the next box. The column head **SOURCE** refers to the source of variation, and for each source, **DF** refers to the degrees of freedom, **SS** to the sum of squares, **MS** to the mean square, and **F** to the $F$ statistic comparing the treatment mean square to the error mean square. Table 14.3 is the ANOVA summary table corresponding to the analysis of variance data for Example 14.4 obtained from the SAS printout.

**TABLE 14.2 ANOVA Summary Table for Example 14.4**

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Paint types | 2 | 198,772 | 99,386 | 3.48 |
| Error | 27 | 770,671 | 28,543 | |
| Total | 29 | 969,443 | | |

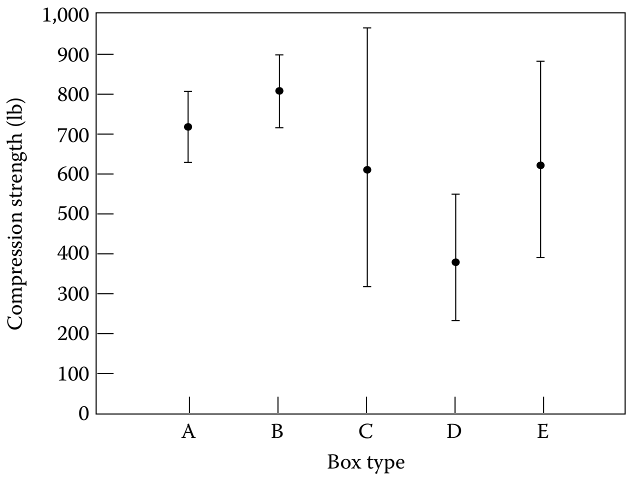
### ANOVA Summary Table for a Completely Randomized Design

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Treatments | $p - 1$ | SST | MST | MST/MSE |
| Error | $n - p$ | SSE | MSE | |
| Total | $n - 1$ | SS(Total) | | |

Once differences among treatment means are established in ANOVA, it is often important to rank the means from lowest to highest. Two useful methods for ranking treatment means in ANOVA are presented in Section 14.8.

## Applied Exercises

14.1 *Strength of fiberboard boxes.* The *Journal of Testing and Evaluation* (July 1992) published an investigation of the mean compression strength of corrugated fiberboard shipping containers. Comparisons were made for boxes of five different sizes: A, B, C, D, and E. Twenty identical boxes of each size were tested and the peak compression strength (pounds) recorded for each box. The accompanying figure shows the sample means for the five box types as well as the variation around each sample mean.




*Source*: Singh, S. P., et al. "Compression of single-wall corrugated shipping containers using fixed and floating test platens." *Journal of Testing and Evaluation*, Vol. 20, No. 4, July 1992, p. 319 (Figure 3).

a. Explain why the data are collected as a completely randomized design.
b. Refer to box types B and D. Based on the graph, does it appear that the mean compression strengths of these two box types are significantly different? Explain.
c. Based on the graph, does it appear that the mean compression strengths of all five box types are significantly different? Explain.

14.2 *Properties of cemented soils.* Refer to the *Bulletin of Engineering Geology and the Environment* (Vol. 69, 2010) study of cemented sandy soils, Exercise 13.8 (p. 727). Recall that the researchers applied one of three different sampling methods (rotary core, metal tube, or plastic tube) to randomly selected soil specimens, then measured the effective stress level (Newtons per meters-squared) of each specimen. Each method was applied to 10 soil specimens—a total of 30 measurements in all. For this completely randomized design, use a random number generator to assign the sampling methods to the soil specimens. List the soil specimens that are to analyzed by each method.

14.3 *A new dental bonding agent. Trends in Biomaterials & Artificial Organs* (Jan. 2003) published a study of a new bonding adhesive for teeth. The new adhesive (called "Smartbond") has been developed to eliminate the necessity of a dry field. In one portion of the study, 30 extracted teeth were bonded with Smartbond and each was randomly assigned one of three different bonding times: 1 hour, 24 hours, or 48 hours. At the end of the bonding period, the breaking strength (in MPa) of each tooth was determined. The data were analyzed using analysis of variance in order to determine if true mean breaking strength of the new adhesive differs depending on the length of bonding time.

a. Identify the experimental units, treatments, and response variable for this completely randomized design.
b. Set up the null and alternative hypothesis for the ANOVA.
c. Find the rejection region for the test using $\alpha = .01$.
d. The test results were $F = 61.62$ and $p$-value $\approx 0$. Give the appropriate conclusion for the test.
e. What conditions are required for the test results to be valid?

14.4 *Robots trained to behave like ants.* Robotics researchers investigated whether robots could be trained to behave like ants in an ant colony (*Nature*, Aug. 2000). Robots were trained and randomly assigned to "colonies" (i.e., groups) consisting of 3, 6, 9, or 12 robots. The robots were assigned the task of foraging for "food" and to recruit another robot when they identified a resource-rich area. One goal of the experiment was to compare the mean energy expended (per robot) of the four different colony sizes.

a. What type of experimental design was employed?
b. Identify the treatments and the dependent variable.
c. Set up the null and alternative hypotheses of the test.
d. The following ANOVA results were reported: $F = 7.70$, numerator df $= 3$, denominator df $= 56$, $p$-value $< .001$. Conduct the test at a significance level of $\alpha = .05$ and interpret the result.

14.5 *Whales entangled in fishing gear.* Entanglement of marine mammals (e.g., whales) in fishing gear is considered a significant threat to the species. A study published in *Marine Mammal Science* (April 2010) investigated the type of net most likely to entangle a certain species of whale inhabiting the East Sea of Korea. A sample of 207 entanglements of whales in the area formed the data for the study. These entanglements were caused by one of three types of fishing gear: set nets, pots, and gill nets. One of the variables investigated was body length (in meters) of the entangled whale.

a. Set up the null and alternative hypotheses for determining whether the average body length of entangled whales differs for the three types of fishing gear.
b. An ANOVA $F$-test yielded the following results: $F = 34.81$, $p$-value $< .0001$. Interpret the results for $\alpha = .05$.

14.6 *Performance of a bus depot.* The performances of public bus depots in India were evaluated and ranked in the *International Journal of Engineering Science and Technology* (February 2011). A survey was administered to 150 customers selected randomly and independently at each of three different bus depots (Depot 1, Depot 2, and Depot 3); thus, the total sample consisted of 450 bus customers. Based on responses to 16 different items (e.g., bus punctuality, seat comfort, luggage service, etc.), a performance score (out of 100 total points) was calculated for each customer. The average performance scores were compared across the three bus depots using an analysis of variance. The ANOVA $F$-test resulted in a $p$-value of .0001.

a. Give details (experimental units, dependent variable, factor, treatments) on the experimental design utilized in this study.
b. The researchers concluded that the "mean customer performance scores differed across the three bus depots at a 95% confidence level". Do you agree?

14.7 *Evaluation of flexography printing plates.* Flexography is a printing process used in the packaging industry. The process is popular since it is cost-effective and can be used to print on a variety of surfaces (e.g., paperboard, foil, and plastic). A study was conducted to determine if flexography exposure time has an impact on the quality of the printing (*Journal of Graphic Engineering and Design,* Vol. 3, 2012). Four different exposure times were studied: 8, 10, 12, and 14 minutes. A sample of 36 print images were collected at each exposure time level, for a total of 144 print images. The measure of print quality used was dot area (hundreds of dots per square millimeter). The data were subjected to an analysis of variance, with partial results shown in the next table.

| Source | df | SS | MS | F | $p$-value |
|---|---|---|---|---|---|
| Exposure | 3 | .010 | --- | --- | <.001 |
| Error | 140 | .029 | --- | | |
| Total | 143 | .039 | | | |

a. Compute the missing entries in the ANOVA table.
b. Is there sufficient evidence to indicate that mean dot area differs depending on the exposure time? Use $\alpha = .05$.

**GASTURBINE**

14.8 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of gas turbines augmented with high-pressure inlet fogging, Exercise 8.39 (p. 399). Recall that the researchers classified gas turbines into three categories—traditional, advanced, and aeroderivative—and measured the heat rate (kilojoules per kilowatt per hour) of sampled gas turbines of each type. The data in the **GASTURBINE** file was analyzed using a completely randomized design ANOVA; the results are shown in the MINITAB printout below.

a. State the null and alternative hypothesis for the ANOVA $F$ test.
b. Give the value of the test statistic and corresponding $p$-value.
c. Is there sufficient evidence of differences among the mean heat rates of the three gas turbine types? Test using $\alpha = .01$.

14.9 *High-strength aluminum alloys.* Refer to the *JOM* (Jan. 2003) study of a new high-strength aluminum alloy for use in antisubmarine aircraft, tankers, and long-range bombers, Exercise 7.43 (p. 320). Recall that the new alloy is obtained by applying a retrogression and reaging (RAA)

### MINITAB Output for Exercise 14.8

**One-way ANOVA: HEATRATE versus ENGINE**

```
Source  DF         SS         MS      F      P
ENGINE   2   55360022   27680011  15.74  0.000
Error   64  112537186    1758394
Total   66  167897208

S = 1326   R-Sq = 32.97%   R-Sq(adj) = 30.88%


                               Individual 95% CIs For Mean Based on
                               Pooled StDev
Level         N   Mean  StDev  ---+---------+---------+---------+------
Advanced     21   9764    639  (---*----)
Aeroderiv     7  12312   2652                   (---------*--------)
Traditional  39  11544   1279                 (--*---)
                               ---+---------+---------+---------+------
                               9600     10800     12000     13200

Pooled StDev = 1326
```

heat treatment to the current strongest aluminum alloy. Strength tests conducted on three specimens of the new RAA alloy and three specimens of the current strongest alloy yielded the yield strength results (measured in megapascals, MPa) shown in the table.

**ALLOY2**

| *Current* (1) | *RAA* (2) |
|---|---|
| 604 | 662 |
| 595 | 624 |
| 580 | 637 |

a. Give the linear model appropriate for analyzing the data using regression.
b. Fit the model, part **a**, to the data and conduct the analysis. Summarize the results in an ANOVA table.
c. Calculate MST for the data using the ANOVA formulas. What type of variability is measured by this quantity? Does this value agree with MST in the ANOVA table, part **b**?
d. Calculate MSE for the data using the ANOVA formulas. What type of variability is measured by this quantity? Does this value agree with MSE in the ANOVA table, part **b**?
e. How many degrees of freedom are associated with MST?
f. How many degrees of freedom are associated with MSE?
g. Compute the test statistic appropriate for testing $H_0$: $\mu_1 = \mu_2$ against the alternative that the two treatment means differ, using a significance level of $\alpha = .05$. (Compare the value to the test statistic obtained using regression in part **b**.)
h. Specify the rejection region, using a significance level of $\alpha = .05$.
i. State the proper conclusion in the words of the problem.
j. Use the independent samples Student's $T$ test of Section 8.7 to test $H_0$: $\mu_1 = \mu_2$ against the alternative hypothesis $H_a$: $\mu_1 \neq \mu_2$. Test using $\alpha = .05$.
k. It can be shown (proof omitted) that an $F$ statistic with $\nu_1 = 1$ numerator degree of freedom and $\nu_2$ denominator degrees of freedom is equal to $T^2$, where $T$ is a Student's $T$ statistic based on $\nu_2$ degrees of freedom. Square the value of $T$ calculated in part **j**, and show that it is equal to the value of $F$ calculated in part **g**.
l. Is the analysis of variance $F$ test for comparing two population means a one- or a two-tailed test of $H_0$: $\mu_1 = \mu_2$? (*Hint*: Although the $T$ test can be used to test for either $H_a$: $\mu_1 > \mu_2$ or $H_a$: $\mu_1 < \mu_2$, the alternative hypothesis for the $F$ test is $H_a$: The two means are different.)

14.10 *Virtual reality-based rehabilitation systems.* Hand rehabilitation systems that use virtual reality (VR) technology have been developed for patients with disabled hands and arms. In *Robotica* (Vol. 22, 2004), mechanical and systems engineers assessed the effectiveness of display devices for three VR-based hand rehabilitation systems. System A employs a projector to manipulate VR simulation, System B uses a desktop computer monitor, and System C uses a head-mounted display. Twelve nondisabled, right-handed male subjects were randomly assigned to the three VR systems, four subjects in each group. Each subject was required to carry out a "pick-and-place" procedure using the VR system. The variable of interest to the engineers was collision frequency, measured as the number of collisions between moved objects. One objective of the study was to compare the mean collision frequencies of subjects in the three VR-based hand rehabilitation systems.

a. Identify the experimental units for this study.
b. Identify the treatments.
c. Identify the dependent variable.
d. Explain why a completely randomized design was employed to collect the data.
e. Propose a regression model that will allow you to compare the means.
f. Give the null and alternative hypotheses for the ANOVA $F$ test.
g. The ANOVA $F$ test was found to be nonsignificant at $\alpha = .05$. Give a practical interpretation of this result.

14.11 *Estimating the age of glacial drifts.* Refer to the *American Journal of Science* (Jan. 2005) study of the chemical makeup of buried tills (glacial drifts) in Wisconsin, Exercise 2.22 (p. 38). The ratio of the elements aluminum (Al) and beryllium (Be) in sediment is related to the duration of burial. Recall that the Al/Be ratios for a sample of 26 buried till specimens were determined. The till specimens were obtained from five different boreholes (labeled UMRB-1, UMRB-2, UMRB-3, SWRA, and SD). The data are shown in the table. Conduct an analysis of variance of the data. Is there sufficient evidence to indicate differences among the mean Al/Be ratios for the five boreholes? Test using $\alpha = .10$.

**TILLRATIO**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *UMRB-1:* | 3.75 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 |
| *UMRB-2:* | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 | |
| *UMRB-3:* | 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 |
| *SWRA:* | 2.73 | 2.95 | 2.25 | | | | |
| *SD:* | 2.73 | 2.55 | 3.06 | | | | |

*Source*: Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

14.12 *Soil scouring and overturned trees.* Trees that grow in flood plains are susceptible to overturning. This is typically due to floodwaters exposing the tree roots (called soil *scouring*). Environmental engineers at Saitama University (Japan) investigated the impact of soil scouring on the characteristics of overturned and uprooted trees (*Landscape Ecology Engineering,* January 2013). Tree pulling experiments were conducted in the floodplains of the Komagama river. Trees were randomly selected to be uprooted in each of three areas that had different scouring

### MINITAB Output for Exercise 14.12

**One-way ANOVA: MOMENT versus SCOURING**

```
Source     DF      SS     MS     F      P
SCOURING    2   528.5  264.3  5.40  0.021
Error      12   586.9   48.9
Total      14  1115.4

S = 6.993   R-Sq = 47.38%   R-Sq(adj) = 38.61%
```

conditions: no scouring (NS), shallow scouring (SS), and deep scouring (DS). During the uprooting of the trees, the maximum resistive bending moment at the trunk base (kiloNewton-meters) was measured. Simulated data for five medium-sized trees selected at each area are shown in the table below. A MINITAB printout of the analysis is provided above. Interpret the results. Does soil scouring impact the mean maximum resistive bending moment at the tree trunk base?

**SCOURING**

| None | Shallow | Deep |
|---|---|---|
| 23.68 | 11.13 | 4.27 |
| 8.88 | 29.19 | 2.36 |
| 7.52 | 13.66 | 8.48 |
| 25.89 | 20.47 | 12.09 |
| 22.58 | 23.24 | 3.46 |

14.13 *Effect of scopolamine on memory.* The drug scopolamine is often used as a sedative to induce sleep in patients. In *Behavioral Neuroscience* (Feb. 2004), medical researchers examined scopolamine's effects on memory for word-pair associates. A total of 28 human subjects, recruited from a university community, were given a list of related word pairs to memorize. For every word pair in the list (e.g., robber-jail), there was an associated word pair with the same first word but a different second word (e.g., robber-police). The subjects were then randomly divided into three treatment groups. Group 1 subjects were administered an injection of scopolamine, group 2 subjects were given an injection of glycopyrrolate (an active placebo), and group 3 subjects were not given any drug. Four hours later, subjects were shown 12 word pairs from the list and tested on how many of the associated word pairs they could recall. The data on number of pairs recalled (simulated based on summary information provided in the research article) are listed in the next table. Prior to the analysis, the researchers theorized that the mean number of word pairs recalled for the scopolamine subjects (group 1) would be less than the corresponding means for the other two groups.

a. Explain why this is a completely randomized design.
b. Identify the treatments and response variable.
c. Find the sample means for the three groups. Is this sufficient information to support the researchers' theory? Explain.
d. Conduct the ANOVA $F$ test on the data. Is there sufficient evidence (at $\alpha = .05$) to conclude that the mean number of word pairs recalled differs among the three treatment groups?

**SCOPOLAMINE**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Group 1 (Scopolamine):* | 5 | 8 | 8 | 6 | 6 | 6 | 6 | 8 | 6 | 4 | 5 | 6 |
| *Group 2 (Placebo):* | 8 | 10 | 12 | 10 | 9 | 7 | 9 | 10 | | | | |
| *Group 3 (No drug):* | 8 | 9 | 11 | 12 | 11 | 10 | 12 | 12 | | | | |

14.14 *Effects of temperature on ethanol production.* A low cost and highly productive bio-fuel production method is high-temperature fermentation. However, heat stress can inhibit the amount of ethanol produced during the process. In *Engineering Life Sciences* (March 2013), biochemical engineers carried out a series of experiments to assess the effect of temperature on ethanol production during fermentation. The maximum inhibitory concentration of ethanol (grams per liter) was measured at four different temperatures (30º, 35º, 40º, and 45º Celsius). The experiment was replicated 3 times, with the data shown in the table. (Note: The data is simulated based on information provided in the journal article.) Do the data indicate that high temperatures inhibit mean concentration of ethanol? Test using $\alpha = .10$.

**FERMENT**

| Temperature | | | |
|---|---|---|---|
| 30º | 35º | 40º | 45º |
| 103.3 | 101.7 | 97.2 | 55.0 |
| 103.4 | 102.0 | 96.9 | 56.4 |
| 101.0 | 101.1 | 96.2 | 54.9 |

14.15 *"Creep" in concrete.* Unlike most other commonly used engineering materials, concrete experiences a characteristic marked increase in "creep" when it is heated for the first time under load. To investigate this phenomenon, a study of the thermal strain behavior of concrete was conducted (*Magazine of Concrete Research*, Dec. 1985). Concrete specimens were prepared and a constant load applied to each. The test specimens were then heated to a specified temperature at a rate of 1°C per minute, with the specimens randomly assigned to one of five temperature settings (100°, 200°, 300°, 400°, and 500°C). For each specimen, the difference between the free (unloaded) thermal strain and load-induced thermal strain, called the *total thermal strain*, was calculated. The sample size, mean, and standard deviation of the total thermal strain values for each temperature setting are given in the table on p. 758. (*Note*: Thermal strain is recorded in units $\times 10^6$.)

| Temperature | Number of Specimens | Mean | Standard Deviation |
|---|---|---|---|
| 100 | 16 | 52 | 55 |
| 200 | 16 | 112 | 108 |
| 300 | 16 | 143 | 127 |
| 400 | 16 | 186 | 136 |
| 500 | 14 | 257 | 178 |

*Source*: Khoury, G. A., Grainger, B. N., and Sullivan, P. J. E. "Strain of concrete during first heating to 600°C under load." *Magazine of Concrete Research*, Vol. 37, No. 133, Dec. 1985, p. 198 (Table 2).

a. Find the totals for each temperature and the total for all 78 strain measurements. Then compute CM and SST.
b. Since we do not know the value of $\sum_{i=1}^{n} y_i^2$, calculate SSE using the pooled method:

$$\begin{aligned}\text{SSE} &= \sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2 \\ &\quad + \cdots + \sum_{i=1}^{n_5}(y_{i5} - \bar{y}_5)^2 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \\ &\quad + \cdots + (n_5 - 1)s_5^2\end{aligned}$$

c. Find SS(Total).
d. Construct an analysis of variance table for the data.
e. Is there sufficient evidence to indicate that heating temperature affects the mean total thermal strain of concrete? Test using $\alpha = .01$.

**OILSPILL**

14.16 *Hull failures of oil tankers.* Refer to the *Marine Technology* (Jan. 1995) study of major ocean oil spills by tanker vessels, Exercise 2.85 (p. 74). The spillage amounts (thousands of metric tons) and cause of accident for 48 tankers are saved in the **OILSPILL** file. (*Note*: Delete the two tankers with oil spills of unknown causes.) Conduct an analysis of variance (at $\alpha = .01$) to compare the mean spillage amounts for the four accident types: (1) collision, (2) grounding, (3) fire/explosion, and (4) hull failure. Interpret your results.

## 14.4 Randomized Block Designs

Randomized block design is a commonly used noise-reducing design. Recall (Definition 13.9) that a randomized block design employs groups of homogeneous experimental units (matched as closely as possible) to compare the means of the populations associated with $p$ treatments. The general layout of a randomized block design is shown in Figure 14.9. Note that there are $b$ blocks of relatively homogeneous experimental units. Since each treatment must be represented in each block, the blocks each

**FIGURE 14.9**
General form of a randomized block design (treatment is denoted by $T_p$)

Block

| 1 | 2 | 3 | ... | $b$ |
|---|---|---|---|---|
| $T_1$ | $T_1$ | $T_1$ | ... | $T_1$ |
| $T_2$ | $T_2$ | $T_2$ | ... | $T_2$ |
| $T_3$ | $T_3$ | $T_3$ | ... | $T_3$ |
| $\vdots$ | $\vdots$ | $\vdots$ | ... | $\vdots$ |
| $T_p$ | $T_p$ | $T_p$ | ... | $T_p$ |

contain $p$ experimental units. Although Figure 14.9 shows the $p$ treatments in order within the blocks, in practice they would be assigned to the experimental units in random order (hence the name **randomized block design**).

The complete model for a randomized block design contains $(p - 1)$ dummy variables for treatments and $(b - 1)$ dummy variables for blocks. Therefore, the total number of terms in the model, excluding $\beta_0$, is $(p - 1) + (b - 1) = p + b - 2$, as shown here.

*Complete model:*

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}}_{\text{Treatment effects}} + \underbrace{\beta_p x_p + \cdots + \beta_{p+b-2} x_{p+b-2}}_{\text{Block effects}}$$

where

$$x_1 = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if treatment 3} \\ 0 & \text{if not} \end{cases} \cdots \quad x_{p-1} = \begin{cases} 1 & \text{if treatment } p \\ 0 & \text{if not} \end{cases}$$

$$x_p = \begin{cases} 1 & \text{if block 2} \\ 0 & \text{if not} \end{cases} \quad x_{p+1} = \begin{cases} 1 & \text{if block 3} \\ 0 & \text{if not} \end{cases} \cdots \quad x_{p+b-2} = \begin{cases} 1 & \text{if block } b \\ 0 & \text{if not} \end{cases}$$

Note that the model does *not* include terms for treatment–block interaction. The reasons are twofold. First, the addition of these terms would leave 0 degrees of freedom for estimating $\sigma^2$. Second, the failure of the mean difference between a pair of treatments to remain the same from block to block is, by definition, experimental error. In other words, in a randomized block design, treatment–block interaction and experimental error are synonymous.

The primary objective of the analysis is to compare the $p$ treatment means, $\mu_1, \mu_2, \ldots, \mu_p$. That is, we want to test the null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_p$$

Recall (Section 13.3) that this is equivalent to testing whether all the treatment parameters in the complete model are equal to 0, i.e.,

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

To perform this test using regression, we drop the treatment terms and fit the reduced model:

*Reduced model for testing treatments:*

$$E(y) = \beta_0 + \underbrace{\beta_p x_p + \beta_{p+1} x_{p+1} + \cdots + \beta_{p+b-2} x_{p+b-2}}_{\text{Block effects}}$$

Then we compare the SSEs for the two models, $\text{SSE}_\text{R}$ and $\text{SSE}_\text{C}$, using the "partial" $F$ statistic:

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/\text{Number of } \beta\text{'s tested}}{\text{MSE}_\text{C}} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{MSE}_\text{C}}$$

A significant $F$ value implies that the treatment means differ.

Occasionally, experimenters want to determine whether blocking was effective in removing the extraneous source of variation, i.e., whether there is evidence of a difference among block means. In fact, if there are no differences among block means, the experimenter will lose information by blocking because blocking reduces the number of degrees of freedom associated with the estimated variance of the model, $s^2$. If blocking is *not* effective in reducing the variability, then the block parameters in the complete model will all equal 0 (i.e., there will be no differences among block means). Thus, we want to test

$$H_0: \quad \beta_p = \beta_{p+1} = \cdots = \beta_{p+b-2} = 0$$

Another reduced model, with the block $\beta$'s dropped, is fit:

*Reduced model for testing blocks*:

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}}_{\text{Treatment effects}}$$

The SSE for this second reduced model is compared to the SSE for the complete model in the usual fashion. A significant $F$ test implies that blocking is effective in removing (or reducing) the targeted extraneous source of variation.

These two tests are summarized in the following boxes.

### Models and ANOVA *F* Tests for a Randomized Block Design with *p* Treatments and *b* Blocks

*Complete model:*

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}}^{(p\,-\,1)\text{ treatment terms}} + \overbrace{\beta_p x_p + \cdots + \beta_{p+b-2} x_{p+b-2}}^{(b\,-\,1)\text{ block terms}}$$

where

$$x_1 = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{if not} \end{cases} \quad \cdots \quad x_{p-1} = \begin{cases} 1 & \text{if treatment } p \\ 0 & \text{if not} \end{cases}$$

$$x_p = \begin{cases} 1 & \text{if block 2} \\ 0 & \text{if not} \end{cases} \quad \cdots \quad x_{p+b-2} = \begin{cases} 1 & \text{if block } b \\ 0 & \text{if not} \end{cases}$$

#### TEST FOR COMPARING TREATMENT MEANS

$H_0$: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

(i.e., $H_0$: The $p$ treatment means are equal)

$H_a$: At least one of the $\beta$ parameters listed in $H_0$ differs from 0

(i.e., $H_a$: At least two treatment means differ)

*Reduced model*: $E(y) = \beta_0 + \beta_p x_p + \cdots + \beta_{p+b-2} x_{p+b-2}$

*Test statistic:*

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{SSE}_\text{C}/(n-p-b+1)} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{MSE}_\text{C}}$$

where

$\text{SSE}_\text{R}$ = SSE for reduced model
$\text{SSE}_\text{C}$ = SSE for complete model
$\text{MSE}_\text{C}$ = MSE for complete model

*Rejection region:* $F > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F$ is based on $\nu_1 = (p-1)$ and $\nu_2 = (n-p-b+1)$ degrees of freedom, and $F_c$ is the computed value of the test statistic.

TEST FOR COMPARING BLOCK MEANS

*Assumptions:*

$H_0$: $\beta_p = \beta_{p+1} = \cdots = \beta_{p+b-2} = 0$

(i.e., $H_0$: The $b$ block means are equal)

$H_a$: At least one of the $\beta$ parameters listed in $H_0$ differs from 0
(i.e., $H_a$: At least two block means differ)

*Reduced model:* $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1}x_{p-1}$

*Test statistic:*
$$F_c = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(b-1)}{\text{SSE}_\text{C}/(n-p-b+1)} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(b-1)}{\text{MSE}_\text{C}}$$

*Rejection region:* $F_c > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F$ is based on $\nu_1 = (b-1)$ and $\nu_2 = (n-p-b+1)$ degrees of freedom

*Assumptions:*

1. The probability distribution of the difference between any pair of treatment observations within a block is approximately normal.
2. The variance of the difference is constant and the same for all pairs of observations.

## Example 14.5

Randomized Block Design: Comparing Mean Engineer Cost Estimates

Prior to submitting a bid for a construction job, cost engineers prepare a detailed analysis of the estimated labor and materials costs required to complete the job. This estimate will depend on the engineer who performs the analysis. An overly large estimate will reduce the chance of acceptance of a company's bid price, whereas an estimate that is too low will reduce the profit or even cause the company to lose money on the job. A company that employs three job cost engineers wanted to compare the mean level of the engineers' estimates. This was done by having each engineer estimate the cost of the same four jobs. The data (in hundreds of thousands of dollars) are shown in Table 14.3.

a. Perform an analysis of variance on the data, and test to determine whether there is sufficient evidence to indicate differences among treatment means. Test using $\alpha = .05$.
b. Test to determine whether blocking on jobs was successful in reducing the job-to-job variation in the estimates. Use $\alpha = .05$.

Solution

a. The data for this experiment were collected according to a randomized block design because we would expect estimates of the same job to be more nearly alike than estimates between jobs. Thus, the experiment involves three treatments (engineers) and four blocks (jobs).

COSTENG

**TABLE 14.3 Data for the Randomized Block Design of Example 14.5**

| | | Job | | | | Treatment Means |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| | 1 | 4.6 | 6.2 | 5.0 | 6.6 | 5.60 |
| **Engineer** | 2 | 4.9 | 6.3 | 5.4 | 6.8 | 5.85 |
| | 3 | 4.4 | 5.9 | 5.4 | 6.3 | 5.50 |
| **Block Means** | | 4.63 | 6.13 | 5.27 | 6.57 | |

The complete model for this design is

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Treatments (engineers)}} + \underbrace{\beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5}_{\text{Blocks (jobs)}}$$

where

$y$ = Cost estimate

$$x_1 = \begin{cases} 1 & \text{if engineer 2} \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if engineer 3} \\ 0 & \text{if not} \end{cases}$$

Base level = Engineer 1

$$x_3 = \begin{cases} 1 & \text{if block 2} \\ 0 & \text{if not} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{if block 3} \\ 0 & \text{if not} \end{cases} \qquad x_5 = \begin{cases} 1 & \text{if block 4} \\ 0 & \text{if not} \end{cases}$$

Base level = Block 1

The SAS printout for the complete model is shown in Figure 14.10. Note that $\text{SSE}_\text{C} = .18667$ and $\text{MSE}_\text{C} = .03111$ (shaded on the printout).

To test for differences among treatment means, we will test

$$H_0\text{:} \quad \mu_1 = \mu_2 = \mu_3$$

where $\mu_i$ = mean cost estimate of engineer $i$. This is equivalent to testing

$$H_0\text{:} \quad \beta_1 = \beta_2 = 0$$

in the complete model. To proceed, we fit the reduced model

$$E(y) = \beta_0 + \underbrace{\beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5}_{\text{Blocks (jobs)}}$$

The SAS printout for this reduced model is shown in Figure 14.11. Note that $\text{SSE}_\text{R} = .44667$ (shaded on the printout).

The remaining elements of the test follow:

*Test statistic:*

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(p-1)}{\text{MSE}_\text{C}} = \frac{(.44667 - .18667)/2}{.03111} = 4.18$$

*Rejection region:* $F > 5.14$, where $F_{.05} = 5.14$ (obtained from Table 10, Appendix B) is based on $\nu_1 = (p-1) = 2$ df and $\nu_2 = (n - p - b + 1) = 6$ df

*Conclusion:* Since $F = 4.18$ is less than the critical value, 5.14, there is insufficient evidence, at the $\alpha = .05$ level of significance, to indicate the differences among the mean estimates for the three cost engineers.

As an option, SAS will conduct this nested model $F$ test. The test statistic, $F = 4.18$, is highlighted in the middle of the SAS complete model printout, Figure 14.10. The $p$-value of the test (also highlighted) is $p$-value $= .0730$. Since this value exceeds $\alpha = .05$, our conclusion is confirmed—there is insufficient evidence to reject $H_0$.

b. To test for the effectiveness of blocking on jobs, we test

$$H_0\text{:} \quad \beta_3 = \beta_4 = \beta_5 = 0$$

in the complete model specified in part **a**. The reduced model is

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Treatments (engineers)}}$$

Dependent Variable: COST

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 7.02333 | 1.40467 | 45.15 | 0.0001 |
| Error | 6 | 0.18667 | 0.03111 | | |
| Corrected Total | 11 | 7.21000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.17638 | R-Square | 0.9741 |
| Dependent Mean | 5.65000 | Adj R-Sq | 0.9525 |
| Coeff Var | 3.12183 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 4.58333 | 0.12472 | 36.75 | <.0001 |
| X1 | 1 | 0.25000 | 0.12472 | 2.00 | 0.0919 |
| X2 | 1 | -0.10000 | 0.12472 | -0.80 | 0.4533 |
| X3 | 1 | 1.50000 | 0.14402 | 10.42 | <.0001 |
| X4 | 1 | 0.63333 | 0.14402 | 4.40 | 0.0046 |
| X5 | 1 | 1.93333 | 0.14402 | 13.42 | <.0001 |

Test ENGINEER Results for Dependent Variable COST

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 2 | 0.13000 | 4.18 | 0.0730 |
| Denominator | 6 | 0.03111 | | |

Test JOBS Results for Dependent Variable COST

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 3 | 2.25444 | 72.46 | <.0001 |
| Denominator | 6 | 0.03111 | | |

**FIGURE 14.10**
SAS regression printout for randomized block design complete model, Example 14.5

The SAS printout for this second reduced model is shown in Figure 14.12. Note that $\text{SSE}_\text{R} = 6.95$ (shaded on the printout). The elements of the test follow.

*Test statistic:*

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(b-1)}{\text{MSE}_\text{C}} = \frac{(6.95 - .18667)/3}{.03111} = 72.46$$

*Rejection region:* $F > 4.76$, where $F_{.05} = 4.76$ (from Table 10, Appendix B) is based on $\nu_1 = (b-1) = 3$ df and $\nu_2 = (n-p-b+1) = 6$ df.

*Conclusion:* Since $F = 72.46$ exceeds the critical value, 4.76, there is sufficient evidence (at $\alpha = .05$) to indicate the differences among the block (job) means. It appears that blocking on jobs was effective in reducing the job-to-job variation in cost estimates.

Dependent Variable: COST

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 6.76333 | 2.25444 | 40.38 | <.0001 |
| Error | 8 | 0.44667 | 0.05583 | | |
| Corrected Total | 11 | 7.21000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.23629 | R-Square | 0.9380 |
| Dependent Mean | 5.65000 | Adj R-Sq | 0.9148 |
| Coeff Var | 4.18214 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 4.63333 | 0.13642 | 33.96 | <.0001 |
| X3 | 1 | 1.50000 | 0.19293 | 7.77 | <.0001 |
| X4 | 1 | 0.63333 | 0.19293 | 3.28 | 0.0111 |
| X5 | 1 | 1.93333 | 0.19293 | 10.02 | <.0001 |

**FIGURE 14.11**
SAS regression printout for randomized block design reduced model for testing treatments

Dependent Variable: COST

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 0.26000 | 0.13000 | 0.17 | 0.8477 |
| Error | 9 | 6.95000 | 0.77222 | | |
| Corrected Total | 11 | 7.21000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.87876 | R-Square | 0.0361 |
| Dependent Mean | 5.65000 | Adj R-Sq | -0.1781 |
| Coeff Var | 15.55331 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 5.60000 | 0.43938 | 12.75 | <.0001 |
| X1 | 1 | 0.25000 | 0.62138 | 0.40 | 0.6968 |
| X2 | 1 | -0.10000 | 0.62138 | -0.16 | 0.8757 |

**FIGURE 14.12**
SAS regression printout for randomized block design reduced model for testing blocks

We also requested SAS to perform this nested model $F$ test for blocks. The results, $F = 72.46$ and $p$-value $< .0001$, are shaded at the bottom of the SAS complete model printout, Figure 14.9. The small $p$-value confirms our conclusion; there is sufficient evidence (at $\alpha = .05$) to reject $H_0$.

*Caution*: The result of the test for the equality of block means must be interpreted with care, especially when the calculated value of the $F$ test statistic does not fall in the rejection region. This does not necessarily imply that the block means are the same, i.e., that blocking is unimportant. Reaching this conclusion would be equivalent to accepting the null hypothesis, a practice we have carefully avoided because of the unknown probability of committing a Type II error (that is, of accepting $H_0$ when $H_a$ is true). In other words, even when a test for block differences is inconclusive, we may still want to use the randomized block design in similar future experiments. If the experimenter believes that the experimental units are more homogeneous within blocks than among blocks, he or she should use the randomized block design regardless of whether the test comparing the block means shows them to be different.

The traditional analysis of variance approach to analyzing the data collected from a randomized block design is similar to the completely randomized design. The partitioning of SS(Total) for the randomized block design is most easily seen by examining Figure 14.13. Note that SS(Total) is now partitioned into *three* parts:

$$\mathrm{SS(Total)} = \mathrm{SSB} + \mathrm{SST} + \mathrm{SSE}$$

The formulas for calculating SST and SSB follow the same pattern as the formula for calculating SST for the completely randomized design.

From these quantities, we obtain mean square for treatments, MST, mean square for blocks, MSB, and mean square for error, MSE, as shown in the box. The test statistics are

$$F = \frac{\mathrm{MST}}{\mathrm{MSE}} \quad \text{for testing treatments}$$

$$F = \frac{\mathrm{MSB}}{\mathrm{MSE}} \quad \text{for testing blocks}$$

These $F$ values are equivalent to the "partial" $F$ statistics of the regression approach.

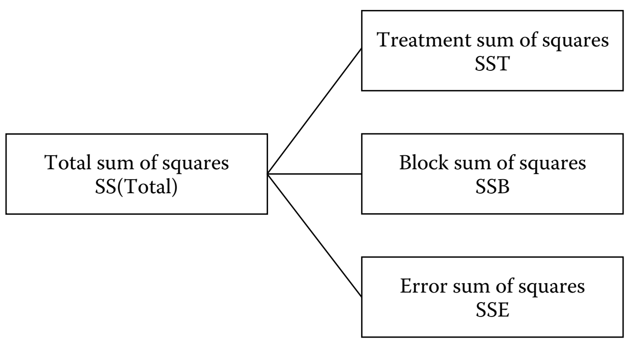



**FIGURE 14.13**
Partitioning of the total sum of squares for the randomized block design

### ANOVA Computing Formulas for a Randomized Block Design

$$\sum_{i=1}^{n} y_i = \text{Sum of all } n \text{ measurements}$$

$$\sum_{i=1}^{n} y_i^2 = \text{Sum of squares of all } n \text{ measurements}$$

$$\begin{aligned}
\text{CM} &= \text{Correction for mean} \\
&= \frac{(\text{Total of all measurements})^2}{\text{Total number of measurements}} = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\
\text{SS(Total)} &= \text{Total sum of squares} \\
&= \text{Sum of squares of all measurements} - \text{CM} \\
&= \sum_{i=1}^{n} y_i^2 - \text{CM} \\
\text{SST} &= \text{Sum of squares for treatments} \\
&= \left(\begin{array}{c}\text{Sum of squares of treatment totals with} \\ \text{each square divided by } b\text{, the number of} \\ \text{measurements for that treatment}\end{array}\right) - \text{CM} \\
&= \frac{T_1^2}{b} + \frac{T_2^2}{b} + \cdots + \frac{T_p^2}{b} - \text{CM} \\
\text{SSB} &= \text{Sum of squares for blocks} \\
&= \left(\begin{array}{c}\text{Sum of squares for block totals with} \\ \text{each square divided by } p\text{, the number} \\ \text{of measurements in that block}\end{array}\right) - \text{CM} \\
&= \frac{B_1^2}{p} + \frac{B_2^2}{p} + \cdots + \frac{B_b^2}{p} - \text{CM} \\
\text{SSE} &= \text{Sum of squares for error} \\
&= \text{SS(Total)} - \text{SST} - \text{SSB} \\
\text{MST} &= \text{Mean square for treatments} \\
&= \frac{\text{SST}}{p - 1} \\
\text{MSB} &= \text{Mean square for blocks} \\
&= \frac{\text{SSB}}{b - 1} \\
\text{MSE} &= \text{Mean square for error} \\
&= \frac{\text{SSE}}{n - p - b + 1} \\
F &= \frac{\text{MST}}{\text{MSE}} \quad \text{for testing treatments} \\
F &= \frac{\text{MSB}}{\text{MSE}} \quad \text{for testing blocks}
\end{aligned}$$

**Example 14.6**

Randomized Block Design: ANOVA

Refer to Example 14.5. Perform an analysis of variance of the data in Table 14.4 using the ANOVA approach.

Solution

Rather than perform the calculations by hand (again, we leave this as an exercise for the student), we utilize a statistical software package. The SPSS printout of the ANOVA is displayed in Figure 14.13. The $F$ value for testing treatments, $F = 4.179$, and the $F$ value for testing blocks, $F = 72.464$, are both shaded on the printout. Note that these values are identical to the $F$ values computed using the regression approach, Example 14.5 and the $p$-values of the tests (also shaded) lead to the same conclusions. For example, the $p$-value for the test of treatment differences, $p = .073$, exceeds $\alpha = .05$; thus, there is insufficient evidence of differences among the treatment means.

**FIGURE 14.14**
SPSS ANOVA printout for randomized block design

**Tests of Between-Subjects Effects**

Dependent Variable: COST

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 7.023[a] | 5 | 1.405 | 45.150 | .000 |
| Intercept | 383.070 | 1 | 383.070 | 12312.964 | .000 |
| ENGINEER | .260 | 2 | .130 | 4.179 | .073 |
| JOB | 6.763 | 3 | 2.254 | 72.464 | .000 |
| Error | .187 | 6 | 3.111E-02 | | |
| Total | 390.280 | 12 | | | |
| Corrected Total | 7.210 | 11 | | | |

a. R Squared = .974 (Adjusted R Squared = .953)

As with a completely randomized design, the sources of variation and their respective degrees of freedom, sums of squares, and mean squares for a randomized block design are summarized in an analysis of variance table. The format of an ANOVA table for a randomized block design is shown in the next box; the ANOVA table for the data for Table 14.4 is shown in Table 14.5. (These quantities were obtained from the printout, Figure 14.13.) Note that the degrees of freedom for the three sources of variation, treatments, blocks, and error, sum to the degrees of freedom for SS(Total). Similarly, the sums of squares for the three sources will always sum to SS(Total).

## General Format of ANOVA Table for a Randomized Block Design

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Treatments | $p - 1$ | SST | MST | MST/MSE |
| Blocks | $b - 1$ | SSB | MSB | MSB/MSE |
| Error | $n - p - b + 1$ | SSE | MSE | |
| Total | $n - 1$ | SS(Total) | | |

There is one very important point to note when you block the treatments in an experiment. Recall from Section 13.3 that the block effects cancel. This fact enables us to calculate confidence intervals for the difference between treatment means. But, if a sample treatment mean is used to estimate a *single treatment mean*, the block effects do not cancel. *Therefore, the only way that you can obtain an unbiased estimate of a single treatment mean (and corresponding confidence interval) in a blocked design is to randomly select the blocks from a large collection (population) of blocks and to treat the block*

**TABLE 14.4 ANOVA Summary Table for Example 14.6**

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Treatments (engineers) | 2 | .260 | .130 | 4.18 |
| Blocks (jobs) | 3 | 6.763 | 2.254 | 72.46 |
| Error | 6 | .187 | .031 | |
| Total | 11 | 7.210 | | |

*effect as a second random component, in addition to random error*. Designs that contain two or more random components are called *nested designs* and are beyond the scope of this text. For more information on this topic, consult the references for this chapter.

## Applied Exercises

14.17 *Forecasting electrical consumption.* Two different methods of forecasting monthly electrical consumption were compared and the results published in *Applied Mathematics and Computation* (Vol. 186, 2007). The two methods were Artificial Neural Networks (ANN) and Time Series Regression (TSR). Forecasts were made using each method for each of four months. These forecasts were also compared to the actual monthly consumption values. A layout of the design is shown in the next table. The researchers want to compare the mean electrical consumption values of the ANN forecast, TSR forecast and Actual consumption.

| *Month* | ANN Forecast | TSR Forecast | Actual Consumption |
|---|---|---|---|
| 1 | -- | -- | -- |
| 2 | -- | -- | -- |
| 3 | -- | -- | -- |
| 4 | -- | -- | -- |
| Sample Mean | 13.480 | 13.260 | 13.475 |

a. Identify the experimental design employed in the study.
b. A partial ANOVA table for the study is provided below. Fill in the missing entries.
c. Use the information in the table to conduct the appropriate ANOVA F-test using $\alpha = .05$. State your conclusion in the words of the problem.

| *Source* | *df* | *SS* | *MS* | *F-value* | *p-value* |
|---|---|---|---|---|---|
| Forecast Method | ---- | ---- | .195 | 2.83 | .08 |
| Month | 3 | ---- | 10.780 | ---- | $<.01$ |
| Error | ---- | .414 | .069 | | |
| Total | 11 | 33.144 | | | |

14.18 *Solar energy generation along highways.* The potential of solar panels on roofs built above national highways as a source of solar energy was investigated in the *International Journal of Energy and Environmental Engineering* (December 2013). Computer simulation was used to estimate the monthly solar energy (kilowatt hours) generated from solar panels installed across a 200-kilometer stretch of highway in India. Each month, the simulation was run under each of four conditions: single-layer solar panels, double-layer solar panels 1 meter apart, double-layer solar panels 2 meters apart, and double-layer solar panels 3 meters apart. The data for 12 months are shown in the table. In order to compare the mean solar energy values generated by the four panel configurations, a randomized block design ANOVA was conducted. A MINITAB printout of the analysis is provided on p. 769.

**SOLARPANEL**

| | Single | Double Layer | | |
|---|---|---|---|---|
| Month | Layer | 1-meter | 2-meters | 3-meters |
| January | 7,308 | 8,917 | 9,875 | 10,196 |
| February | 6,984 | 8,658 | 9,862 | 9,765 |
| March | 7,874 | 9,227 | 11,092 | 11,861 |
| April | 7,328 | 7,930 | 9,287 | 10,343 |
| May | 7,089 | 7,605 | 8,422 | 9,110 |
| June | 5,730 | 6,350 | 7,069 | 7,536 |
| July | 4,531 | 5,120 | 5,783 | 6,179 |
| August | 4,587 | 5,171 | 5,933 | 6,422 |
| September | 5,985 | 6,862 | 8,208 | 8,925 |
| October | 7,051 | 8,608 | 10,008 | 10,239 |
| November | 6,724 | 8,264 | 9,238 | 9,334 |
| December | 6,883 | 8,297 | 9,144 | 9,808 |

*Source:* Sharma, P. & Harinarayana, T. "Solar energy generation potential along national highways", *International Journal of Energy and Environmental Engineering,* Vol. 49, No. 1, Dec. 2013 (Table 3).

### MINITAB Output for Exercise 14.18

**Two-way ANOVA: Energy versus Condition, Month**

```
Source     DF         SS        MS       F      P
Condition   3   49730750  16576917  115.54  0.000
Month      11   90618107   8238010   57.42  0.000
Error      33    4734730    143477
Total      47  145083587

S = 378.8   R-Sq = 96.74%   R-Sq(adj) = 95.35%
```

a. Identify the dependent variable, treatments, and blocks for this experiment.

b. Give the equation of the regression model used to analyze the data.

c. In terms of the model parameters, what null hypothesis would you test in order to compare the mean solar energy values generated by the four panel configurations?

d. Carry out the test, part c, using regression. Give the $F$-value and associated $p$-value.

e. Do the results, part d, agree with the results shown on the ANOVA MINITAB printout.

f. What conclusion can you draw from the analysis?

14.19 *Repairing pipeline cracks.* The Perth (Australia) Metropolitan Water Authority recently completed construction of a land pipeline for transporting domestic wastewaters from a primary treatment plant. During construction, the cement mortar lining of the pipeline was tested for cracking to determine whether autogenous healing will seal the cracks. Otherwise, expensive epoxy filling repairs would be necessary (*Proceedings of the Institute of Civil Engineers*, Apr. 1986). After cracks were observed in the pipeline, it was kept full of water for a period of 14 weeks. At each of 12 crack locations, crack widths were measured (in millimeters) after the 2nd, 6th, and 14th weeks of the wet period, as shown in the table. The data were subjected to an ANOVA using SAS. The SAS printout is shown below. Conduct a test to determine whether the mean crack widths differ for the four time periods. Test using $\alpha = .05$.

**CRACKPIPE**

| | Crack Width After Wetting | | | | | Crack Width After Wetting | | | |
|---|---|---|---|---|---|---|---|---|---|
| Crack Location | 0 Weeks | 2 Weeks | 6 Weeks | 14 Weeks | Crack Location | 0 Weeks | 2 Weeks | 6 Weeks | 14 Weeks |
| 1 | .50 | .20 | .10 | .10 | 7 | .90 | .25 | .05 | .05 |
| 2 | .40 | .20 | .10 | .10 | 8 | 1.00 | .30 | .05 | .10 |
| 3 | .60 | .30 | .15 | .10 | 9 | .70 | .25 | .10 | .10 |
| 4 | .80 | .40 | .10 | .10 | 10 | .60 | .25 | .10 | .05 |
| 5 | .80 | .30 | .05 | .05 | 11 | .30 | .15 | .10 | .05 |
| 6 | 1.00 | .40 | .05 | .05 | 12 | .30 | .14 | .05 | .05 |

*Source*: Cox, B. G., and Kelsall, K. J. "Construction of Cape Peron Ocean Outlet, Perth, Western Australia." *Proceedings of the Institute of Civil Engineers*, Part 1, Vol. 80, Apr. 1986, p. 479 (Table 1).

### SAS output for Exercise 14.19

```
                              The ANOVA Procedure

Dependent Variable: WIDTH

                                        Sum of
   Source                    DF        Squares    Mean Square    F Value    Pr > F

   Model                     14     2.96188333     0.21156310      13.71    <.0001

   Error                     33     0.50930833     0.01543359

   Corrected Total           47     3.47119167

            R-Square     Coeff Var     Root MSE    WIDTH Mean

            0.853276      46.08296     0.124232      0.269583

   Source                    DF       Anova SS    Mean Square    F Value    Pr > F

   PERIOD                     3     2.68489167     0.89496389      57.99    <.0001
   LOCATION                  11     0.27699167     0.02518106       1.63    0.1354
```

14.20 *Exposure to low-frequency sound.* Infrasound describes sound frequencies below the audibility range of the human ear. Refer to the *Journal of Low Frequency Noise, Vibration and Active Control* (Mar. 2004) study of the physiological effects of infrasound, Exercise 7.53 (p. 327). In the experiment, one group of five university students (Group A) was exposed to infrasound at 4 hertz and 120 decibels for 1 hour, and a second group of five students (Group B) was exposed to infrasound at 2 hertz and 110 decibels. The heart rate (beats/minute) of each student was measured both before and after infrasound exposure. The experimental data are provided in the table. To determine the impact of infrasound, the researchers compared the mean heart rate before exposure to the mean heart rate after exposure.

a. Analyze the data for Group A students using an ANOVA for a randomized block design. Conduct the ANOVA test of interest using $\alpha = .05$.
b. Repeat part **a** for Group B students.
c. In Exercise 7.49, you analyzed the data using a paired-difference $T$ test. Show that the results are equivalent to the randomized block ANOVA. (*Hint*: Show that $F = T^2$, where $F$ is the test statistic, part **a**, and $T$ is the paired-difference test statistic.)

**INFRASOUND2**

| Group A Students | Before Exposure | After Exposure | Group B Students | Before Exposure | After Exposure |
|---|---|---|---|---|---|
| A1 | 70 | 70 | B1 | 73 | 79 |
| A2 | 69 | 80 | B2 | 68 | 60 |
| A3 | 76 | 84 | B3 | 61 | 69 |
| A4 | 77 | 86 | B4 | 72 | 77 |
| A5 | 64 | 76 | B5 | 61 | 66 |

*Source:* Qibai, C. Y. H., and Shi, H. "An investigation on the physiological and psychological effects of infrasound on persons." *Journal of Low Frequency Noise, Vibration and Active Control*, Vol. 23, No. 1, March 2004 (Tables I–IV).

14.21 *Effect of scale deposition on well flow performance.* Oil wells can suffer from scale deposits, leading to a reduction in well flow performance. Consequently, it is important for oil well managers to periodically assess the damage caused by scale deposits. In the *Journal of Petroleum & Gas Engineering* (April 2013) researchers compared four different computer software products designed to assess scale deposit damage. These software products were (1) Kappa Saphir, (2) EPS Pansystem, (3) MS Office Excel, and (4) Mellitah B. V. Ten oil wells (all suffering from scale deposits of varying degrees) were randomly selected from all oil wells in the field of interest. Scale deposit damage (called "skin factor") was measured for each well using all four software products. The data are shown in the next table. The objective of the study is to compare the mean skin factor values determined by the four software products.

**SKINFACTOR**

| Well | Kappa Saphir | EPS Pansystem | MS Office Excel | Mellitah B.V. |
|---|---|---|---|---|
| 1 | 39.15 | 37.77 | 44.48 | 16.80 |
| 2 | 12.92 | 13.21 | 18.34 | 12.50 |
| 3 | 6.84 | 7.02 | 19.21 | 7.00 |
| 4 | 4.13 | 4.77 | 11.70 | 4.13 |
| 5 | 2.59 | 1.96 | 9.25 | 2.40 |
| 6 | 281.43 | 281.74 | 317.40 | 287.60 |
| 7 | 192.78 | 192.16 | 181.44 | 193.50 |
| 8 | 138.23 | 140.84 | 154.65 | 140.00 |
| 9 | 54.21 | 56.86 | 77.43 | 57.30 |
| 10 | 45.65 | 45.01 | 49.37 | 42.00 |

*Source:* Rahuma, K.M., et al. "Comparison between spreadsheet and specialized programs in calculating the effect of scale deposition on the well flow performance", *Journal of Petroleum & Gas Engineering,* Vol. 4, No. 4, April 2013 (Table 2).

a. Explain why the data has been collected as a randomized block design.
b. Identify the dependent variable, treatments and blocks in the study.
c. Form the appropriate ANOVA table for the analysis.
d. Is there any evidence to indicate that the mean skin factor values determined by the four software products differ? Test using $\alpha = .01$.

14.22 *Stress in cows prior to slaughter.* What is the level of stress (if any) that cows undergo prior to being slaughtered? To answer this question, researchers designed an experiment involving cows bred in Normandy, France. (*Applied Animal Behaviour Science,* June 2010.) The heart rate (beats per minute) of a cow was measured at four different pre-slaughter phases—(1) first phase of visual contact with pen mates, (2) initial isolation from pen mates for prepping, (3) restoration of visual contact with pen mates, and (4) first contact with human prior to slaughter. Data for eight cows (simulated from information provided in the article) are shown in the table on p. 771. The researchers analyzed the data using an analysis of variance for a randomized block design. Their objective was to determine whether the mean heart rate of cows differed in the four pre-slaughter phases.

a. Identify the treatments and blocks for this experimental design.
b. Conduct the appropriate analysis using a statistical software package. Summarize the results in an ANOVA table.
c. Is there evidence of differences among the mean heart rates of cows in the four pre-slaughter phases? Test using $\alpha = .05$.
d. If warranted, conduct a multiple comparisons procedure to rank the four treatment means. Use an experimentwise error rate of $\alpha = .05$.

## Data for Exercise 14.22

**COWSTRESS**

| | PHASE | | | |
|---|---|---|---|---|
| COW | 1 | 2 | 3 | 4 |
| 1 | 124 | 124 | 109 | 107 |
| 2 | 100 | 98 | 98 | 99 |
| 3 | 103 | 98 | 100 | 106 |
| 4 | 94 | 91 | 98 | 95 |
| 5 | 122 | 109 | 114 | 115 |
| 6 | 103 | 92 | 100 | 106 |
| 7 | 98 | 80 | 99 | 103 |
| 8 | 120 | 84 | 107 | 110 |

14.23 *Containers designed to cool citrus fruit.* Prior to shipping and during storage, citrus fruit stacked on pallets are susceptible to damage from high temperatures. Consequently, containers have been designed to keep the fruit cool. The *Journal of Food Engineering* (September 2013) published an article that investigated the cooling performance of an existing fruit container design (Standard) and two new container designs (Supervent and Ecopack). Both the Standard and Supervent containers arrange the fruit in three rows, while the Ecopack container uses only two rows. Pallets of oranges were randomly divided into three groups. One group was stored using the Standard container design, one using the Supervent design, and one using the Ecopack design. Since oranges in the first row of the container tend to stay cooler than those in the back rows, the researchers employed a randomized block design, with rows representing the blocks and container design representing the treatments. The response variable of interest was the half-cooling time, measured as the time (in minutes) required to reduce the temperature difference between the fruit and cooling air by half. Half-cooling times were measured for each row of fruit for each design. The data is shown in the accompanying table. Note that there is no data for Row 3 of the Ecopack design; this is because the Ecopack container utilizes only two rows of fruit. Consequently, the design is *unbalanced* and the ANOVA formulas shown on p. 766 are not applicable. However, the regression approach to ANOVA will yield the correct analysis. Conduct the appropriate analysis of variance and state your conclusion using $\alpha = .10$.

**COOLING**

| | Standard | Supervent | Ecopack |
|---|---|---|---|
| Row 1 | 116 | 93 | 115 |
| Row 2 | 181 | 139 | 164 |
| Row 3 | 247 | 176 | |

14.24 *Evaluating lead-free solders.* Traditionally, solders used in electronics assembly are made with lead. Due to numerous environmental hazards associated with lead solders (e.g., groundwater contamination and breathing in of fine lead-bearing particles), engineers are developing lead-free solders. In *Soldering & Surface Mount Technology* (Vol. 13, 2001), researchers compared the traditional tin–lead alloy solder to three lead-free alloys: tin–silver, tin–copper, and tin–silver–copper. A measure of plastic hardening ($Nm/m^2$) was obtained for each solder type at each of six different temperatures. The data are given in the table.

**LEADSOLDER**

| Temperature | Tin–Lead | Tin–Silver | Tin–Copper | Tin–Silver–Copper |
|---|---|---|---|---|
| 23°C | 50.1 | 33.0 | 14.9 | 41.0 |
| 50°C | 24.6 | 27.7 | 10.5 | 20.7 |
| 75°C | 23.1 | 10.7 | 9.3 | 17.1 |
| 100°C | 1.8 | 9.0 | 8.8 | 8.7 |
| 125°C | 1.1 | 4.9 | 5.4 | 7.1 |
| 150°C | 0.3 | 3.2 | 5.0 | 4.9 |

*Source:* Harrison, M. R., Vincent, J. H., and Steen, H. A. H. "Lead-free reflow soldering for electronics assembly." *Soldering & Surface Mount Technology*, Vol. 13, No. 3, 2001 (Table X).

a. Explain why the data should be analyzed using a randomized block design ANOVA.
b. Form a summary ANOVA table for the analysis.
c. Do you detect differences in the mean plastic hardening values for the four solder types? Test using $\alpha = .10$.

14.25 *Light to dark transition of genes.* Refer to the *Journal of Bacteriology* (July 2002) study of the sensitivity of bacteria genes to light, Exercise 8.49 (p. 406). Recall that scientists isolated 103 genes of the bacterium responsible for photosynthesis and respiration. Each gene was grown to midexponential phase in a growth incubator in "full light," then exposed to three alternative light/dark conditions: "full dark" (lights extinguished for 24 hours), "transient light" (lights turned back on for 90 minutes), and "transient dark" (lights turned back off for an additional 90 minutes). At the end of each light/dark condition, the standardized growth measurement was determined for each of the 103 genes. The complete data set is saved in the **GENEDARK** file. (Data for the first 10 genes are shown in the table on p. 772.) Assume that the goal of the experiment is to compare the mean standardized growth measurements for the three light/dark conditions.

a. Write a linear model appropriate for analyzing the data.
b. In terms of the $\beta$ parameters of the model, part **a**, give the null hypothesis for comparing the light/dark condition means.
c. Using a statistical software package, conduct the test, part **c**. Interpret the results at $\alpha = .05$.

**Data for Exercise 14.25**

**GENEDARK**

(*First 10 observations shown*)

| Gene ID | Full-dark | TR-light | TR-dark |
|---|---|---|---|
| SLR2067 | −0.00562 | 1.40989 | −1.28569 |
| SLR1986 | −0.68372 | 1.83097 | −0.68723 |
| SSR3383 | −0.25468 | −0.79794 | −0.39719 |
| SLL0928 | −0.18712 | −1.20901 | −1.18618 |
| SLR0335 | −0.20620 | 1.71404 | −0.73029 |
| SLR1459 | −0.53477 | 2.14156 | −0.33174 |
| SLL1326 | −0.06291 | 1.03623 | 0.30392 |
| SLR1329 | −0.85178 | −0.21490 | 0.44545 |
| SLL1327 | 0.63588 | 1.42608 | −0.13664 |
| SLL1325 | −0.69866 | 1.93104 | −0.24820 |

*Source:* Gill, R. T., et al. "Genome-wide dynamic transcriptional profiling of the light to dark transition in *Synechocystis Sp.* PCC6803," *Journal of Bacteriology*, Vol. 184, No. 13, July 2002.

14.26 *Automated handling system for garments.* A plant that manufactures denim jeans in the United Kingdom recently introduced a computerized automated handling system. The new system delivers garments to the assembly line operators by means of an overhead conveyor. While the automated system minimizes operator handling time, it inhibits operators from working ahead and taking breaks from their machine. A study in *New Technology, Work, and Employment* (July 2001) investigated the impact of the new handling system on worker absentee rates at the jeans plant. One theory is that the mean absentee rate will vary by day of the week, as operators decide to indulge in one-day absences to relieve work pressure. Nine weeks were randomly selected and the absentee rate (percentage of workers absent) determined for each day (Monday through Friday) of the work week. The data are listed in the table below. Conduct a complete analysis of the data to determine whether the mean absentee rate differs across the five days of the work week.

**JEANS**

| Week | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| 1 | 5.3 | 0.6 | 1.9 | 1.3 | 1.6 |
| 2 | 12.9 | 9.4 | 2.6 | 0.4 | 0.5 |
| 3 | 0.8 | 0.8 | 5.7 | 0.4 | 1.4 |
| 4 | 2.6 | 0.0 | 4.5 | 10.2 | 4.5 |
| 5 | 23.5 | 9.6 | 11.3 | 13.6 | 14.1 |
| 6 | 9.1 | 4.5 | 7.5 | 2.1 | 9.3 |
| 7 | 11.1 | 4.2 | 4.1 | 4.2 | 4.1 |
| 8 | 9.5 | 7.1 | 4.5 | 9.1 | 12.9 |
| 9 | 4.8 | 5.2 | 10.0 | 6.9 | 9.0 |

*Source:* Boggis, J. J. "The eradication of leisure." *New Technology, Work, and Employment*, Volume 16, Number 2, July 2001 (Table 3).

## 14.5 Two-Factor Factorial Experiments

In Section 13.4, we learned that factorial experiments are volume-increasing designs conducted to investigate the effect of two or more independent variables (factors) on the mean value of the response $y$. In this section, we focus on the analysis of two-factor factorial experiments.

For example, suppose we want to relate the mean number of defects on a finished item—say, a new desk top—to two factors, type of nozzle for the varnish spray gun and length of spraying time. Suppose further that we want to investigate the mean number of defects per desk for three types (three levels) of nozzles ($N_1$, $N_2$, and $N_3$) and for two lengths (two levels) of spraying time ($S_1$ and $S_2$). If we choose the treatments for the experiment to include all combinations of the three levels of nozzle type with the two levels of spraying time, i.e., we observe the number of defects for the factor–level combinations $N_1S_1$, $N_1S_2$, $N_2S_1$, $N_2S_2$, $N_3S_1$, $N_3S_2$, our design is called a **complete 3 × 2 factorial experiment**. Note that the design will contain $3 \times 2 = 6$ treatments.

Factorial experiments, you will recall, are useful methods for selecting treatments because they permit us to make inferences about factor interactions. The complete model for the $3 \times 2$ factorial experiment contains $(3 - 1) = 2$ main effect terms for nozzles, $(2 - 1) = 1$ main effect term for spray time, and $(3 - 1)(2 - 1) = 2$ nozzle–spray time interaction terms:

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\substack{\text{Nozzle} \\ \text{main effects}}} + \underbrace{\beta_3 x_3}_{\substack{\text{Spray time} \\ \text{main effect}}} + \underbrace{\beta_4 x_1 x_2 + \beta_5 x_1 x_3}_{\substack{\text{Nozzle} \times \text{spray time} \\ \text{interaction}}}$$

The independent variables (factors) in the model can be either quantitative or qualitative. If they are quantitative, the main effects are represented by terms such as $x$, $x^2$, $x^3$, etc.; if qualitative, the main effects are represented by dummy variables. In our $3 \times 2$ factorial experiment, nozzle type is qualitative and spraying time is quantitative; hence, the $x$ variables in the model are defined as follows:

$$x_1 = \begin{cases} 1 & \text{if nozzle } N_1 \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if nozzle } N_2 \\ 0 & \text{if not} \end{cases} \qquad \text{Base level} = N_3$$

$$x_3 = \text{Length of spraying time (in minutes)}$$

Note that the model for the $3 \times 2$ factorial contains $3 \times 2 = 6$ $\beta$ parameters. If we observe only a single value of the response $y$ for each of the $3 \times 2 = 6$ treatments, then $n = 6$ and df(Error) for the complete model is $(n - 6) = 0$. Consequently, for a factorial experiment, *the number r of observations per factor–level combination (i.e., the number of replications of the factorial experiment) must always be 2 or more*. Otherwise, no degrees of freedom are available for estimating $\sigma^2$.

To test for factor interaction, we drop the interaction terms and fit the reduced model:

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\substack{\text{Main effect}\\ \text{nozzle}}} + \underbrace{\beta_3 x_3}_{\substack{\text{Main effect}\\ \text{spray time}}}$$

The null hypothesis of no interaction, $H_0$: $\beta_4 = \beta_5 = 0$, is tested by comparing the SSEs for the two models in a partial $F$ statistic. This test for interaction is summarized, in general, in the accompanying box.

## Models and ANOVA *F* Test for Interaction in a Two-Factor Factorial Experiment with Factor *A* at *a* Levels and Factor *B* at *b* Levels

*Complete model:*

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \cdots + \beta_{a-1} x_{a-1}}^{\text{Main effect } A \text{ terms}} + \overbrace{\beta_a x_a + \cdots + \beta_{a+b-2} x_{a+b-2}}^{\text{Main effect } B \text{ terms}}$$

$$+ \overbrace{\beta_{a+b-1} x_1 x_a + \beta_{a+b} x_1 x_{a+1} + \cdots + \beta_{ab-1} x_{a-1} x_{a+b-2}}^{AB \text{ interaction terms}}$$

where*

$$x_1 = \begin{cases} 1 & \text{if level 2 of factor } A \ldots \\ 0 & \text{if not} \end{cases}$$

$$x_{a-1} = \begin{cases} 1 & \text{if level } a \text{ of factor } A \ldots \\ 0 & \text{if not} \end{cases}$$

$$x_a = \begin{cases} 1 & \text{if level 2 of factor } B \ldots \\ 0 & \text{if not} \end{cases}$$

$$x_{a+b-2} = \begin{cases} 1 & \text{if level } b \text{ of factor } B \ldots \\ 0 & \text{if not} \end{cases}$$

$H_0$: $\beta_{a+b-1} = \beta_{a+b} = \cdots = \beta_{ab-1} = 0$
(i.e., $H_0$: No interaction between factors $A$ and $B$)

$H_a$: At least one of the $\beta$ parameters listed in $H_0$ differs from 0
(i.e., $H_a$: Factors $A$ and $B$ interact)

*Reduced model:*

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \cdots + \beta_{a-1}x_{a-1}}^{\text{Main effect } A \text{ terms}} + \overbrace{\beta_a x_a + \cdots + \beta_{a+b-2}x_{a+b-2}}^{\text{Main effect } B \text{ terms}}$$

*Test statistic:*

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/[(a-1)(b-1)]}{\text{SSE}_\text{C}/[ab(r-1)]} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/[(a-1)(b-1)]}{\text{MSE}_\text{C}}$$

where

$\text{SSE}_\text{R}$ = SSE for reduced model
$\text{SSE}_\text{C}$ = SSE for complete model
$\text{MSE}_\text{C}$ = MSE for complete model
$r$ = Number of replications (i.e., number of $y$ measurements per cell of the $a \times b$ factorial)

Rejection region: $F > F_\alpha$,
$p$-value: $P(F > F_c)$
where $F$ is based on $\nu_1 = (a-1)(b-1)$ and $\nu_2 = ab(r-1)$ df, and $F_c$ is the computed value of the test statistic.

*Assumptions:* 1. The population probability distribution of the observations for any factor–level combination is approximately normal.
2. The variance of the probability distribution is constant and the same for all factor–level combinations.

**Note:* The independent variables, $x_1, x_2, \ldots, x_{a+b-2}$, are defined for an experiment in which both factors represent *qualitative* variables. When a factor is *quantitative*, you may choose to represent the main effects with quantitative terms such as $x$, $x^2$, $x^3$, and so forth.

Tests for factor main effects are conducted in a similar manner. The main effect terms of interest are dropped from the complete model and the reduced model is fit. The SSEs for the two models are compared in the usual fashion.

Before we work through a numerical example of an analysis of variance for a factorial experiment, we need to understand the practical significance of the tests for factor interaction and factor main effects. We illustrate these concepts in Example 14.7.

## Example 14.7

**Factorial Experiment: Illustrating Factor Interaction**

A company that stamps gaskets out of sheets of rubber, plastic, and other materials wants to compare the mean number of gaskets produced per hour for two different types of stamping machines. Practically, the manufacturer wants to determine whether one machine is more productive than the other and, even more important, whether one machine is more productive in making rubber gaskets whereas the other is more productive in making plastic gaskets. To answer these questions, the manufacturer decides to conduct a $2 \times 3$ factorial experiment using three types of gasket material, $B_1$, $B_2$, and $B_3$, with each of the two types of stamping machines, $A_1$ and $A_2$. Each machine is operated for three 1-hour time periods for each of the gasket materials, with the eighteen 1-hour time periods assigned to the six machine–material combinations in random order. (The purpose of the randomization is to eliminate the possibility that uncontrolled environmental factors might bias the results.) Suppose we have calculated and plotted the six treatment means. Two hypothetical plots of the six means are shown in Figures 14.15a and 14.15b. The three means for stamping machine $A_1$ are connected by solid line segments and the corresponding three means for machine $A_2$ by dashed line segments. What do these plots imply about the productivity of the two stamping machines?

**Solution**

Figure 14.15a suggests that machine $A_1$ produces a larger number of gaskets per hour, regardless of the gasket material, and is therefore superior to machine $A_2$. On the average, machine $A_1$ stamps more cork ($B_1$) gaskets per hour than rubber or plastic, but the *difference* in the mean numbers of gaskets produced by the two machines remains

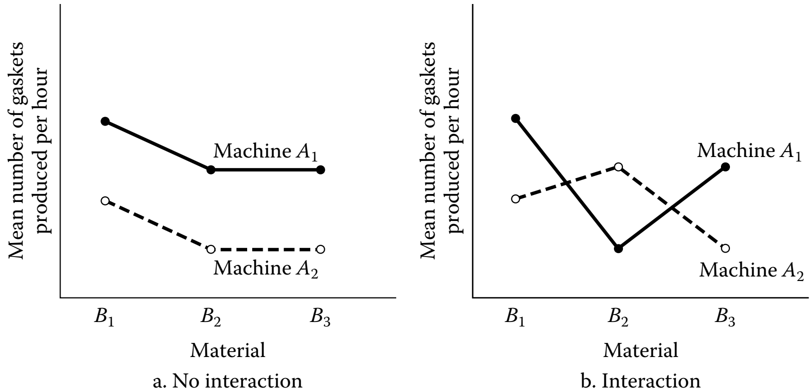



**FIGURE 14.15**
Hypothetical plot of the means for the six machine–material combinations

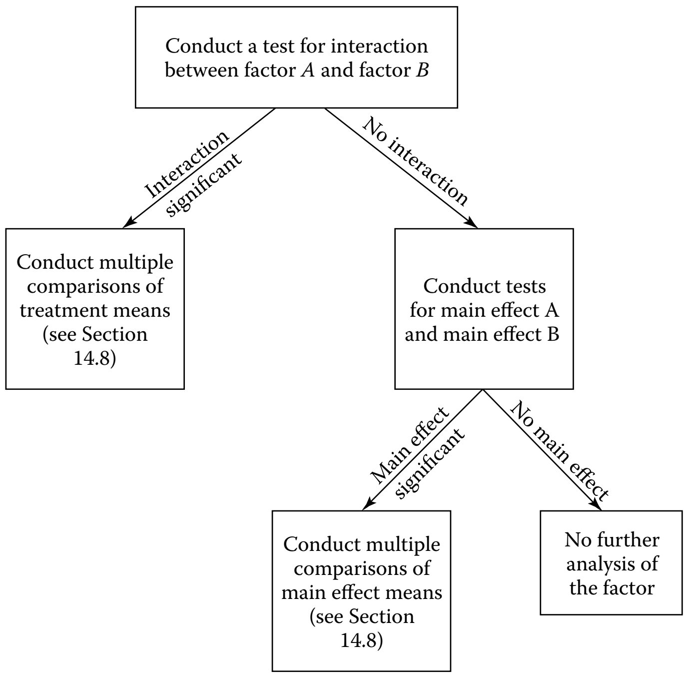
approximately the same, regardless of the gasket material. Thus, the difference in the mean number of gaskets produced by the two machines is *independent* of the gasket material used in the stamping process.

In contrast to Figure 14.15a, Figure 14.15b shows the productivity for machine $A_1$ to be greater than for machine $A_2$ when the gasket material is cork ($B_1$) or plastic ($B_3$). But the means are reversed for rubber ($B_2$) gasket material. For this material, machine $A_2$ produces, on the average, more gaskets per hour than machine $A_1$. Thus, Figure 14.14b illustrates a situation where the difference in the mean number of gaskets produced by the two machines *depends* on gasket material. When this situation occurs, we say that the factors *interact*. Thus, one of the most important objectives of a factorial experiment is to detect factor interaction if it exists.

**Definition 14.2**

In a factorial experiment, when the difference in the mean levels of factor *A* depends on the different levels of factor *B*, we say that the factors *A* and *B* **interact**. If the difference is independent of the levels of *B*, then there is **no interaction** between *A* and *B*.

**FIGURE 14.16**
Testing Guidelines for a Two-Factor Factorial Experiment



Tests for main effects are relevant only when no interaction exists between factors. Generally, the test for interaction is performed first. (See Figure 14.16.) *If there is evidence of factor interaction, then we will not perform the tests on the main effects.* Rather, we will want to focus attention on the individual cell (treatment) means, perhaps locating one that is the largest or the smallest.

## Example 14.8

3 × 3 Factorial Design: Regression Approach

A manufacturer whose daily supply of raw materials is variable and limited can use the material to produce two different products in various proportions. The profit per unit of raw material obtained by producing each of the two products depends on the length of a product's manufacturing run and, hence, on the amount of raw material assigned to it. Other factors, such as worker productivity and machine breakdown, affect the profit per unit as well, but their net effect on profit is random and uncontrollable. The manufacturer has conducted an experiment to investigate the effect of the level of supply of raw materials, $S$, and the ratio of its assignment, $R$, to the two product manufacturing lines on the profit $y$ per unit of raw material. The ultimate goal would be able to choose the best ratio $R$ to match each day's supply of raw materials, $S$. The levels of supply of the raw material chosen for the experiment were 15, 18, and 21 tons; the levels of the ratio of allocation to the two product lines were $\frac{1}{2}$, 1, and 2. The response was the profit (in dollars) per unit of raw material supply obtained from a single day's production. Three replications of a complete $3 \times 3$ factorial experiment were conducted in a random sequence (i.e., a completely randomized design). The data for the 27 days are shown in Table 14.6.

a. Write the complete model for the experiment.

b. Do the data present sufficient evidence to indicate an interaction between supply $S$ and ratio $R$? Use $\alpha = .05$.

c. Based on the result, part **b**, should we perform tests for main effects?

**RAWMATERIAL**

**TABLE 14.5 Data for Example 14.8**

| | | Raw Material Supply, tons (S) | | |
|---|---|---|---|---|
| | | **15** | **18** | **21** |
| **Ratio of** | $\frac{1}{2}$ | 23, 20, 21 | 22, 19, 20 | 19, 18, 21 |
| **Raw Material** | **1** | 22, 20, 19 | 24, 25, 22 | 20, 19, 22 |
| **Allocation (R)** | **2** | 18, 18, 16 | 21, 23, 20 | 20, 22, 24 |

Solution

a. Both factors, supply and ratio, are set at three levels. As such, we require two dummy variables for each factor. (The number of main effect terms will be one less than the number of levels for a factor.) Consequently, the complete factorial model for this 3 × 3 factorial experiment is

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Supply main effects}} + \underbrace{\beta_3 x_3 + \beta_4 x_4}_{\text{Ratio main effects}}$$

$$+ \underbrace{\beta_5 x_1 x_3 + \beta_6 x_1 x_4 + \beta_7 x_2 x_3 + \beta_8 x_2 x_4}_{\text{Supply–Ratio interaction}} + \varepsilon$$

where

$$x_1 = \begin{cases} 1 \text{ if Supply is 15 tons} \\ 0 \text{ if not} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if Supply is 18 tons} \\ 0 \text{ if not} \end{cases}$$

$$(\text{Supply base level} = 21 \text{ tons})$$

$$x_3 = \begin{cases} 1 \text{ if Ratio is 1:2} \\ 0 \text{ if not} \end{cases}$$

$$x_4 = \begin{cases} 1 \text{ if Ratio is 1:1} \\ 0 \text{ if not} \end{cases}$$

(Ratio base level $= 2{:}1$)

Note that the interaction terms for the model are constructed by taking the products of the various main effect terms, one from each factor. For example, we included terms involving the products of $x_1$ with $x_3$ and $x_4$. The remaining interaction terms were formed by multiplying $x_2$ by $x_3$ and by $x_4$.

b. To test the null hypothesis that supply and ratio do not interact, we must test the null hypothesis that the interaction terms are not needed in the linear model of part **a**:

$$H_0\text{:}\quad \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

This requires that we fit the reduced model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and perform the partial $F$ test outlined in Section 12.8. The test statistic is

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/4}{\text{MSE}_\text{C}}$$

where

$$\begin{aligned} \text{SSE}_\text{C} &= \text{SSE for complete model} \\ \text{MSE}_\text{C} &= \text{MSE for complete model} \\ \text{SSE}_\text{R} &= \text{SSE for reduced model} \end{aligned}$$

The complete model of part **a** and the reduced model presented here were fit to the data in Table 14.5 using SAS. The SAS printouts are displayed in Figures 14.17a and 14.17b. The pertinent quantities, shaded on the printout, are

$$\begin{aligned} \text{SSE}_\text{C} &= 43.33333 \quad \text{(see Figure 14.17a)} \\ \text{MSE}_\text{C} &= 2.40741 \quad \text{(see Figure 14.17a)} \\ \text{SSE}_\text{R} &= 89.55556 \quad \text{(see Figure 14.17b)} \end{aligned}$$

Substituting these values into the formula for the test statistic, we obtain

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/4}{\text{MSE}_\text{C}} = \frac{(89.55556 - 43.33333)/4}{2.40741} = 4.80$$

This "partial" $F$ value is shaded at the bottom of the SAS printout, Figure 14.17a, as is the $p$-value of the test, .0082. Since $\alpha = .05$ exceeds the $p$-value, we reject $H_0$ and conclude that supply and ratio interact.

c. The presence of interaction tells you that the mean profit depends on the particular combination of levels of supply $S$ and ratio $R$. Consequently, there is little point in checking to see whether the means differ for the three levels of supply or whether they differ for the three levels of ratio (i.e., we will not perform the tests for main effects). For example, the supply level that gave the highest mean profit (over all levels of $R$) might not be the same Supply–Ratio level combination that produces the largest mean profit per unit of raw material.

The traditional analysis of variance approach to analyzing a complete two-factor factorial with factor $A$ at $a$ levels and factor $B$ at $b$ levels utilizes the fact that the total sum of squares, SS(Total), can be partitioned into four parts, SS($A$), SS($B$), SS($AB$), and SSE (see Figure 14.18). The first two sums of squares, SS($A$) and SS($B$), are called **main effect sums of squares** to distinguish them from the **interaction sum of squares**, SS($AB$).

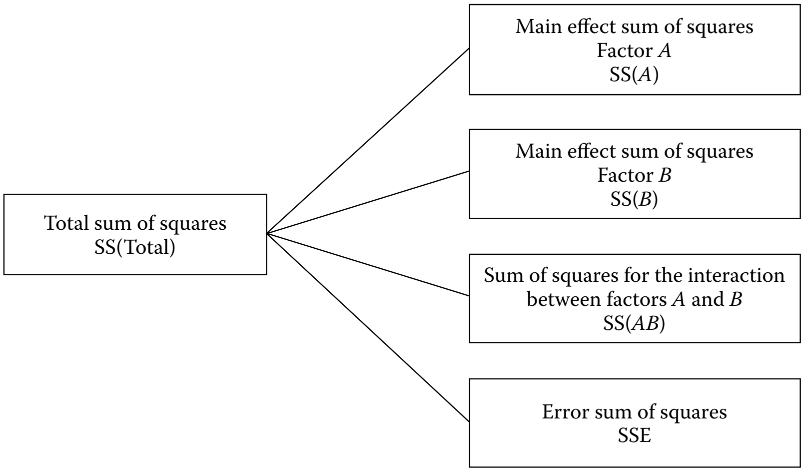
```
                    Dependent Variable: PROFIT

           Number of Observations Read          27
           Number of Observations Used          27

                      Analysis of Variance

                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F

Model                     8     74.66667        9.33333       3.88    0.0081
Error                    18     43.33333        2.40741
Corrected Total          26    118.00000

             Root MSE              1.55158    R-Square     0.6328
             Dependent Mean       20.66667    Adj R-Sq     0.4696
             Coeff Var             7.50766

                      Parameter Estimates

                         Parameter       Standard
     Variable     DF      Estimate          Error    t Value    Pr > |t|

     Intercept     1      22.00000        0.89581      24.56      <.0001
     X1            1      -2.66667        1.26686      -2.10      0.0496
     X2            1      -1.66667        1.26686      -1.32      0.2048
     X3            1      -4.66667        1.26686      -3.68      0.0017
     X4            1      -0.66667        1.26686      -0.53      0.6051
     X1X3          1       6.66667        1.79161       3.72      0.0016
     X1X4          1       1.66667        1.79161       0.93      0.3645
     X2X3          1       4.66667        1.79161       2.60      0.0179
     X2X4          1       4.00000        1.79161       2.23      0.0385
        Test INTERACTION Results for Dependent Variable PROFIT

                                      Mean
        Source             DF       Square    F Value    Pr > F

        Numerator           4     11.55556       4.80    0.0082
        Denominator        18      2.40741
```

**FIGURE 14.17a**
SAS Regression Printout for Complete Factorial Model

```
                    Dependent Variable: PROFIT

           Number of Observations Read          27
           Number of Observations Used          27

                      Analysis of Variance

                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F

Model                     4     28.44444        7.11111       1.75    0.1757
Error                    22     89.55556        4.07071
Corrected Total          26    118.00000

             Root MSE              2.01760    R-Square     0.2411
             Dependent Mean       20.66667    Adj R-Sq     0.1031
             Coeff Var             9.76258

                      Parameter Estimates

                         Parameter       Standard
     Variable     DF      Estimate          Error    t Value    Pr > |t|

     Intercept     1      20.11111        0.86824      23.16      <.0001
     X1            1       0.11111        0.95111       0.12      0.9081
     X2            1       1.22222        0.95111       1.29      0.2121
     X3            1      -0.88889        0.95111      -0.93      0.3601
     X4            1       1.22222        0.95111       1.29      0.2121
```

**FIGURE 14.17b**
SAS Regression Printout for Reduced (Main Effects) Factorial Model

**FIGURE 14.18**
Partitioning of the total sum of squares for a complete two-factor factorial experiment




Since the sums of squares and the degrees of freedom for the analysis of variance are additive, the analysis of variance table appears as shown in the following box. Note that the $F$ statistics for testing factor main effects and factor interaction are obtained by dividing the appropriate mean square by MSE. The numerator df for the test of interest will equal the df of the source of variation being tested; the denominator df will always equal df(Error). These $F$ tests are equivalent to the $F$ tests obtained by fitting complete and reduced models in regression.*

For completeness, the formulas for calculating the ANOVA sums of squares for a complete two-factor factorial experiment are given in the box on p. 781.

### ANOVA Table for an $a \times b$ Factorial Design with $r$ Observations per Cell (Note: $n = abr$)

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Main effects $A$ | $(a - 1)$ | SS($A$) | $\text{MS}(A) = \text{SS}(A)/(a - 1)$ | MS($A$)/MSE |
| Main effects $B$ | $(b - 1)$ | SS($B$) | $\text{MS}(B) = \text{SS}(B)/(b - 1)$ | MS($B$)/MSE |
| $AB$ interaction | $(a - 1) \times (b - 1)$ | SS($AB$) | $\text{MS}(AB) = \text{SS}(AB)/[(a - 1)(b - 1)]$ | MS($AB$)/MSE |
| Error | $ab(r - 1)$ | SSE | $\text{MSE} = \text{SSE}/[ab(r - 1)]$ | |
| Total | $abr - 1$ | SS(Total) | | |

**Example 14.9**

**3 × 3 Factorial Design: ANOVA Approach**

Refer to Example 14.8.

a. Construct an ANOVA summary table for the analysis.
b. Conduct the test for supply × ratio interaction using the traditional analysis of variance approach.
c. Illustrate the nature of the interaction by plotting the sample mean profits (as in Figure 14.15). Interpret the results.

*The ANOVA $F$ tests for main effects shown in the box are equivalent to those of the regression approach only when the reduced model includes interaction terms. Since we usually test for main effects only after determining that interaction is nonsignificant, some statisticians favor dropping the interaction terms from both the complete and reduced models prior to conducting the main effect tests. For example, to test for main effect $A$, the complete model includes terms for main effects $A$ and $B$, whereas the reduced model includes terms for main effect $B$ only. To obtain the equivalent result using the ANOVA approach, the sum of squares for $AB$ interaction and error are "pooled" and a new MSE computed, where

$$\text{MSE} = \frac{\text{SS}(AB) + \text{SSE}}{n - a - b + 1}$$

Solution

a. Although the formulas given in the box are straightforward, they can become quite tedious to use. Therefore, we resort to a statistical software package to conduct the ANOVA. A MINITAB printout of the ANOVA is displayed in Figure 14.19. The summary ANOVA table is highlighted on the printout.

b. To test the hypothesis that supply and ratio interact, we use the test statistic

$$F = \frac{\text{MS(SR)}}{\text{MSE}} = \frac{11.56}{2.41} = 4.80 \quad \text{(shown on the MINITAB printout in the "Interaction" row)}$$

**Two-way ANOVA: PROFIT versus SUPPLY, RATIO**

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| SUPPLY | 2 | 20.222 | 10.1111 | 4.20 | 0.032 |
| RATIO | 2 | 8.222 | 4.1111 | 1.71 | 0.209 |
| Interaction | 4 | 46.222 | 11.5556 | 4.80 | 0.008 |
| Error | 18 | 43.333 | 2.4074 | | |
| Total | 26 | 118.000 | | | |

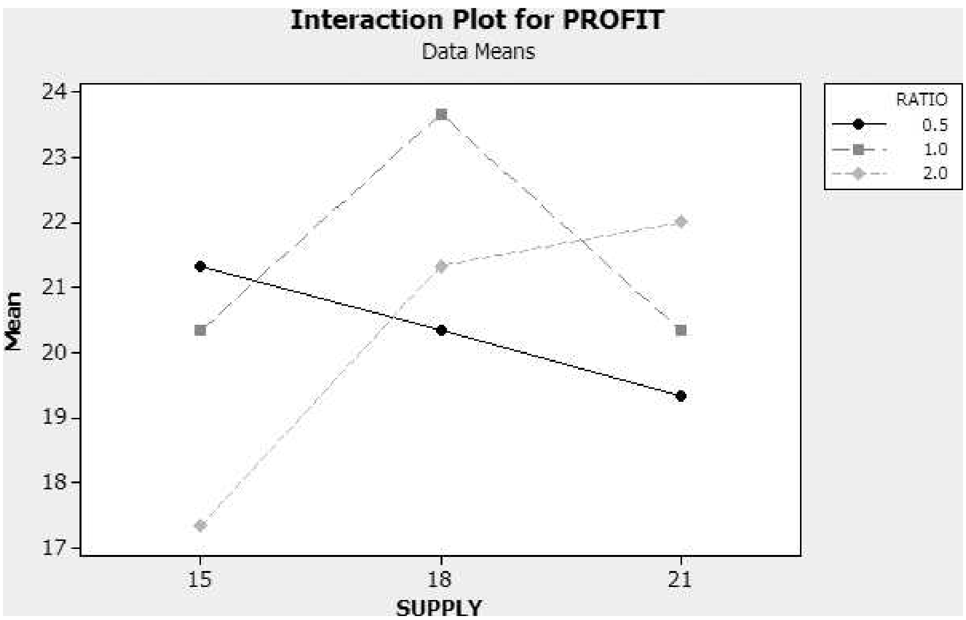
S = 1.552   R-Sq = 63.28%   R-Sq(adj) = 46.96%

**FIGURE 14.19**
MINITAB ANOVA printout for complete factorial design

Note that this value is identical to the test statistic obtained in Example 14.8 using regression. The $p$-value of the test (also shaded on the MINITAB printout) is .008. Since this value is less than the selected value of $\alpha = .05$, we conclude that supply and ratio interact.

c. A MINITAB plot of the sample profit means, shown in Figure 12.20, illustrates the nature of the Supply $\times$ Ratio interaction. From the graph, you can see that the difference between the profit means for any two levels of Ratio (e.g., for $R = .5$ and $R = 2$) is not the same for the different levels of Supply. For example, at $S = 15$, the mean is largest for $R = .5$ and smallest for $R = 2$; however, at $S = 21$, the mean is largest for $R = 2$ and smallest for $R = .5$. Consequently, the ratio of raw material allocation which yields the greatest profit will depend on the supply available.

**FIGURE 14.20**
MINITAB Plot of Profit Means Illustrating Interaction



### ANOVA Computing Formulas for a Two-Factor Factorial Experiment

$$\begin{aligned}
\text{CM} &= \text{Correction for the mean} \\
&= \frac{(\text{Total of all } n \text{ measurements})^2}{n} \\
&= \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\
\text{SS(Total)} &= \text{Total sum of squares} \\
&= \text{Sum of squares of all } n \text{ measurements} - \text{CM} \\
&= \sum_{i=1}^{n} y_i^2 - \text{CM} \\
\text{SS}(A) &= \text{Sum of squares for main effects, independent variable 1} \\
&= \left(\begin{array}{c}\text{Sum of squares of the totals } A_1, A_2, \ldots, A_a \\ \text{divided by the number of measurements} \\ \text{in a single total, namely, } br\end{array}\right) - \text{CM} \\
&= \frac{\sum_{i=1}^{a} A_i^2}{br} - \text{CM} \\
\text{SS}(B) &= \text{Sum of squares for main effects, independent variable 2} \\
&= \left(\begin{array}{c}\text{Sum of squares of the totals } B_1, B_2, \ldots, B_b \\ \text{divided by the number of measurements} \\ \text{in a single total, namely, } ar\end{array}\right) - \text{CM} \\
&= \frac{\sum_{i=1}^{b} B_i^2}{ar} - \text{CM} \\
\text{SS}(AB) &= \text{Sum of squares for } AB \text{ interaction} \\
&= \left(\begin{array}{c}\text{Sum of squares of the cell totals} \\ AB_{11}, AB_{12}, \ldots, AB_{ab} \text{ divided} \\ \text{by the number of measurements} \\ \text{in a single total, namely, } r\end{array}\right) - \text{SS}(A) - \text{SS}(B) - \text{CM} \\
&= \frac{\sum_{j=1}^{b}\sum_{i=1}^{a} AB_{ij}^2}{r} - \text{SS}(A) - \text{SS}(B) - \text{CM} \\
\text{MS}(A) &= \text{SS}(A)/(a-1) \\
\text{MS}(B) &= \text{SS}(B)/(b-1) \\
\text{MS}(A \times B) &= \text{SS}(AB)/(a-1)(b-1) \\
F &= \text{MS}(A)/\text{MSE for testing main effect } A \\
F &= \text{MS}(B)/\text{MSE for testing main effect } B \\
F &= \text{MS}(A \times B)/\text{MSE for testing interaction}
\end{aligned}$$

where

$a$ = Number of levels of independent variable 1

$b$ = Number of levels of independent variable 2

$r$ = Number of measurements for each pair of levels of independent variables 1 and 2

$n$ = Total number of measurements
$= a \times b \times r$
$A_i$ = Total of all measurements of independent variable 1 at level $i$ $(i = 1, 2, \ldots, a)$
$B_j$ = Total of all measurements of independent variable 2 at level $j$ $(j = 1, 2, \ldots, b)$
$AB_{ij}$ = Total of all measurements at the $i$th level of independent variable 1 and at the $j$th level of independent variable 2 $(i = 1, 2, \ldots, a;$ $j = 1, 2, \ldots, b)$

Throughout this chapter we have presented two methods for analyzing data from a designed experiment: the regression approach and the traditional ANOVA approach. In a factorial experiment, the two methods yield identical results when both factors are qualitative; however, regression will provide more information when at least one of the factors is quantitative and if we represent the main effects with quantitative terms like $x$, $x^2$, and so on. For example, the analysis of variance in Example 14.9 enables us to estimate the mean profit per unit of supply for *only* the nine combinations of supply–ratio levels. It will not permit us to estimate the mean response for some other combination of levels of the independent variables not included among the nine used in the factorial experiment. Alternatively, the prediction equation obtained from the regression analysis with quantitative terms enables us to estimate the mean profit per unit of supply when ($S = 17$, $R = 1$). We could not obtain this estimate from the analysis of variance in Example 14.9.

The prediction equation found by regression analysis also contributes other information not provided by traditional analysis of variance. For example, we might wish to estimate the rate of change in the mean profit, $E(y)$, for unit changes in $S$, $R$, or both. Or, we might want to determine whether the third- and fourth-order terms in the complete model really contribute additional information for the prediction of profit, $y$.

We illustrate some of these applications in the final three examples of this section.

## Example 14.10

**Factorial Design Model with Quantitative Factors**

Refer to the $3 \times 3$ factorial design and the data of Example 14.8. Since both factors, Supply and Ratio, are quantitative in nature, we can represent the main effects of the complete factorial model using quantitative terms such as $x$, $x^2$, $x^3$, etc., rather than dummy variables. Like with dummy variables, the number of quantitative main effects will be one less than the number of levels for a quantitative factor. The logic follows from our discussion about estimating model parameters in Section 11.11. At two levels, the quantitative main effect is $x$; at three levels, the quantitative main effects are $x$ and $x^2$.

a. Specify the complete model for the factorial design using quantitative main effects for Supply and Ratio.

b. Fit the model to the data in Table 14.5 and show that the $F$ test for interaction produced on the printout is equivalent to the corresponding test produced using dummy variables for main effects.

**Solution**

a. Now both Supply (15, 18, and 21 tons) and Ratio (1/2, 1, and 2) are set at 3 levels; consequently, each factor will have two quantitative main effects. If we let $x_1$ represent the actual level of Supply of raw material (in tons) and let $x_2$ represent the actual level of Ratio of allocation (e.g., 1/2, 1, and 2), then the main effects are $x_1$ and $x_1^2$ for Supply and $x_2$ and $x_2^2$ for Ratio. Consequently, the complete factorial model for mean profit, $E(y)$, is

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_1^2}_{\substack{\text{Supply} \\ \text{main effects}}} + \underbrace{\beta_3 x_2 + \beta_4 x_2^2}_{\substack{\text{Ratio} \\ \text{main effects}}}$$
$$+ \underbrace{\beta_5 x_1 x_2 + \beta_6 x_1 x_2^2 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_2^2}_{\text{Supply} \times \text{Ratio interaction}}$$

```
                         Dependent Variable: PROFIT

                            Analysis of Variance

                                      Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F

Model                     8         74.66667        9.33333       3.88    0.0081
Error                    18         43.33333        2.40741
Corrected Total          26        118.00000

             Root MSE                1.55158    R-Square     0.6328
             Dependent Mean         20.66667    Adj R-Sq     0.4696
             Coeff Var               7.50766

                            Parameter Estimates

                           Parameter        Standard
Variable        DF          Estimate           Error    t Value    Pr > |t|

Intercept        1         245.33333       130.49665       1.88      0.0764
SUPPLY           1         -25.07407        14.71842      -1.70      0.1057
SUPPSQ           1           0.67901         0.40837       1.66      0.1137
RATIO            1        -534.33333       252.45535      -2.12      0.0485
RATSQ            1         192.66667        97.17011       1.98      0.0629
RAT_SUPP         1          60.55556        28.47387       2.13      0.0475
S_RATSQ          1         -22.14815        10.95960      -2.02      0.0584
R_SUPPSQ         1          -1.66667         0.79003      -2.11      0.0492
RSQ_SSQ          1           0.61728         0.30408       2.03      0.0574

    Test INTERACT Results for Dependent Variable PROFIT

                                       Mean
Source                DF              Square    F Value    Pr > F

Numerator              4            11.55556       4.80    0.0082
Denominator           18             2.40741

    Test HIGHORDR Results for Dependent Variable PROFIT

                                       Mean
Source                DF              Square    F Value    Pr > F

Numerator              3             3.71958       1.55    0.2373
Denominator           18             2.40741
```

**FIGURE 14.21**
SAS regression printout for complete factorial model with quantitative main effects

Note that the number of terms (main effects and interactions) in the model is equivalent to the dummy variable model of Example 14.8.

b. The SAS printout for the complete model, part **a**, is shown in Figure 14.21. First, note that SSE = 43.33333 and MSE = 2.40741 (highlighted) are equivalent to the corresponding values shown on the printout for the dummy variable model, Figure 14.17a. Second, the partial $F$ value ($F = 4.80$) for testing the null hypothesis of no interaction ($H_0$: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$), highlighted in the middle of the printout, is equivalent to the corresponding test shown on Figure 14.17a. Thus, whether you conduct the test for factor interaction using regression with dummy variables, regression with quantitative main effects, or with the traditional ANOVA approach, the results will be identical.

**Example 14.11**

Testing Higher-order Terms in Factorial Design Model

Do the data provide sufficient information to indicate that third- and fourth-order terms in the complete factorial model given in Example 14.10 contribute information for the prediction of $y$? Use $\alpha = .05$.

Solution If the response to the question is yes, then at least one of the parameters, $\beta_6$, $\beta_7$, or $\beta_8$, of the complete factorial model differs from 0 (i.e., they are needed in the model). Consequently, the null hypothesis is

$$H_0\text{:}\quad \beta_6 = \beta_7 = \beta_8 = 0$$

and the alternative hypothesis is

$H_a$: At least one of the three $\beta$'s is nonzero.

To test this hypothesis, we compute the drop in SSE between the appropriate reduced and complete model.

For this application the complete model is the complete factorial model of Example 14.10:

$$\textit{Complete model}\text{:}\quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 x_1 x_2^2 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_2^2$$

The reduced model is the complete model, minus the third- and fourth-order terms; i.e., the reduced model is the second-order model:

$$\textit{Reduced model:}\quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

Recall (from Figure 14.21) that the SSE and MSE for the complete model are $\text{SSE}_\text{C} = 43.3333$ and $\text{MSE}_\text{C} = 2.4074$. A SAS printout of the regression analysis of the reduced model is shown in Figure 14.22. The SSE for the reduced model (shaded) is $\text{SSE}_\text{R} = 54.49206$.

Consequently, the test statistic required to conduct the test is

*Test statistic:*

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{C})/(\text{number of }\beta\text{'s tested})}{\text{MSE}_\text{C}} = \frac{(54.49206 - 43.3333)/3}{2.4074} = 1.55$$

Dependent Variable: PROFIT

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 63.50794 | 12.70159 | 4.89 | 0.0040 |
| Error | 21 | 54.49206 | 2.59486 | | |
| Corrected Total | 26 | 118.00000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.61086 | R-Square | 0.5382 |
| Dependent Mean | 20.66667 | Adj R-Sq | 0.4283 |
| Coeff Var | 7.79447 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -27.81481 | 23.80152 | -1.17 | 0.2557 |
| SUPPLY | 1 | 5.94444 | 2.64418 | 2.25 | 0.0354 |
| RATIO | 1 | -7.76190 | 5.04523 | -1.54 | 0.1389 |
| RAT_SUPP | 1 | 0.74603 | 0.20295 | 3.68 | 0.0014 |
| SUPPSQ | 1 | -0.18519 | 0.07307 | -2.53 | 0.0193 |
| RATSQ | 1 | -2.29630 | 1.33939 | -1.71 | 0.1012 |

**FIGURE 14.22**
SAS regression printout for reduced (second-order) factorial model

This "partial" $F$ value can also be obtained using SAS options and is given at the bottom of the SAS complete model printout, Figure 14.21 (p. 783). The $p$-value of the test (also highlighted) is .2373.

*Conclusion*: Since $\alpha = .05$ is less than $p$-value $= .2373$, we cannot reject the null hypothesis that $\beta_6 = \beta_7 = \beta_8 = 0$. That is, there is insufficient evidence (at $\alpha = .05$) to indicate that the third- and fourth-order terms associated with $\beta_6$, $\beta_7$, and $\beta_8$ contribute information for the prediction of $y$. Since the complete factorial model contributes no more information about $y$ than the reduced (second-order) model, we recommend using the second-order model in practice.

## Example 14.12

**Finding Confidence Intervals for Treatment Means**

Use the second-order model of Example 14.11 and find a 95% confidence interval for the mean profit per unit supply of raw material when $S = 17$ and $R = 1$.

**Solution**

The portion of the SAS printout for the second-order model with 95% confidence intervals for $E(y)$ is shown in Figure 14.23.

The confidence interval for $E(y)$ when $S = 17$ and $R = 1$ is shaded in the last row of the printout. You can see that the interval is (20.97, 23.72). Thus, we estimate (with confidence coefficient equal to .95) that the mean profit per unit of supply will lie between \$20.97 and \$23.72 when $S = 17$ tons and $R = 1$. Beyond this immediate result, you will note that this example illustrates the power and versatility of a regression analysis. In particular, there is no way to obtain this estimate from the analysis of variance in Example 14.9. However, a computerized regression package can be easily programmed to include the confidence interval automatically.

| Obs | SUPPLY | RATIO | Dep Var PROFIT | Predicted Value | Std Error Mean Predict | 95% CL Mean | | Residual |
|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 0.5 | 23.0000 | 20.8254 | 0.8033 | 19.1549 | 22.4959 | 2.1746 |
| 2 | 15 | 0.5 | 20.0000 | 20.8254 | 0.8033 | 19.1549 | 22.4959 | -0.8254 |
| 3 | 15 | 0.5 | 21.0000 | 20.8254 | 0.8033 | 19.1549 | 22.4959 | 0.1746 |
| 4 | 18 | 0.5 | 22.0000 | 21.4444 | 0.6932 | 20.0029 | 22.8860 | 0.5556 |
| 5 | 18 | 0.5 | 19.0000 | 21.4444 | 0.6932 | 20.0029 | 22.8860 | -2.4444 |
| 6 | 18 | 0.5 | 20.0000 | 21.4444 | 0.6932 | 20.0029 | 22.8860 | -1.4444 |
| 7 | 21 | 0.5 | 19.0000 | 18.7302 | 0.8033 | 17.0596 | 20.4007 | 0.2698 |
| 8 | 21 | 0.5 | 18.0000 | 18.7302 | 0.8033 | 17.0596 | 20.4007 | -0.7302 |
| 9 | 21 | 0.5 | 21.0000 | 18.7302 | 0.8033 | 17.0596 | 20.4007 | 2.2698 |
| 10 | 15 | 1 | 22.0000 | 20.8175 | 0.7006 | 19.3605 | 22.2744 | 1.1825 |
| 11 | 15 | 1 | 20.0000 | 20.8175 | 0.7006 | 19.3605 | 22.2744 | -0.8175 |
| 12 | 15 | 1 | 19.0000 | 20.8175 | 0.7006 | 19.3605 | 22.2744 | -1.8175 |
| 13 | 18 | 1 | 24.0000 | 22.5556 | 0.6932 | 21.1140 | 23.9971 | 1.4444 |
| 14 | 18 | 1 | 25.0000 | 22.5556 | 0.6932 | 21.1140 | 23.9971 | 2.4444 |
| 15 | 18 | 1 | 22.0000 | 22.5556 | 0.6932 | 21.1140 | 23.9971 | -0.5556 |
| 16 | 21 | 1 | 20.0000 | 20.9603 | 0.7006 | 19.5034 | 22.4173 | -0.9603 |
| 17 | 21 | 1 | 19.0000 | 20.9603 | 0.7006 | 19.5034 | 22.4173 | -1.9603 |
| 18 | 21 | 1 | 22.0000 | 20.9603 | 0.7006 | 19.5034 | 22.4173 | 1.0397 |
| 19 | 15 | 2 | 18.0000 | 17.3571 | 0.8590 | 15.5707 | 19.1436 | 0.6429 |
| 20 | 15 | 2 | 18.0000 | 17.3571 | 0.8590 | 15.5707 | 19.1436 | 0.6429 |
| 21 | 15 | 2 | 16.0000 | 17.3571 | 0.8590 | 15.5707 | 19.1436 | -1.3571 |
| 22 | 18 | 2 | 21.0000 | 21.3333 | 0.6932 | 19.8917 | 22.7749 | -0.3333 |
| 23 | 18 | 2 | 23.0000 | 21.3333 | 0.6932 | 19.8917 | 22.7749 | 1.6667 |
| 24 | 18 | 2 | 20.0000 | 21.3333 | 0.6932 | 19.8917 | 22.7749 | -1.3333 |
| 25 | 21 | 2 | 20.0000 | 21.9762 | 0.8590 | 20.1897 | 23.7627 | -1.9762 |
| 26 | 21 | 2 | 22.0000 | 21.9762 | 0.8590 | 20.1897 | 23.7627 | 0.0238 |
| 27 | 21 | 2 | 24.0000 | 21.9762 | 0.8590 | 20.1897 | 23.7627 | 2.0238 |
| 28 | 17 | 1 | . | 22.3466 | 0.6625 | 20.9687 | 23.7244 | . |

**FIGURE 14.23**
SAS regression printout for reduced (second-order) factorial model

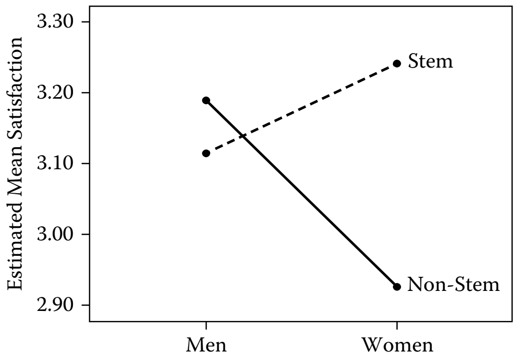
## Applied Exercises

14.27 *Egg shell quality in laying hens.* Introducing calcium into a hen's diet can improve the shell quality of the eggs laid. One way to do this is with a limestone diet. In *Animal Feed Science and Technology* (June 2010) researchers investigated the effect of hen's age and limestone diet on egg shell quality. Two different diets were studied—fine limestone (FL) and coarse limestone (CL). Hens were classified as either younger hens (24-36 weeks old) or older hens (56-68 weeks old). The study used 120 younger hens and 120 older hens. Within each age group, half the hens were fed a fine limestone diet and the other half a coarse limestone diet. Thus, there were 60 hens in each of the four combinations of age and diet. The characteristics of the eggs produced from the laying hens were recorded, including shell thickness.

a. Identify the type of experimental design employed by the researchers.
b. Identify the factors and the factor levels (treatments) for this design.
c. Identify the experimental unit.
d. Identify the dependent variable.
e. The researchers found no evidence of factor interaction. Interpret this result, practically.
f. The researchers found no evidence of a main effect for hen's age. Interpret this result, practically.
g. The researchers found statistical evidence of a main effect for limestone diet. Interpret this result, practically. (Note: The mean shell thickness for eggs produced by hens on a CL diet was larger than the corresponding mean for hens on a FL diet.)

14.28 *Job satisfaction of STEM Faculty.* Are university faculty in Science, Technology, Engineering, and Math (STEM) disciplines more satisfied with their job than non-STEM faculty? And if so, does this difference vary depending on gender? These were some of the questions of interest in a study published in the *Journal of Women and Minorities in Science and Engineering* (Vol. 18, 2012). A sample of 215 faculty at a large public university participated in a survey. One question asked the degree to which the faculty was satisfied with university policies and procedures. Responses were recorded on a 5-point numerical scale, with 1 = Strongly Disagree to 5 = Strongly Agree. Each participant was categorized according to gender (male or female) and discipline (STEM or non-STEM). Thus, a $2 \times 2$ factorial design was utilized.

a. Identify the treatments for this experiment.
b. For this study, what does it mean to say that discipline and gender interact?
c. A plot of the treatment means is shown above. Based on this graph only, would you say that discipline and gender interact?
d. Construct a partial ANOVA table for this study. (Give the sources of variation and degrees of freedom.)
e. The journal article reported the $F$-test for interaction as $F = 4.10$ with $p$-value $= .04$. Interpret these results.




14.29 *Baker's versus brewer's yeast.* Refer to the *Electronic Journal of Biotechnology* (Dec. 15, 2003) study of two yeast extracts, baker's yeast and brewer's yeast, Exercise 13.10 (p. 732). Recall that samples of both yeast extracts were prepared at four different temperatures (45°, 48°, 51°, and 54°C), and the autolysis yield (recorded as a percentage) was measured for each of the yeast–temperature combinations. Thus, a $2 \times 4$ factorial design was employed, with extract at two levels (baker's yeast and brewer's yeast) and temperature at four levels.

a. For this design, give a practical explanation of an interaction between yeast extract and temperature. Illustrate with a graph.
b. Redraw the graph, part **a**, if no interaction exists.
c. Write the complete model for mean autolysis yield, $E(y)$, appropriate for analyzing the data.
d. Give the null and alternative hypotheses for testing interaction. Explain how you would conduct this test using regression.
e. An ANOVA resulted in a $p$-value of .0027 for interaction. Interpret this result, practically, using $\alpha = .01$.
f. Give the null and alternative hypotheses for testing the main effects of yeast extract and temperature using regression. Explain how you would conduct these main effect tests using regression.
g. Explain why the tests, part **f**, should not be conducted.

14.30 *Virtual reality-based rehabilitation systems.* Refer to the *Robotica* (Vol. 22, 2004) study of the effectiveness of display devices for three virtual reality (VR)-based hand rehabilitation systems, Exercise 14.10 (p. 756). Display device A is a projector, device B is a desktop computer monitor, and device C is a head-mounted display. Recall that 12 nondisabled, right-handed male subjects were randomly assigned to the three VR display devices, four subjects in each group. Additionally, within each group two subjects were randomly assigned to use an auxiliary lateral image and two subjects were not. Consequently, a $3 \times 2$ factorial design was employed, with VR display device at three levels (A, B, or C) and auxiliary lateral image at two levels (yes or no). Each subject carried out a "pick-and-place" procedure using the assigned VR system, and the collision frequency (number of collisions between moved objects) was measured.

a. Give the sources of variation and associated degrees of freedom in an ANOVA summary table for this design.
b. Write the complete model for analyzing these data. The degrees of freedom for each source, part **a**, should correspond with the number of dummy variable terms for that source in the model.
c. The factorial ANOVA resulted in the following $p$-values: display main effect (.045), auxiliary lateral image main effect (.003), and interaction (.411). Interpret, practically, these results. Use $\alpha = .05$ for each test you conduct.

14.31 *Impact of cutting tool material on cutting force.* The use of coated cutting tools to machine various materials is considered state-of-the art technology. The impact of cutting tool coating and material on cutting force was investigated in an article published in the *International Journal of Engineering and Applied Sciences* (Vol. 7, 2011). Five different steel cutting tool materials were compared: (1) uncoated CBN-High, (2) CBN-High coated with TiN alloy, (3) CBN-Low coated with TiN alloy, (4) CBN-Low coated with TiAIN alloy, and (5) mixed ceramic. Each cutting tool was run at three different speeds—100, 140, and 200 meters per minute—and two observations on cutting feed force (Newtons) were obtained for each run. The data (simulated from information provided in the journal article) are shown in the table. The objective of the experiment is to assess the impact of cutting tool type and cutting speed on feed force.

**CUTTING**

| | 100 m/min | 140 m/min | 200 m/min |
|---|---|---|---|
| Uncoated CBN-H | 73, 81 | 64, 58 | 57, 56 |
| CBN-H w/TiN | 79, 85 | 50, 41 | 103, 97 |
| CBN-L w/TiN | 77, 81 | 76, 71 | 80, 78 |
| CBN-L w/TiAIN | 99, 102 | 125, 132 | 131, 122 |
| Mixed Ceramic | 31, 42 | 26, 20 | 35, 26 |

a. For this study, identify the experimental design, factors, treatments, experimental units, and response variable.
b. Give the equation of the complete model for analyzing the data.
c. In terms of the model parameters, what null hypothesis would you test to determine whether the impact of cutting tool on feed force depends on cutting speed.
d. Use regression to carry out the test, part **c**. What do you conclude?
e. An SPSS printout of the ANOVA of the data is displayed below. Verify that the test results, part **d**, agree with the information provided on the printout.
f. Do you recommend conducting tests for main effects? Why or why not?

14.32 *Mussel settlement patterns on algae.* Mussel larvae are in great abundance in the drift material that washes up on Ninety Mile Beach in New Zealand. These larvae tend to settle on algae. Environmentalists at the University of Auckland investigated the impact of algae type on the abundance of mussel larvae in drift material. (*Malacologia*, Feb. 8, 2002.) Drift material from three different wash-up events on Ninety Mile Beach were collected; for each wash-up, the algae was separated into four strata—coarse-branching, medium-branching, fine-branching, and hydroid algae. Two samples were randomly selected for each of the $3 \times 4 = 12$ event/strata combinations, and the mussel density (percent per square centimeter) was measured for each. The data were analyzed as a complete $3 \times 4$ factorial design. The ANOVA summary table is shown here.

| Source | df | $F$ | $p$-Value |
|---|---|---|---|
| Event | 2 | .35 | >.05 |
| Strata | 3 | 217.33 | >.05 |
| Interaction | 6 | 1.91 | >.05 |
| Error | 12 | | |
| Total | 23 | | |

a. Identify the factors (and levels) in this experiment.
b. How many treatments are included in the experiment?
c. How many replications are included in the experiment?

**SPSS Output for Exercise 14.31**

**Tests of Between-Subjects Effects**

Dependent Variable: FORCE

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 28813.867[a] | 14 | 2058.133 | 91.337 | .000 |
| Intercept | 161040.133 | 1 | 161040.133 | 7146.751 | .000 |
| TOOL | 24065.867 | 4 | 6016.467 | 267.003 | .000 |
| SPEED | 789.267 | 2 | 394.633 | 17.513 | .000 |
| TOOL * SPEED | 3958.733 | 8 | 494.842 | 21.960 | .000 |
| Error | 338.000 | 15 | 22.533 | | |
| Total | 190192.000 | 30 | | | |
| Corrected Total | 29151.867 | 29 | | | |

a. R Squared = .988 (Adjusted R Squared = .978)

d. What is the total sample size for the experiment?
e. What is the response variable measured?
f. Which ANOVA $F$ test should be conducted first? Conduct this test (at $\alpha = .05$) and interpret the results.
g. If appropriate, conduct the $F$ tests (at $\alpha = .05$) for the main effects. Interpret the results.

14.33 *Baiting traps to maximize beetles catch.* A field experiment was conducted to compare the effectiveness of different traps for catching beetles (*Journal of Chemical Ecology,* Vol. 94, 2011). Paraffin traps were baited either with or without linalool. Also, the color of the trap was varied as green or yellow. Seven traps for each combination of bait and color—a total of 28 traps—were set 1 meter from the ground in a random grid pattern. After 5 days, the total number of beetles captured by each trap was determined. The data (simulated from information provided in the journal article) are displayed in the following table. The researchers are investigating the effect of bait type and color on the mean number of beetles captured by the traps. Analyze the data for the researchers. What do you conclude?

**BEETLES**

| | Yellow | Green |
|---|---|---|
| With Linalool | 17, 22, 13, 15, 14, 18, 11 | 4, 5, 0, 2, 3, 2, 0 |
| Without Linalool | 29, 10, 6, 5, 12, 11, 13 | 1, 0, 2, 0, 0, 0, 1 |

14.34 *Strength of solder joints.* The chemical element antimony is sometimes added to tin–lead solder to replace the more expensive tin and to reduce the cost of soldering. A factorial experiment was conducted to determine how antimony affects the strength of the tin–lead solder joint (*Journal of Materials Science*, May 1986). Tin–lead solder specimens were prepared using one of four possible cooling methods (water-quenched, WQ; oil-quenched, OQ; air-blown, AB; and furnace-cooled, FC) and with one of four possible amounts of antimony (0%, 3%, 5%, and 10%) added to the composition. Three solder joints were randomly assigned to each of the $4 \times 4 = 16$ treatments and the shear strength of each measured. The experimental results, shown in the next table, were subjected to an ANOVA.

a. Construct an ANOVA summary table for the experiment.
b. Conduct a test to determine whether the two factors, amount of antimony and cooling method, interact. Use $\alpha = .01$.
c. Interpret the result obtained in part **b**.
d. If appropriate, conduct the tests for main effects. Use $\alpha = .01$.

**ANTIMONY**

| Amount of Antimony % weight | Cooling Method | Shear Strength MPa |
|---|---|---|
| 0 | WQ | 17.6, 19.5, 18.3 |
| 0 | OQ | 20.0, 24.3, 21.9 |
| 0 | AB | 18.3, 19.8, 22.9 |
| 0 | FC | 19.4, 19.8, 20.3 |
| 3 | WQ | 18.6, 19.5, 19.0 |
| 3 | OQ | 20.0, 20.9, 20.4 |
| 3 | AB | 21.7, 22.9, 22.1 |
| 3 | FC | 19.0, 20.9, 19.9 |
| 5 | WQ | 22.3, 19.5, 20.5 |
| 5 | OQ | 20.9, 22.9, 20.6 |
| 5 | AB | 22.9, 19.7, 21.6 |
| 5 | FC | 19.6, 16.4, 20.5 |
| 10 | WQ | 15.2, 17.1, 16.6 |
| 10 | OQ | 16.4, 19.0, 18.1 |
| 10 | AB | 15.8, 17.3, 17.1 |
| 10 | FC | 16.4, 17.6, 17.6 |

*Source*: Tomlinson, W. J., and Cooper, G. A. "Fracture mechanism of brass/Sn-Pb-Sb solder joints and the effect of production variables on the joint strength." *Journal of Materials Science*, Vol. 21, No. 5, May 1986, p. 1731 (Table II). Copyright 1986 Chapman and Hall.

14.35 *Mowing effects on highway right-of-way.* A vegetation height of greater than 30 centimeters on a highway right-of-way is generally considered a safety hazard to drivers. How often and at what height should the right-of-way be mowed in order to maintain a safe environment? This was the question of interest in an article published in the *Landscape Ecology Journal* (Jan. 2013). The researchers designed an experiment to estimate the effects of mowing frequency and mowing height on the mean height of vegetation in the highway right-of-way. Mowing frequency was set at three levels—once, twice, or three times per year. Mowing height of the equipment was also set at three levels—5, 10, or 20 centimeters. A sample of 36 plots of land along a highway right-of-way were selected, and each was randomly assigned to receive one of the $3 \times 3 = 9$ mowing frequency/mowing height treatments. The design was balanced so that each treatment was applied to 4 plots of land. At the end of the year, the vegetation height (in centimeters) was measured for each plot. Simulated data are shown in the table (p. 789). Conduct an analysis of variance of the data. What inferences can you make about the effects of mowing frequency and mowing height on vegetation height?

### Data for Exercise 14.35

**MOW**

| Mow Height | Mow Frequency | Vegetation Height (cm) |
|---|---|---|
| 5 | 1 | 19.3, 17.3, 16.7, 15.0 |
| 10 | 1 | 16.0, 15.6, 16.9, 15.0 |
| 20 | 1 | 16.7, 17.9, 15.9, 13.7 |
| 5 | 2 | 22.4, 20.8, 24.5, 21.7 |
| 10 | 2 | 23.9, 23.6, 23.8, 21.7 |
| 20 | 2 | 24.7, 26.3, 27.2, 26.4 |
| 5 | 3 | 18.6, 17.9, 16.1, 19.4 |
| 10 | 3 | 22.2, 25.6, 21.8, 23.6 |
| 20 | 3 | 27.0, 25.3, 23.8, 28.0 |

14.36 *Detecting early part failure.* A trade-off study regarding the inspection and test of transformer parts was conducted by the quality department of a major defense contractor. The investigation was structured to examine the effects of varying inspection levels and incoming test times to detect early part failure or fatigue. The levels of inspection selected were full military inspection ($A$), reduced military specification level ($B$), and commercial grade ($C$). Operational burn-in test times chosen for this study were at 1-hour increments from 1 hour to 9 hours. The response was failures per thousand pieces obtained from samples taken from lot sizes inspected to a specified level and burned-in over a prescribed time length. Three replications were randomly sequenced under each condition, making this a complete $3 \times 9$ factorial experiment (a total of 81 observations). The data for the study (shown in the table below) were subjected to an ANOVA using SAS. The SAS printout follows. Analyze and interpret the results.

**BURNIN**

| Burn-in (hours) | Inspection Levels: Full Military Specification, *A* | | | Reduced Military Specification, *B* | | | Commercial, *C* | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.60 | 7.50 | 7.67 | 7.70 | 7.10 | 7.20 | 6.16 | 6.13 | 6.21 |
| 2 | 6.54 | 7.46 | 6.84 | 5.85 | 6.15 | 6.15 | 6.21 | 5.50 | 5.64 |
| 3 | 6.53 | 5.85 | 6.38 | 5.30 | 5.60 | 5.80 | 5.41 | 5.45 | 5.35 |
| 4 | 5.66 | 5.98 | 5.37 | 5.38 | 5.27 | 5.29 | 5.68 | 5.47 | 5.84 |
| 5 | 5.00 | 5.27 | 5.39 | 4.85 | 4.99 | 4.98 | 5.65 | 6.00 | 6.15 |
| 6 | 4.20 | 3.60 | 4.20 | 4.50 | 4.56 | 4.50 | 6.70 | 6.72 | 6.54 |
| 7 | 3.66 | 3.92 | 4.22 | 3.97 | 3.90 | 3.84 | 7.90 | 7.47 | 7.70 |
| 8 | 3.76 | 3.68 | 3.80 | 4.37 | 3.86 | 4.46 | 8.40 | 8.60 | 7.90 |
| 9 | 3.46 | 3.55 | 3.45 | 5.25 | 5.63 | 5.25 | 8.82 | 9.76 | 9.52 |

*Source*: La Nuez, Danny, College of Business, former graduate student, University of South Florida.

### SAS Output for Exercise 14.36

```
                              The ANOVA Procedure

Dependent Variable: FAILURES

                                            Sum of
      Source                      DF       Squares     Mean Square    F Value    Pr > F

      Model                       26   168.6120667       6.4850795     101.31    <.0001

      Error                       54     3.4565333       0.0640099

      Corrected Total             80   172.0686000

                   R-Square     Coeff Var      Root MSE    FAILURES Mean

                   0.979912      4.405990      0.253002         5.742222

      Source                      DF      Anova SS     Mean Square    F Value    Pr > F

      BURNIN                       8   27.97440000      3.49680000      54.63    <.0001
      INSLEVEL                     2   43.08411852     21.54205926     336.54    <.0001
      BURNIN*INSLEVEL             16   97.55354815      6.09709676      95.25    <.0001
```

14.37 *Combustion rate of graphite.* As part of a study on the rate of combustion of artificial graphite in humid air flow, researchers conducted an experiment to investigate oxygen diffusivity through a water vapor mixture. A 3 × 9 factorial experiment was conducted with mole fraction of water ($H_2O$) at three levels and temperature of the nitrogen–water mixture at nine levels. The data are shown in the table.

**WATERVAPOR**

| Temperature K | Mole Fraction of $H_2O$ .0022 | .017 | .08 |
|---|---|---|---|
| 1,000 | 1.68 | 1.69 | 1.72 |
| 1,100 | 1.98 | 1.99 | 2.02 |
| 1,200 | 2.30 | 2.31 | 2.35 |
| 1,300 | 2.64 | 2.65 | 2.70 |
| 1,400 | 3.00 | 3.01 | 3.06 |
| 1,500 | 3.38 | 3.39 | 3.45 |
| 1,600 | 3.78 | 3.79 | 3.85 |
| 1,700 | 4.19 | 4.21 | 4.27 |
| 1,800 | 4.63 | 4.64 | 4.71 |

*Source*: Matsui, K., Tsuji, H., and Makino, A. "The effects of water vapor concentration on the rate of combustion of an artificial graphite in humid air flow." *Combustion and Flame*, Vol. 50, 1983, pp. 107–118. Copyright 1983 by The Combustion Institute. Reprinted by permission of Elsevier Science Publishing Co., Inc.

a. Explain why the traditional analysis of variance (using the ANOVA formulas) is inappropriate for the analysis of these data.
b. Plot the data to determine if a first- or second-order model for mean oxygen diffusivity, $E(y)$, is more appropriate.
c. Write an interaction model relating mean oxygen diffusivity, $E(y)$, to temperature $x_1$ (in hundreds) and mole fraction $x_2$ (in thousandths).
d. Suppose that temperature and mole fraction do not interact. What does this imply about the relationship between $E(y)$ and $x_1$ and $x_2$?
e. Do the data provide sufficient information to indicate that temperature and mole fraction of $H_2O$ interact? Use the accompanying MINITAB printout to conduct the test using $\alpha = .05$.
f. Give the least-squares prediction equation for $E(y)$.
g. Substitute into the prediction equation to predict the mean diffusivity when the temperature of the process is 1,300 K and the mole fraction of water is .017.
h. Locate the 95% confidence interval for mean diffusivity when the temperature of the process is 1,300 K and the mole fraction of water is .017 on the MINITAB printout. Interpret the result.

**MINITAB Output for Exercise 14.37**

```
The regression equation is
OXYDIFF = - 2.10 + 0.00368 TEMP - 0.24 MOLE + 0.00073 TEMPMOLE

Predictor        Coef     SE Coef        T       P
Constant     -2.09528     0.09035   -23.19   0.000
TEMP       0.00368411  0.00006347    58.05   0.000
MOLE           -0.238       1.913    -0.12   0.902
TEMPMOLE     0.000733    0.001344     0.55   0.591

S = 0.06081     R-Sq = 99.7%     R-Sq(adj) = 99.6%

Analysis of Variance

Source            DF        SS        MS        F       P
Regression         3   24.7733    8.2578  2233.31   0.000
Residual Error    23    0.0850    0.0037
Total             26   24.8583


Predicted Values for New Observations

New Obs     Fit    SE Fit        95.0% CI              95.0% PI
1        2.7062    0.0139   ( 2.6774,  2.7350)   ( 2.5772,  2.8352)

Values of Predictors for New Observations

New Obs       TEMP      MOLE  TEMPMOLE
1             1300    0.0170      22.1
```

## 14.6 More Complex Factorial Designs (*Optional*)

In this optional section, we present some useful factorial designs that are more complex than the basic two-factor factorial of Section 14.5. These designs fall under the general category of a ***k*-way classification of data**. A $k$-way classification of data arises when we run all combinations of the levels of $k$ independent variables. These independent variables can be factors or blocks.

For example, consider a replicated $2 \times 3 \times 3$ factorial experiment in which the $2 \times 3 \times 3 = 18$ treatments are assigned to the experimental units according to a completely randomized design. Since every combination of the three factors (a total of 18) is examined, the design is often called a three-way classification of data. Similarly, a $k$-way classification of data would result if we randomly assign the treatments of a $(k - 1)$-factor factorial experiment to the experimental units of a randomized block design. For example, if we assigned the $2 \times 3 = 6$ treatments of a complete $2 \times 3$ factorial experiment to blocks containing six experimental units each, the data would be arranged in a three-way classification, i.e., according to the two factors and the blocks.

The formulas required for calculating the sums of squares for main effects and interactions for an analysis of variance for a $k$-way classification of data are complicated and, therefore, are not given here. If you are interested in the computational formulas, see the references. As with the designs in the previous three sections, we provide the appropriate linear model for these more complex designs and use either regression or the standard ANOVA output of a statistical software package to analyze the data.

**Example 14.13**

3-Factor Factorial Design and Model

Consider a $2 \times 3 \times 3$ factorial experiment with qualitative factors and $r = 3$ experimental units randomly assigned to each treatment.

a. Write the appropriate linear model for the design.

b. Indicate the sources of variation and their associated degrees of freedom in a partial ANOVA table.

Solution

a. Denote the three qualitative factors as $A$, $B$, and $C$, with $A$ at two levels and $B$ and $C$ at three levels. Then the linear model for the experiment will contain one parameter corresponding to main effects for $A$, two each for $B$ and $C$, $(1)(2) = 2$ each for the $AB$ and $AC$ interactions, $(2)(2) = 4$ for the $BC$ interaction, and $(1)(2)(2) = 4$ for the three-way $ABC$ interaction. Three-way interaction terms measure the failure of two-way interaction effects to remain the same from one level to another level of the third factor. The model is

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1}_{\text{A main effect}} + \underbrace{\beta_2 x_2 + \beta_3 x_3}_{\text{B main effects}} + \underbrace{\beta_4 x_4 + \beta_5 x_5}_{\text{C main effects}}$$

$$+ \underbrace{\beta_6 x_1 x_2 + \beta_7 x_1 x_3}_{\text{A}\times\text{B interaction}} + \underbrace{\beta_8 x_1 x_4 + \beta_9 x_1 x_5}_{\text{A}\times\text{C interaction}}$$

$$+ \underbrace{\beta_{10} x_2 x_4 + \beta_{11} x_2 x_5 + \beta_{12} x_3 x_4 + \beta_{13} x_3 x_5}_{\text{B}\times\text{C interaction}}$$

$$+ \underbrace{\beta_{14} x_1 x_2 x_4 + \beta_{15} x_1 x_3 x_4 + \beta_{16} x_1 x_2 x_5 + \beta_{17} x_1 x_3 x_5}_{\text{A}\times\text{B}\times\text{C interaction}}$$

**TABLE 14.6 Partial ANOVA Table for Example 14.13**

| Source | df |
|---|---|
| Main effect $A$ | 1 |
| Main effect $B$ | 2 |
| Main effect $C$ | 2 |
| $AB$ interaction | 2 |
| $AC$ interaction | 2 |
| $BC$ interaction | 4 |
| $ABC$ interaction | 4 |
| Error | 36 |
| Total | 53 |

where

$$x_1 = \begin{cases} 1 & \text{if level 1 of } A \\ 0 & \text{if level 2 of } A \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if level 1 of } B \\ 0 & \text{if not} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if level 2 of } B \\ 0 & \text{if not} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{if level 1 of } C \\ 0 & \text{if not} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if level 2 of } C \\ 0 & \text{if not} \end{cases}$$

b. The sources of variation and the respective degrees of freedom corresponding to these sets of parameters are shown in Table 14.6.

The degrees of freedom for SS(Total) will always equal $(n - 1)$—that is, $n$ minus 1 degree of freedom for $\beta_0$. Since the degrees of freedom for all sources must sum to the degrees of freedom for SS(Total), it follows that the degrees of freedom for Error will equal the degrees of freedom for SS(Total), minus the sum of the degrees of freedom for main effects and interactions, i.e., $(n - 1) - 17$. Our experiment will contain three observations for each of the $2 \times 3 \times 3 = 18$ treatments; therefore, $n = (18)(3) = 54$, and the degrees of freedom for Error will equal $53 - 17 = 36$.

When analyzing the data from a more complex factorial experiment, the ANOVA table will not only include the degrees of freedom for each source of variation, but also the associated mean squares, values of the $F$ test statistics, and their observed significance levels. Each $F$ statistic would represent the ratio of the source mean square to $\text{MSE} = s^2$. We illustrate with two practical examples.

## Example 14.14

ANOVA for a 3-Factor Factorial Design

A transistor manufacturer conducted an experiment to investigate the effects of three factors on productivity (measured in thousands of dollars of items produced) per 40-hour week. The factors were as follows:

1. Length of work week (two levels): five consecutive 8-hour days or four consecutive 10-hour days
2. Shift (two levels): day or evening shift
3. Number of coffee breaks (three levels): 0, 1, or 2

The experiment was conducted over a 24-week period with the $2 \times 2 \times 3 = 12$ treatments assigned in a random manner to the 24 weeks. The data for this completely randomized design are shown in Table 14.7. Perform an analysis of variance for the data.

Solution

The data were subjected to an analysis of variance. The SAS printout is shown in Figure 14.24. Pertinent sections of the SAS printout are boxed and numbered, as follows:

1. The value of SS(Total), shown in the **Corrected Total** row of box 1, is 1,091.833333. The degrees of freedom associated with this quantity is equal to $(n - 1) = (24 - 1) = 23$. Box 1 gives the partitioning (the analysis of variance) of this quantity into two sources of variation. The first source, **Model**, corresponds to the 11 parameters (all except $\beta_0$) in the model. The second source is **Error**. The

TRANSISTOR1

**TABLE 14.7 Data for Example 14.14**

| | | Day Shift Coffee Breaks | | | Night Shift Coffee Breaks | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 0 | 1 | 2 |
| **Length** | **4 days** | 94 | 105 | 96 | 90 | 102 | 103 |
| **of Work** | | 97 | 106 | 91 | 89 | 97 | 98 |
| **Week** | **5 days** | 96 | 100 | 82 | 81 | 90 | 94 |
| | | 92 | 103 | 88 | 84 | 92 | 96 |

The ANOVA Procedure

Dependent Variable: PRODUCT

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 1009.833333 | 91.803030 | 13.43 | <.0001 |
| Error | 12 | 82.000000 | 6.833333 | | |
| Corrected Total | 23 | 1091.833333 | | | |

| R-Square | Coeff Var | Root MSE | PRODUCT Mean |
|---|---|---|---|
| 0.924897 | 2.768647 | 2.614065 | 94.41667 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| SHIFT | 1 | 48.1666667 | 48.1666667 | 7.05 | 0.0210 |
| DAYS | 1 | 204.1666667 | 204.1666667 | 29.88 | 0.0001 |
| SHIFT*DAYS | 1 | 8.1666667 | 8.1666667 | 1.20 | 0.2958 |
| BREAKS | 2 | 334.0833333 | 167.0416667 | 24.45 | <.0001 |
| SHIFT*BREAKS | 2 | 385.5833333 | 192.7916667 | 28.21 | <.0001 |
| BREAKS*DAYS | 2 | 8.0833333 | 4.0416667 | 0.59 | 0.5689 |
| SHIFT*BREAKS*DAYS | 2 | 21.5833333 | 10.7916667 | 1.58 | 0.2461 |

**FIGURE 14.24**
SAS ANOVA printout for $2 \times 2 \times 3$ factorial

degrees of freedom, sums of squares, and mean squares for these quantities are shown in their respective columns. For example, MSE = 6.833333. The $F$ statistic for testing

$$H_0\colon \quad \beta_1 = \beta_2 = \cdots = \beta_{11} = 0$$

is based on $\nu_1 = 11$ and $\nu_2 = 12$ degrees of freedom and is shown on the printout as $F = 13.43$. The observed significance level, shown under **Pr > F**, is less than .0001. This small observed significance level presents ample evidence to indicate that at least one of the three independent variables—shifts, number of days in a working week, or number of coffee breaks per day—contributes information for the prediction of mean productivity.

2. To determine which sets of parameters are actually contributing information for the prediction of $y$, we examine the breakdown (box 2) of SS(Model) into components corresponding to the sets of parameters for main effects SHIFT, DAYS, and BREAKS, and two-way interactions, SHIFT*DAYS, SHIFT*BREAKS, and BREAKS*DAYS. The last **Model** source of variation corresponds to the set of all three-way SHIFT*BREAKS*DAYS parameters. Note that the degrees of freedom for these sources sum to 11, the number of degrees of freedom for **Model**. Similarly, the sum of the component sums of squares is equal to SS(Model). Box 2 does not give the mean squares associated with the sources, but it does give the $F$ values associated with testing hypotheses concerning the set of parameters associated with each source. It also gives the observed significance levels of these tests. You can see that there is ample evidence to indicate the presence of a SHIFT*BREAKS interaction. The $F$ tests associated with all three main effect parameter sets are also statistically significant at the $\alpha = .05$ level of significance. The practical implication of these results is that there is evidence to indicate that all three independent variables, shift, number of work days per week,

and number of coffee breaks per day, contribute information for the prediction of productivity. The presence of a SHIFT*BREAKS interaction means that the effect of the number of breaks on productivity is not the same from shift to shift. Thus, the specific number of coffee breaks that might achieve maximum productivity on the shift might be different from the number of breaks that will achieve maximum productivity on the other shift.

3. Box 3 gives the value of $s = \sqrt{\text{MSE}} = 2.614065$. This value will be used to construct confidence intervals for pairwise comparisons of the 12 treatment means. (Details of the procedure are provided in Section 14.8.)
4. Box 4 gives the value of $R^2$, a measure of how well the model fits the experimental data. It is of value primarily when the number of degrees of freedom for error is large—say, at least 5 or 6. The larger the number of degrees of freedom for error, the greater will be its practical importance. The value of $R^2$ for this analysis, .924897, indicates that the model provides a fairly good fit to the data. It also suggests that the model could be improved by adding new predictor variables or, possibly, by including higher-order terms in the variables originally included in the model.

## Example 14.15

Mixed Design: Factorial Laid Out in Blocks

In a manufacturing process, a plastic rod is produced by heating a granular plastic to a molten state and then extruding it under pressure through a nozzle. An experiment was conducted to investigate the effect of two factors, extrusion temperature (°F) and pressure (pounds per square inch), on the rate of extrusion (inches per second) of the molded rod. A complete $2 \times 2$ factorial experiment (that is, with each factor at two levels) was conducted. Three batches of granular plastic were used for the experiment, with each batch (viewed as a block) divided into four equal parts. The four portions of granular plastic for a given batch were randomly assigned to the four treatments; this was repeated for each of the three batches, resulting in a $2 \times 2$ factorial experiment laid out in three blocks. (This is called a *mixed design,* since it has the features of both a factorial experiment and a randomized block design.) The data are shown in Table 14.8. Perform an analysis of variance for this mixed design.

Solution

This experiment consists of a three-way classification of the data corresponding to batches (blocks), pressure, and temperature. The analysis of variance for this $2 \times 2$ factorial experiment (four treatments) laid out in a randomized block design (three blocks) yields the sources and degrees of freedom shown in Table 14.9.

The linear model for the experiment is

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1}^{\substack{\text{Main}\\\text{effect}\\P}} + \overbrace{\beta_2 x_2}^{\substack{\text{Main}\\\text{effect}\\T}} + \overbrace{\beta_3 x_1 x_2}^{\substack{\text{PT}\\\text{inter-}\\\text{action}}} + \overbrace{\beta_4 x_3 + \beta_5 x_4}^{\substack{\text{Block}\\\text{terms}}}$$

where

$$x_1 = \text{Pressure} \qquad x_2 = \text{Temperature}$$

$$x_3 = \begin{cases} 1 & \text{if block 2} \\ 0 & \text{otherwise} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{if block 3} \\ 0 & \text{otherwise} \end{cases}$$

RODMOLD

**TABLE 14.8 Data for Example 14.15**

| | | Batch (block) 1 | | 2 | | 3 | |
|---|---|---|---|---|---|---|---|
| | | Pressure | | Pressure | | Pressure | |
| | | 40 | 60 | 40 | 60 | 40 | 60 |
| **Temperature** | **200°** | 1.35 | 1.74 | 1.31 | 1.67 | 1.40 | 1.86 |
| | **300°** | 2.48 | 3.63 | 2.29 | 3.30 | 2.14 | 3.27 |

**TABLE 14.9 Table of Sources and Degrees of Freedom for Example 14.15**

| Source | df |
|---|---|
| Pressure ($P$) | 1 |
| Temperature ($T$) | 1 |
| Blocks | 2 |
| Pressure–temperature interaction | 1 |
| Error | 6 |
| Total | 11 |

The SPSS printout for the analysis of variance is shown in Figure 14.25. The $F$ test for the overall model is highly significant ($p$-value $=$ .000). Thus, there is ample evidence to indicate differences among the block means, or the treatment means, or both. Proceeding to the breakdown of the model sources, you can see that the values of the $F$ statistics for pressure, temperature, and the temperature–pressure interaction are all highly significant (that is, their observed significance levels are very small). Therefore, all of the terms ($\beta_1x_1$, $\beta_2x_2$, and $\beta_3x_1x_2$) contribute information for the prediction of $y$.

The treatments in the experiment were assigned according to a randomized block design. Thus, we expected the extrusion of the plastic to vary from batch to batch. Because the $F$ test for testing differences among block means was not statistically significant ($p$-value $=$ .265), there is insufficient evidence to indicate a difference in the mean extrusion of the plastic from batch to batch. Blocking does not appear to have increased the amount of information in the experiment.

**Tests of Between-Subjects Effects**

Dependent Variable: RATE

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 7.149[a] | 5 | 1.430 | 83.226 | .000 |
| Intercept | 58.256 | 1 | 58.256 | 3390.818 | .000 |
| PRESSURE | 1.687 | 1 | 1.687 | 98.222 | .000 |
| TEMP | 5.044 | 1 | 5.044 | 293.590 | .000 |
| PRESSURE * TEMP | .361 | 1 | .361 | 20.985 | .004 |
| BATCH | 5.732E-02 | 2 | 2.866E-02 | 1.668 | .265 |
| Error | .103 | 6 | 1.718E-02 | | |
| Total | 65.509 | 12 | | | |
| Corrected Total | 7.252 | 11 | | | |

a. R Squared = .986 (Adjusted R Squared = .974)

**FIGURE 14.25**
SPSS ANOVA printout for Example 14.15

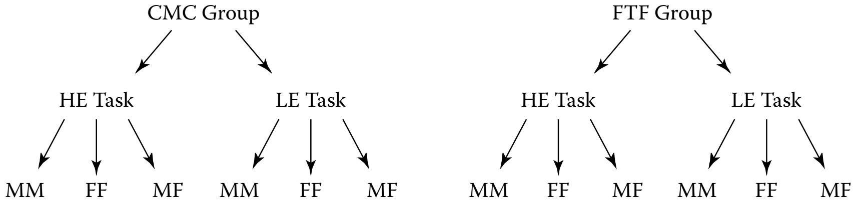
Many other complex designs, such as fractional factorials, Latin square designs, and incomplete blocks designs, fall under the general $k$-way classification of data. Consult the references for the layout of these designs and the linear models appropriate for analyzing them.

## Applied Exercises

14.38 *Computer-mediated communication study.* Computer-mediated communication (CMC) is a form of interaction that heavily involves technology (e.g., instant messaging, e-mail). Refer to the *Journal of Computer-Mediated Communication* (Apr. 2004) study to compare relational intimacy in people interacting via CMC to people meeting face-to-face (FTF), Exercise 8.42 (p. 401). Recall that participants were 48 undergraduate students, of which half were randomly assigned to the CMC group (communicating with the "chat" mode of instant-messaging software) and half assigned to the FTF group (meeting in a conference room). Subjects within each group were randomly assigned to either a high equivocality (HE) or low equivocality (LE) task that required communication with their group members. In addition, the researchers counterbalanced gender, so that each group–task combination had an equal number of females and males; these subjects were then divided into male–male pairs, female–female pairs, and male–female pairs. Consequently, there were two pairs of subjects assigned to each of the 2 (groups) $\times$ 2 ( tasks ) $\times$ 3 (gender pairs) $=$ 12 treatments. A layout of the design is shown on p. 796. The variable of interest, relational intimacy score, was measured (on a 7-point scale) for each subject pair.

a. Write the complete model appropriate for the 2 $\times$ 2 $\times$ 3 factorial design.

**Design Layout for Exercise 14.38**




b. Give the sources of variation and associated degrees of freedom for an ANOVA table for this design.
c. The researchers found no significant three-way interaction. Interpret this result practically.
d. The researchers found a significant two-way interaction between group and task. Interpret this result practically.
e. The researchers found no significant main effect or interactions for gender pair. Interpret this result practically.

14.39 *Flotation of sulfured copper materials.* The *Brazilian Journal of Chemical Engineering* (Vol. 22, 2005) published a study to compare two foaming agents in the flotation of sulfured copper materials process. The two agents were surface-active bio oil (SABO) and pine oil (PO). A $2 \times 2 \times 2 \times 2$ factorial design was used to investigate the effect of four factors on the percentage of copper in the flotation concentrate. The four factors are: foaming agent (SABO or PO), agent-to-mineral mass ratio (low or high), collector-to-mineral mass ratio (low or high), and liquid-to-solid ratio (low or high). Percentage copper measurements $(y)$ were obtained for each of the $2 \times 2 \times 2 \times 2 = 16$ treatments. The data are listed in the table.

**FOAM**

| Agent-to-Mineral Mass Ratio | Collector-to-Mineral Mass Ratio | Liquid-to-Solid Ratio | % Copper | |
|---|---|---|---|---|
| | | | SABO | PO |
| L | L | L | 6.11 | 6.96 |
| H | L | L | 6.17 | 7.31 |
| L | H | L | 6.60 | 7.37 |
| H | H | L | 7.15 | 7.52 |
| L | L | H | 6.24 | 7.17 |
| H | L | H | 6.98 | 7.48 |
| L | H | H | 7.19 | 7.57 |
| H | H | H | 7.59 | 7.78 |

*Source:* Brossard, L. E., et al. "The surface-active bio oil solution in sulfured copper mineral benefit." *Brazilian Journal of Chemical Engineering*, Vol. 22, No. 1, 2005 (Table 3).

a. Write the complete model appropriate for the $2 \times 2 \times 2 \times 2$ factorial design.
b. Note that there is no replication in the experiment. (That is, there is only one observation for each of the 16 treatments.) How will this impact the analysis of the model, part **a**?
c. Write a model for $E(y)$ that includes only main effects and two-way interaction terms.
d. Fit the model, part **c**, to the data. Give the least-squares prediction equation.
e. Conduct tests (at $\alpha = .05$) for the interaction terms. Interpret the results.
f. Is it advisable to conduct any main effect tests? If so, perform the analysis. If not, explain why.

14.40 *Whiteness of bond paper.* An experiment was conducted to investigate the effects of three factors—paper stock, bleaching compound, and coating type—on the whiteness of fine bond paper. Three paper stocks (factor $A$), four types of bleaches (factor $B$), and two types of coatings (factor $C$) were used for the experiment. Six paper specimens were prepared for each of the $3 \times 4 \times 2$ stock–bleach–coating combinations and a measure of whiteness was recorded.

a. Construct an analysis of variance table showing the sources of variation and the respective degrees of freedom.
b. Suppose MSE $= .14$, MS$(AB) = .39$, and the mean square for all interactions combined is .73. Do the data provide sufficient evidence to indicate any interactions among the three factors? Test using $\alpha = .05$.
c. Do the data present sufficient evidence to indicate an $AB$ interaction? Test using $\alpha = .05$. From a practical point of view, what is the significance of an $AB$ interaction?
d. Suppose SS$(A) = 2.35$, SS$(B) = 2.71$, and SS$(C) = .72$. Find SS(Total). Then find $R^2$ and interpret its value.

14.41 *Grafting cardanol onto natural rubber.* Chemical modification of natural rubber is used to increase the rubber's resistance to weathering. Cardanol, an agricultural byproduct of the cashew industry, is often grafted onto natural rubber for this purpose. Chemical engineers investigated a new method of grafting cardanol onto natural rubber and reported the results in *Industrial & Engineering Chemical Research* (May 1, 2013). An experiment was designed to assess the effect of four factors on grafting efficiency (measured as a percentage). The four factors and their levels are: Initiator concentration (IC)—1, 2, or 3 parts per hundred rubber (phr); Cardanol concentration (CC)—5, 10, or 15 phr; Reaction temperature—35, 50, or 65 ºC; and, Reaction time—6, 8, or 10 hours.

a. Consider a full $3 \times 3 \times 3 \times 3$ factorial design. How many treatments are investigated in this design? List them.
b. Give the equation of the complete model required to conduct the $3 \times 3 \times 3 \times 3$ factorial ANOVA.

c. Construct a partial ANOVA table for this design, giving the sources of variation and associated degrees of freedom. Assume that the experiment has 2 replications.

d. The researchers, with limited resources, did not run a full factorial. Rather, they ran an orthogonal fractional factorial design with a single replication that required only 9 treatments (runs). These treatments and associated responses (grafting efficiencies) are shown in the accompanying table. With this design, is it possible to investigate all factor interactions?

e. Use the regression approach to fit a main effects only model to the data. Which factors appear to have an effect on mean grafting efficiency?

**CARDANOL**

| Run | Treatment IC | CC | Temp | Time | Efficiency |
|---|---|---|---|---|---|
| 1 | 1 | 5 | 35 | 6 | 81.94 |
| 2 | 1 | 10 | 50 | 8 | 52.38 |
| 3 | 1 | 15 | 65 | 10 | 54.62 |
| 4 | 2 | 5 | 50 | 10 | 84.92 |
| 5 | 2 | 10 | 65 | 6 | 78.93 |
| 6 | 2 | 15 | 35 | 8 | 36.47 |
| 7 | 3 | 5 | 65 | 8 | 67.79 |
| 8 | 3 | 10 | 35 | 10 | 43.96 |
| 9 | 3 | 15 | 50 | 6 | 42.85 |

*Source:* Mohapatra, S. & Nando, G.B. "Chemical Modification of Natural Rubber in the Latex Stage by Grafting Cardanol, a Waste from the Cashew Industry and a Renewable Resource", *Industrial & Engineering Chemical Research,* Vol. 52, No. 17, May 1, 2013 (Tables 2 and 3).

**MEDWIRE**

14.42 *Quality of manufactured medical wires.* The factors that impact the quality of manufactured medical wires used in cardiovascular devices was investigated in *Quality Engineering* (Vol. 25, 2013). A complete 3-factor factorial design, with each factor at 2 levels, was employed. The three factors (levels) are: Machine type (I or II), Reduction angle (narrow or wide), and Bearing length (short or long). The dependent variable of interest was the ratio of load to tensile strength. The experiment was replicated 3 times, with the data saved in the **MEDWIRE** file. A MINITAB printout of the analysis of variance is shown below. Fully interpret the results. As part of your answer, make a statement about whether or not the factors impact load-to-tensile-strength ratio independently.

14.43 *High-strength nickel alloys.* In increasingly severe oil well environments, oil producers are interested in high-strength nickel alloys that are corrosion-resistant. Since nickel alloys are especially susceptible to hydrogen embrittlement, an experiment was conducted to compare the yield strengths of nickel alloy tensile specimens cathodically charged in a 4% sulfuric acid solution saturated with carbon disulfide, a hydrogen recombination poison. Two alloys were combined: inconel alloy (75% nickel composition) and incoloy (30% nickel composition). The alloys were tested under two material conditions (cold-rolled and cold-drawn), each at three different charging times (0, 25, and 50 days). Thus, a $2 \times 2 \times 3$ factorial experiment was conducted, with alloy type at two levels, material condition at two levels, and charging time at three levels. Two hydrogen-charged tensile specimens were prepared for each of the $2 \times 2 \times 3 = 12$ factor–level combinations. Their yield strengths (kilograms per square inch) are recorded in the table on p. 798.

**MINITAB Output for Exercise 14.42**

```
Analysis of Variance for RATIO, using Adjusted SS for Tests

Source                DF  Seq SS  Adj SS  Adj MS      F      P
MACHINE                1  0.2301  0.2301  0.2301   4.09  0.060
ANGLE                  1  4.1750  4.1750  4.1750  74.24  0.000
LENGTH                 1  0.3197  0.3197  0.3197   5.69  0.030
MACHINE*ANGLE          1  2.8635  2.8635  2.8635  50.92  0.000
MACHINE*LENGTH         1  0.1426  0.1426  0.1426   2.54  0.131
ANGLE*LENGTH           1  0.6048  0.6048  0.6048  10.76  0.005
MACHINE*ANGLE*LENGTH   1  0.4401  0.4401  0.4401   7.83  0.013
Error                 16  0.8997  0.8997  0.0562
Total                 23  9.6756

S = 0.237136   R-Sq = 90.70%   R-Sq(adj) = 86.63%
```

### Data for Exercise 14.43

**NICKEL**

| | | Alloy Type | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Inconel | | | | Incoloy | | | |
| | | Cold-rolled | | Cold-drawn | | Cold-rolled | | Cold-drawn | |
| *Charging* | *0 days* | 53.4 | 52.6 | 47.1 | 49.3 | 50.6 | 49.9 | 30.9 | 31.4 |
| *Time* | *25 days* | 55.2 | 55.7 | 50.8 | 51.4 | 51.6 | 53.2 | 31.7 | 33.3 |
| | *50 days* | 51.0 | 50.5 | 45.2 | 44.0 | 50.5 | 50.2 | 29.7 | 28.1 |

### SAS Output for Exercise 14.43

```
                              The ANOVA Procedure

Dependent Variable: YIELD

                                          Sum of
     Source                      DF      Squares     Mean Square    F Value    Pr > F

     Model                       11  1931.734583      175.612235     258.73    <.0001

     Error                       12     8.145000        0.678750

     Corrected Total             23  1939.879583


                R-Square     Coeff Var      Root MSE    YIELD Mean

                0.995801      1.801942      0.823863      45.72083


     Source                      DF     Anova SS     Mean Square    F Value    Pr > F

     ALLOY                        1  552.0004167     552.0004167     813.26    <.0001
     MATCOND                      1  956.3437500     956.3437500    1408.98    <.0001
     ALLOY*MATCOND                1  339.7537500     339.7537500     500.56    <.0001
     TIME                         2   71.0408333      35.5204167      52.33    <.0001
     ALLOY*TIME                   2    7.9858333       3.9929167       5.88    0.0166
     MATCOND*TIME                 2    4.1725000       2.0862500       3.07    0.0836
     ALLOY*MATCOND*TIME           2    0.4375000       0.2187500       0.32    0.7306
```

a. The SAS analysis of variance printout for the data is shown above. Is there evidence of any interactions among the three factors? Test using $\alpha = .05$. (*Note*: This means that you must test all the interaction parameters. The drop in SSE appropriate for the test would be the sum of all interaction sums of squares.)

b. Now examine the $F$ tests shown on the printout for the individual interactions. Which, if any, of the interactions are statistically significant at the .05 level of significance?

14.44 *High-strength nickel alloys (continued).* Refer to Exercise 14.43. Since charging time is a quantitative factor, we could plot the strength $y$ versus charging time $x_1$ for each of the four combinations of alloy type and material condition. This suggests that a prediction equation relating mean strength $E(y)$ to charging time $x_1$ may be useful. Consider the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3 + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_2 x_3 + \beta_9 x_1^2 x_2 + \beta_{10} x_1^2 x_3 + \beta_{11} x_1^2 x_2 x_3$$

where

$$x_1 = \text{Charging time}$$
$$x_2 = \begin{cases} 1 & \text{if inconel alloy} \\ 0 & \text{if incoloy alloy} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{if cold-rolled} \\ 0 & \text{if cold-drawn} \end{cases}$$

a. Using this model, give the relationship between mean strength $E(y)$ and charging time $x_1$ for cold-drawn incoloy alloy.

b. Using this model, give the relationship between mean strength $E(y)$ and charging time $x_1$ for cold-drawn inconel alloy.

c. Using this model, give the relationship between mean strength $E(y)$ and charging time $x_1$ for cold-rolled inconel alloy.

d. Fit the model to the data and find the prediction equation.

e. Refer to part **d**. Find the prediction equations for each of the four combinations of alloy type and material condition.

**MINITAB Output for Exercise 14.46**

```
Analysis of Variance for YIELD, using Adjusted SS for Tests

Source          DF   Seq SS   Adj SS   Adj MS        F       P
PRESSURE         1   208.33   208.33   208.33   122.95   0.000
TEMP             1   363.00   363.00   363.00   214.23   0.000
PRESSURE*TEMP    1    27.00    27.00    27.00    15.93   0.007
WEEK             2    15.17    15.17     7.58     4.48   0.065
Error            6    10.17    10.17     1.69
Total           11   623.67


S = 1.30171   R-Sq = 98.37%   R-Sq(adj) = 97.01%
```

f. Refer to part **d**. Plot the data points for each of the four combinations of alloy type and material condition. Graph the respective prediction equations.

14.45 *High-strength nickel alloys (continued).* Refer to Exercises 14.43–14.44. If the relationship between mean strength $E(y)$ and charging time $x_1$ is the same for all four combinations of alloy type and material condition, the appropriate model for $E(y)$ is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Fit this model to the data. Use the regression results, together with the information in the printout of Exercise 14.36, to decide whether the data provide sufficient evidence to indicate differences among the second-order models relating $E(y)$ to $x_1$ for the four categories of alloy type and material condition. Test using $\alpha = .05$.

14.46 *Investigating the yield of a chemical.* A $2 \times 2$ factorial experiment was conducted for each of 3 weeks to determine the effect of two factors, temperature and pressure, on the yield of a chemical. Temperature was set at 300° and 500°. The pressure maintained in the reactor was set at 100 and 200 pounds per square inch. Four days were randomly selected within each week, and the four factor-level combinations were randomly assigned to them. The yield data for the $2 \times 2$ factorial experiment, laid out in three blocks of time, are shown in the accompanying table. The MINITAB printout for the analysis of variance is shown above.

**CHEMICAL**

| | | Week 1 Temperature | | Week 2 Temperature | | Week 3 Temperature | |
|---|---|---|---|---|---|---|---|
| | | 300 | 500 | 300 | 500 | 300 | 500 |
| *Pressure* | *100* | 64 | 73 | 65 | 72 | 62 | 70 |
| | *200* | 69 | 81 | 71 | 85 | 67 | 83 |

a. What type of design was used for this experiment?
b. Construct an analysis of variance table showing all sources and their respective degrees of freedom.
c. Why does the analysis of variance table not include sources for the interaction of weeks with temperature and pressure?
d. Do the data provide sufficient evidence to indicate an interaction between temperature and pressure? Give the $p$-value for the test. What is the practical significance of this result?
e. Was blocking in time useful in increasing the amount of information in the experiment? That is, do the data provide sufficient evidence to indicate differences among the block means? Give the $p$-value for the test.

## 14.7 Nested Sampling Designs (*Optional*)

The random error $\varepsilon$ in an ANOVA model is intended to represent the contribution of many variables (most of them unknown) that affect the response variable $y$. We hope that the net effect of these variables on the response will assume the properties described in the assumptions listed in Section 11.2. Sometimes the random sources of variation that enter into the sum of squares for error can be partitioned into two or more sources. The following example illustrates this situation.

Suppose a pharmaceutical manufacturer wants to estimate the mean potency of a batch of an antibiotic. The potency reading produced by a piece of equipment will vary from observation to observation as a result of at least two sources of random error. Antibiotic that is being produced in a vat is not a homogeneous substance; the potency varies slightly from one location in the batch to another. In addition, the

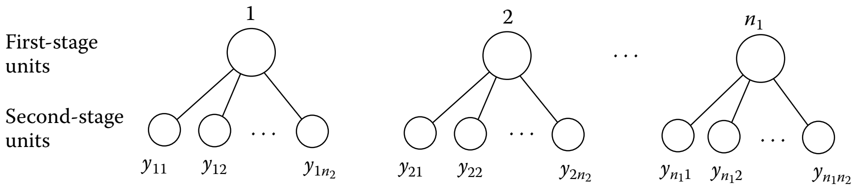



**FIGURE 14.26**
Diagrammatic representation of a two-stage nested sampling design

potency reading produced in the measurement process will vary from observation to observation because of equipment error. Thus, repeated measurements on the same specimen vary from one reading to another.

One way to separate and to estimate the magnitudes of these two sources of variation is to perform the sampling in two stages. First, we randomly select $n_1$ specimens from the batch. Then we measure the potency of each specimen $n_2$ times. Because $n_2$ second-stage sampling units are obtained from each first-stage or **primary unit** (see Figure 14.26), the sampling procedure is called a **nested sampling design**. It is also referred to as **subsampling**—that is, sampling within a sample.

**Definition 14.3**

A **two-stage nested sampling design** involves the random selection of $n_1$ first-stage (primary) units from a population. Subsamples of $n_2$ second-stage units are then randomly selected from within each primary unit.

Nested sampling can be expanded to any number of stages. For example, suppose that after the equipment reacts to a specimen's potency, an operator must reset a gauge before taking an individual reading. Thus, repeated readings of the equipment's reaction to a specimen will vary from one observation to another as a result of the operator's recalibration process. The magnitude of this third source of sampling error can be evaluated using a three-stage sampling design. In addition to the two stages previously described, for each measurement produced by the equipment's reaction to a specimen, the operator would be required to recalibrate and read the meter $n_3$ times. This three-stage nested sampling experiment is shown diagrammatically in Figure 14.27.

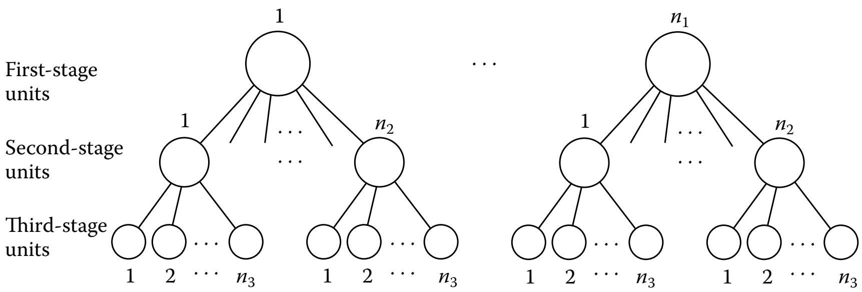



**FIGURE 14.27**
Diagrammatic representation of a three-stage nested sampling design

The probabilistic models for the ANOVA designs presented in previous sections are called *fixed-effect models* since the levels of all components in the model (e.g., treatments, factor interaction) other than the random error term are set or "fixed" prior to observing the response $y$. Conversely, models for nested sampling designs contain more than one *random component;* hence, they are called *nested (or random effects) models*. Models for nested designs and the corresponding ANOVAs are presented in this optional section.

## Two-Stage Nested Sampling Designs

Consider a two-stage nested sampling design consisting of $n_2$ second-stage units for each of $n_1$ first-stage units. Since each second-stage unit will yield one observation, the experiment will yield $n = n_1n_2$ values of the response variable $y$.

We will let $y_{ij}$ denote the observation on the $j$th second-stage unit $(j = 1, 2, \ldots, n_2)$ within the $i$th first-stage unit $(i = 1, 2, \ldots, n_1)$. The probabilistic model that we will use to describe this response is shown in the next box. From a practical point of view, this model implies that $y$ is equal to a constant, $\mu$, plus two random components, $\alpha_i$ and $\varepsilon_{ij}$. The response associated with every second-stage unit within the same first-stage unit $i$ will read higher or lower than $\mu$ by the same random amount, $\alpha_i$. The response $y_{ij}$ associated with each second-stage unit will also be larger or smaller than $(\mu + \alpha_i)$ by an amount $\varepsilon_{ij}$. This random error will vary from one second-stage unit to another.

Because $y_{ij}$ is equal to a constant $(\mu)$ plus the sum of two normally distributed random variables, it follows that $y_{ij}$ is a normally distributed random variable with mean and variance

$$E(y_{ij}) = \mu + E(\alpha_i) + E(\varepsilon_{ij}) = \mu + 0 + 0 = \mu$$

$$\begin{aligned} V(y_{ij}) &= V(\alpha_i) + V(\varepsilon_{ij}) + 2\text{Cov}(\varepsilon_{ij}, \alpha_i) = \sigma_\alpha^2 + \sigma^2 + 0 \\ &= \sigma_\alpha^2 + \sigma^2 \end{aligned}$$

### The Probabilistic Model for a Two-Stage Nested Sampling Design

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad (i = 1, 2, \ldots, n_1; \quad j = 1, 2, \ldots, n_2)$$

where $\alpha_i$ and $\varepsilon_{ij}$ are independent normally distributed random variables with

$$E(\alpha_i) = E(\varepsilon_{ij}) = 0$$

$$V(\alpha_i) = \sigma_\alpha^2$$

$$V(\varepsilon_{ij}) = \sigma^2$$

In addition, every pair of values, $\alpha_i$ and $\alpha_j$ $(i \neq j)$, are independent. Similarly, pairs of values of $\varepsilon$ are independent.

Note that although all the random components of the model are independent of one another, the $y$ values within the same first-stage unit will be correlated. To illustrate, the correlation between two observations from the $i$th first-stage unit is

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ik}) &= E\{[y_{ij} - E(y_{ij})][y_{ik} - E(y_{ik})]\} \\ &= E[(\mu + \alpha_i + \varepsilon_{ij} - \mu)(\mu + \alpha_i + \varepsilon_{ik} - \mu)] \\ &= E[(\alpha_i + \varepsilon_{ij})(\alpha_i + \varepsilon_{ik})] \\ &= E(\alpha_i^2 + \alpha_i\varepsilon_{ij} + \alpha_i\varepsilon_{ik} + \varepsilon_{ij}\varepsilon_{ik}) \\ &= E(\alpha_i^2) + E(\alpha_i\varepsilon_{ij}) + E(\alpha_i\varepsilon_{ik}) + E(\varepsilon_{ij}\varepsilon_{ik}) \end{aligned}$$

The last three expectations, which are covariances, equal 0 because the random components of the model are assumed to be independent. Then, since $E(\alpha_i) = 0$, it follows that $E(\alpha_i^2) = \sigma_i^2$ and the covariance between two $y$ values in the same first-stage unit is

$$\text{Cov}(y_{ij}, y_{ik}) = \sigma_\alpha^2$$

The analysis of variance for a nested sampling design partitions SS(Total) into two parts (see Table 14.10), one measuring the variability *between* the first-stage means and the second measuring the variability of the $y$ values *within* the individual first-stage units.

The objectives of the analysis of variance are to obtain estimates of $\sigma_\alpha^2$ and $\sigma^2$ and to determine whether $\sigma_\alpha^2 > 0$—that is, whether the variation among the first-stage ($A$)

**TABLE 14.10 An Analysis of Variance Table for a Two-Stage Nested Sampling Design**

| Source | df | SS | MS | $E$(MS) | $F$ |
|---|---|---|---|---|---|
| First stage: $A$ | $n_1 - 1$ | SS($A$) | MS($A$) | $\sigma^2 + n_2\sigma_\alpha^2$ | MS($A$)/MS($B$ in $A$) |
| Second stage: $B$ within $A$ | $n_1(n_2 - 1)$ | SS($B$ in $A$) | MS($B$ in $A$) | $\sigma^2$ | |
| Total | $n_1 n_2 - 1$ | SS(Total) | | | |

units exceeds the variation of the $y$ values within first-stage units. The expected values of MS($A$) and MS($B$ within $A$) are shown in the $E$(MS) column of Table 14.10. Unbiased estimates of $\sigma_\alpha^2$ and $\sigma^2$ can be obtained from these mean squares. In addition, it can be shown (proof omitted) that when $\sigma_\alpha^2 = 0$,

$$F = \frac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$$

is an $F$ statistic with $\nu_1 = n_1 - 1$ and $\nu_2 = n_1(n_2 - 1)$ degrees of freedom. The test of $H_0$: $\sigma_\alpha^2 = 0$ against $H_a$: $\sigma_\alpha^2 > 0$ is conducted in the same manner as the $F$ tests of the previous sections.

The notation for an analysis of variance for a two-stage nested sampling design, the formulas for computing the mean squares, and the $F$ test are shown in the accompanying boxes. When you examine the formulas for calculating the sums of squares, note their similarity to the corresponding formulas for the analysis of variance of a replicated two-factor factorial experiment. If the first-stage units are viewed as one direction of classification, then the main effect sum of squares for this direction is SS($A$). Then SS($B$ in $A$) can be calculated as SS($B$ in $A$) = SS(Total) − SS($A$).

## Notation for the Analysis of Variance of a Two-Stage Nested Sampling Design

$y_{ij}$ = Observation on the $j$th second-stage unit within the $i$th first-stage unit

$n_1$ = Number of first-stage units

$n_2$ = Number of second-stage units

$n = n_1 n_2$ = Total number of observations

$A_i$ = Total of all observations in the $i$th first-stage unit

$\overline{A}_i$ = Mean of the $n_2$ observations in the $i$th first-stage unit

$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} y_{ij}$ = Total of all $n$ observations

$\bar{y}$ = Mean of all $n$ observations

## Analysis of Variance $F$ Test for a Two-Stage Nested Sampling Design

$H_0$: $\sigma_\alpha^2 = 0$

$H_a$: $\sigma_\alpha^2 > 0$

*Test statistic*: $F_c = \dfrac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$

*Rejection region*: $F_c > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F_\alpha$ is the tabulated value for an $F$ statistic with $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$ degrees of freedom

Although these formulas are easy to use, the calculations can become quite tedious. Therefore, we will use a statistical software package to perform the computations in the examples.

### Calculation Formulas for a Two-Stage Nested Sampling Design

$$\text{CM} = \text{Correction for the mean}$$

$$= \frac{(\text{Total of all observations})^2}{n} = \frac{\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} y_{ij}\right)^2}{n}$$

$$\text{SS(Total)} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(y_{ij} - \bar{y})^2$$

$$= (\text{Sum of squares of all observations}) - \text{CM}$$

$$= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} y_{ij}^2 - \text{CM}$$

$$\text{SS}(A) = n_2 \sum_{i=1}^{n_1}(\bar{A}_i - \bar{y})^2 = \sum_{i=1}^{n_1}\frac{A_i^2}{n_2} - \text{CM}$$

$$\text{SS}(B \text{ in } A) = \text{SS(Total)} - \text{SS}(A)$$

$$\text{MS}(A) = \frac{\text{SS}(A)}{n_1 - 1}$$

$$\text{MS}(B \text{ in } A) = \frac{\text{SS}(B \text{ in } A)}{n_1(n_2 - 1)}$$

**Example 14.16**
Two-Stage Nested Design ANOVA

The compressive strength of concrete depends on the proportion of water mixed with the cement, the mixing time, the thoroughness of the mixing process, and so on. Even though these variables are presumed fixed at values that will produce maximum compressive strength, they vary slightly from batch to batch and the compressive strength of the concrete varies accordingly. A state highway department conducted an experiment to compare the strength variation between batches to the strength variation of concrete specimens prepared within the same batch. Five concrete specimens were prepared for each of six batches. The compressive strength measurements (in thousands of pounds per square inch) are shown in Table 14.11. Perform an analysis of variance on the data and test $H_0$: $\sigma_\alpha^2 = 0$ against $H_a$: $\sigma_\alpha^2 > 0$, i.e., whether the batch-to-batch variation exceeds the within-batch variation.

Solution

The SAS printout for the nested-design analysis of variance is shown in Figure 14.28. The ANOVA summary table giving the breakdown of SS(Total) is highlighted on the printout.

**CONCRETE3**

**TABLE 14.11 Compressive Strength Measurements for Concrete in Example 14.16**

| | Batch | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| | 5.01 | 4.74 | 4.99 | 5.64 | 5.07 | 5.90 |
| | 4.61 | 4.41 | 4.55 | 5.02 | 4.93 | 5.27 |
| | 5.22 | 4.98 | 4.87 | 4.89 | 4.81 | 5.65 |
| | 4.93 | 4.26 | 4.19 | 5.51 | 5.19 | 4.96 |
| | 5.37 | 4.80 | 4.77 | 5.17 | 5.48 | 5.39 |
| Totals | 25.14 | 23.19 | 23.37 | 26.23 | 25.48 | 27.17 |

Nested Random Effects Analysis of Variance for Variable STRENGTH

| Variance Source | DF | Sum of Squares | F Value | Pr > F | Error Term | Mean Square | Variance Component | Percent of Total |
|---|---|---|---|---|---|---|---|---|
| Total | 29 | 4.745987 | | | | 0.163655 | 0.174665 | 100.0000 |
| BATCH | 5 | 2.469747 | 5.21 | 0.0022 | Error | 0.493949 | 0.079821 | 45.6997 |
| Error | 24 | 2.276240 | | | | 0.094843 | 0.094843 | 54.3003 |

| | |
|---|---|
| STRENGTH Mean | 5.01933333 |
| Standard Error of STRENGTH Mean | 0.12831593 |

**FIGURE 14.28**
SAS printout for nested ANOVA of Example 14.16

The highlighted table shows the partitioning of SS(Total) into two sources of variation, **Batch** and **Error**. The portion corresponding to **Error** is always associated with the variation in the units of the last stage of a nested sampling design. Thus, for a two-stage design, **Error** corresponds to specimen ($B$) within batch ($A$). Thus, the $F$ value for **Batch** represents the $F$ value for testing $H_0$: $\sigma_\alpha^2 = 0$. Now, the tabulated value of $F_\alpha$ for $\alpha = .05$ with $\nu_1 = 5$ and $\nu_2 = 24$ degrees of freedom (given in Table 10 of Appendix B) is $F_{.05} = 2.62$. Since the computed value of $F$ exceeds this value, there is evidence to indicate that $H_a$: $\sigma_\alpha^2 > 0$, is true, i.e., that the variation between batches exceeds the variation within batches. Note that the same conclusion can be reached by observing that the $p$-value of the test (shaded on the printout) is .0022.

## Three-Stage Nested Sampling Designs

We now assume that we have a three-stage sampling design containing $n_1$ first-stage units, $n_2$ second-stage units per first-stage unit, and $n_3$ third-stage units per second-stage unit. The total number of observations for this experiment is then $n = n_1 n_2 n_3$. The probabilistic model for a response obtained from a three-stage nested sampling design contains three random components, which represent the variation between first-, second-, and third-stage sampling units. We will let $y_{ijk}$ denote the response on the $k$th third-stage unit within the $j$th second-stage and the $i$th first-stage units. The model for $y_{ijk}$ is shown in the box.

### The Probabilistic Model for a Three-Stage Nested Sampling Design

$$y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}$$

where $\alpha_i$, $\gamma_{ij}$, and $\varepsilon_{ijk}$ are independent, normally distributed random variables with

$$E(\alpha_i) = E(\gamma_{ij}) = E(\varepsilon_{ijk}) = 0$$

$$V(\alpha_i) = \sigma_\alpha^2$$

$$V(\gamma_{ij}) = \sigma_\gamma^2$$

$$V(\varepsilon_{ijk}) = \sigma^2$$

In addition, every pair of values, $\alpha_i$ and $\alpha_j$ $(i \neq j)$, are independent. Similarly, pairs of values of $\gamma$ and $\varepsilon$ are also independent.

The analysis of variance for a three-stage nested sampling design is an extension of the two-stage analysis. Before giving the computational formulas, we will examine the analysis of variance table shown in Table 14.12.

**TABLE 14.12 Analysis of Variance Table for a Three-Stage Nested Sampling Design**

| Source | df | SS | MS | $E$(MS) | $F$ |
|---|---|---|---|---|---|
| First stage ($A$) | $n_1 - 1$ | SS($A$) | $\dfrac{\text{SS}(A)}{n_1 - 1}$ | $\sigma^2 + n_3\sigma_\gamma^2 + n_2n_3\sigma_\alpha^2$ | $\dfrac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$ |
| Second stage ($B$ within $A$) | $n_1(n_2 - 1)$ | SS($B$ in $A$) | $\dfrac{\text{SS}(B \text{ in } A)}{n_1(n_2 - 1)}$ | $\sigma^2 + n_3\sigma_\gamma^2$ | $\dfrac{\text{MS}(B \text{ in } A)}{\text{MS}(C \text{ in } B)}$ |
| Third stage ($C$ within $B$) | $n_1n_2(n_3 - 1)$ | SS($C$ in $B$) | $\dfrac{\text{SS}(C \text{ in } B)}{n_1n_2(n_3 - 1)}$ | $\sigma^2$ | |
| Total | $n_1n_2n_3 - 1$ | SS(Total) | | | |

When certain assumptions are made concerning $\sigma_\alpha^2$ and $\sigma_\gamma^2$, the ratios of mean squares are $F$ statistics with degrees of freedom, $\nu_1$ and $\nu_2$, corresponding to the numerator and denominator mean squares, respectively. For example, if $\sigma_\gamma^2 = 0$, then $E[\text{MS}(B \text{ in } A)] = E[\text{MS}(C \text{ in } B)]$ and

$$F = \frac{\text{MS}(B \text{ in } A)}{\text{MS}(C \text{ in } B)}$$

has an $F$ distribution with $\nu_1 = n_1(n_2 - 1)$ and $\nu_2 = n_1n_2(n_3 - 1)$ degrees of freedom. This statistic is used to test $H_0$: $\sigma_\gamma^2 = 0$ against $H_a$: $\sigma_\gamma^2 > 0$.

Similarly, if $\sigma_\alpha^2 = 0$, then $E[\text{MS}(A)] = E[\text{MS}(B \text{ in } A)]$ and

$$F = \frac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$$

has an $F$ distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_1(n_2 - 1)$ degrees of freedom. This statistic is used to test $H_0$: $\sigma_\alpha^2 = 0$ against $H_a$: $\sigma_\alpha^2 > 0$.

The notation used in the analysis of variance for a three-stage nested sampling design, the computational formulas, and statistical tests are shown in the accompanying boxes.

## Notation for the Analysis of Variance of a Three-Stage Nested Sampling Design

$y_{ijk}$ = Observation on the $k$th third-stage unit within the $j$th second-stage and the $i$th first-stage unit

$n_1$ = Number of first-stage units

$n_2$ = Number of second-stage units

$n_3$ = Number of third-stage units

$n = n_1n_2n_3$ = Total number of observations

$A_i$ = Total of all observations in the $i$th first-stage unit

$\bar{A}_i$ = Mean of all observations in the $i$th first-stage unit

$B_{ij}$ = Total of all observations in the $j$th second-stage unit within $i$th first-stage unit

$\bar{B}_{ij}$ = Mean of all observations in the $j$th second-stage unit within $i$th first-stage unit

$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} y_{ijk}$ = Total of all $n$ observations

$\bar{y}$ = Mean of all $n$ observations

### Analysis of Variance $F$ Tests for a Three-Stage Nested Sampling Design

#### A Test for First-Stage Variation

$H_0$: $\sigma_\alpha^2 = 0$

$H_a$: $\sigma_\alpha^2 > 0$

*Test statistic*: $F = \dfrac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$

*Rejection region*: $F > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F_\alpha$ is the tabulated value for an $F$ statistic that possesses $\nu_1 = n_1 - 1$ and $\nu_2 = n_1(n_2 - 1)$ degrees of freedom, and $F_c$ is the computed value of the test statistic.

#### A Test for Second-Stage Variation

$H_0$: $\sigma_\gamma^2 = 0$

$H_a$: $\sigma_\gamma^2 > 0$

*Test statistic*: $F = \dfrac{\text{MS}(B \text{ in } A)}{\text{MS}(C \text{ in } B)}$

*Rejection region*: $F > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F_\alpha$ is the tabulated value for an $F$ statistic that possesses $\nu_1 = n_1(n_2 - 1)$ and $\nu_2 = n_1 n_2(n_3 - 1)$ degrees of freedom, and $F_c$ is the computed value of the test statistic.

### Calculation Formulas for a Three-Stage Nested Sampling Design

$$\text{CM} = \text{Correction for the mean}$$

$$= \frac{(\text{Total of all observations})^2}{n} = \frac{\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} y_{ijk}\right)^2}{n}$$

$$\text{SS(Total)} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} (y_{ijk} - \bar{y})^2$$

$$= (\text{Sum of squares of all observations}) - \text{CM}$$

$$= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} y_{ijk}^2 - \text{CM}$$

$$\text{SS}(A) = n_2 n_3 \sum_{i=1}^{n_1} (\bar{A}_i - \bar{y})^2 = \sum_{i=1}^{n_1} \frac{A_i^2}{n_2 n_3} - \text{CM}$$

$$\text{SS}(B \text{ in } A) = n_3 \sum_{j=1}^{n_2} (\bar{B}_{1j} - \bar{A}_1)^2 + n_3 \sum_{j=1}^{n_2} (\bar{B}_{2j} - \bar{A}_2)^2$$

$$+ \cdots + n_3 \sum_{j=1}^{n_2} (\bar{B}_{n_1 j} - \bar{A}_{n_1})^2$$

$$= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{B_{ij}^2}{n_3} - \sum_{i=1}^{n_1} \frac{A_i^2}{n_2 n_3}$$

*Note:* Whenever totals are squared and summed, the divisor is equal to the number of observations in a *single* total. Thus, there are $n_3$ observations in a second-stage total and $n_2 n_3$ in a first-stage total.

$$\text{SS}(C \text{ in } B) = \text{SS}(\text{Total}) - \text{SS}(A) - \text{SS}(B \text{ in } A)$$

$$\text{MS}(A) = \frac{\text{SS}(A)}{n_1 - 1}$$

$$\text{MS}(B \text{ in } A) = \frac{\text{SS}(B \text{ in } A)}{n_1(n_2 - 1)}$$

$$\text{MS}(C \text{ in } B) = \frac{\text{SS}(C \text{ in } B)}{n_1 n_2(n_3 - 1)}$$

## Example 14.17

Three-Stage Nested Design ANOVA

One job of computer scientists is to evaluate computer hardware and software systems. Computer performance evaluation for software involves monitoring the CPU times of processed jobs. In addition to job-to-job variability, the CPU time will vary depending on the day on which the job is submitted and the initiator (a hardware device that initiates job processing) on which the job runs. A three-stage nested sampling experiment was conducted to compare the three sources of variation. On each of five randomly selected days, two randomly selected initiators were monitored. Then four jobs of a particular type were randomly selected from each initiator. The CPU times (in seconds) are shown in Table 14.13. Perform an analysis of variance on the data, and test the following hypotheses:

a. $H_0$: $\sigma_\alpha^2 = 0$ against $H_a$: $\sigma_\alpha^2 > 0$ (i.e., whether the day-to-day variation exceeds the initiators-within-days variation)

b. $H_0$: $\sigma_\gamma^2 = 0$ against $H_a$: $\sigma_\gamma^2 > 0$ (i.e., whether the initiators-within-days variation exceeds the jobs-within-initiators variation)

**Solution**

The SAS printout for the analysis of variance for the three-stage nested design is shown in Figure 14.29. The printout is similar to that for the two-stage nested design. The ANOVA summary table highlighted on the printouts shows the partitioning of SS(Total) into sources of variation due to **Days, Initiators**, and **Error**. For this three-stage nested design, **Days** represents the first-stage source of variation, Days ($A$); **Initiators** represents the second-stage source of variation. Initiators ($B$) within Days ($A$); and **Error** corresponds to Jobs ($C$) in Initiators ($B$), the last-stage source. Note that the error term used to compute the $F$ statistics is given under **Error Term**.

a. The $F$ value for testing $H_0$: $\sigma_\alpha^2 = 0$ against $H_a$: $\sigma_\alpha^2 > 0$ is $F = 3.27$ and the corresponding $p$-value is .1131. Since the $p$-value exceeds $\alpha = .05$, there is insufficient evidence to indicate that $\sigma_\alpha^2 > 0$; that is, we cannot conclude that the variation between days exceeds the variation of initiators within days.

b. The $F$ value for testing $H_0$: $\sigma_\gamma^2 = 0$ against $H_a$: $\sigma_\gamma^2 > 0$ is $F = .77$ and the corresponding $p$-value is .5763. Since the $p$-value exceeds $\alpha = .05$, there is insufficient evidence to indicate that $\sigma_\gamma^2 > 0$. We cannot conclude that initiators-within-days variation exceeds the jobs-within-initiators variation.

CPU

**TABLE 14.13 CPU Times for Example 14.17**

| | | Day | | | | |
|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** |
| **Initiator** | **1** | 5.61 | 1.22 | .89 | 3.69 | 7.61 |
| | | 3.44 | 1.86 | 1.26 | 10.84 | 6.02 |
| | | .66 | .05 | 1.43 | 1.07 | .52 |
| | | .29 | 2.11 | 1.90 | 2.46 | 1.98 |
| | **2** | 8.17 | 1.53 | 6.27 | 15.20 | 2.41 |
| | | .13 | 1.03 | 1.01 | 3.62 | 3.02 |
| | | 4.22 | 3.67 | 2.55 | 10.22 | 1.77 |
| | | 2.50 | 2.29 | 1.52 | 1.83 | 1.38 |

Nested Random Effects Analysis of Variance for Variable CPU

| Variance Source | DF | Sum of Squares | F Value | Pr > F | Error Term | Mean Square | Variance Component | Percent of Total |
|---|---|---|---|---|---|---|---|---|
| Total | 39 | 414.277438 | | | | 10.622498 | 11.485132 | 100.0000 |
| DAY | 4 | 95.275425 | 3.27 | 0.1131 | INIT | 23.818856 | 2.066375 | 17.9917 |
| INIT | 5 | 36.439288 | 0.77 | 0.5763 | Error | 7.287858 | -0.532725 | 0.0000 |
| Error | 30 | 282.562725 | | | | 9.418758 | 9.418758 | 82.0083 |

CPU Mean 3.23125000
Standard Error of CPU Mean 0.77166794

**FIGURE 14.29**
SAS printout for nested ANOVA of Example 14.17

More complex nested sampling designs (e.g., those involving factorial experiments and interaction effects) are beyond the scope of this text. Consult the references for more information on these complex, but useful, designs.

## Applied Exercises

14.47 *Density of black clay and mud.* Large highwall failures at a strip mine in Queensland, Australia, occur by the sliding of soft, black bands of clay, called black clay planes, near the base of the highwall. A study was conducted to determine whether the chemical and mineralogical properties of the black clay planes are similar to mudstone (*Engineering Geology*, Oct. 1985). Black clay and mudstone specimens were randomly selected at each of three randomly selected sites within the siltstone faces in the ramp area of the mine. The densities of the specimens (in kilograms per cubic meter) are recorded in the table. An SAS printout of the nested ANOVA follows.

a. How many first-stage observations were included in the sample?
b. How many second-stage units were selected per first-stage unit?
c. Give the total number of observations obtained in the sample.
d. Write the probabilistic model for this sampling design.
e. Find estimates of $\sigma_\alpha^2$ and $\sigma^2$ on the printout.
f. Conduct a test to determine whether the variation in black clay and mudstone specimen densities between sites exceeds the variation within sites. Use $\alpha = .10$.

**CLAYMUD**

| Site 1 | Site 2 | Site 3 |
|---|---|---|
| 2.06 | 2.09 | 2.07 |
| 1.84 | 2.03 | 2.04 |
| 2.47 | 2.01 | 1.90 |
| 2.12 | 2.04 | 2.00 |
| 2.00 | 2.41 | 2.64 |

*Source:* Seedsman, R. W., and Emerson, W. W. "The formation of planes of weakness in the highwall at Goonyella Mine, Queensland, Australia." *Engineering Geology*, Vol. 22, No. 2, Oct. 1985, p. 164 (Table I).

### SAS Output for Exercise 14.47

Nested Random Effects Analysis of Variance for Variable DENSITY

| Variance Source | DF | Sum of Squares | F Value | Pr > F | Error Term | Mean Square | Variance Component | Percent of Total |
|---|---|---|---|---|---|---|---|---|
| Total | 14 | 0.672173 | | | | 0.048012 | 0.055800 | 100.0000 |
| SITE | 2 | 0.002573 | 0.02 | 0.9772 | Error | 0.001287 | -0.010903 | 0.0000 |
| Error | 12 | 0.669600 | | | | 0.055800 | 0.055800 | 100.0000 |

DENSITY Mean 2.11466667
Standard Error of DENSITY Mean 0.00926163

14.48 *Porosity of paper.* A two-stage nested sampling design was used to collect data to estimate the mean porosity of paper emerging from a paper machine. Ten patches of paper were randomly selected from the end of the paper roll, and four porosity readings were made on each. The data are shown in the following table.

**PAPER**

| Paper Patch | Porosity Readings | | | |
|---|---|---|---|---|
| 1 | 974 | 978 | 976 | 975 |
| 2 | 981 | 985 | 978 | 986 |
| 3 | 1,014 | 1,012 | 1,018 | 1,010 |
| 4 | 990 | 996 | 989 | 988 |
| 5 | 1,012 | 1,009 | 1,011 | 1,012 |
| 6 | 978 | 980 | 974 | 982 |
| 7 | 988 | 979 | 986 | 983 |
| 8 | 1,004 | 1,001 | 1,008 | 1,008 |
| 9 | 989 | 984 | 982 | 983 |
| 10 | 999 | 1,002 | 998 | 1,003 |

a. Perform an analysis of variance on the data. Give the ANOVA summary table.
b. Obtain the estimates of $\sigma_\alpha^2 = 0$ and $\sigma^2$.
c. Do the data provide sufficient evidence to indicate that the variation in porosity between patches exceeds the variation of porosity within patches? Test using $\alpha = .05$.

14.49 *Nested sampling at DuPont.* Quality control engineers at DuPont utilize nested sampling schemes to determine the percentage of a product shipped that conforms to specifications.* First, a random sample of $n_1$ production lots is selected; then, a random sample of $n_2$ batches is selected from each production lot. Finally, $n_3$ shipping lots are randomly selected from each batch for inspection. Suppose $n_1 = 10$, $n_2 = 5$, and $n_3 = 20$. Give the sources and degrees of freedom for an analysis of variance for the nested sampling design.

14.50 *Sulfur content of mined coal.* An experiment was conducted to estimate the mean level of sulfur content in coal produced by a particular mine. Five days were randomly selected and identified as coal sampling days. On each day, five coal cars were randomly selected and portions of coal were removed from each. Two specimens were prepared from each portion and analyzed for sulfur content. The data are shown in the accompanying table; a MINITAB printout of the analysis follows.

a. Construct an analysis of variance table to display the results.
b. Do the data provide sufficient evidence to indicate that the variation in sulfur content between days exceeds the variation within days? Test using $\alpha = .05$.

*Henderson, R. K. "On Making the Transition from Inspection to Process Control." Paper presented at Joint Statistical Meetings, American Statistical Association and Biometric Society, August 1986, Chicago, IL.

**COALMINE**

| | | Day | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| | *1* | .107 | .091 | .110 | .088 | .089 |
| | | .105 | .089 | .113 | .092 | .088 |
| | *2* | .104 | .093 | .108 | .091 | .087 |
| | | .103 | .090 | .110 | .093 | .089 |
| *Coal Cars* | *3* | .101 | .092 | .111 | .092 | .092 |
| *Within Days* | | .099 | .093 | .108 | .089 | .090 |
| | *4* | .106 | .091 | .106 | .088 | .091 |
| | | .105 | .091 | .108 | .087 | .090 |
| | *5* | .108 | .092 | .106 | .091 | .086 |
| | | .104 | .090 | .109 | .088 | .089 |

**MINITAB Output for Exercise 14.50**

**Nested ANOVA: SULFUR versus DAY, CAR**

```
Analysis of Variance for SULFUR

Source   DF       SS       MS          F       P
DAY       4   0.0034   0.0008    135.300   0.000
CAR      20   0.0001   0.0000      2.288   0.026
Error    25   0.0001   0.0000
Total    49   0.0036
```

c. Do the data provide sufficient evidence to indicate that the variation of sulfur content between cars within a day exceeds the variation within the coal specimens? Test using $\alpha = .05$.

14.51 *Resistivity of silicon crystals.* An experiment was conducted to monitor the resistivity of silicon monocrystals. The original data were collected according to a two-stage nested sampling design in which random samples of eight crystals were selected from among 30 lots. The measured resistivity of the crystals is recorded in the accompanying table for five of these lots.

a. Construct an ANOVA summary table for the nested design.

**CRYSTALS**

| Lot | Measured Values of Resistivity | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.8 | 2.7 | 2.3 | 2.6 | 2.7 | 2.3 | 2.7 | 2.7 |
| 2 | 3.0 | 3.0 | 2.8 | 2.4 | 3.0 | 3.2 | 2.9 | 2.4 |
| 3 | 2.4 | 2.3 | 2.4 | 2.9 | 2.4 | 2.4 | 2.3 | 2.3 |
| 4 | 3.1 | 2.9 | 3.0 | 3.0 | 2.6 | 3.0 | 2.9 | 3.0 |
| 5 | 3.1 | 3.3 | 2.9 | 2.5 | 2.5 | 3.1 | 2.5 | 3.0 |

*Source:* Hoshide, M. "Optimization of lot size for quality assurance of silicon wafers." *Reports of Statistical Application Research*, Union of Japanese Scientists and Engineers, Vol. 19, No. 1, 1972, pp. 8–21.

b. Let $\sigma_B^2$ and $\sigma_W^2$ represent the components of between- and within-lot variances, respectively, of the resistivity readings. Obtain the estimates of $\sigma_B^2$ and $\sigma_W^2$.

c. Do the data provide sufficient evidence to indicate that the variation in resistivity between lots exceeds the variation within lots? Test using $\alpha = .05$.

14.52 *Characteristics of paper stock.* The strength of paper depends upon the length and other characteristics of the wood fiber stock entering the paper machine. Consequently, as the source of the fiber stock varies over time, we expect the strength of the produced paper to vary also. To test this theory, 6 days were randomly selected from within a 4-month period of time. On each of these days, an end-of-the-roll paper patch was selected from each of three randomly selected rolls. Two strength tests were conducted on each of the 18 patches of paper. The strength measurements (pounds per square inch) are shown in the table.

a. Perform an analysis of variance for the data using the formulas provided in this section. Construct an analysis of variance table to display the results.

b. Do the data provide sufficient evidence to indicate that the variation in paper strength between days exceeds the variation within days? Test using $\alpha = .05$.

c. Do the data provide sufficient evidence to indicate that the variation in strength from roll to roll exceeds the variation between strength tests within a roll? Test using $\alpha = .05$.

**WOODFIBER**

| | | Day 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| | *1* | 20.7 | 22.1 | 19.0 | 20.6 | 23.2 | 20.7 |
| | | 19.3 | 20.4 | 19.9 | 18.9 | 22.5 | 18.5 |
| *Rolls* | *2* | 21.2 | 21.6 | 18.8 | 19.8 | 24.2 | 19.6 |
| *Within Days* | | 20.1 | 22.5 | 19.3 | 20.1 | 22.9 | 21.3 |
| | *3* | 19.9 | 20.9 | 20.2 | 20.7 | 23.4 | 20.0 |
| | | 20.5 | 22.1 | 19.4 | 19.2 | 24.6 | 18.6 |

## 14.8 Multiple Comparisons of Treatment Means

Many practical experiments are conducted to determine the largest (or the smallest) mean in a set. For example, suppose that a chemist has developed five chemical solutions for removing a corrosive substance from a metal fitting. The chemist would then want to determine the solution that will remove the greatest amount of the corrosive substance from the fitting in a single application. Similarly, a production engineer might want to determine which among six machines or which among three foremen achieves the highest mean productivity per hour. A mechanical engineer might want to choose one engine, from among five, that is most efficient, and so on.

Once differences among, say, five treatment means have been detected in an ANOVA, choosing the treatment with the largest mean might appear to be a simple matter. We could, for example, obtain the sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_5$, and compare them by constructing a $(1 - \alpha)100\%$ confidence interval for the difference between each pair of treatment means. However, there is a problem associated with this procedure: **A confidence interval for $\mu_i - \mu_j$, with its corresponding value of $\alpha$, is valid only when the two treatments ($i$ and $j$) to be compared are selected prior to experimentation**. After you have looked at the data, you cannot use a confidence interval to compare the treatments for the largest and smallest sample means because they will always be farther apart, on the average, than any pair of treatments selected at random. Furthermore, if you construct a series of confidence intervals, each with a chance $\alpha$ of indicating a difference between a pair of means if in fact no difference exists, then the risk of making *at least one* Type I error in a series of inferences will be larger than the value of $\alpha$ specified for a single interval.

There are a number of procedures for comparing and ranking a group of treatment means. A popular method, known as **Tukey's method**, utilizes the Studentized range

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{s/\sqrt{n}}$$

(where $\bar{y}_{\max}$ and $\bar{y}_{\min}$ are the largest and smallest sample means, respectively) to determine whether the difference in any pair of sample means implies a difference in the corresponding treatment means. The logic behind this **multiple comparisons procedure** is that if we determine a critical value for the difference between the largest and smallest sample means, $|\bar{y}_{\max} - \bar{y}_{\min}|$, one that implies a difference in their respective treatment means, then any other pair of sample means that differ by as much as or more than this critical value would also imply a difference in the corresponding treatment means. Tukey's (1949) procedure selects this critical distance, $\omega$, so that the probability of making one or more Type I errors (concluding that a difference exists between a pair of treatment means if, in fact, they are identical) is $\alpha$. Therefore, the risk of making a Type I error applies to the whole procedure, i.e., to the comparisons of all pairs of means in the experiment, rather than to a single comparison. Consequently, the value of $\alpha$ selected by the researchers is called an **experimentwise error rate** (in contrast to a **comparisonwise error rate**).

Tukey's procedure relies on the assumption that the $p$ sample means are based on independent random samples, *each containing an equal number* $n_t$ *of observations*. Then if $s = \sqrt{\text{MSE}}$ is the computed standard deviation for the analysis, the distance $\omega$ is

$$\omega = q(p, \nu)\frac{s}{\sqrt{n_t}}$$

The tabulated statistic $q_\alpha(p, \nu)$ is the critical value of the Studentized range, the value that locates $\alpha$ in the upper tail of the $q$ distribution. This critical value depends on $\alpha$, the number of treatment means involved in the comparison, and $\nu$, the number of degrees of freedom associated with MSE, as shown in the box. Values of $q(p, \nu)$ for $\alpha = .05$ and $\alpha = .01$ are given in Tables 13 and 14 respectively, of Appendix B.

### Tukey's Multiple Comparisons Procedure: Equal Sample Sizes

1. Select the desired experimentwise error rate, $\alpha$.
2. Calculate

$$\omega = q_\alpha(p, \nu)\frac{s}{\sqrt{n_t}}$$

where

$p$ = Number of sample means

$s = \sqrt{\text{MSE}}$

$\nu$ = Number of degrees of freedom associated with MSE

$n_t$ = Number of observations in each of the $p$ samples

$q_\alpha(p, \nu)$ = Critical value of the Studentized range (Tables 13 and 14 of Appendix B)

3. Calculate and rank the $p$ sample means.
4. Place a bar over those pairs of treatment means that differ by less than $\omega$. Any pair of treatments not connected by an overbar (i.e., differing by more than $\omega$) implies a difference in the corresponding population means.

*Note:* The confidence level associated with all inferences drawn from the analysis is $(1 - \alpha)$.

**Example 14.18**

**Ranking Treatment Means: Tukey's Procedure**

Refer to the ANOVA for the completely randomized design, Example 14.3 (p. 750). Recall that, at $\alpha = .05$, we rejected the null hypothesis of no differences among the mean times until abrasion for the three paint types. Use Tukey's method to compare the three treatment means.

**Solution**

*Step 1* For this analysis, we will select an experimentwise error rate of $\alpha = .05$.

*Step 2* From previous examples, we have $p = 3$ treatments, $\nu = 27$ df for error, $s = \sqrt{\text{MSE}} = 168.95$, and $n_t = 10$ observations per treatment. The critical value of the Studentized range (obtained from Table 13, Appendix B) is $q_{.05}(3, 27) = 3.5$. Substituting these values into the formula for $\omega$, we obtain

$$\omega = q_{.05}(3, 27)\left(\frac{s}{\sqrt{n_t}}\right) = 3.5\left(\frac{168.95}{\sqrt{10}}\right) = 187.0$$

*Step 3* The sample means for the three paint types (obtained from Table 14.2) are

$$\bar{y}_1 = 229.6 \qquad \bar{y}_2 = 309.8 \qquad \bar{y}_3 = 427.8$$

*Step 4* Based on the critical difference $\omega = 187$, the three treatment means are ranked as follows:

$$\begin{array}{llll} \text{Sample Means:} & \overline{229.6 \quad 309.9} & & \\ & & \overline{309.9 \quad 427.8} & \\ \text{Treatments:} & \text{Type 1} & \text{Type 2} & \text{Type 3} \end{array}$$

This same information can be obtained using a statistical software package. The SAS printout of the Tukey analysis is shown in Figure 14.30. Tukey's critical difference, $\omega = 187.33$, is shaded on the printout. (This value differs slightly from our calculated value because of rounding.) Note that SAS lists the treatment means vertically in descending order. Treatment means connected by the same letter (A, B, C, etc.) are *not* significantly different.

From this information we infer that the mean wear time for paint type 3 is significantly larger than the mean wear time for paint type 1, since $\bar{y}_3$ exceeds $\bar{y}_1$ by more than the critical value. However, the treatment pairs (1, 2) and (2, 3) are connected by a bar (or the same letter) since neither $(\bar{y}_2 - \bar{y}_1)$ nor $(\bar{y}_3 - \bar{y}_2)$ exceeds $\omega$. This indicates that the sample means for these pairs of treatments are not significantly different. Practically, these results imply that paint type 3 has the highest mean time until abrasion and paint type 1 has the lowest. The mean for paint type 2, however, is not significantly different from either of the other two means. These inferences are made with an overall confidence level of $(1 - \alpha) = .95$.

```
                              The ANOVA Procedure

                    Tukey's Studentized Range (HSD) Test for WEAR

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                          Type II error rate than REGWQ.


                    Alpha                                     0.05
                    Error Degrees of Freedom                    27
                    Error Mean Square                     28543.37
                    Critical Value of Studentized Range    3.50643
                    Minimum Significant Difference          187.33


               Means with the same letter are not significantly different.


                 Tukey Grouping          Mean       N     TYPE

                              A        427.80      10     3
                              A
                         B    A        309.90      10     2
                         B
                         B             229.60      10     1
```

**FIGURE 14.30**
SAS printout of Tukey rankings of wear means, Example 14.17

Remember that Tukey's multiple comparisons procedure requires the sample sizes associated with the treatments to be equal. This, of course, will be satisfied for the randomized block designs and factorial experiments described in Sections 14.4 and 14.5, respectively. The sample sizes, however, may not be equal in a completely randomized design (Section 14.3). In this case a modification of Tukey's method (sometimes called the **Tukey–Kramer method**) is necessary, as described in the box. The technique requires that the critical difference $\omega_{ij}$ be calculated for each pair of treatments $(i, j)$ in the experiment and pairwise comparisons made based on the appropriate value of $\omega_{ij}$. However, when Tukey's method is used with unequal sample sizes, the value of $\alpha$ selected a priori by the researcher only approximates the true experimentwise error rate. In fact, when applied to unequal sample sizes, the procedure has been found to be more conservative, i.e., less likely to detect differences between pairs of treatment means when they exist, than in the case of equal sample sizes. For this reason, researchers sometimes look to alternative methods of multiple comparisons when the sample sizes are unequal. One such method is **Bonferroni's procedure**.

### Tukey's Approximate Multiple Comparisons Procedure for Unequal Sample Sizes

1. Select the desired experimentwise error rate, $\alpha$.
2. Calculate for each treatment pair $(i, j)$

$$\omega_{ij} = q_{\alpha}(p, \nu)\frac{s}{\sqrt{2}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where

$$\begin{aligned} p &= \text{Number of sample means} \\ s &= \sqrt{\text{MSE}} \\ \nu &= \text{Number of degrees of freedom associated with MSE} \\ n_i &= \text{Number of observations in sample for treatment } i \\ n_j &= \text{Number of observations in sample for treatment } j \\ q_{\alpha}(p, \nu) &= \text{Critical value of the Studentized range (Tables 13} \\ &\phantom{=} \text{ and 14 of Appendix B)} \end{aligned}$$

3. Rank the $p$ sample means and place a bar over any treatment pair $(i, j)$ that differs by less than $\omega_{ij}$. Any pair of sample means not connected by an overbar (i.e., differing by more than $\omega$) implies a difference in the corresponding population means.

*Note:* This procedure is approximate, i.e., the value of $\alpha$ selected by the researcher approximates the true probability of making at least one Type I error.

The Bonferroni approach is based on the following result (proof omitted): If $g$ comparisons are to be made, each with confidence coefficient $1 - \alpha/g$, then the overall probability of making one or more Type I errors (i.e., the experimentwise error rate) is at most $\alpha$. That is, the set of intervals constructed using the Bonferroni method yields an overall confidence level of at least $1 - \alpha$. For example, if you want to construct $g = 2$ confidence intervals with an experimentwise error rate of at most $\alpha = .05$, then each individual interval must be constructed using a confidence level of $1 - .05/2 = .975$.

When applied to pairwise comparisons of treatment means, the Bonferroni technique can be carried out by comparing the difference between two treatment means, $(\bar{y}_i - \bar{y}_j)$, to a critical difference $B_{ij}$, when $B_{ij}$ depends on $n_i$, $n_j$, $\alpha$, MSE, and the total number of treatments to be compared. If the difference between the sample means exceeds the critical difference, there is sufficient evidence to conclude that the population means differ. The steps to follow in carrying out the Bonferroni multiple comparisons procedure are described in the box.

### Bonferroni Multiple Comparisons Procedure for Pairwise Comparisons of Treatment Means

1. Select the experimentwise error rate, $\alpha$.
2. Calculate $B_{ij}$ for each treatment pair $(i, j)$:

$$B_{ij} = (t_{\alpha^*/2})(s)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where

$p$ = Number of sample (treatment) means in the experiment

$g$ = Number of pairwise comparisons [*Note*: If all pairwise comparisons are to be made, then $g = p(p-1)/2$.]

$\alpha^* = \alpha/g$ = Comparisonwise error rate

$s = \sqrt{\text{MSE}}$

$n_i$ = Number of observations in sample for treatment $i$

$n_j$ = Number of observations in sample for treatment $j$

$\nu$ = Number of degrees of freedom associated with MSE

$t_{\alpha^*/2}$ = Critical value of $T$ distribution with $\nu$ df and tail area $\alpha^*/2$ (Table 7, Appendix B)

3. Calculate and rank the sample means.

Place a bar over any treatment pair $(i, j)$ that differs by less than $B_{ij}$. Any pair of means not connected by an overbar implies a difference in the corresponding population means.

*Note:* The level of confidence associated with all inferences drawn from the analysis is at least $(1 - \alpha)$.

### Example 14.19

**Ranking Treatment Means: Bonferroni's Method**

Refer to the rankings of the mean wear times for three paint types, Example 14.18.

a. Use Bonferroni's method to compare the three treatment means.
b. Compare the results, part **a**, with Tukey's procedure.

**Solution**

a. We will follow the three steps outlined in the box.

*Step 1* As in Example 14.18, we will select an experimentwise error rate of $\alpha = .05$.

*Step 2* For $p = 3$ treatments, the number of pairwise comparisons is

$$g = p(p-1)/2 = 3(2)/2 = 3$$

Hence, the adjusted $\alpha$ level (i.e., comparisonwise error rate) is $\alpha^* = \alpha/g = .05/3 \approx .017$. The critical value of the Student's $T$ statistic with $\nu = 27$ df (obtained using a statistical software package) is $t_{\alpha/2} = t_{(.017)/2} = t_{.0083} \approx 2.55$. From Example 14.17, we have $s = \sqrt{\text{MSE}} = 168.95$, and $n_i = 10$ observations per treatment. Substituting these values into the formula for $B$, we obtain

$$B = t_{.0083}(s)\sqrt{(1/n_i) + (1/n_j)} = 2.55(168.95)\sqrt{2/10} \approx 192.7$$

```
                              The ANOVA Procedure

                       Bonferroni (Dunn) t Tests for WEAR

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                         Type II error rate than REGWQ.

                    Alpha                                0.05
                    Error Degrees of Freedom               27
                    Error Mean Square                28543.37
                    Critical Value of t               2.55246
                    Minimum Significant Difference     192.85

              Means with the same letter are not significantly different.

                 Bon Grouping          Mean       N     TYPE

                            A        427.80      10     3
                            A
                       B    A        309.90      10     2
                       B
                       B             229.60      10     1
```

**FIGURE 14.31**
SAS printout of Bonferroni rankings of wear means, Example 14.19

*Step 3* Based on the critical difference $B = 192.7$, the three treatment means are ranked as follows:

| | | | |
|---|---|---|---|
| Sample Means: | 229.6 | 309.9 | 427.8 |
| Treatments: | Type 1 | Type 2 | Type 3 |

(*Note*: These results are shown on the SAS printout, Figure 14.31.)

b. The conclusion reached by Bonferroni's method is identical to Tukey's method: At an overall significance level of .05, (1) the mean wear for paint type 3 is significantly higher than the corresponding mean for paint type 1, and (2) the mean for paint type 2 is not significantly different from either of the other two means. Note, however, that the Bonferroni critical difference, $B = 192.7$, is larger than the Tukey critical difference, $\omega = 187.33$, obtained in Example 14.17. Thus, for this example, the Tukey method will be able to detect smaller differences in the treatment means than Bonferroni's method using the same comparisonwise error rate $\alpha$.

The result, Example 14.19b, reveals that the Bonferroni method produces wider confidence intervals on the differences between treatment means than Tukey's method. This will be true, in general, whenever the sample sizes are the same for the treatments. Consequently, *Tukey's method is the superior multiple comparisons procedure for balanced ANOVA designs* (i.e., designs with the same sample size per treatment). However, with unequal $n$'s, the Bonferroni critical difference will usually be smaller than the Tukey critical difference. Hence, *the Bonferroni method is preferred over Tukey's method when the design is unbalanced* (i.e., when the sample sizes for the treatments are unequal). Keep in mind that the exact $T$ value needed to calculate the Bonferroni critical difference may not be available in the $T$ tables provided in most texts. If you do not have access to a software package that provides this information, you will have to estimate its value.

In general, multiple comparisons of treatment means should be performed only as a follow-up analysis to the ANOVA, i.e., only after we have conducted the appropriate analysis of variance $F$ test(s) and determined that sufficient evidence exists of differences among the treatment means. Be wary of conducting multiple comparisons when the ANOVA $F$ test indicates no evidence of a difference among a small number of treatment means—this may lead to confusing and contradictory results.*

### Warning

In practice, it is advisable to avoid conducting multiple comparisons of a small number of treatment means when the corresponding ANOVA $F$ test is nonsignificant; otherwise, confusing and contradictory results may occur.

## Applied Exercises

14.53 *Robots trained to behave like ants.* Refer to the *Nature* (Aug. 2000) study of robots trained to behave like ants, Exercise 14.4 (p. 754). Multiple comparisons of mean energy expended for the four colony sizes were conducted using an experimentwise error rate of .05. The results are summarized in the table.

| *Sample Mean:* | .97 | .95 | .93 | .80 |
|---|---|---|---|---|
| *Group Size:* | 3 | 6 | 9 | 12 |

a. How many pairwise comparisons are conducted in this analysis?
b. Interpret the results shown in the table.

14.54 *Whales entangled in fishing gear.* Refer to the *Marine Mammal Science* (April 2010) investigation of whales entangled by fishing gear, Exercise 14.5 (p. 754). The mean body lengths (meters) of whales entangled in each of the three types of fishing gear (set nets, pots, and gill nets) are reported below. Tukey's method was used to conduct multiple comparisons of the means with an experiment wise error rate of .01. Based on the results, which type of fishing gear will entangle the shortest whales, on average? The longest whales, on average?

| Mean Length: | 4.45 | 5.28 | 5.63 |
|---|---|---|---|
| Fishing Gear: | Set nets | Gill nets | Pots |

14.55 *Performance of a bus depot.* Refer to the *International Journal of Engineering Science and Technology* (February, 2011) study of public bus depot performance, Exercise 14.6 (p. 755). Recall that 150 customers provided overall performance ratings at each of three different bus depots (Depot 1, Depot 2, and Depot 3). The average performance scores were determined to be significantly different at $\alpha = .05$ using an ANOVA $F$ test. The sample mean performance scores were reported as $\bar{x}_1 = 67.17$, $\bar{x}_2 = 58.95$, and $\bar{x}_3 = 44.49$. The researchers employed the Bonferroni method to rank the three performance means using an experimentwise error rate of .05. Adjusted 95% confidence intervals for the differences between each pair of treatment means are shown in the next table. Use this information to rank the mean performance scores at the three bus depots.

| Comparison | Adjusted 95% CI |
|---|---|
| $(\mu_1 - \mu_2)$ | (1.50, 14.94) |
| $(\mu_1 - \mu_3)$ | (15.96, 29.40) |
| $(\mu_2 - \mu_3)$ | (7.74, 21.18) |

14.56 *Evaluation of flexography printing plates.* Refer to the *Journal of Graphic Engineering and Design* (Vol. 3, 2012) study of the quality of flexography printing, Exercise 14.7 (p. 755). Recall that four different exposure times were studied—8, 10, 12, and 14 minutes—and that the measure of print quality used was dot area (hundreds of dots per square millimeter). Tukey's multiple comparisons procedure (at an experiment wise error rate of .05) was used to rank the mean dot areas of the four exposure times. The results are summarized below. Which exposure time yields the highest mean dot area? Lowest?

| Mean Dot Area: | .571 | .582 | .588 | .594 |
|---|---|---|---|---|
| Exposure Time: (minutes) | 12 | 10 | 14 | 8 |

14.57 *Chemical properties of whole wheat breads.* Whole wheat breads contain a high amount of phytic acid, which tends to lower the absorption of nutrient minerals. The *Journal of Agricultural and Food Chemistry* (Jan. 2005) published the results of a study to determine if sourdough can increase the solubility of whole wheat bread. Four types of bread were prepared from whole meal flour: (1) yeast added, (2) sourdough added, (3) no yeast or sourdough added (control), and (4) lactic acid added. Data were collected on the soluble magnesium level (percent of total magnesium) during fermentation for dough samples of each bread type and analyzed using a one-way ANOVA. The four mean soluble magnesium means were compared

*When a large number of treatments are to be compared, a borderline, nonsignificant $F$ value (e.g., $.05 < p\text{-value} < .10$) may mask differences between some of the means. In this situation, it is better to ignore the $F$ test and proceed directly to a multiple comparisons procedure.

in pairwise fashion using Bonferroni's method. The results are summarized in the table.

| *Mean:* | 7% | 12.5% | 22% | 27.5% |
|---|---|---|---|---|
| *Bread Type:* | Control | Yeast | Lactate | Sourdough |

a. How many pairwise comparisons are made in the Bonferroni analysis?
b. Which treatment(s) yielded the significantly highest mean soluble magnesium level? The lowest?
c. The experimentwise error rate for the analysis was .05. Interpret this value.

**GASTURBINE**

14.58 *Coding method for gas turbines.* Refer to Exercise 14.8 (p. 755). Use Bonferroni's method to compare the mean heat rates of the three gas turbine engines. Use $\alpha = .06$.

**TILLRATIO**

14.59 *Estimating the age of glacial drifts.* Refer to Exercise 14.11 (p. 756). Use a multiple comparisons procedure to compare the mean Al/Be ratios for the five boreholes. Identify the means that appear to differ. Use $\alpha = .05$.

**SCOPOLAMINE**

14.60 *Effect of scopolamine on memory.* Refer to the *Behavioral Neuroscience* (Feb. 2004) study of the drug scopolamine's effects on memory for word-pair associates, Exercise 14.13 (p. 757). Recall that the researchers theorized that the mean number of word pairs recalled for the scopolamine subjects (group 1) would be less than the corresponding means for the placebo subjects (group 2) and the no-drug subjects (group 3). Conduct multiple comparisons of the three means (using an experimentwise error rate of .05). Do the results support the researchers' theory? Explain.

**CRACKPIPE**

14.61 *Repairing pipeline cracks.* Refer to Exercise 14.19 (p. 769). Use a multiple comparisons procedure to compare the mean crack widths for the four wetting periods. Identify the means that appear to differ. Use $\alpha = .05$.

14.62 *Baker's versus brewer's yeast.* Refer to the *Electronic Journal of Biotechnology* (Dec. 15, 2003) study to compare the yeast extracts baker's yeast and brewer's yeast, Exercise 14.29 (p. 786). Recall that a $2 \times 4$ factorial design was employed, with yeast extract at two levels and temperature at four levels. Multiple comparisons of the four temperature means were conducted for each of the two yeast extracts. Interpret the results shown below.

| | | | | | |
|---|---|---|---|---|---|
| *Baker's Yeast:* | Mean yield (%): | 41.1 | 47.5 | 48.6 | 50.3 |
| | Temperature (°C): | 54 | 45 | 48 | 51 |
| *Brewer's Yeast:* | Mean yield (%): | 39.4 | 47.3 | 49.2 | 49.6 |
| | Temperature (°C): | 54 | 51 | 48 | 45 |

**ANTIMONY**

14.63 *Strength of solder joints.* Refer to Exercise 14.34 (p. 788). Use a multiple comparisons procedure to compare the mean shear strengths for the four antimony amounts. Identify the means that appear to differ. Use $\alpha = .01$.

**MOW**

14.64 *Mowing effects on highway right-of-way.* Refer to the *Landscape Ecology Journal* (Jan. 2013) study of mowing effects on vegetation in highway rights-of-way, Exercise 14.35 (p. 788). Recall that a $3 \times 3$ factorial design was employed to estimate the effects of mowing frequency and mowing height on the mean height of vegetation. The researchers detected evidence of interaction between the two factors, mowing frequency (once, twice, or three times per year) and mowing height of the equipment (5, 10, or 20 centimeters). Consequently, they did not rank the mowing frequency means independent of mowing height, and vice-versa. Rather, the researchers ranked all $3 \times 3 = 9$ treatment means in order to determine which treatments yield the lowest and highest mean vegetation height. Use a multiple comparisons method to carry out this analysis at an experimentwise error rate of .05.

## 14.9 Checking ANOVA Assumptions

For each of the experiments and designs discussed in this chapter, we listed in the relevant boxes the assumptions underlying the analysis in the terminology of ANOVA. For example, in the box on p. 749, the assumptions for a completely randomized design are that (1) the $p$ probability distributions of the response $y$ corresponding to the $p$ treatments are normal and (2) the population variances of the $p$ treatments are equal. Similarly, for randomized block designs and factorial designs, the data for the treatments must come from normal probability distributions with equal variances.

These assumptions are equivalent to those required for a regression analysis (see Section 11.2). The reason, of course, is that the probabilistic model for the response $y$ that underlies each design is the familiar general linear regression model of Chapter 11. A brief overview of the techniques available for checking the ANOVA assumptions follows.

### Detecting Nonnormal Populations

1. For each treatment, construct a histogram, stem-and-leaf display, or normal probability plot for the response $y$. Look for highly skewed distributions. (*Note:* For relatively large samples, e.g., 20 or more observations per treatment, ANOVA, like regression, is **robust** with respect to the normality assumption. That is, slight departures from normality will have little impact on the validity of the inferences derived from the analysis.) If the sample size for each treatment is small, then these graphs will probably be of limited use.
2. Formal statistical tests of normality (such as the **Anderson–Darling test, Shapiro–Wilk test**, or **Kolmogorov–Smirnov test**) are also available. The null hypothesis is that the probability distribution of the response $y$ is normal. These tests, however, are sensitive to slight departures from normality. Since in most scientific applications the normality assumption will not be satisfied exactly, these tests will likely result in a rejection of the null hypothesis and, consequently, are of limited use in practice. Consult the references for more information on these formal tests.
3. If the distribution of the response departs greatly from normality, a **normalizing transformation** may be necessary. For example, for highly skewed distributions, transformations on the response $y$ such as $\log(y)$ or $\sqrt{y}$ tend to "normalize" the data since these functions "pull" the observations in the tail of the distribution back toward the mean.

### Detecting Unequal Variances

1. For each treatment, construct a box plot or frequency (dot) plot for $y$ and look for differences in spread (variability). If the variability of the response in each plot is about the same, then the assumption of equal variances is likely to be satisfied. (*Note*: ANOVA is robust with respect to unequal variances for **balanced designs**, i.e., designs with equal sample sizes for each treatment.)
2. When the sample sizes are small for each treatment, only a few points are graphed on the frequency plots, making it difficult to detect differences in variation. In this situation, you may want to use one of several formal statistical tests of homogeneity of variances that are available. For $p$ treatments, the null hypothesis is $H_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$, where $\sigma_i^2$ is the population variance of the response $y$ corresponding to the $i$th treatment. If all $p$ populations are approximately normal, **Bartlett's test for homogeneity of variances** can be applied. Bartlett's test works well when the data come from normal (or near normal) distributions. The results, however, can be misleading for nonnormal data. In situations where the response is clearly not normally distributed, **Levene's test** is more appropriate. The elements of these tests are shown in the accompanying boxes. Note that Bartlett's test statistic depends on whether the sample sizes are equal or unequal.
3. When unequal variances are detected, use one of the **variance-stabilizing transformations** of the response $y$ discussed in Section 11.10.

#### Bartlett's Test of Homogeneity of Variance

$H_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$

$H_a$: At least two variances differ.

*Test statistic* (equal sample sizes):

$$B = \frac{(n-1)\left[p \ln \bar{s}^2 - \sum \ln s_i^2\right]}{1 + \dfrac{p+1}{3p(n-1)}}$$

where

$$n = n_1 = n_2 = \cdots = n_p$$

$s_i^2$ = Sample variance for sample $i$

$\bar{s}^2$ = Average of the $p$ sample variances $= \left(\dfrac{\sum s_i^2}{p}\right)$

$\ln x$ = Natural logarithm (i.e., log to the base $e$) of the quantity $x$

*Test statistic* (unequal sample sizes):

$$B = \frac{\left[\sum(n_i - 1)\right] \ln \bar{s}^2 - \sum(n_i - 1) \ln s_i^2}{1 + \dfrac{1}{3(p-1)}\left\{\sum \dfrac{1}{(n_i - 1)} - \dfrac{1}{\sum(n_i - 1)}\right\}}$$

where

$n_i$ = Sample size for sample $i$

$s_i^2$ = Sample variance for sample $i$

$\bar{s}^2$ = Weighted average of the $p$ sample variances $= \dfrac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)}$

$\ln x$ = Natural logarithm (i.e., log to the base $e$) of the quantity $x$

*Rejection region*: $B > \chi^2_\alpha$,

$p$-value: $P(\chi^2 > B)$

where $\chi^2_\alpha$ locates an area $\alpha$ in the upper tail of a $\chi^2$ distribution with $(p - 1)$ degrees of freedom

*Assumptions*:
1. Independent random samples are selected from the $p$ populations.
2. All $p$ populations are normally distributed.

## Levene's Test of Homogeneity of Variance

$H_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$

$H_a$: At least two variances differ

*Test statistic*: $F = \text{MST}/\text{MSE}$

where MST and MSE are obtained from an ANOVA with $p$ treatments conducted on the transformed response variable $y_i^* = |y_i - \text{Med}_p|$, and $\text{Med}_p$ is the median of the response $y$ values for treatment $p$.

*Rejection region*: $F > F_\alpha$,

$p$-value: $P(F > F_c)$

where $F_\alpha$ locates an area $\alpha$ in the upper tail of an $F$ distribution with $\nu_1 = (p - 1)$ df and $\nu_2 = (n - p)$ df, and $F_c$ is the computed value of the test statistic.

*Assumptions*:
1. Independent random samples are selected from the $p$ treatment populations.
2. The response variable $y$ is a continuous random variable.

### Example 14.20

**Checking ANOVA Assumptions**

Refer to the ANOVA for the completely randomized design, Example 14.3. Recall that we found differences among the mean wear times for the three paint types. Check to see if the ANOVA assumptions are satisfied for this analysis.

**Solution**

First, we'll check the assumption of normality. For this design, there are only seven observations per treatment (class); consequently, constructing graphs (e.g., histograms or stem-and-leaf plots) for each treatment will not be very informative. Alternatively, we can combine the data for the three treatments and form a histogram for all 30 observations in the data set. A MINITAB normal probability plot for the response variable, wear time, is shown in Figure 14.32. The points fall in an approximate straight line. The result of a test for normality of the data is also shown (highlighted) in Figure 14.32. Since the $p$-value of the test exceeds .10, there is insufficient evidence (at $\alpha = .05$) to conclude that the data are nonnormal. Consequently, it appears that the wear times come from a normal distribution.

Next, we check the assumption of equal variances. MINITAB dot plots for wear times are displayed in Figure 14.33. Note that the variability of the response in each plot is about the same; thus, the assumption of equal variances appears to be satisfied. To formally test the hypothesis, $H_0$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, we conduct both Bartlett's and Levene's test for homogeneity of variances. Rather than use the computing formulas shown in the boxes, we resort to a statistical software package. The MINITAB printout of the test results is shown in Figure 14.34. The $p$-values for both tests are shaded on the printout. Since both $p$-values exceed at $\alpha = .05$, there is insufficient evidence to reject the null hypothesis of equal variances. Therefore, it appears that the assumption of equal variance is satisfied also.

**FIGURE 14.32**
MINITAB normal probability plot and normality test, Example 14.20

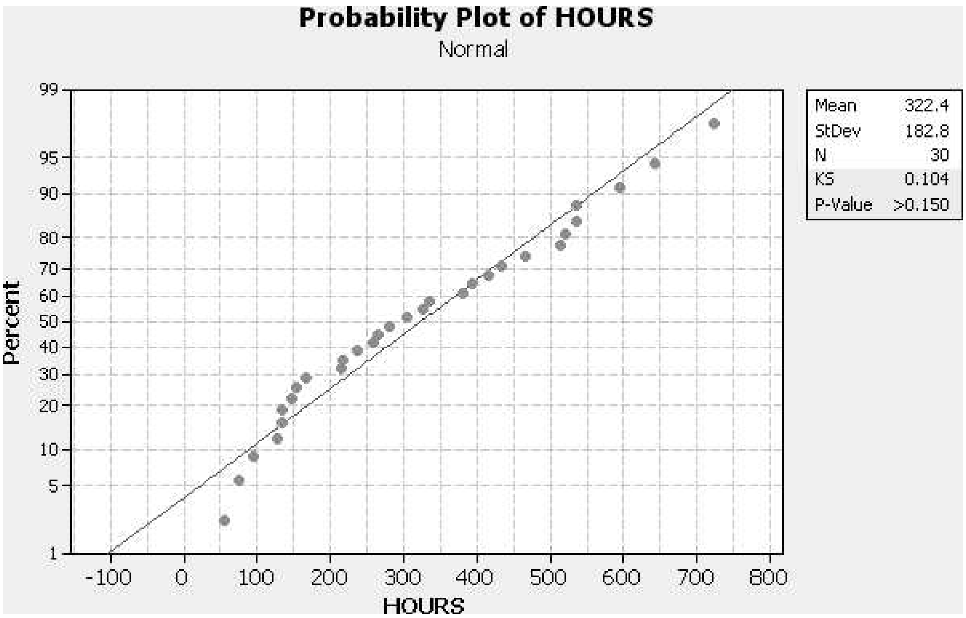



**FIGURE 14.33**
MINITAB dot plot, Example 14.20

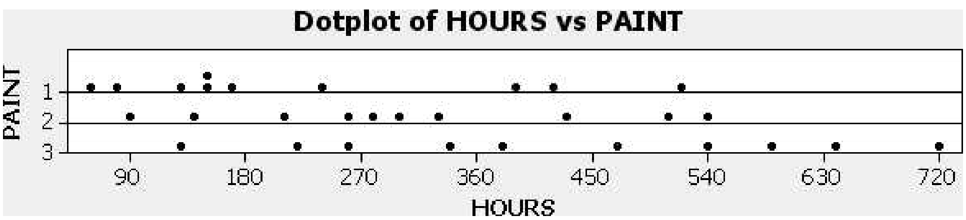
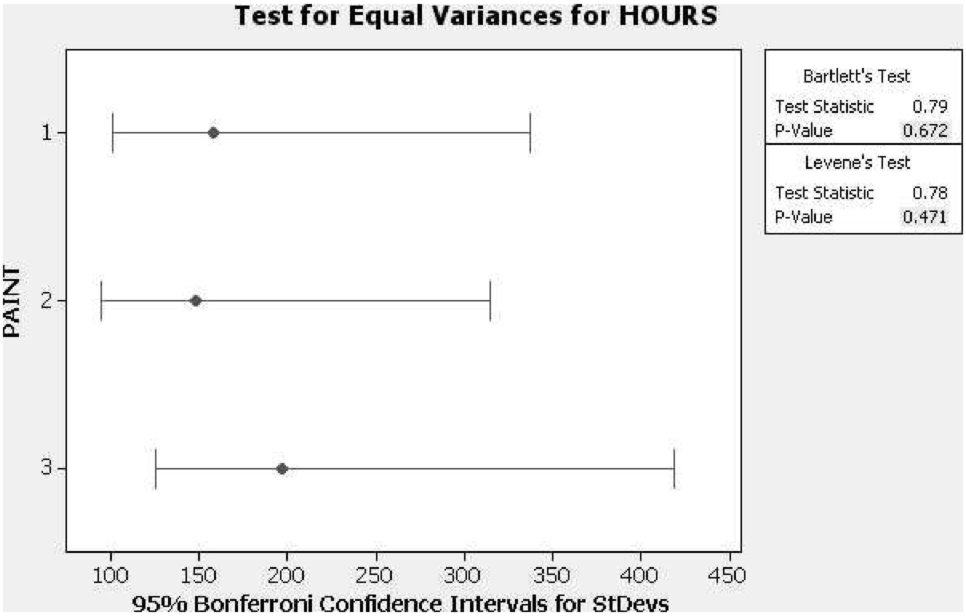


**FIGURE 14.34**
MINITAB test for homogeneity of variances, Example 14.20




In most scientific applications, the assumptions will not be satisfied exactly. These analysis of variance procedures are flexible, however, in the sense that slight departures from the assumptions will not significantly affect the analysis or the validity of the resulting inferences. On the other hand, gross violations of the assumptions (e.g., a nonconstant variance) will cast doubt on the validity of the inferences. Therefore, you should make it standard practice to verify that the assumptions are (approximately) satisfied.

## Applied Exercises

**GASTURBINE**

14.65 *Cooling method for gas turbines.* Check the assumptions for the completely randomized ANOVA of Exercise 14.8 (p. 755).

**TILLRATIO**

14.66 *Estimating the age of glacial drifts.* Check the assumptions for the completely randomized ANOVA of Exercise 14.11 (p. 756).

**SCOPOLAMINE**

14.67 *Effect of scopolamine on memory.* Check the assumptions for the completely randomized ANOVA of Exercise 14.13 (p. 757).

**GENEDARK**

14.68 *Light to dark transition of genes.* Check the assumptions for the randomized block ANOVA of Exercise 14.25 (p. 771).

**ANTIMONY**

14.69 *Strength of solder joints.* Check the assumptions for the factorial design ANOVA of Exercise 14.34 (p. 788).

**BURNIN**

14.70 *Detecting early part failure.* Check the assumptions for the factorial design ANOVA of Exercise 14.36 (p. 789).

## *STATISTICS IN ACTION REVISTED*

### Pollutants at a Housing Development—A Case of Mishandling Small Samples

We now return to the case of the Florida land developer who blamed the failure of his housing plan on the discovery of pollutants (PAH) at the site, and who filed suit against two industries that produced PAH waste materials as part of their industrial processes. Soil specimens were collected at each of four locations: 7 at the housing development site, 8 at Industry A, 5 at Industry B, and 2 at Industry C. Two different molecular diagnostic ratios for measuring level of PAH in soil were determined for each soil specimen. These data are displayed in Table SIA14.1. Recall that the objective is to compare the mean PAH ratios at the four different locations.

**PAH**

**TABLE SIA14.1 Data on PAH ratios at Four Sites**

| Soil Specimen | SITE | RATIO PAH1 | PAH2 |
|---|---|---|---|
| 1 | Development | 0.620 | 1.040 |
| 2 | Development | 0.630 | 1.020 |
| 3 | Development | 0.660 | 1.070 |
| 4 | Development | 0.670 | 1.180 |
| 5 | Development | 0.610 | 1.020 |
| 6 | Development | 0.670 | 1.090 |
| 7 | Development | 0.660 | 1.100 |
| 8 | IndustryA | 0.620 | 0.950 |
| 9 | IndustryA | 0.660 | 1.090 |
| 10 | IndustryA | 0.700 | 0.960 |
| 11 | IndustryA | 0.560 | 0.970 |
| 12 | IndustryA | 0.560 | 1.000 |
| 13 | IndustryA | 0.570 | 1.030 |
| 14 | IndustryA | 0.600 | 0.970 |
| 15 | IndustryA | 0.580 | 1.015 |
| 16 | IndustryB | 0.770 | 1.130 |
| 17 | IndustryB | 0.720 | 1.110 |
| 18 | IndustryB | 0.560 | 0.980 |
| 19 | IndustryB | 0.705 | 1.130 |
| 20 | IndustryB | 0.670 | 1.140 |
| 21 | IndustryC | 0.675 | 1.115 |
| 22 | IndustryC | 0.650 | 1.060 |

*Source:* Info Tech, Inc. (For confidentiality purposes, data values have been altered.)

Since the soil samples were obtained independently from the four different sites, we can treat the data as coming from a completely randomized design. There are two different response (dependent) variables: PAH Ratio 1 and PAH Ratio 2. The design employs a single factor (independent variable): Site (or location). The four levels of Site (Industry A, Industry B, Industry C, and Development) represent the treatments in the experiment. Then, the appropriate null and alternative hypotheses are:

$H_0$: $\mu_A = \mu_B = \mu_C = \mu_D$ $H_a$: At least two of the means, $\mu_A$, $\mu_B$, $\mu_C$, $\mu_D$, are different

*A Flawed Analysis of the Data*

The biochemical expert hired by Industry A chose to analyze the data using a series of $T$-tests for comparing two means. That is, he conducted a two-sample $T$-test (Section 8.7) for each possible pair of sites—Industry A vs. Industry B, Industry A vs. Industry C, Industry A vs. Development, Industry B vs. Industry C, Industry B vs. Development, and Industry C vs. Development. The results of these 6 $T$-tests for the second PAH ratio variable are shown in the MINITAB printouts, Figures SIA14.1a-f.

Recall (from Section 8.7) that each of the $T$-tests is a test of the null hypothesis, $H_0$: $\mu_i = \mu_j$, where $\mu_i$ and $\mu_j$ represent the two population means being compared. The biochemical expert conducted each test

**FIGURE SIA14.1**
MINITAB Output for Two-Sample *T*-tests to Compare PAH2 Ratio Means

**a. Development Site vs. Industry A**

```
Two-sample T for PAH2

SITE         N    Mean   StDev  SE Mean
Development  7  1.0743  0.0565    0.021
IndustryA    8  0.9981  0.0464    0.016


Difference = mu (Development) - mu (IndustryA)
Estimate for difference:  0.0762
95% CI for difference:  (0.0188, 0.1336)
T-Test of difference = 0 (vs not =): T-Value = 2.87  P-Value = 0.013  DF = 13
Both use Pooled StDev = 0.0513
```

**b. Development Site vs. Industry B**

```
Two-sample T for PAH2

SITE         N    Mean   StDev  SE Mean
Development  7  1.0743  0.0565    0.021
IndustryB    5  1.0980  0.0669    0.030


Difference = mu (Development) - mu (IndustryB)
Estimate for difference:  -0.0237
95% CI for difference:  (-0.1031, 0.0557)
T-Test of difference = 0 (vs not =): T-Value = -0.67  P-Value = 0.521  DF = 10
Both use Pooled StDev = 0.0609
```

**c. Development Site vs. Industry C**

```
Two-sample T for PAH2

SITE         N    Mean   StDev  SE Mean
Development  7  1.0743  0.0565    0.021
IndustryC    2  1.0875  0.0389    0.027


Difference = mu (Development) - mu (IndustryC)
Estimate for difference:  -0.0132
95% CI for difference:  (-0.1163, 0.0898)
T-Test of difference = 0 (vs not =): T-Value = -0.30  P-Value = 0.771  DF = 7
Both use Pooled StDev = 0.0544
```

**d. Industry A vs. Industry B**

```
Two-sample T for PAH2

SITE       N    Mean   StDev  SE Mean
IndustryA  8  0.9981  0.0464    0.016
IndustryB  5  1.0980  0.0669    0.030


Difference = mu (IndustryA) - mu (IndustryB)
Estimate for difference:  -0.0999
95% CI for difference:  (-0.1686, -0.0312)
T-Test of difference = 0 (vs not =): T-Value = -3.20  P-Value = 0.008  DF = 11
Both use Pooled StDev = 0.0548
```

using a significance level of $\alpha = .05$. Comparing $\alpha$ to the *p*-value of each test (highlighted on Figure SIA14.1), the expert concluded the following:

(1) The mean PAH2 ratio at **Industry A** is *statistically different* than the corresponding mean at **Industry B** since *p*-value = .008 (see Figure SIA14.1d)
(2) The mean PAH2 ratio at the **development** site is *statistically different* than the corresponding mean at **Industry A** since *p*-value = .013 (see Figure SIA14.1a)
(3) The mean PAH2 ratio at the **development** site is *not statistically different* than the corresponding mean at **Industry B** since *p*-value = .521 (see Figure SIA14.1b)

**FIGURE SIA14.1 (*continued*)**

**e. Industry A vs. Industry C**

```
Two-sample T for PAH2

SITE        N    Mean   StDev  SE Mean
IndustryA   8  0.9981  0.0464    0.016
IndustryC   2  1.0875  0.0389    0.027


Difference = mu (IndustryA) - mu (IndustryC)
Estimate for difference:  -0.0894
95% CI for difference:  (-0.1724, -0.0063)
T-Test of difference = 0 (vs not =): T-Value = -2.48  P-Value = 0.038  DF = 8
Both use Pooled StDev = 0.0456
```

**f. Industry B vs. Industry C**

```
Two-sample T for PAH2

SITE        N    Mean   StDev  SE Mean
IndustryB   5  1.0980  0.0669    0.030
IndustryC   2  1.0875  0.0389    0.027


Difference = mu (IndustryB) - mu (IndustryC)
Estimate for difference:  0.0105
95% CI for difference:  (-0.1234, 0.1444)
T-Test of difference = 0 (vs not =): T-Value = 0.20  P-Value = 0.848  DF = 5
Both use Pooled StDev = 0.0623
```

The last two inferences led the expert to argue that the source of the PAH contamination at the housing development site is more likely to have been derived from Industry A than from Industry B.

The statistician, hired to rebut this testimony, argued that the analysis was flawed. To see why, consider the fact that the biochemical expert conducted 6 independent $T$-tests on the data, each using $\alpha =$ P(Type I error) $= .05$. Now, the probability of the expert concluding that a difference in means exists when, in fact there is no difference (i.e., the probability of committing a Type I error) is .05 for any individual test. However, the expert drew his final conclusion based on the results of all six tests. It can be shown (proof omitted) that the probability of committing at least one Type I error—called the *overall Type I error rate*—when six tests are conducted at $\alpha = .05$ is approximately .265. In other words, there is more than a one in four chance that the expert erroneously concludes that a difference in means exists when there is actually no difference. This error rate is unacceptably high.

A second problem with the testimony of the biochemical expert is that of "accepting the null hypothesis". When the expert's test failed to show a significant difference in means, he declared the mean PAH ratio at the two sites being compared to be "statistically indistinguishable", implying that the population means are equal. By accepting the null hypothesis of equal means, the expert is failing to account for the possibility of a Type II error (i.e., the error of accepting $H_0$ when $H_0$ is false). As we discussed in Chapter 8, the probability of a Type II error, $\beta$, is typically unknown and is not controlled for in the series of two-sample $t$-tests and is likely to be particularly large with the very small samples collected at the sites.

### *A Statistically Valid Analysis of the Data*

Based on our discussion in this chapter, the appropriate way to analyze the data is with an analysis of variance (ANOVA). Since there is a single null hypothesis tested in an ANOVA, the probability of making a Type I error (i.e., the probability of concluding that the means differ when, in fact, they are the same) is simply $\alpha = .05$. A MINITAB printout of the ANOVA results for both dependent variables, PAH Ratio 1 and PAH Ratio 2, are shown in Figures SIA14.2a-b.

The $F$-value and $p$-value of each test are highlighted on the printouts. For the first PAH ratio, $p$-value $= .083$. Consequently, at $\alpha = .05$ there is *insufficient evidence* of differences in the mean PAH ratios in the population of soil samples collected at the four sites. This result contradicts the conclusions drawn by conducting a series of independent samples $t$-tests on the data. Now, the $p$-value for the second

**FIGURE SIA14.2**

**a. Dependent Variable = PAH1**

**One-way ANOVA: PAH1 versus SITE**

```
Source  DF       SS       MS     F      P
SITE     3  0.02041  0.00680  2.61  0.083
Error   18  0.04697  0.00261
Total   21  0.06738

S = 0.05108   R-Sq = 30.29%   R-Sq(adj) = 18.67%


                                    Individual 95% CIs For Mean Based on
                                    Pooled StDev
Level        N     Mean    StDev  ------+---------+---------+---------+---
Development  7  0.64571  0.02507         (-------*-------)
IndustryA    8  0.60625  0.05097  (------*-------)
IndustryB    5  0.68500  0.07858               (---------*---------)
IndustryC    2  0.66250  0.01768     (--------------*---------------)
                                  ------+---------+---------+---------+---
                                      0.600     0.650     0.700     0.750

Pooled StDev = 0.05108
```

**b. Dependent Variable = PAH2**

**One-way ANOVA: PAH2 versus SITE**

```
Source  DF       SS       MS     F      P
SITE     3  0.03977  0.01326  4.45  0.017
Error   18  0.05366  0.00298
Total   21  0.09343

S = 0.05460   R-Sq = 42.56%   R-Sq(adj) = 32.99%


Level        N    Mean   StDev
Development  7  1.0743  0.0565
IndustryA    8  0.9981  0.0464
IndustryB    5  1.0980  0.0669
IndustryC    2  1.0875  0.0389

             Individual 95% CIs For Mean Based on Pooled StDev
Level          +---------+---------+---------+---------
Development                (------*------)
IndustryA      (-----*------)
IndustryB                    (--------*--------)
IndustryC             (------------*-------------)
               +---------+---------+---------+---------
             0.960     1.020     1.080     1.140

Pooled StDev = 0.0546
```

PAH ratio ($p = .017$) indicates that there are some differences in the four PAH ratio means. To determine which sites have significantly different means, a follow-up analysis is required. This will involve ranking the means while controlling the overall Type I error rate.

Since the sample sizes associated with the four sites are not equal, and because we desire pair-wise comparisons of the means, the method with the highest power (i.e., the one with the greatest chance of detecting a difference when differences actually exist) is the Bonferroni multiple comparisons method. Also, this method explicitly controls the comparison-wise error rate (i.e., the overall Type I error rate).

For this problem there are four treatments (sites). Consequently, there are $c = p(p - 1)/2 = 4(3)/2 = 6$ comparisons of interest. Using the symbol $\mu_j$ to represent the population mean PAH ratio at site $j$, the 6 comparisons we desire are: $(\mu_A - \mu_B)$, $(\mu_A - \mu_C)$, $(\mu_A - \mu_D)$, $(\mu_B - \mu_C)$, $(\mu_B - \mu_C)$, and $(\mu_C - \mu_D)$. We used MINITAB to perform the multiple comparisons for the data saved in the **PAH** file. The results for the two dependent variables, both using an experimentwise error rate of .05, are shown in Figure SIA14.3a-b. Based on the confidence intervals for the differences in means, MINITAB determines which means are significantly different. Treatments with the same letter in the "Grouping" column are not significantly different.

For the first PAH ratio, all four sites have the same letter (see Figure SIA14.3a). Consequently, none of the four PAH ratio means differ significantly. Of course, this result is consistent with the ANOVA $F$ test conducted earlier. The results for the second PAH ratio are shown in Figure SIA14.3b. You can see that the development site, Industry B, and Industry C do not have significantly different means. Similarly, the development site, Industry C, and Industry A do not have significantly different means. The only two sites found to have significantly different mean PAH2 ratios are Industry A and Industry B (since they do not have the same "Grouping" letter). These inferences can be made with an overall 5% chance of a Type I error.

**FIGURE SIA14.3**
MINITAB Output for Multiple Comparisons of PAH Ratio Means

**a. Dependent Variable = PAH1**

```
Grouping Information Using Bonferroni Method and 95.0% Confidence for PAH1

INDUSTRY     N    Mean  Grouping
IndustryB    5  0.6850  A
IndustryC    2  0.6625  A
Development  7  0.6457  A
IndustryA    8  0.6063  A

Means that do not share a letter are significantly different.
```

**b. Dependent Variable = PAH2**

```
Grouping Information Using Bonferroni Method and 95.0% Confidence for PAH2

INDUSTRY     N    Mean  Grouping
IndustryB    5  1.0980  A
IndustryC    2  1.0875  A B
Development  7  1.0743  A B
IndustryA    8  0.9981    B

Means that do not share a letter are significantly different.
```

The expert statistician used these results to conclude that although the two industries in question, Industry A and Industry B, have PAH2 ratio means that are significantly different, neither mean is significantly different from either the housing development site mean or the Industry C mean. Consequently, based on the available data it is impossible to determine which industry (A or B) was most likely to have contaminated the development site. In fact, according to the statistician's court testimony, "the results provide clear evidence that these samples are simply too small to make a reliable determination about the sites' similarity or dissimilarity with respect to [PAH] diagnostic ratios." The statistician went on to conclude that "the small samples relied upon by [the biochemical expert] shed no light on the issue of whether [Industry A or Industry B] are similar or dissimilar to the [development] site . . . ."

*Concluding Note:* The trial judge ultimately decided that the biochemist's statistical analyses and his opinions based on them would be excluded from the evidence used to decide the case. As of this date, the issue of responsibility for the pollution has still not been decided.

# Quick Review

(*Note:* Items marked with an asterisk (*) are from the optional sections in this chapter.)

## Key Terms

Analysis of variance (ANOVA) 743
Anderson–Darling test 818
Balanced design 818
Bartlett's test of variances 818
Bonferroni multiple comparisons procedure 814
Comparisonwise error rate 811
Complete factorial experiment 000
Completely randomized design 860
Experimentwise error rate 811
Factor interaction 775
*Fixed effects 800
**k*-way classification 791
Kolmogorov–Smirnov test 818
Levene's test of variances 818
Mean square for error 751
Mean square for treatments 751
Multiple comparisons of means 811
*Nested sampling design 800
Normalizing transformation 818
*Primary unit 800
Randomized block design 759
*Random effects 804
Robust method 818
Shapiro–Wilk test 818
*Subsampling 800
Sum of squares for error 748
Sum of squares for interaction 748
Sum of squares for main effects 781
Sum of squares for treatment 745
*Three-stage nested design 807
Tukey-Kramer method 813
Tukey multiple comparisons procedure 813
*Two-stage nested design 803
Variance-stabilizing transformation 818

## Key Formulas

Completely randomized design:

$F = \dfrac{\text{MST}}{\text{MSE}}$ Testing treatments 753

Randomized block design:

$F = \dfrac{\text{MST}}{\text{MSE}}$ Testing treatments 766

$F = \dfrac{\text{MSB}}{\text{MSE}}$ Testing blocks 766

Factorial design with 2 factors:

$F = \dfrac{\text{MS}(A)}{\text{MSE}}$ Testing main effect $A$ 781

$F = \dfrac{\text{MS}(B)}{\text{MSE}}$ Testing main effect $B$ 781

$F = \dfrac{\text{MS}(AB)}{\text{MSE}}$ Testing $A \times B$ interaction 781

*Two-Stage Nested Design:

$F = \dfrac{\text{MS}(A)}{\text{MS}(B \text{ in } A)}$ Testing first-stage factor $A$ 803

*Three-Stage Nested Design:

$F = \text{MS}(A)/\text{MS}(B \text{ in } A)$ Testing first-stage factor $A$ 806

$F = \text{MS}(B \text{ in } A)/\text{MS}(C \text{ in } B)$ Testing second-stage factor $B$ 806

Tukey's multiple comparisons:

| | |
|---|---|
| $\omega = q_\alpha(p, \nu)\dfrac{s}{\sqrt{n_t}}$ | Critical difference for equal sample sizes per treatment, $n_t$ where $q_\alpha(p, \nu)$ is the standardized range value for $p$ means, with $\nu$ degrees of freedom and significance level $\alpha$ 811 |
| $\omega_{ij} = q_\alpha(p, \nu)\dfrac{s}{\sqrt{2}}\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}$ | Critical difference for unequal sample sizes $n_i$ and $n_j$ 813 |

Bonferroni multiple comparisons:

| | |
|---|---|
| $B_{ij} = (t_{\alpha^*/2})(s)\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}$ | Critical difference for sample sizes $n_i$ and $n_j$ 814 |
| | where $\alpha^* = \alpha/g =$ comparisonwise error rate<br>$\alpha =$ experimentwise error rate<br>$g = p(p-1)/2 =$ number of pairwise comparisons for $p$ means |

## LANGUAGE LAB

| Symbol | Description |
|---|---|
| ANOVA | Analysis of variance |
| SST | Sum of Squares for Treatments (i.e., the variation among treatment means) |
| SSE | Sum of Squares for Error (i.e., the variability around the treatment means due to sampling error) |
| MST | Mean Square for Treatments |
| MSE | Mean Square for Error (an estimate of $\sigma^2$) |
| SSB | Sum of Squares for Blocks |
| MSB | Mean Square for Blocks |
| $a \times b$ factorial | Two-factor factorial experiment with one factor at $a$ levels and the other at $b$ levels (thus, there are $a \times b$ treatments in the experiment) |
| SS($A$) | Sum of Squares for Factor $A$ |
| MS($A$) | Mean Square for Factor $A$ |
| SS($B$) | Sum of Squares for Factor $B$ |
| MS($B$) | Mean Square for Factor $B$ |
| SS($AB$) | Sum of Squares for $A \times B$ interaction |
| MS($AB$) | Mean Square for $A \times B$ interaction |
| *SS($B$ in $A$) | Sum of Squares for Factor $B$ nested within Factor $A$ |
| *MS($B$ in $A$) | Mean Square for Factor $B$ nested within Factor $A$ |

## Chapter Summary Notes

- A **balanced design** is one where the sample sizes for each treatment are equal.
- Conditions required for a valid **ANOVA *F* test in a completely randomized design**: (1) all $p$ treatment populations are approximately normal, (2) $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$.
- Conditions required for valid **ANOVA *F* tests in a randomized block design**: (1) all treatment–block populations are approximately normal, (2) all treatment–block populations have the same variance.
- Conditions required for valid **ANOVA *F* tests in a factorial design**: (1) all treatment populations are approximately normal, (2) all treatment populations have the same variance.

- ANOVA is a **robust method**—slight to moderate departures from normality do not have an impact on the validity of the results.
- The **experimentwise error rate** is the risk of making at least one Type I error when making multiple comparisons in an ANOVA.
- **Multiple comparisons methods** for controlling the experimentwise error rate: *Tukey* and *Bonferroni*.
- **Tukey's method** is appropriate when (1) the treatment sample sizes are equal and (2) pairwise comparisons of treatment means are desired.
- **Bonferroni's method** is appropriate when (1) the treatment sample sizes are equal or unequal and (2) pairwise comparisons of treatment means are desired.
- **Tests for main effects** in a factorial design are only appropriate if the **test for interaction is nonsignificant**.

## Supplementary Applied Exercises

(Note: Starred (*) exercises are from the optional sections in this chapter.)

14.71 *Safety of nuclear power plants.* An article in the *American Journal of Political Science* (Jan. 1998) examined the attitudes of three groups of professionals that influence U.S. policy. Random samples of 100 scientists, 100 journalists, and 100 government officials were asked about the safety of nuclear power plants. Responses were made on a 7-point scale, where 1 = very unsafe and 7 = very safe. The mean safety scores for the groups are: scientists, 4.1; journalists, 3.7; government officials, 4.2.

a. Identify the response variable for this study.
b. How many treatments are included in this study? Describe them.
c. Specify the null and alternative hypotheses that should be used to investigate whether there are differences in the attitudes of scientists, journalists, and government officials regarding the safety of nuclear power plants.
d. The MSE for the sample data is 2.355. At least how large must MST be in order to reject the null hypothesis of the test of part **a** using $\alpha = .05$?
e. If the MST = 11.280, what is the approximate $p$-value of the test of part **a**?

14.72 *Flexible work schedules.* Refer to the completely randomized design of Exercise 13.27 (p. 740). Recall that the researchers want to compare the mean job satisfaction rating of workers using three types of work scheduling: flextime, staggered starting hours, and fixed hours. Use the random number table (Table 1 in Appendix B) to randomly assign the workers to the three work schedules.

14.73 *Computerized speech recognition.* Speech recognition technology has advanced to the point that it is now possible to communicate with a computer through verbal commands. A study was conducted to evaluate the value of speech recognition in human interactions with computer systems (*Special Interest Group on Computer-Human Interaction Bulletin*, July 1993). A sample of 45 subjects was randomly divided into three groups (15 subjects per group), and each subject was asked to perform tasks on a basic voice mail system. A different interface was employed in each group: (1) touch-tone, (2) human operator, or (3) simulated speech recognition. One of the variables measured was overall time (in seconds) to perform the assigned tasks. An analysis was conducted to compare the mean overall performance times of the three groups.

a. Identify the experimental design employed in this study.
b. Propose a regression model that will allow you to compare the three means.
c. In terms of means, give the appropriate null hypothesis to be tested.
d. In terms of the $\beta$'s of the model, part **b**, give the appropriate null hypothesis to be tested.
e. The sample mean performance times for the three groups are given below. Despite differences among the sample means, the null hypothesis of part **c** could not be rejected at $\alpha = .05$. Explain how this is possible.

| Group | Mean Performance Time (seconds) |
|---|---|
| Touch-tone | 1,400 |
| Human operator | 1,030 |
| Speech recognition | 1,040 |

14.74 *Hazardous organic solvents.* The *Journal of Hazardous Materials* (July 1995) published the results of a study of the chemical properties of three different types of hazardous organic solvents used to clean metal parts: aromatics, chloroalkanes, and esters. One variable studied was sorption rate, measured as mole percentage. Independent samples of solvents from each type were tested and their sorption rates were recorded, as shown in the next table.

a. Construct an ANOVA table for the data.
b. Is there evidence of differences among the mean sorption rates of the three organic solvent types? Test using $\alpha = .10$.

## Data for Exercise 14.74

**SORPRATE**

| Aromatics | | Chloroalkanes | | Esters | | |
|---|---|---|---|---|---|---|
| 1.06 | .95 | 1.58 | 1.12 | .29 | .43 | .06 |
| .79 | .65 | 1.45 | .91 | .06 | .51 | .09 |
| .82 | 1.15 | .57 | .83 | .44 | .10 | .17 |
| .89 | 1.12 | 1.16 | .43 | .61 | .34 | .60 |
| 1.05 | | | | .55 | .53 | .17 |

*Source*: Reprinted from *Journal of Hazardous Materials*, Vol. 42, No. 2, J. D. Ortego *et al*., "A review of polymeric geosynthetics used in hazardous waste facilities." p. 142 (Table 9), July 1995, Elsevier Science-NL., Sara Burgerhartstraat 25, 1055 KV Amsterdam. The Netherlands.

14.75 *Bonding agent study.* An evaluation of diffusion bonding of zircaloy components is performed. The main objective is to determine which of three elements—nickel, iron, or copper—is the best bonding agent. A series of zircaloy components are bonded using each of the possible bonding agents. Since there is a great deal of variation in components machined from different ingots, a randomized block design is used, blocking on the ingots. A pair of components from each ingot are bonded together using each of the three agents, and the pressure (in units of 1,000 pounds per square inch) required to separate the bonded components is measured. The data are shown in the accompanying table, followed by a partial ANOVA summary table.

**INGOT2**

| | Bonding Agent | | |
|---|---|---|---|
| Ingot | Nickel | Iron | Copper |
| 1 | 67.0 | 71.9 | 72.2 |
| 2 | 67.5 | 68.8 | 66.4 |
| 3 | 76.0 | 82.6 | 74.5 |
| 4 | 72.7 | 78.1 | 67.3 |
| 5 | 73.1 | 74.2 | 73.2 |
| 6 | 65.8 | 70.8 | 68.7 |
| 7 | 75.6 | 84.9 | 69.0 |

| Source | df | SS | MS | *F* |
|---|---|---|---|---|
| Agent | 2 | 131.90 | — | — |
| Ingot | 6 | 268.29 | — | — |
| Error | 12 | 124.46 | — | |
| Total | 20 | 524.65 | | |

a. Identify the treatments in this experiment.
b. Identify the blocks in this experiment.
c. Is there evidence of a difference in pressure required to separate the components among the three bonding agents? Use $\alpha = .05$.

14.76 *Drift ratio study of buildings.* A commonly used index to estimate the reliability of a building subjected to lateral loads is the drift ratio. Sophisticated computer programs such as STAAD-III have been developed to estimate the drift ratio based on variables such as beam stiffness, column stiffness, story height, moment of inertia, and so on. Civil engineers at the State University of New York at Buffalo and the University of Central Florida performed an experiment to compare drift ratio estimates using STAAD-III with the estimates produced by a new, simpler microcomputer program called DRIFT (*Microcomputers in Civil Engineering*, 1993). Data for a 21-story building were used as input to the programs. Two runs were made with STAAD-III: Run 1 considered axial deformation of the building columns, and run 2 neglected this information. The goal of the analysis is to compare the mean drift ratios (where drift is measured as lateral displacement) estimated by the three computer runs (the two STAAD-III runs and DRIFT). The lateral displacements (in inches) estimated by the three programs are recorded in the next table for each of five building levels (1, 5, 10, 15, and 21). A MINITAB printout of the analysis of variance for the data is shown above.

## MINITAB Output for Exercise 14.76

**Two-way ANOVA: DRIFT versus PROGRAM, LEVEL**

```
Analysis of Variance for DRIFT
Source        DF        SS        MS        F        P
PROGRAM        2    0.4664    0.2332     4.79    0.043
LEVEL          4   52.1812   13.0453   267.74    0.000
Error          8    0.3898    0.0487
Total         14   53.0374
```

**STAAD**

| Level | STAAD-III(1) | STAAD-III(2) | Drift |
|---|---|---|---|
| 1 | .17 | .16 | .16 |
| 5 | 1.35 | 1.26 | 1.27 |
| 10 | 3.04 | 2.76 | 2.77 |
| 15 | 4.54 | 3.98 | 3.99 |
| 21 | 5.94 | 4.99 | 5.00 |

*Source*: Valles, R. E., et al. "Simplified drift evaluation of wall-frame structures," *Microcomputers in Civil Engineering*, Vol. 8, 1993, p. 000 (Table 2).

a. Identify the treatments in the experiment.
b. Because lateral displacement will vary greatly across building levels (floors), a randomized block design will be used to reduce the level-to-level variation in drift. Explain, diagrammatically, the setup of the design if all 21 levels are to be included in the study.
c. Using the information in the printout, compare the mean drift ratios estimated by the three programs.

14.77 *Acid rain study. Acid rain* is considered by some environmentalists to be the nation's most serious environmental problem. It is formed by the combination of water vapor in clouds with nitrogen oxide and sulfuric dioxide emissions from the burning of coal, oil, and natural gas. The acidity of rain in central and northern Florida consistently ranges

**ACIDRAIN**

| | | April 3 Acid Rain pH | | June 16 Acid Rain pH | | June 30 Acid Rain pH | |
|---|---|---|---|---|---|---|---|
| | | 3.7 | 4.5 | 3.7 | 4.5 | 3.7 | 4.5 |
| | *0–15 cm* | 5.33 | 5.33 | 5.47 | 5.47 | 5.20 | 5.13 |
| *Soil Depth* | *15–30 cm* | 5.27 | 5.03 | 5.50 | 5.53 | 5.33 | 5.20 |
| | *30–46 cm* | 5.37 | 5.40 | 5.80 | 5.60 | 5.33 | 5.17 |

*Source:* "Acid rain linked to growth of coal-fired power." *Florida Agricultural Research* 83, Vol. 2, No. 1, Winter 1983.

from 4.5 to 5 on the pH scale, a decidedly acid condition. To determine the effects of acid rain on the acidity of soils in a natural ecosystem, engineers at the University of Florida's Institute of Food and Agricultural Sciences irrigated experimental plots near Gainesville, Florida, with acid rain at two pH levels, 3.7 and 4.5. The acidity of the soil was then measured at three different depths, 0–15, 15–30, and 30–46 centimeters. Tests were conducted during three different time periods. The resulting soil pH values are shown in the table above. Treat the experiment as a $2 \times 3$ factorial laid out in three blocks, where the factors are acid rain at two pH levels and soil depth at three levels, and the blocks are the three time periods.

a. Is there evidence of an interaction between pH level of acid rain and soil depth? Test using $\alpha = .05$.
b. Conduct a test to determine whether blocking over time was effective in removing an extraneous source of variation. Use $\alpha = .05$.

14.78 *A new method of seedling production.* In *Ecological Engineering* (Feb. 2004), a new methodology for tree seedling production (called aqua-forest system) was compared to a conventional tree nursery method. The new method was applied to four plants grown in a clean-water creek (Treatment T1) and four plants grown in a polluted-water creek (Treatment T2), while the conventional method (Treatment T3) was applied to four plants raised in a tree nursery. Thus, the experimental design was completely randomized with three treatments and four replicates (plants) per treatment. One dependent variable of interest was the ratio of shoot weight to root weight. Tukey's multiple comparisons of the three treatment means yielded the following results at an experimentwise error rate of .05:

| | | | |
|---|---|---|---|
| *Mean Shoot/ Root Ratio:* | 1.50 | $\overline{2.31}$ | $\overline{3.29}$ |
| *Treatment:* | T1 | T2 | T3 |

a. How many pairwise comparisons are made in the Tukey analysis?
b. Which treatment(s) yielded the significantly highest mean shoot/root ratio? The lowest?

14.79 *"Wayfinding" experiment.* What is the optimal method of directing newcomers to a specific location in a complex building? Researchers at Ball State University (Indiana) investigated this "wayfinding" problem and reported their results in *Human Factors* (Mar. 1993). Subjects met in a starting room on a multilevel building and were asked to locate the "goal" room as quickly as possible. (Some of the subjects were provided directional aids, whereas others were not.) Upon reaching their destination, the subjects returned to the starting room and were given a second room to locate. (One of the goal rooms was located in the east end of the building, the other in the west end.) The experimentally controlled variables in the study were aid type at three levels (signs, map, no aid) and room order at two levels (east/west, west/east). Subjects were randomly assigned to each of the $3 \times 2 = 6$ experimental conditions; the travel time (in seconds) was recorded. The results of the analysis of the east room data for this $3 \times 2$ factorial design are provided in the accompanying table. Interpret the results.

| Source | df | MS | *F* | *p*-Value |
|---|---|---|---|---|
| Aid type | 2 | 511,323.06 | 76.67 | $<.0001$ |
| Room order | 1 | 13,005.08 | 1.95 | $>.10$ |
| Aid $\times$ Order | 2 | 8,573.13 | 1.29 | $>.10$ |
| Error | 46 | 6,668.94 | | |

*Source*: Butler, D. L., et al. "Wayfinding by newcomers in a complex building." *Human Factors*, Vol. 35, No. 1, Mar. 1993, p. 163 (Table 2).

14.80 *Steam explosion experiment.* The steam explosion of peat renders fermentable carbohydrates that have a number of potentially important industrial uses. A study of the steam explosion process was initiated to determine the optimum conditions for the release of fermentable carbohydrate (*Biotechnology and Bioengineering*, Feb. 1986). Triplicate samples of peat were treated for .5, 1.0, 2.0, 3.0, and 5.0 minutes at 170°, 200°, and 215°C, in the steam explosion process. Thus, the experiment consists of two factors—temperature at three levels and treatment time at five levels. The accompanying table gives the percentage of carbohydrate solubilized for each of the $3 \times 5 = 15$ peat samples.

**STEAM**

| Temperature °C | Time minutes | Carbohydrate Solubilized % |
|---|---|---|
| 170 | .5 | 1.3 |
| 170 | 1.0 | 1.8 |
| 170 | 2.0 | 3.2 |
| 170 | 3.0 | 4.9 |
| 170 | 5.0 | 11.7 |
| 200 | .5 | 9.2 |
| 200 | 1.0 | 17.3 |
| 200 | 2.0 | 18.1 |
| 200 | 3.0 | 18.1 |
| 200 | 5.0 | 18.8 |
| 215 | .5 | 12.4 |
| 215 | 1.0 | 20.4 |
| 215 | 2.0 | 17.3 |
| 215 | 3.0 | 16.0 |
| 215 | 5.0 | 15.3 |

*Source:* Forsberg, C. W., et al. "The release of fermentable carbohydrate from peat by steam explosion and its use in the microbial production of solvents." *Biotechnology and Bioengineering*, Vol. 28, No. 2, Feb. 1986, p. 179 (Table I). Copyright 1986.

a. What type of experimental design was employed?
b. Explain why the traditional analysis of variance formulas are inappropriate for the analysis of these data.
c. Write a second-order model relating mean amount of carbohydrate solubilized, $E(y)$, to temperature ($x_1$) and time ($x_2$).
d. Explain how you could test the hypothesis that the two factors, temperature ($x_1$) and time ($x_2$), interact.
e. If you have access to a statistical software package, fit the model and perform the test for interaction.

14.81 *Oil drill bit comparison.* As oil drilling costs rise at unprecedented rates, the task of measuring drilling performance becomes essential to a successful oil company. One method of lowering drilling costs is to increase drilling speed. Researchers at Cities Service Co. have developed a drill bit, called the PD-1, which they believe penetrates rock at a faster rate than any other bit on the market. It is decided to compare the speed of the PD-1 with the two fastest drill bits known, the IADC 1-2-6 and the IADC 5-1-7, at 12 drilling locations in Texas. Four drilling sites were randomly assigned to each bit, and the rate of penetration (RoP) in feet per hour (fph) was recorded after drilling 3,000 feet at each site. The data are given in the table. Can Cities Service Co. conclude that the mean RoP differs for at least two of the three drill bits? Test at $\alpha = .05$. If appropriate, rank the treatment means using a multiple comparisons procedure.

**DRILLBIT**

| PD-1 | IADC 1-2-6 | IADC 5-1-7 |
|---|---|---|
| 35.2 | 25.8 | 14.7 |
| 30.1 | 29.7 | 28.9 |
| 37.6 | 26.6 | 23.3 |
| 34.3 | 30.1 | 16.2 |

14.82 *Traits of collared lemmings.* Many temperate-zone animal species exhibit physiological and morphological changes when the hours of daylight begin to decrease during autumn months. A study was conducted to investigate the "short day" traits of collared lemmings (*The Journal of Experimental Zoology*, Sept. 1993). A total of 124 lemmings were bred in a colony maintained with a photoperiod of 22 hours of light per day. At weaning (19 days of age), the lemmings were weighed and randomly assigned to live under one of two photoperiods: 16 hours or less of light per day, more than 16 hours of light per day. (Each group was assigned the same number of males and females.) After 10 weeks, the lemmings were weighed again. The response variable of interest was the gain in body weight (measured in grams) over the 10-week experimental period. The researchers analyzed the data using an ANOVA for a $2 \times 2$ factorial design, where the two factors are photoperiod (at two levels) and gender (at two levels).

a. Construct an ANOVA table for the experiment, listing the sources of variation and associated degrees of freedom.
b. Give the models that will enable the researchers to test for photoperiod by gender interaction.
c. The $F$ test for interaction was not significant. Interpret this result practically.
d. The $p$-values for testing for photoperiod and gender main effects were both smaller than .001. Interpret these results practically.

14.83 *Removing water from paper.* The percentage of water removed from paper as it passes through a dryer depends on the temperature of the dryer and the speed of the paper passing through it. A laboratory experiment was conducted to investigate the relationship between dryer temperature $T$ at three levels and exposure time $E$ (which is related to speed). A $3 \times 3$ factorial experiment was conducted with temperatures at 100°, 120°, and 140°F and for exposure time $T$ at 10, 20, and 30 seconds. Four paper specimens were prepared for each condition. The data (percentages of water removed) are shown in the table below.

**PAPER2**

| | | Temperature ($T$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 100 | | 120 | | 140 | |
| | 10 | 24 | 26 | 33 | 33 | 45 | 49 |
| | | 21 | 25 | 36 | 32 | 44 | 45 |
| *Exposure Time (E)* | 20 | 39 | 34 | 51 | 50 | 67 | 64 |
| | | 37 | 40 | 47 | 52 | 68 | 65 |
| | 30 | 58 | 55 | 75 | 71 | 89 | 87 |
| | | 56 | 53 | 70 | 73 | 86 | 83 |

a. Perform an analysis of variance for the data and construct an analysis of variance table.

b. Do the data provide sufficient evidence to indicate that temperature and time interact? Test using $\alpha = .05$. What is the practical significance of this test?

c. Fit the second-order model

$$E(y) = \beta_0 + \beta_1 E + \beta_2 T + \beta_3 (E \times T) + \beta_4 E^2 + \beta_5 T^2$$

to the data. Give the prediction equation.

d. Estimate the mean percentage of water removed when $T = 120$ and $E = 20$. Why does this value differ from the sample mean of the four observations obtained for this factor–level combination?

e. Find and interpret the 95% confidence interval for the mean percentage of water removed when $T = 140$ and $E = 30$.

14.84 *Coal ash study.* The data shown in the table below are the results of an experiment conducted to investigate the effect of three factors on the percentage of ash in coal.

The three factors, each at four levels, were

*Type of coal* (factor $A$): Mojiri, Michel, Kairan, and Metallurgical coke

*Maximum particle size* (factor $B$): 246, 147, 74, and 48 microns

*Weight of selected coal specimen* (factor $C$): 1 gram, 100 milligrams, 20 milligrams, and 5 milligrams

Three specimens were prepared for each of the $4 \times 4 \times 4 = 64$ factor–level combinations, yielding three replications of a complete $4 \times 4 \times 4$ factorial experiment.

a. Set up an analysis of variance table showing the sources and degrees of freedom for each.

**COALASH**

| Sample Replication | | $A_1$ Mojiri $x_1$ | $x_2$ | $x_3$ | $A_2$ Michel $x_1$ | $x_2$ | $x_3$ | $A_3$ Kairan $x_1$ | $x_2$ | $x_3$ | $A_4$ Met. Coke $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $B_1$ | $C1$ | 7.30 | 7.35 | 7.42 | 10.69 | 10.58 | 10.72 | 12.20 | 12.27 | 12.23 | 9.99 | 10.02 | 9.95 |
| | $C2$ | 6.84 | 6.07 | 6.91 | 10.26 | 10.35 | 10.42 | 11.85 | 11.85 | 12.05 | 9.45 | 9.86 | 9.78 |
| | $C3$ | 7.05 | 6.49 | 7.24 | 10.61 | 10.08 | 10.31 | 12.34 | 11.74 | 11.44 | 9.76 | 9.79 | 9.77 |
| | $C4$ | 6.75 | 5.62 | 7.24 | 10.66 | 10.61 | 10.01 | 12.22 | 11.68 | 12.09 | 9.92 | 10.17 | 10.50 |
| $B_2$ | $C1$ | 7.56 | 7.44 | 7.51 | 10.86 | 10.88 | 10.90 | 12.47 | 12.42 | 12.44 | 9.87 | 9.81 | 9.79 |
| | $C2$ | 7.10 | 7.37 | 7.32 | 10.45 | 10.62 | 10.87 | 12.47 | 12.28 | 12.04 | 9.46 | 9.60 | 9.62 |
| | $C3$ | 7.41 | 7.60 | 7.49 | 10.85 | 10.89 | 10.61 | 12.33 | 12.35 | 12.40 | 9.97 | 9.77 | 9.76 |
| | $C4$ | 7.29 | 7.62 | 7.43 | 10.68 | 11.58 | 10.60 | 12.04 | 12.21 | 12.51 | 9.76 | 10.10 | 9.61 |
| $B_3$ | $C1$ | 7.51 | 7.64 | 7.58 | 10.30 | 10.68 | 10.73 | 12.42 | 12.41 | 12.39 | 9.97 | 10.02 | 10.01 |
| | $C2$ | 7.36 | 7.50 | 7.21 | 10.33 | 10.50 | 10.64 | 12.05 | 12.30 | 12.20 | 9.78 | 10.02 | 9.91 |
| | $C3$ | 7.56 | 7.55 | 7.47 | 10.73 | 10.75 | 10.84 | 12.44 | 12.30 | 12.26 | 9.88 | 9.90 | 10.06 |
| | $C4$ | 7.71 | 7.67 | 7.76 | 10.92 | 10.80 | 10.79 | 12.11 | 12.02 | 12.26 | 9.77 | 9.74 | 9.69 |
| $B_4$ | $C1$ | 7.45 | 7.49 | 7.47 | 10.85 | 10.89 | 10.85 | 12.23 | 12.30 | 12.17 | 10.06 | 10.07 | 10.11 |
| | $C2$ | 7.15 | 7.68 | 7.18 | 10.37 | 10.79 | 10.71 | 11.52 | 12.17 | 11.82 | 9.71 | 9.86 | 9.78 |
| | $C3$ | 7.60 | 7.55 | 6.61 | 10.82 | 10.82 | 10.88 | 12.40 | 11.99 | 12.17 | 10.13 | 9.93 | 10.01 |
| | $C4$ | 8.06 | 7.05 | 7.57 | 11.26 | 10.56 | 10.31 | 11.96 | 11.87 | 12.06 | 10.01 | 9.98 | 9.84 |

*Source*: Fujimori, T., and Ishikawa, K. "Sampling error on taking analysis-sample of coal after the last stage of a reduction process." *Reports of Statistical Application Research*, Union of Japanese Scientists and Engineers, Vol. 19, No. 4, 1972, pp. 22–32.

b. Do the data provide evidence of any interactions among the factors? Test using $\alpha = .05$.

c. Does the mean level of coal ash obtained in the analysis depend on the weight of the coal specimen? Test using $\alpha = .05$.

d. Find 95% confidence intervals for the difference in the mean ash content between Mojiri and Michel coal at each of the four levels of maximum particle size.

14.85 *Glass transition temperature.* A chemist has run an experiment to study the effect of four treatments on the glass transition temperature (in degrees Kelvin) of a particular polymer compound. Raw material used to make this polymer is bought in small batches. The material is thought to be fairly uniform within a batch but variable between batches. Therefore, each treatment was run on samples from each batch with the results shown in the table.

a. Do the data provide sufficient evidence to indicate a difference in mean temperature among the four treatments? Use $\alpha = .05$.

b. Is there sufficient evidence to indicate a difference in mean temperature among the three batches? Use $\alpha = .05$.

c. If the experiment were to be conducted again in the future, would you recommend any changes in the design of the experiment?

**POLYMER**

| | | Treatment | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | *1* | 576 | 584 | 562 | 543 |
| *Batch* | 2 | 515 | 563 | 522 | 536 |
| | *3* | 562 | 555 | 550 | 530 |

14.86 *Student use of the computer lab.* A computer lab at the University of Oklahoma is open 24 hours a day, 7 days a week. In *Production and Inventory Management Journal* (3rd Quarter, 1999), S. Barman investigated whether computer usage differed significantly (1) among the 7 days of the week and (2) among the 24 hours of the day. Using student log-on records, data on hourly student loads (number of users per hour) were collected during a 7-week period. A factorial ANOVA was employed with the results presented in the table.

| Source | df | SS | MS | $F$ | $p$-Value |
|---|---|---|---|---|---|
| Day | 6 | 18732.13 | 3122.02 | 68.39 | .0001 |
| Time | 23 | 164629.86 | 7157.82 | 156.80 | .0001 |
| Day × Time | 138 | 7685.22 | 55.69 | 1.22 | .0527 |

*Source*: Barman, S. "A statistical analysis of the attendance pattern of a computer laboratory." *Production and Inventory Management Journal*, 3rd Quarter, 1999, pp. 26–30.

a. Is this an observational or a designed experiment? Explain.

b. What are the two factors of the experiment and how many levels of each factor are used?

c. This is an $a \times b$ factorial experiment. What are $a$ and $b$?

d. Specify the null and alternative hypotheses that should be used to test for an interaction effect between the two factors of the study.

e. Conduct the test of part **d** using $\alpha = .05$. Interpret your result in the context of the problem.

f. If appropriate, conduct main effects tests for both day and time. Use $\alpha = .05$. Interpret your results in the context of the problem.

14.87 *Light output experiment.* A $2 \times 2 \times 2 \times 2 = 2^4$ factorial experiment was conducted to investigate the effect of four factors on the light output, $y$, of flashbulbs. Two observations were taken for each of the factorial treatments. The factors are: amount of foil contained in a bulb (100 and 120 milligrams); speed of sealing machine (1.2 and 1.3 revolutions per minute); shift (day or night); machine operator (A or B). The data for the two replications of the $2^4$ factorial experiment are shown below.

To simplify computations, let

$$x_1 = \frac{\text{Amount} - 110}{10} \qquad x_2 = \frac{\text{Speed} - 1.25}{.05}$$

so that $x_1$ and $x_2$ will take values $-1$ and $+1$. Also,

$$x_3 = \begin{cases} -1 & \text{if night shift} \\ 1 & \text{if day shift} \end{cases} \qquad x_4 = \begin{cases} -1 & \text{if operator B} \\ 1 & \text{if operator A} \end{cases}$$

**FLASHBULB**

| | | Amount of Foil | | | |
|---|---|---|---|---|---|
| | | 100 milligrams | | 120 milligrams | |
| | | Speed of Machine | | | |
| | | 1.2 rpm | 1.3 rpm | 1.2 rpm | 1.3 rpm |
| *Day* | *Operator B* | 6; 5 | 5; 4 | 16; 14 | 13; 14 |
| *Shift* | *Operator A* | 7; 5 | 6; 5 | 16; 17 | 16; 15 |
| *Night* | *Operator B* | 8; 6 | 7; 5 | 15; 14 | 17; 14 |
| *Shift* | *Operator A* | 5; 4 | 4; 3 | 15; 13 | 13; 14 |

a. Do the data provide sufficient evidence to indicate that any of the factors contribute information for the prediction of $y$? Give the results of a statistical test to support your answer.
b. Identify the factors that appear to affect the amount of light $y$ in the flashbulbs.
c. Give the complete factorial model for $y$. (*Hint*: For a factorial experiment with four factors, the complete model includes main effects for each factor, two-way cross-product terms, three-way cross-product terms, and four-way cross-product terms.)
d. How many degrees of freedom will be available for estimating $\sigma^2$?

14.88 *Speech recognition device.* Refer to the *Human Factors* (Apr. 1990) study of recognizer accuracy at three levels (90%, 95%, and 99%) and vocabulary size at three levels (75%, 87.5%, and 100%) on the performance of a computerized speech recognizer, Exercise 12.26 (p. 661). The data on task completion times (minutes) were subjected to an analysis of variance for a $3 \times 3$ factorial design. The $F$ test for accuracy $\times$ vocabulary interaction resulted in a $p$-value less than .0003.

a. Interpret the result of the test for interaction.
b. As a follow-up to the test for interaction, the mean task completion times for the three levels of accuracy were compared under each level of vocabulary. Do you agree with this method of analysis? Explain.
c. Refer to part **b**. Turkey's multiple comparison method was used to compare the three accuracy means within each level of vocabulary at an experimentwise error rate of $\alpha = .05$. The results are summarized here. Interpret these results.

| | Accuracy Level | | |
|---|---|---|---|
| Vocabulary Size | 99% | 95% | 90% |
| 75% | 15.49 | 19.29 | 22.19 |
| 87.5% | 12.77 | 14.31 | 16.48 |
| 100% | 8.67 | 9.68 | 11.88 |

*Source*: Casali, S. P., Williges, B. H., and Dryden, R. D. "Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance." *Human Factors*, Vol. 32, No. 2, April 1990, p. 190 (Figure 2).

14.89 *Strength of mortar used in steel pipe.* Aroni and Fletcher (1979) presented data on the compressive and tensile strength of mortar used to line steel water pipelines. They noted that mortar strength is expected to increase as the curing time of the mortar increases from 7 to 28 days. The compressive and tensile strength means and standard deviations, each based on the testing of samples of $n = 50$ specimens, are shown in the accompanying table.

| Compressive Strength Curing Time | | Tensile Strength Curing Time | |
|---|---|---|---|
| 7 days | 28 days | 7 days | 28 days |
| $\bar{y}_1 = 8{,}477$ | $\bar{y}_2 = 10{,}404$ | $\bar{y}_1 = 621$ | $\bar{y}_2 = 737$ |
| $s_1 = 820$ | $s_2 = 928$ | $s_1 = 48$ | $s_2 = 55$ |
| $n_1 = 50$ | $n_2 = 50$ | $n_1 = 50$ | $n_2 = 50$ |

*Source:* Aroni, S., and Fletcher, G. "Observations on mortar lining of steel pipelines." *Transportation Engineering Journal*, Nov. 1979.

a. Refer to the compressive strength data and regard the two curing times as treatments. Find the total for all $n = 100$ observations. Then find CM and calculate SST.
b. Find SSE.
c. Find SS(Total).
d. Construct an analysis of variance table for the results of parts **a–c**.
e. Suppose the researchers want to estimate the mean compressive strength of the mortar mix using a simple linear regression model to relate mean compressive strength $E(y)$ to curing time $x$ over the time interval from 7 to 28 days. Explain why the least-squares line will pass through the points $(7, \bar{y}_1)$ and $(28, \bar{y}_2)$.
f. Find the least-squares line.
g. Use the prediction equation and the value of SSE found in part **b** to find a 95% confidence interval for the mean compressive strength at $x = 20$ days.
h. Find $r^2$ and interpret its value.

14.90 *Sintering metal study.* An experiment was conducted to determine the effect of sintering time (two levels) on the compressive strength of two different metals. Five test specimens were sintered for each metal at each of the two sintering times. The data (in thousands of pounds per square inch) are shown in the accompanying table.

**SINTIME**

| | | Sintering Time | | | | | |
|---|---|---|---|---|---|---|---|
| | | 100 minutes | | | 200 minutes | | |
| | *1* | 17.1 | 16.5 | 14.9 | 19.4 | 18.9 | 20.1 |
| *Metal* | | 15.2 | 16.7 | | 17.2 | 20.7 | |
| | *2* | 12.3 | 13.8 | 10.8 | 15.6 | 17.2 | 16.7 |
| | | 11.6 | 12.1 | | 16.1 | 18.3 | |

a. Perform an analysis of variance for the data, and construct an analysis of variance table.
b. What is the practical significance of an interaction between sintering time and metal type?
c. Do the data provide sufficient evidence to indicate an interaction between sintering time and metal type? Test using $\alpha = .05$.

**ROACH**

14.91 *On the Trail of the Cockroach.* Is the navigational behavior of cockroaches scavenging for food random or linked to a chemical trail? In an attempt to answer this question, an entomologist designed an experiment to test a cockroach's ability to follow a trail of its fecal material (*Explore*, Research at the University of Florida, Fall 1998). A methanol extract from roach feces—called a pheromone—was used to create chemical trail on a strip of white chromatography paper at the bottom of a plastic container. German cockroaches were released into the container at the beginning of the trail, one at a time, and a video surveillance camera was used to monitor the roach's movements.

In addition to the trail containing the fecal extract (the treatment), a trail using methanol only was created. This second trail served as a "control" to compare back against the treated trail. Since the entomologist also wanted to determine if trail-following ability differed among cockroaches of different age, sex, and reproductive status, four roach groups were utilized in the experiment: adult males, adult females, gravid (pregnant) females, and nymphs (immatures). Twenty roaches of each type were randomly assigned to the treatment trail and 10 of each type were randomly assigned to the control trail. Thus, a total of 120 roaches were used in the experiment. The design is a factorial with two factors—Trail (extract or control) and Group (adult males, adult females, gravid females, or nymphs). The response (dependent) variable of interest was the average trail deviation (measured in "pixels," where 1 pixel equals approximately 2 centimeters). The data for each of the 120 cockroaches in the study are stored in the **ROACH** file.

The entomologist wants to determine whether cockroaches in different age–sex groups differ in their ability to follow either the extract trail or the control trail. In other words, how do the two factors, age–sex group and trail type, impact the mean trail deviation of cockroaches? Answer this question by conducting a two-way factorial analysis of variance on the data. Fully interpret the results.

CHAPTER 15

# Nonparametric Statistics

**OBJECTIVE**

*To present some statistical tests that require fewer, or less stringent, assumptions than the methods of Chapters 8 and 10–14*

**CONTENTS**



*STATISTICS IN ACTION*

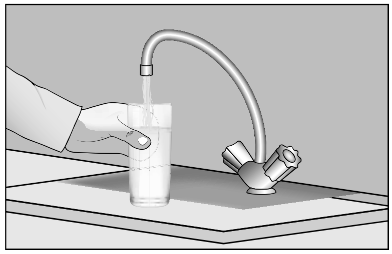
How Vulnerable Are New Hampshire Wells to Groundwater Contamination?

## *STATISTICS IN ACTION*

### How Vulnerable Are New Hampshire Wells to Groundwater Contamination?

Methyl tertbutyl ether (commonly known as MTBE) is a volatile, flammable, colorless liquid manufactured by the chemical reaction of methanol and isobutylene. MTBE was first produced in the United States as a lead fuel additive (octane booster) in 1979 and then as an oxygenate in reformulated fuel in the 1990s. Unfortunately, MTBE was introduced into water-supply aquifers by leaking underground storage tanks at gasoline stations, thus contaminating the drinking water. Consequently, by late 2006 most (but not all) American gasoline retailers had ceased using MTBE as an oxygenate, and accordingly, U.S. production has declined. Despite the reduction in production, there is no federal standard for MTBE in public water supplies; therefore, the chemical remains a dangerous pollutant, especially in states like New Hampshire that mandate the use of reformulated gasoline.

A study published in *Environmental Science & Technology* (Jan. 2005) investigated the risk of exposure to MTBE through drinking water in New Hampshire. In particular, the study reported on the factors related to MTBE contamination in public and private New Hampshire wells. Data were collected on a sample of 223 wells. These data are saved in the **MTBE** file (part of which you analyzed in Exercise 2.19). One of the variables measured was MTBE level (micrograms per liter) in the well water. An MTBE value exceeding .2 microgram per liter on the measuring instrument is a detectable level of MTBE. Of the 223 wells, 70 had detectable levels of MTBE. (Although the other wells are below the detection limit of the measuring device, the MTBE values for these wells are recorded as .2 rather than 0.) The other variables in the data set are described in Table S1A15.1.



How contaminated are these New Hampshire wells? Is the level of MTBE contamination different for the two classes of wells? For the two types of aquifers? What environmental factors are related to the MTBE level of a groundwater well? These are just a few of the research questions addressed in the study.

The researchers applied several nonparametric methods to the data in order to answer the research questions. We demonstrate the use of this methodology in *Statistics in Action Revisited* (p. 876).

***Data Set:*** **MTBE**

**TABLE SIA15.1 Variables Measured in the MTBE Contamination Study**

| Variable Name | Type | Description | Units of Measurement, or Levels |
|---|---|---|---|
| CLASS | QL | Class of well | Public or Private |
| AQUIFER | QL | Type of aquifer | Bedrock or Unconsolidated |
| DETECTION | QL | MTBE detection status | Below limit or Detect |
| MTBE | QN | MTBE level | micrograms per liter |
| PH | QN | pH level | standard pH unit |
| DISSOXY | QN | Dissolved oxygen | milligrams per liter |
| DEPTH | QN | Well depth | meters |
| DISTANCE | QN | Distance to underground storage tank | meters |
| INDUSTRY | QN | Industries in proximity | Percent of industrial land within 500 meters of well |

## 15.1 Introduction: Distribution-Free Tests

The confidence interval and testing procedures developed in earlier chapters all involve making inferences about population parameters. Consequently, they are often referred to as **parametric statistical tests**. Many of these parametric methods (e.g., the small-sample $T$ test of Chapter 8 or the ANOVA $F$ test of Chapter 14) rely on the assumption that the data are sampled from a normally distributed population. When the data are normal, these tests are *most powerful*. That is, the use of these parametric tests maximizes power—the probability of the researcher correctly rejecting the null hypothesis.

Consider a population of data that is decidedly nonnormal. For example, the distribution might be very flat, peaked, or strongly skewed to the right or left (see Figure 15.1). Applying the small-sample $T$ test to such a data set may result in serious consequences. Since the normality assumption is clearly violated, the results of the $T$ test are unreliable: (1) The probability of a Type I error (i.e., rejecting $H_0$ when it is true) may be larger than the value of $\alpha$ selected; and (2) the power of the test, $1 - \beta$, is not maximized.

A host of *nonparametric* techniques are available for analyzing data that do not follow a normal distribution. Nonparametric tests do not depend on the distribution of the sampled population; thus, they are called *distribution-free tests*. Also, nonparametric methods focus on the location of the probability distribution of the population, rather than on specific parameters of the population, such as the mean (hence, the name "nonparametrics").

**Definition 15.1**

**Distribution-free tests** are statistical tests that do not rely on any underlying assumptions about the probability distribution of the sampled population.

**Definition 15.2**

The branch of inferential statistics devoted to distribution-free tests is called **nonparametrics**.

Nonparametric tests are also appropriate when the data are nonnumerical in nature but can be ranked.* For example, when taste-testing foods or in other types of consumer product evaluations, we can say we like product A better than product B, and B better than C, but we cannot obtain exact quantitative values for the respective measurements. Nonparametric tests based on the ranks of measurements are called *rank tests*.

**Definition 15.3**

Nonparametric statistics (or tests) based on the ranks of measurements are called **rank statistics** (or **rank tests**).

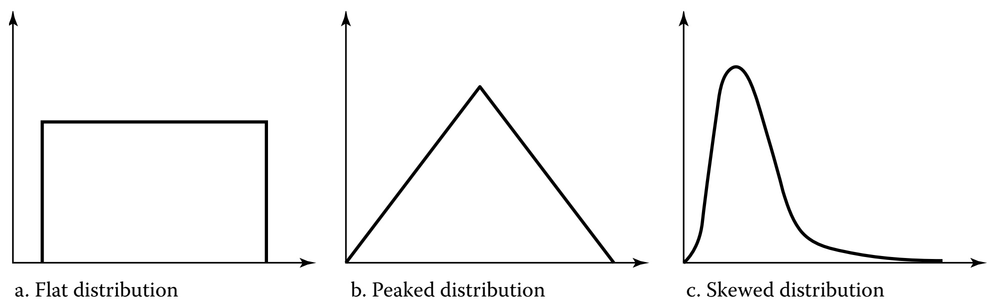



**FIGURE 15.1**
Some nonnormal distributions for which the $t$ statistic is invalid

*Qualitative data that can be ranked in order of magnitude are called *ordinal* data.

In this chapter, we present several useful nonparametric methods. Keep in mind that these nonparametric tests are more powerful than their corresponding parametric counterparts in those situations where either the data are nonnormal or the data are ranked.

In Section 15.2, we develop a test to make inferences about the central tendency of a single population. In Sections 15.3 and 15.5, we present rank statistics for comparing two or more probability distributions using independent samples. In Sections 15.4 and 15.6, the matched-pairs and randomized block designs are used to make nonparametric comparisons of populations. Finally, in Section 15.7, we present a nonparametric measure of correlation between two variables.

## 15.2 Testing for Location of a Single Population

Recall from Section 8.5 that small-sample procedures for testing a hypothesis about a population mean require that the population have an approximately normal distribution. For situations in which we collect a small sample from a decidedly nonnormal population (e.g., one of the populations shown in Figure 15.1), the $T$ test is not valid, and we must resort to a nonparametric procedure. The simplest nonparametric technique to apply in this situation is the **sign test**. The sign test is specifically designed for testing hypotheses about the median of any continuous population. Like the mean, the median is a measure of the center, or location, of the distribution; consequently, the sign test is sometimes referred to as a **test for location**.

Let $y_1, y_2, \ldots, y_n$ be a random sample from a population with unknown median $\tau$. Suppose we want to test the null hypothesis $H_0$: $\tau = 100$ against the one-sided alternative $H_a$: $\tau > 100$. From Definition 1.8 we know that the median is a number such that half the area under the probability distribution lies to the left of $\tau$ and half lies to the right (see Figure 15.2). Therefore, the probability that a $y$ value selected from the population is larger than $\tau$ is .5, i.e., $P(y_i > \tau) = .5$. If, in fact, the null hypothesis is true, then we should expect to observe approximately half the sample $y$ values greater than $\tau = 100$.

The sign test utilizes the test statistic $S$, where

$$S = \text{Number of } y_i\text{'s that exceed 100}$$

Notice that $S$ depends only on the *sign* (positive or negative) of the difference between each sample value $y_i$ and 100. That is, we are simply counting the number of positive (+) signs among the sample differences $(y_i - 100)$. If $S$ is "too large" (i.e., if we observe an unusually large number of $y_i$'s exceeding 100), then we will reject $H_0$ in favor of the alternative $H_a$: $\tau > 100$.

The rejection region for the sign test is derived as follows. Let each sample difference $(y_i - 100)$ denote the outcome of a single trial in an experiment consisting of $n$ identical trials. If we call a positive difference a "success" and a negative difference a "failure," then $S$ is the number of successes in $n$ trials. Under $H_0$, the probability of observing a success on any one trial is

$$p = P(\text{Success}) = P(y_i - 100 > 0) = P(y_i > 100) = .5$$

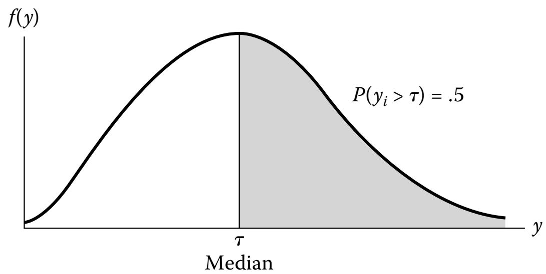



**FIGURE 15.2**
Location of the population median, $\tau$

Since the trials are independent, the properties of a binomial experiment, listed in Chapter 4, are satisfied. Therefore, $S$ has a binomial distribution with parameters $n$ and $p = .5$. We can use this fact to calculate the observed significance level ($p$-value) of the sign test, as illustrated in the following example.

**Example 15.1**

Application of the Sign Test

Bacteria are a most important component of microbial ecosystems in sewage treatment plants. Water management engineers have determined that the percentages of active bacteria in sewage specimens collected at a particular plant have a distribution with a median of 40 percent. If the median percentage is larger than 40, then adjustments in the sewage treatment process must be made. The percentages of active bacteria in a random sample of 10 sewage specimens are given in Table 15.1.

Do the data provide sufficient evidence to indicate that the median percentage of active bacteria in sewage specimens is greater than 40? Test using $\alpha = .05$.

**BACTERIA**

**TABLE 15.1 Active Bacteria Percentages**

| | | | | |
|---|---|---|---|---|
| 41 | 33 | 43 | 52 | 46 |
| 37 | 44 | 49 | 53 | 30 |

Solution

We want to test

$$H_0: \quad \tau = 40$$
$$H_a: \quad \tau > 40$$

using the sign test. The test statistic is

$$S = \text{Number of } y_i\text{'s in the sample that exceed 40} = 7$$

where $S$ has a binomial distribution with parameters $n = 10$ and $p = .5$.

From Definition 8.4, the observed significance level ($p$-value) of the test is the probability that we observe a value of the test statistic $S$ that is at least as contradictory to the null hypothesis as the computed value. For this one-sided case, the $p$-value is the probability that we observe a test statistic value greater than or equal to $S = 7$. We find this probability using statistical software or the cumulative binomial table for $n = 10$ and $p = .5$ in Table 2 of Appendix B. If $x$ has a binomial distribution with $n = 10$ and $p = .5$, then the $p$-value of the test is

$$p\text{-value} = P(x \geq S) = P(x \geq 7) = 1 - P(x \leq 6) = 1 - .828 = .172$$

This observed significance level is also shown (shaded) on the MINITAB printout of the analysis shown in Figure 15.3. Since the $p$-value, .1719, is larger than $\alpha = .05$, we cannot reject the null hypothesis. That is, there is insufficient evidence to indicate that the median percentage of active bacteria in sewage specimens exceeds 40.

**Sign Test for Median: ACTBAC**

```
Sign test of median =  40.00 versus > 40.00

         N  Below  Equal  Above       P  Median
ACTBAC  10      3      0      7  0.1719   43.50
```

**FIGURE 15.3**
MINITAB sign test for Example 15.1

A summary of the sign test for both one-sided and two-sided alternatives is provided in the next box.

For a two-tailed test, you may calculate the test statistic as either

$$S_1 = \text{Number of } y_i\text{'s greater than } \tau_0 = \text{Number of successes in } n \text{ trials}$$

or

$$S_2 = \text{Number of } y_i\text{'s less than } \tau_0 = \text{Number of failures in } n \text{ trials}$$

Note that $S_1 + S_2 = n$; therefore, $S_2 = n - S_1$. In either case, the $p$-value of the test is double the corresponding one-sided $p$-value. To simplify matters, we suggest using the larger of $S_1$ and $S_2$ as the test statistic and calculating the $p$-value as shown in the box.

### Sign Test for a Population Median, $\tau$

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\tau = \tau_0$ | $H_0$: $\tau = \tau_0$ |
| $H_a$: $\tau > \tau_0$ [or, $H_a$: $\tau < \tau_0$] | $H_a$: $\tau \neq \tau_0$ |
| *Test statistic*: | *Test statistic*: |
| $S =$ Number of sample observations greater than $\tau_0$ [or, $S =$ Number of sample observations less than $\tau_0$] | $S =$ larger of $S_1$ and $S_2$ where $S_1 =$ Number of sample observations greater than $\tau_0$; $S_2 =$ Number of sample observations less than $\tau_0$ |

[*Note:* Eliminate observations from the analysis that are exactly equal to the hypothesized median, $\tau_0$, and reduce the sample size accordingly.]

| *Observed significance level*: | *Observed significance level*: |
|---|---|
| $p$-value $= P(x \geq S)$ | $p$-value $= 2P(x \geq S)$ |

where $x$ has a binomial distribution with parameters $n$ and $p = .5$.

*Rejection region*: Reject $H_0$ if $\alpha > p$-value.

*Assumption*: The sample is randomly selected from a continuous probability distribution. (*Note*: No assumptions have to be made about the shape of the probability distribution.)

Recall from Section 6.10 that a normal distribution with mean $\mu = np$ and $\sigma = \sqrt{npq}$ can be used to approximate the binomial distribution for large $n$. When $p = .5$, the normal approximation performs reasonably well even for $n$ as small as 10 (see Example 6.23). Thus, for $n \geq 10$, we can conduct the sign test using the familiar standard normal $Z$ statistic of Chapter 8. This large-sample sign test is summarized in the box. Review Chapter 8 for example of how to apply this test.

### Sign Test Based on a Large Sample ($n \geq 10$)

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\tau = \tau_0$ | $H_0$: $\tau = \tau_0$ |
| $H_a$: $\tau > \tau_0$ [or, $H_a$: $\tau < \tau_0$] | $H_a$: $\tau \neq \tau_0$ |

*Test statistic*: $$Z_c = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{S - .5n}{\sqrt{(.5)(.5)n}} = \frac{S - .5n}{.5\sqrt{n}}$$

[*Note*: The value of $S$ is calculated as shown in the previous box.]

| *Rejection region*: $Z_c > z_\alpha$ | *Rejection region*: $Z_c > z_{\alpha/2}$ |
|---|---|
| $p$-value: $P(z > Z_c)$ | $p$-value: $2P(z > Z_c)$ |

where tabulated values of $z_\alpha$ and $z_{\alpha/2}$ are given in Table 5 of Appendix B, and $Z_c$ is the computed value of the test statistic.

## Applied Exercises

15.1 *Caffeine in Starbucks' coffee.* Scientists at the University of Florida College of Medicine investigated the level of caffeine in 16-ounce cups of Starbucks' coffee (*Journal of Analytical Toxicology*, Oct. 2003). In one phase of the experiment, cups of Starbucks Breakfast Blend (a mix of Latin American coffees) were purchased on six consecutive days from a single specialty coffee shop. The amount of caffeine in each of the 6 cups (measured in milligrams) is provided in the table.

**STARBUCKS**

| | | | | | |
|---|---|---|---|---|---|
| 564 | 498 | 259 | 303 | 300 | 307 |

a. Suppose the scientists are interested in determining whether the median amount of caffeine in Breakfast Blend coffee exceeds 300 milligrams. Set up the null and alternative hypotheses of interest.

b. How many of the cups in the sample have a caffeine content that exceeds 300 milligrams?

c. Assuming $p = .5$, find the probability that at least 4 of the 6 cups have caffeine amounts that exceed 300 milligrams.

d. Based on the probability, part **c**, what do you conclude about $H_0$ and $H_a$? (Use $\alpha = .05$.)

15.2 *Cheek teeth of extinct primates.* Refer to the *American Journal of Physical Anthropology* (Vol. 142, 2010) study of the characteristics of cheek teeth (e.g., molars) in an extinct primate species, Exercise 2.14 (p. 35). Recall that the researchers measured the dentary depth of molars (in millimeters) for 18 cheek teeth extracted from skulls. These depth measurements are reproduced in the accompanying table. The researchers are interested in the median molar depth of all cheek teeth from this extinct primate species. In particular, they want to know if the population median differs from 15 mm.

a. Specify the null and alternative hypothesis of interest to the researchers.

b. Explain why the sign test is appropriate to apply in this case.

**CHEEKTEETH**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 18.12 | 19.48 | 19.36 | 15.94 | 15.83 | 19.70 | 15.76 | 17.00 | 16.20 |
| 13.96 | 16.55 | 15.70 | 17.83 | 13.25 | 16.12 | 18.13 | 14.02 | 14.04 |

*Source:* Boyer, D.M., Evans, A.R., and Jernvall, J. "Evidence of Dietary Differentiation Among Late Paleocene-Early Eocene Plesiadapids (Mammalia, Primates)", *American Journal of Physical Anthropology*, Vol. 142, 2010. (Table A3.)

### MINITAB Output for Exercise 15.2

**Sign Test for Median: M2Depth**

```
Sign test of median =  15.00 versus not = 15.00

          N  Below  Equal  Above        P  Median
M2Depth  18      4      0     14   0.0309   16.16
```

c. A MINITAB printout of the analysis is shown below. Locate the test statistic on the printout.

d. Find the $p$-value on the printout, and use it to draw a conclusion. Test using $\alpha = .05$.

15.3 *Do social robots walk or roll?* Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study on the current trend in the design of social robots, Exercise 7.33 (p. 313). Recall that in a random sample of social robots obtained through a web search, 28 were built with wheels. The number of wheels on each of the 28 robots are reproduced in the accompanying table.

**ROBOTS**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 3 | 3 | 3 | 6 | 4 | 2 | 2 | 2 | 1 | 3 | 3 | 3 |
| 3 | 4 | 4 | 3 | 2 | 8 | 2 | 2 | 3 | 4 | 3 | 3 | 4 | 2 |

*Source:* Chew, S., et al. "Do social robots walk or roll?", *International Conference on Social Robotics*, Vol. 6414, 2010 (adapted from Figure 2).

a. Suppose you want to test whether the mean number of wheels exceeds 3. There is concern that the robot data do not follow a normal distribution. If so, how will this impact the analysis.?

b. Propose an alternative nonparametric test to analyze the data.

c. Compute the value of the test statistic for the nonparametric test.

d. Find the $p$-value of the test.

e. At $\alpha = .05$, what is the appropriate conclusion?

15.4 *Radioactive lichen.* Refer to the Lichen Radionuclide Baseline Research project to monitor the level of radioactivity in lichen, Exercise 2.15 (p. 36). Recall that University of Alaska researchers collected 9 lichen specimens and measured the amount of the radioactive element cesium-137 (in $\mu$Ci/ml) in each specimen. (The natural logarithms of the data values, are listed in the table on p. 844.) Suppose you want to test whether the median cesium amount in lichen differs from $\tau = .003\ \mu$Ci/ml. Use the accompanying MINITAB printout (p. 844) to conduct the nonparametric test at $\alpha = .10$.

**MINITAB Output for Exercise 15.4**

**Sign Test for Median: CESIUM**

```
Sign test of median =  0.00300 versus not = 0.00300

          N  Below  Equal  Above        P   Median
CESIUM    9      1      0      8   0.0391  0.00783
```

**LICHEN**

| Location | | | |
|---|---|---|---|
| Bethel | −5.50 | −5.00 | |
| Eagle Summit | −4.15 | −4.85 | |
| Moose Pass | −6.05 | | |
| Turnagain Pass | −5.00 | | |
| Wickersham Dome | −4.10 | −4.50 | −4.60 |

*Source:* Lichen Radionuclide Baseline Research project, 2003.

15.5 *Surface roughness of pipe.* Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated interior pipe used in oil fields, Exercise 8.24 (p. 390). The data (in micrometers) for 20 sampled pipe sections are reproduced in the table. Conduct a nonparametric test to determine whether the median surface roughness of coated interior pipe, $\tau$, differs from 2 micrometers. Test using $\alpha = .05$.

**ROUGHPIPE**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.72 | 2.50 | 2.16 | 2.13 | 1.06 | 2.24 | 2.31 | 2.03 | 1.09 | 1.40 |
| 2.57 | 2.64 | 1.26 | 2.05 | 1.19 | 2.13 | 1.27 | 1.51 | 2.41 | 1.95 |

*Source:* Farshad, F., and Pesacreta, T. "Coated pipe interior surface roughness as measured by three scanning probe instruments." *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III).

15.6 *Quality of white shrimp.* In *The American Statistician* (May 2001), the nonparametric sign test was used to analyze data on the quality of white shrimp. One measure of shrimp quality is cohesiveness. Since freshly caught shrimp are usually stored on ice, there is concern that cohesiveness will deteriorate after storage. For a sample of 20 newly caught white shrimp, cohesiveness was measured both before storage and after storage on ice for two weeks. The difference in the cohesiveness measurements (before minus after) was obtained for each shrimp. If storage has no effect on cohesiveness, the population median of the differences will be 0. If cohesiveness deteriorates after storage, the population median of the differences will be positive.

a. Set up the null and alternative hypotheses to test whether cohesiveness will deteriorate after storage.
b. In the sample of 20 shrimp, there were 13 positive differences. Use this value to find the *p*-value of the test.
c. Make the appropriate conclusion (in the words of the problem) if $\alpha = .05$.

15.7 *Ammonia in car exhaust.* Refer to the *Environmental Science & Technology* (Sept. 1, 2000) study of ammonia levels near the exit ramp of a San Francisco highway tunnel, Exercise 2.30 (p. 43). The daily ammonia concentrations (parts per million) on eight randomly selected days during afternoon drive-time are reproduced in the table. Suppose you want to determine if the median daily ammonia concentration for all afternoon drive-time days exceeds 1.5 ppm.

**AMMONIA**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.53 | 1.50 | 1.37 | 1.51 | 1.55 | 1.42 | 1.41 | 1.48 |

a. Set up the null and alternative hypotheses for the test.
b. Find the value of the test statistic.
c. Find the *p*-value of the test.
d. Give the appropriate conclusion (in the words of the problem) if $\alpha = .05$.

15.8 *Characteristics of a rock fall.* Refer to the *Environmental Geology* (Vol. 58, 2009) simulation study of how far a block from a collapsing rock wall will bounce down a soil slope, Exercise 2.29 (p. 43). Recall that the variable of interest was *rebound length* (measured in meters) of the falling block. Based on the depth, location, and angle of block-soil impact marks left on the slope from an actual rock fall, the following 13 rebound lengths (meters) were estimated. Consider the following statement: "In all similar rock falls, half of the rebound lengths will exceed 10 meters." Is this statement supported by the sample data? Test using $\alpha = .10$.

**ROCKFALL**

| | | | | | | |
|---|---|---|---|---|---|---|
| 10.94 | 13.71 | 11.38 | 7.26 | 17.83 | 11.92 | 11.87 |
| 5.44 | 13.35 | 4.90 | 5.85 | 5.10 | 6.77 | |

*Source:* Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy", *Environmental Geology*, Vol. 58, 2009. (Table 2.)

15.9 *Radon exposure in Egyptian tombs.* Refer to the *Radiation Protection Dosimetry* (December 2010) study of radon exposure in Egyptian tombs, Exercise 7.28 (p. 312). The radon levels — measured in becquerels per cubic meter (Bq/m$^3$) — in the inner chambers of a sample of 12 tombs are reproduced in the table. Recall that for safety purposes, the Egypt Tourism Authority (ETA) temporarily closes the tombs if the level of radon exposure in the tombs rises too high, say 6,000 Bq/m$^3$. Conduct a nonparametric test to determine if the true median level of radon exposure in the tombs is less than 6,000 Bq/m3. Use $\alpha = .10$. Should the tombs be closed?

**TOMBS**

| | | | | | |
|---|---|---|---|---|---|
| 50 | 910 | 180 | 580 | 7800 | 4000 |
| 390 | 12100 | 3400 | 1300 | 11900 | 1100 |

## Theoretical Exercises

15.10 Suppose we want to test $H_0$: $\tau = \tau_0$ against $H_a$: $\tau \neq \tau_0$ using the sign test, where

$$S_1 = \text{Number of sample observations greater than } \tau_0$$

and

$$S_2 = \text{Number of sample observations less than } \tau_0$$

Show that $P(S_1 \geq c) = P(S_2 \leq n - c)$, where $0 \leq c \leq n$.

15.11 Refer to the two-tailed test of Exercise 15.10. Use the result of that exercise to show that the observed significance level for the test is

$$p\text{-value} = 2P(S_1 \geq c)$$

## 15.3 Comparing Two Populations: Independent Random Samples

Suppose two independent random samples are to be used to compare two populations and the $T$ test of Chapter 8 is inappropriate for making the comparison. Either we are unwilling to make assumptions about the form of the underlying probability distributions, or we are unable to obtain exact values of the sample measurements. For either of these situations, if the data can be ordered, we could apply a test to compare the medians of the two populations. A more powerful nonparametric test, however, is one that compares entire probability distributions and not just the medians. This test, called the **Wilcoxon rank sum test**, tests the null hypothesis that the probability distributions associated with the two populations are equivalent against the alternative hypothesis that one population probability distribution is shifted to the right (or left) of the other.

For example, suppose that an experiment is conducted to compare the ratings of a technical writing software package by two groups of people—word-processing operators who must use the package and computer programming specialists who are trained to develop computer software. Independent random samples of $n_1 = 7$ word-processing operators and $n_2 = 7$ programming specialists were selected for the experiment. Each was asked to rate the package on a scale from 1 to 100, with 100 denoting the best rating. After the data were recorded, the 14 ratings were ranked in order of magnitude, 1 for the smallest and 14 for the largest. Tied observations (if they occur) are assigned ranks equal to the average of the ranks of the tied observations. For example, if the second and third ranked observations were tied, each would be assigned the rank 2.5. The data for the experiment and their ranks are shown in Table 15.2.

**TECHWRITE**

**TABLE 15.2 Ratings of a Technical Writing Software Package**

| *Word-Processing Operators* | | *Programming Specialists* | |
|---|---|---|---|
| Rating | Rank | Rating | Rank |
| 35 | 5 | 45 | 7 |
| 50 | 8 | 60 | 10 |
| 25 | 3 | 40 | 6 |
| 55 | 9 | 90 | 13 |
| 10 | 1 | 65 | 11 |
| 30 | 4 | 85 | 12 |
| 20 | 2 | 95 | 14 |
| $n_1 = 7$ | $T_1 = 32$ | $n_2 = 7$ | $T_2 = 73$ |

The Wilcoxon rank sum test is based on the sums of the ranks (called **rank sums**) for the two samples. The logic is that if the null hypothesis

$H_0$: The two population probability distributions are identical is true, then any one ranking of the $n = n_1 + n_2$ observations is just as likely as any other. Then, for equal sample sizes, we would expect the rank sums, $T_1$ and $T_2$, to be nearly equal.

In contrast, if the one-sided alternative hypothesis

$H_a$: Probability distribution for population 1 is shifted to the right of that for population 2

is true, then, for equal sample sizes, we would expect the rank sum $T_1$ to be larger than the rank sum $T_2$. In fact, it can be shown (proof omitted) that, regardless of the sample sizes $n_1$ and $n_2$,

$$T_1 + T_2 = \frac{n(n+1)}{2}$$

where $n = n_1 + n_2$. Therefore, as $T_2$ becomes smaller, $T_1$ will become larger and we would reject $H_0$ and accept $H_a$ for large values of $T_1$. A summary of the Wilcoxon rank sum test for independent random samples is shown in the box.

In Example 15.3, we will illustrate the procedure for finding the rejection region for a specified value of $\alpha$. First, we will use the rejection regions provided in Table 15 of Appendix B and a computer printout to compare the programmer and word-processor operator ratings of Table 15.2.

## Wilcoxon Rank Sum Test for a Shift in Population Locations: Independent Random Samples*

Let $D_1$ and $D_2$ represent the relative frequency distributions for populations 1 and 2, respectively.

*One-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted to the right of $D_2$
(or $H_a$: $D_1$ is shifted to the left of $D_2$)

*Two-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted either to the left or to the right of $D_2$

Rank the $n_1 + n_2$ observations in the two samples from the smallest (rank 1) to the largest (rank $n_1 + n_2$). Calculate $T_1$ and $T_2$, the rank sums associated with sample 1 and sample 2, respectively. Then calculate the test statistic.

*One-Tailed Test*

*Test statistic*:

$T_1$, if $n_1 < n_2$; $T_2$, if $n_2 < n_1$
(Either rank sum can be used if $n_1 = n_2$.)

*Rejection region*:

$T_1$: $T_1 \geq T_U$ [or $T_1 \leq T_L$ ]

$T_2$: $T_2 \leq T_L$ [or $T_2 \geq T_U$ ]

*Two-Tailed Test*

*Test statistic*:

$T_1$, if $n_1 < n_2$; $T_2$, if $n_2 < n_1$
(Either rank sum can be used if $n_1 = n_2$.) We will denote this rank sum as $T$.

*Rejection region*:

$T \leq T_L$ or $T \geq T_U$

where $T_L$ and $T_U$ are obtained from Table 15, Appendix B.

*Note:* Tied observations are assigned ranks equal to the average of the ranks that would have been assigned to the observations had they not been tied.

*Another statistic used for comparing two populations based on independent random samples is the *Mann–Whitney U statistic*. The $U$ statistic is a simple function of the rank sums. It can be shown that the Wilcoxon rank sum test and the Mann–Whitney $U$ test are equivalent.

## Example 15.2

Two-Tailed Wilcoxon Rank Sum Test

Refer to the data in Table 15.2.

a. Use Table 15 of Appendix B to test the null hypothesis that the probability distributions of the operator and programmer ratings are identical against the alternative hypothesis that one of the distributions is shifted to the right of the other. Test using $\alpha = .05$.

b. Find the $p$-value of the test and interpret the result.

**Solution**

a. We can use either rank sum as the test statistic for this two-tailed test and we will reject $H_0$ if that rank sum, say $T_1$, is very small or very large—that is, if $T_1 \leq T_L$ or $T_1 \geq T_U$. The tabulated values of $T_L$ and $T_U$, the lower- and upper-tail values of the rank sum distribution, are given in Table 15 of Appendix B. The critical values of the rank sum for a one-tailed test with $\alpha = .025$ and for a two-tailed test with $\alpha = .05$ are given in Table 15a, which is reproduced in Table 15.3. Table 15b of Appendix B gives the critical values, $T_L$ and $T_U$, for a one-tailed test with $\alpha = .05$ and for a two-tailed test with $\alpha = .10$. Examining Table 15.3, you will find that the critical values (shaded) corresponding to $n_1 = n_2 = 7$ are $T_L = 37$ and $T_U = 68$. Therefore, for $\alpha = .05$, we will reject $H_0$ if

$$T_1 \leq 37 \quad \text{or} \quad T_1 \geq 68$$

Since the observed value of the test statistic, $T_1 = 32$ (calculated in Table 15.2), is less than 37, we reject the hypothesis that the distributions of ratings are identical. There is sufficient evidence to indicate that one of the distributions is shifted to the right of the other.

b. An SPSS printout of the identical analysis is shown in Figure 15.4. Both rank sums are shaded on the printout, as well as the two-tailed observed significance level ($p$-value) for the Wilcoxon rank sum test. Since the $p$-value, .009, is less than $\alpha = .05$, we reach the same conclusion as in part a—namely, reject $H_0$ and conclude that the probability distributions have different locations.

## Example 15.3

One-Tailed Wilcoxon Rank Sum Test

Suppose that the alternative hypothesis in Example 15.2 had implied a one-tailed test. For example, suppose that we wanted to test $H_0$ against the alternative

$H_a$: Distribution 2 is shifted to the right of distribution 1

Locate the rejection region for the test using $\alpha = .025$.

**Solution**

We can use either $T_1$ or $T_2$ as the test statistic; small values of $T_1$ and large values of $T_2$ support the alternative hypothesis. If we use $T_1$ as the test statistic, we will reject $H_0$ if $T \leq T_L$ where $T_L$ is the lower-tail value of the rank sum, given in Table 15.3, for $n_1 = n_2 = 7$. This value is 37. Therefore, the rejection region for the one-tailed test with $\alpha = .025$ is $T_1 \leq 37$.

If we choose $T_2$ as the test statistic, we would reject $H_0$ if $T_2$ is large, say, $T_2 \geq T_U$. The value of $T_U$ given in Table 15.3 for $n_1 = n_2 = 7$ is $T_U = 68$. The two tests are equivalent.

**TABLE 15.3 A Partial Reproduction of Table 15 of Appendix B**

| $n_2$ \ $n_1$ | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ |
| 3 | 5 | 16 | 6 | 18 | 6 | 21 | 7 | 23 | 7 | 26 | 8 | 28 | 8 | 31 |
| 4 | 6 | 18 | 11 | 25 | 12 | 28 | 12 | 32 | 13 | 35 | 14 | 38 | 15 | 41 |
| 5 | 6 | 21 | 12 | 28 | 18 | 37 | 19 | 41 | 20 | 45 | 21 | 49 | 22 | 53 |
| 6 | 7 | 23 | 12 | 32 | 19 | 41 | 26 | 52 | 28 | 56 | 29 | 61 | 31 | 65 |
| 7 | 7 | 26 | 13 | 35 | 20 | 45 | 28 | 56 | 37 | 68 | 39 | 73 | 41 | 78 |
| 8 | 8 | 28 | 14 | 38 | 21 | 49 | 29 | 61 | 39 | 73 | 49 | 87 | 51 | 93 |
| 9 | 8 | 31 | 15 | 41 | 22 | 53 | 31 | 65 | 41 | 78 | 51 | 93 | 63 | 108 |
| 10 | 9 | 33 | 16 | 44 | 24 | 56 | 32 | 70 | 43 | 83 | 54 | 98 | 66 | 114 |

**FIGURE 15.4**
SPSS Wilcoxon rank sum test for Example 15.2

**Ranks**

| | USERGRP | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| RATING | WP | 7 | 4.57 | 32.00 |
| | PROG | 7 | 10.43 | 73.00 |
| | Total | 14 | | |

**Test Statistics**[b]

| | RATING |
|---|---|
| Mann-Whitney U | 4.000 |
| Wilcoxon W | 32.000 |
| Z | -2.619 |
| Asymp. Sig. (2-tailed) | .009 |
| Exact Sig. [2*(1-tailed Sig.)] | .007[a] |

a. Not corrected for ties.

b. Grouping Variable: USERGRP

## Example 15.4

**Finding a Value of the Rank Sum Test Statistic**

Consider a Wilcoxon rank sum test for $n_1 = n_2 = 4$. Find the value of $T_L$ such that $P(T_1 \leq T_L) \approx .05$. This value of $T_L$ would be appropriate for a one-tailed test with $\alpha = .05$.

Solution

To solve this problem, we use the probability methods of Chapter 3. If $H_0$ is true—i.e., if the two population probability distributions are identical—then any one ranking of the $n_1 + n_2 = 8$ observations is as likely as any other and each would represent a simple event for the experiment. For example, suppose that the four observations associated with samples 1 and 2 are denoted as $y_{11}, y_{12}, y_{13}, y_{14}$ and $y_{21}, y_{22}, y_{23}, y_{24}$, respectively. One ranking of the data that will produce the smallest possible value for $T_1$ is shown in Table 15.4.

To find the value $T_L$ such that $P(T_1 \leq T_L) \approx .05$, we find, $P(T_1 = 10)$, $P(T_1 = 11), \ldots$, and sum these probabilities until

$$P(T_1 = 10) + P(T_1 = 11) + \cdots + P(T_1 = T_L) \approx .05$$

**TABLE 15.4 One Ranking of the $n_1 + n_2 = 8$ Observations of Example 15.4**

| Sample 1 | | Sample 2 | |
|---|---|---|---|
| Observation | Rank | Observation | Rank |
| $y_{11}$ | 4 | $y_{21}$ | 6 |
| $y_{12}$ | 1 | $y_{22}$ | 5 |
| $y_{13}$ | 3 | $y_{23}$ | 7 |
| $y_{14}$ | 2 | $y_{24}$ | 8 |
| | $T_1 = 10$ | | $T_2 = 26$ |

The number of simple events in the sample space $S$ is equal to the number of ways that you can arrange the integers, 1, 2, . . . , 8—namely, 8!. Since the simple events are equiprobable, the probability of each simple event $E_i$ in the sample space is

$$P(E_i) = \frac{1}{8!}$$

The number of rankings that will result in $T_1 = 10$ is equal to the number of ways that you can arrange the four ranks for sample 1 and the four ranks for sample 2. The number of distinctly different arrangements of one sample of four ranks is 4!. Therefore, the number of ways that the two samples, each containing four ranks, can be arranged is

$$(4!)(4!)$$

Therefore, there will be (4!)(4!) simple events in the event $T_1 = 10$, each with probability $P(E_i) = 1/8!$. Then,

$$P(T_1 = 10) = \frac{4!4!}{8!} = \frac{1}{70} = .0143$$

Next, consider the rank sum $T_1 = 11$. The only way that $T_1$ can equal 11 is if the ranks assigned to sample 1 are 1, 2, 3, and 5. Then,

$$P(T_1 = 11) = \frac{4!4!}{8!} = \frac{1}{70} = .0143$$

and

$$P(T_1 \le 11) = P(T_1 = 10) + P(T_1 = 11) = 2(.0143) = .0286$$

Since this value is less than $\alpha = .05$, we will calculate the probability of observing the next larger rank sum for $T_1$—namely, $T_1 = 12$. We can obtain a rank sum $T_1 = 12$ if either the ranks 1, 2, 3, and 6 or the ranks 1, 2, 4, and 5 are assigned to sample 1. The probability of each of these occurrences is 1/70. Therefore,

$$P(T_1 = 12) = P\{1, 2, 3, 6\} + P\{1, 2, 4, 5\} = \frac{1}{70} + \frac{1}{70} = .0286$$

and

$$P(T_1 \le 12) = P(T_1 = 10) + P(T_1 = 11) + P(T_1 = 12) = .0572$$

Since we want $T_L$ to be the value such that $P(T_1 \le T_L)$ is close to $\alpha = .05$, it follows that $T_L = 12$. This is the tabulated value for $T_L$ given in Table 15b of Appendix B $(\alpha = .05)$.

Like the sign test, the Wilcoxon rank sum test can be conducted using the familiar $Z$ test statistic of Section 8.5 when the samples are large. The following (which we state without proof) leads to a large-sample Wilcoxon rank sum test. It can be shown that the mean and variance of the rank sum $T_1$ are

$$E(T_1) = \frac{n_1 n_2 + n_1(n_1 + 1)}{2}$$

and

$$V(T_1) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Then, when $n_1$ and $n_2$ are large (say, $n_1 > 10$ and $n_2 > 10$), the sampling distribution of

$$Z = \frac{T_1 - E(T_1)}{\sqrt{V(T_1)}} = \frac{T_2 - \left[\frac{n_1 n_2 + n_1(n_1 + 1)}{2}\right]}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

will have, approximately, a standard normal distribution. The procedure is summarized in the box.

**The Wilcoxon Rank Sum Test for Large Samples ($n_1 \geq 10$ and $n_2 \geq 10$)**

Let $D_1$ and $D_2$ represent the relative frequency distributions for populations 1 and 2, respectively.

*One-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_2$ is shifted to the right of $D_2$ (or $H_a$: $D_1$ is shifted to the left of $D_2$)

*Two-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted either to the left or to the right of $D_2$

$$\text{Test statistic:} \quad Z_c = \frac{T_1 - \left[\frac{n_1 n_2 + n_1(n_1 + 1)}{2}\right]}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

*Rejection region:* $Z_c > z_\alpha$ (or $Z_c < -z_a$)

$p$-value: $P(z > Z_c)$ [or $P(z < Z_c)$]

*Rejection region:* $|Z_c| > z_{\alpha/2}$

$p$-value: $2P(z > |Z_c|)$

(*Note:* The sample sizes $n_1$ and $n_2$ must both be at least 10.)

## Applied Exercises

15.12 *Use of text messaging in class.* Is text messaging the preferred option for students' communication with their professor? This was the question of interest in a study published in *Chemical Engineering Education* (Spring 2012). Students in two sections of a chemical engineering class participated in the study. One section (18 students) allowed text messaging in addition to the traditional means of communication with the instructor, such as email, phone, and face-to-face meetings; the other section (20 students) did not permit any text messaging. Both sections were taught by the same instructor. At the end of the semester, students responded to the survey item, "I like to interact with my professors by face-to-face meetings". Possible responses were recorded on a 5-point scale, from 1 = "strongly disagree" to 5 = "strongly agree". The median response for the students in the texting section was 5, while the median response for the students in the non-texting section was 4. The two groups of students were compared using a Wilcoxon rank sum test.

a. Set up the null hypothesis for the test.
b. The observed significance level of the test was reported as $p$-value = .004. Interpret this result, practically. Which group of students has more of a preference for face-to-face meetings with their professor?

15.13 *Bursting strength of bottles.* Polyethylene terephthalate (PET) bottles are used for carbonated beverages. A critical property of PET bottles is their bursting strength (i.e., the pressure at which bottles filled with water burst when pressurized). In the *Journal of Data Science* (May 2003), researchers measured the bursting strength of PET bottles made from two different designs—an old design and a new design. The data (pounds per square inch) for 10 bottles of each design are shown in the table. Suppose you want to compare the distributions of bursting strengths for the two designs.

**PET**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Old Design* | 210 | 212 | 211 | 211 | 190 | 213 | 212 | 211 | 164 | 209 |
| *New Design* | 216 | 217 | 162 | 137 | 219 | 216 | 179 | 153 | 152 | 217 |

a. Rank all 20 observed pressures from smallest to largest, and assign ranks from 1 to 20.
b. Sum the ranks of the observations from the old design.

**MINITAB Output for Exercise 15.14**

```
Mann-Whitney Test and CI: TRAD-HR, AERO-HR

          N  Median
TRAD-HR  39   11183
AERO-HR   7   12414


Point estimate for ETA1-ETA2 is -1125
95.3 Percent CI for ETA1-ETA2 is (-2358,1448)
W = 885.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.3431
The test is significant at 0.3431 (adjusted for ties)
```

c. Sum the ranks of the observations from the new design.
d. Compute the Wilcoxon rank sum statistic
e. Carry out a nonparametric test (at $\alpha = .05$) to compare the distribution of bursting strengths for the two designs.

**GASTURBINE**

15.14 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of gas turbines augmented with high-pressure inlet fogging, Exercise 8.29 (p. 392). The data on engine heat rate (kilojoules per kilowatt per hour) is saved in the **GASTURBINE** file. Recall that the researchers classified gas turbines into three categories: traditional, advanced, and aeroderivative. Suppose you want to compare the heat rate distributions for traditional and aeroderivative turbine engines.

a. Demonstrate that the assumptions required to compare the mean heat rates using a $t$ test are likely to be violated.
b. A MINITAB printout of the nonparametric test to compare the two heat rate distributions is shown above. Interpret the $p$-value of the test shown at the bottom of the printout.

15.15 *Cracking torsion moments of T-beams.* An experiment was conducted to study the effect of reinforced flanges on the torsional capacity of reinforced concrete T-beams (*Journal of the American Concrete Institute*, Jan.–Feb. 1986). Several different types of T-beams were used in the experiment, each type having a different flange width. The beams were tested under combined torsion and bending until failure (i.e., cracking). One variable of interest is the cracking torsion moment at the top of the flange of the T-beam. Cracking torsion moments for eight beams with 70-cm slab widths and eight beams with 100-cm slab widths are recorded here. Is there evidence of a difference in the locations of the cracking torsion moment distributions for the two types of T-beams? Test using $\alpha = .10$.

**TBEAMS**

| | |
|---|---|
| *70-cm:* | 6.00, 7.20, 10.20, 13.20, 11.40, 13.60, 9.20, 11.20 |
| *100-cm:* | 6.80, 9.20, 8.80, 13.20, 11.20, 14.90, 10.20, 11.80 |

15.16 *Patent infringement case.* Refer to the *Chance* (Fall 2002) study of a patent infringement case brought against Intel Corp., Exercise 7.44 (p. 320). Recall that the case rested on whether a patent witness's signature was written on top of key text in a patent notebook or under the key text. Using an X-ray beam, zinc measurements were taken at several spots on the notebook page. The zinc measurements for three notebook locations—on a text line, on a witness line, and on the intersection of the witness and text line—are reproduced in the table.

**PATENT**

| | | | | | | |
|---|---|---|---|---|---|---|
| *Text Line* | .335 | .374 | .440 | | | |
| *Witness Line* | .210 | .262 | .188 | .329 | .439 | .397 |
| *Intersection* | .393 | .353 | .285 | .295 | .319 | |

a. Why might the Student's $T$ procedure you applied in Exercise 7.44 be inappropriate for analyzing this data?
b. Use a nonparametric test (at $\alpha = .05$) to compare the distribution of zinc measurements for the text line with the distribution for the intersection.
c. Use a nonparametric test (at $\alpha = .05$) to compare the distribution of zinc measurements for the witness line with the distribution for the intersection.
d. From the results, parts **b** and **c**, what can you infer about the mean zinc measurements at the three notebook locations?

15.17 *Guided bone regeneration.* A new method of guided bone regeneration was evaluated in the *Journal of Tissue Engineering* (Vol. 3, 2012). The method involves attaching a titanium plate and silicon membrane to the underlying bone using titanium screws. After 1 week the titanium plate is elevated, creating space between it and the silicon tissue and allowing the bone to grow. The study focused on effectiveness of this procedure at 2 months and 4 months after elevation of the titanium plate. The surgical method was applied to the cranial bones of 8 white, male rabbits of the same species and size. The rabbits were randomly divided into two groups (4 rabbits in each group). One group was euthanized after 2 months, the other after 4 months. The new bone formation (measured in millimeters) around the cranial bone was recorded for each rabbit. The data (simulated) are shown in the accompanying table. The researchers conducted a nonparametric analysis of the data in order to compare the new bone formation distributions of the two groups. Carry out the appropriate test at $\alpha = .10$.

**GBONE**

| | | | | |
|---|---|---|---|---|
| **Group 1** (*2 months*): | 104.1, | 34.0, | 62.5, | 73.8 |
| **Group 2** (*4 months*): | 96.7, | 53.6, | 64.4, | 69.7 |

15.18 *Crude oil biodegradation.* Refer to the *Journal of Petroleum Geology* (April 2010) study of the environmental factors associated with biodegradation in crude oil reservoirs, Exercise 2.18 (p. 37). Recall that 16 water specimens were randomly selected from various locations in a reservoir on the floor of a mine. Crude oil was detected in 6 of the specimens, but not in the other 10 specimens. The amount of dioxide (milligrams/liter) in each water specimen is provided in the accompanying table. Is there a tendency for crude oil to be present in water with lower levels of dioxide? Use the appropriate nonparametric test to answer the question.

**BIODEG**

| Dioxide Amount | Crude Oil Present |
|---|---|
| 3.3 | No |
| 0.5 | Yes |
| 1.3 | Yes |
| 0.4 | Yes |
| 0.1 | No |
| 4.0 | No |
| 0.3 | No |
| 0.2 | Yes |
| 2.4 | No |
| 2.4 | No |
| 1.4 | No |
| 0.5 | Yes |
| 0.2 | Yes |
| 4.0 | No |
| 4.0 | No |
| 4.0 | No |

*Source:* Permanyer, A., et al. "Crude oil biodegradation and environmental factors at the Riutort oil shale mine, SE Pyrenees", *Journal of Petroleum Geology*, Vol. 33, No. 2, April 2010 (Table 1).

15.19 *Mineral flotation in water study.* Refer to the *Minerals Engineering* (Vol. 46-47, 2013) study of the impact of calcium and gypsum on the flotation properties of silica in water, Exercise 2.23 (p. 38). Recall that solutions of deionized water were prepared both with and without calcium/gypsum, and the level of flotation of silica in the solution was measured using a variable called zeta potential (measured in millivolts, mV). The data are reproduced below. Does the addition of calcium/gypsum to the solution impact water quality (measured by zeta potential of silica)? Use a large-sample nonparametric test to answer the question.

**SILICA**

Without calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| –47.1 | –53.0 | –50.8 | –54.4 | –57.4 | –49.2 | –51.5 | –50.2 | –46.4 | –49.7 |
| –53.8 | –53.8 | –53.5 | –52.2 | –49.9 | –51.8 | –53.7 | –54.8 | –54.5 | –53.3 |
| –50.6 | –52.9 | –51.2 | –54.5 | –49.7 | –50.2 | –53.2 | –52.9 | –52.8 | –52.1 |
| –50.2 | –50.8 | –56.1 | –51.0 | –55.6 | –50.3 | –57.6 | –50.1 | –54.2 | –50.7 |
| –55.7 | –55.0 | –47.4 | –47.5 | –52.8 | –50.6 | –55.6 | –53.2 | –52.3 | –45.7 |

With calcium/gypsum

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| –9.2 | –11.6 | –10.6 | –8.0 | –10.9 | –10.0 | –11.0 | –10.7 | –13.1 | –11.5 |
| –11.3 | –9.9 | –11.8 | –12.6 | –8.9 | –13.1 | –10.7 | –12.1 | –11.2 | –10.9 |
| –9.1 | –12.1 | –6.8 | –11.5 | –10.4 | –11.5 | –12.1 | –11.3 | –10.7 | –12.4 |
| –11.5 | –11.0 | –7.1 | –12.4 | –11.4 | –9.9 | –8.6 | –13.6 | –10.1 | –11.3 |
| –13.0 | –11.9 | –8.6 | –11.3 | –13.0 | –12.2 | –11.3 | –10.5 | –8.8 | –13.4 |

## Theoretical Exercises

15.20 Use the formula for the sum of an arithmetic progression to show that

$$T_1 + T_2 = \frac{n(n + 1)}{2}$$

for the Wilcoxon rank sum test.

15.21 Show that for the special case where $n_1 = 2$ and $n_2 = 2$, the formula for the expected value of the Wilcoxon rank sum $T_2$ given in this section holds. (*Hint:* List the $(n_1 + n_2)! = 4!$ ways that the ranks can be assigned, and compute $T_2$ for each assignment. Then use the fact that the probability of any assignment is equally likely.)

15.22 Consider the Wilcoxon rank sum $T_1$ for the case where $n_1 = 3$ and $n_2 = 3$. Use the technique outlined in this section to find $T_L$ such that $P(T_1 \leq T_L) \approx .05$.

## 15.4 Comparing Two Populations: Matched-Pairs Design

Nonparametric techniques can also be used to compare two probability distributions when a matched-pairs design (Section 8.8) is used. Recall that a **matched-pairs design** is a randomized block design with $k = 2$ treatments. In this section, we will show how the **Wilcoxon signed ranks test** can be used to test the hypothesis that two population probability distributions are identical against the alternative hypothesis that one is shifted to the right (or left) of the other.

For example, for some paper products, the softness of the paper is an important consideration in determining consumer acceptance. One method of assessing softness is to have judges give softness ratings to samples of the products. Suppose each of 10 judges is given a sample of two products that a company wants to compare. Each judge rates the softness of each product on a scale from 1 to 10, with higher ratings implying a softer product. The results of the experiment are shown in Table 15.5.

Since this is a matched-pairs design, we analyze the differences between measurements within each pair. If almost all of the differences are positive (or negative), we have evidence to indicate that the population probability distributions differ in location—that is, one is shifted to the right or to the left of the other. The nonparametric approach requires us to calculate the ranks of the absolute values of the differences between the measurements (the ranks of the differences after removing any minus signs). Note that tied absolute differences are assigned the average of the ranks they would receive if they were unequal but successive measurements. After the absolute differences are ranked, the sum of the ranks of the positive differences, $T_+$, and the sum of the ranks of the negative differences, $T_-$, are computed.

To test the null hypothesis

$H_0$: The probability distributions of the ratings for products A and B are identical

against the alternative hypothesis

$H_a$: The probability distribution of the ratings for product A is shifted to the right or left of the probability distribution for the ratings for product B

we use the test statistic

$$T = \text{Smaller of the positive and negative rank sums } T_+ \text{ and } T_-$$

SOFTPAPER

**TABLE 15.5 Paper Softness Ratings**

| Judge | Product A | Product B | Difference (A − B) | Absolute Value of Difference | Rank of Absolute Value |
|---|---|---|---|---|---|
| 1 | 6 | 4 | 2 | 2 | 5 |
| 2 | 8 | 5 | 3 | 3 | 7.5 |
| 3 | 4 | 5 | −1 | 1 | 2 |
| 4 | 9 | 8 | 1 | 1 | 2 |
| 5 | 4 | 1 | 3 | 3 | 7.5 |
| 6 | 7 | 9 | −2 | 2 | 5 |
| 7 | 6 | 2 | 4 | 4 | 9 |
| 8 | 5 | 3 | 2 | 2 | 5 |
| 9 | 6 | 7 | −1 | 1 | 2 |
| 10 | 8 | 2 | 6 | 6 | 10 |

$T_+ =$ Sum of positive ranks $= 46$

$T_- =$ Sum of negative ranks $= 9$

### The Wilcoxon Signed Ranks Test: Matched Pairs

Let $D_1$ and $D_2$ represent the relative frequency distributions for populations 1 and 2, respectively.

*One-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted to the right of $D_2$ (or $H_a$: $D_1$ is shifted to the left of $D_2$)

*Two-Tailed Test*

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted either to the left or to the right of $D_2$

Calculate the difference within each of the $n$ matched pairs of observations. Then rank the absolute values of the $n$ differences from the smallest (rank 1) to the highest (rank $n$) and calculate the rank sum $T_-$ of the negative differences and the rank sum $T_+$ of the positive differences.

*One-Tailed Test*

*Test statistic*:

$T_-$, the rank sum of the negative differences (or $T_+$, the rank sum of the positive differences)

*Rejection region*:

$T_- \le T_0$ (or $T_+ \le T_0$)

*Two-Tailed Test*

*Test statistic*:

$T$, the smaller of $T_-$ or $T_+$

*Rejection region*:

$T \le T_0$

where $T_0$ is given in Table 16 of Appendix B

(*Note:* Differences equal to 0 are eliminated and the number $n$ of differences is reduced accordingly. Tied absolute differences receive ranks equal to the average of the ranks they would have received had they not been tied.)

The rejection region for the test includes the smallest values of $T$ and is located so that $P(T \le T_0) = \alpha$ for a one-tailed statistical test and $P(T \le T_0) = \alpha/2$ for a two-tailed test. Values of $T_0$ for $n = 5$ to $n = 50$ pairs are presented in Table 16 of Appendix B. The Wilcoxon signed ranks test is summarized in the box and demonstrated in Example 15.5.

**Example 15.5**

Wilcoxon Signed Ranks Application

Refer to the data shown in Table 15.5. Compare the judges' ratings of products 1 and 2, using a Wilcoxon signed ranks test. For $\alpha = .05$, test

$H_0$: The distributions of product ratings are identical for products 1 and 2 against the alternative hypothesis

$H_a$: The distribution of ratings for one of the products is shifted to the left (or right) of the other distribution—that is, one of the products is rated higher than the other

Solution

The test statistic for this two-tailed test is the smaller rank sum, namely, $T_- = 9$. The rejection region is $T \le T_0$, where values of $T_0$ are given in Table 16 of Appendix B. A portion of this table is reproduced in Table 15.6. Examining Table 15.6 in the column corresponding to a two-tailed test, the row corresponding to $\alpha = .05$, and the column for $n = 10$ pairs, we read $T_0 = 8$. Therefore, we will reject $H_0$ if $T$ is less than or equal to 8. Since the smaller rank sum, $T_- = 9$, is not less than or equal to 8, we cannot reject $H_0$. There is insufficient evidence to indicate a shift in the distributions of ratings for the two products.

The MINITAB printout of the analysis is shown in Figure 15.5. Note that MINITAB uses $T_+ = 46$ as the test statistic. The two-tailed $p$-value of the test (shaded) is .067. Since this value exceeds $\alpha = .05$, our results agree—do not reject $H_0$.

As is the case for the rank sum test for independent samples, the sampling distribution of the signed rank statistic can be approximated by a normal distribution when the number $n$ of paired observations is large (say, $n \ge 25$). The large-sample $Z$ test is summarized in the box.

**TABLE 15.6 A Partial Reproduction of Table 16 of Appendix B**

| One-Tailed | Two-Tailed | $n = 5$ | $n = 6$ | $n = 7$ | $n = 8$ | $n = 9$ | $n = 10$ |
|---|---|---|---|---|---|---|---|
| $\alpha = .05$ | $\alpha = .10$ | 1 | 2 | 4 | 6 | 8 | 11 |
| $\alpha = .025$ | $\alpha = .05$ | | 1 | 2 | 4 | 6 | 8 |
| $\alpha = .01$ | $\alpha = .02$ | | | 0 | 2 | 3 | 5 |
| $\alpha = .005$ | $\alpha = .01$ | | | | 0 | 2 | 3 |
| | | $n = 11$ | $n = 12$ | $n = 13$ | $n = 14$ | $n = 15$ | $n = 16$ |
| $\alpha = .05$ | $\alpha = .10$ | 14 | 17 | 21 | 26 | 30 | 36 |
| $\alpha = .025$ | $\alpha = .05$ | 11 | 14 | 17 | 21 | 25 | 30 |
| $\alpha = .01$ | $\alpha = .02$ | 7 | 10 | 13 | 16 | 20 | 24 |
| $\alpha = .005$ | $\alpha = .01$ | 5 | 7 | 10 | 13 | 16 | 19 |
| | | $n = 17$ | $n = 18$ | $n = 19$ | $n = 20$ | $n = 21$ | $n = 22$ |
| $\alpha = .05$ | $\alpha = .10$ | 41 | 47 | 54 | 60 | 68 | 75 |
| $\alpha = .025$ | $\alpha = .05$ | 35 | 40 | 46 | 52 | 59 | 66 |
| $\alpha = .01$ | $\alpha = .02$ | 28 | 33 | 38 | 43 | 49 | 56 |
| $\alpha = .005$ | $\alpha = .01$ | 23 | 28 | 32 | 37 | 43 | 49 |
| | | $n = 23$ | $n = 24$ | $n = 25$ | $n = 26$ | $n = 27$ | $n = 28$ |
| $\alpha = .05$ | $\alpha = .10$ | 83 | 92 | 101 | 110 | 120 | 130 |
| $\alpha = .025$ | $\alpha = .05$ | 73 | 81 | 90 | 98 | 107 | 117 |
| $\alpha = .01$ | $\alpha = .02$ | 62 | 69 | 77 | 85 | 93 | 102 |
| $\alpha = .005$ | $\alpha = .01$ | 55 | 61 | 68 | 76 | 84 | 92 |

**FIGURE 15.5**
MINITAB Wilcoxon signed ranks test for Example 15.5

**Wilcoxon Signed Rank Test: AminusB**

```
Test of median = 0.000000 versus median not = 0.000000

                 N
                for  Wilcoxon            Estimated
           N   Test  Statistic       P      Median
AminusB   10     10       46.0   0.067       2.000
```

### Wilcoxon Signed Ranks Test for Large Samples ($n \geq 25$)

Let $D_1$ and $D_2$ represent the probability distributions for populations 1 and 2, respectively.

One-Tailed Test

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted to the right of $D_2$
(or $H_a$: $D_1$ is shifted to the left of $D_2$)

Two-Tailed Test

$H_0$: $D_1$ and $D_2$ are identical

$H_a$: $D_1$ is shifted either to the left or to the right of $D_2$

$$\textit{Test statistic:} \quad Z_c = \frac{T_+ - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)]/24}}$$

*Rejection region:* $Z_c > z_\alpha$ (or $Z_c < -z_\alpha$)
$p$-vlaue: $P(z > Z_c)$ [or $P(z < Z_c)$]

*Rejection region:* $|Z_c| > z_{\alpha/2}$
$p$-value: $2P(z > |Z_c|)$

*Assumptions:* The sample size $n$ is greater than or equal to 25. Differences equal to 0 are eliminated and the number $n$ of differences is reduced accordingly. Tied absolute differences receive ranks equal to the average of the ranks they would have received had they not been tied.

The Wilcoxon signed ranks procedure can also be used to test the location of a *single* population. That is, the Wilcoxon signed ranks test can be used as an alternative to the sign test of Section 15.2. For example, suppose we want to test the following hypotheses about a population median:

$H_0$: $\tau = 100$
$H_a$: $\tau > 100$

To conduct the test we calculate the differences $(y_i - 100)$ for the sample. Recall that the sign test depends only on the number of positive differences in the sample. The signed ranks test, on the other hand, requires that we first rank the differences, then sum the ranks of the positive differences. Thus, the Wilcoxon signed ranks test for a single sample is conducted exactly as the signed ranks procedure for matched pairs, except that the differences are calculated by subtracting the hypothesized value of the median from each observation. We summarize the procedure in the next box.

### The Wilcoxon Signed Ranks Test for the Median, $\tau$, of a Single Population

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\tau = \tau_0$ | $H_0$: $\tau = \tau_0$ |
| $H_a$: $\tau > \tau_0$ [or, $H_a$: $\tau < \tau_0$] | $H_a$: $\tau \neq \tau_0$ |
| *Test statistic*: | *Test statistic*: |
| $T_-$, the negative rank sum [or, $T_+$, the positive rank sum] | $T$, the smaller of the positive and negative rank sums, $T_+$ and $T_-$ |

[*Note:* The sample differences are computed as $(y_i - \tau_0)$.]

| *Rejection region*: | *Rejection region*: |
|---|---|
| $T_- \leq T_0$ [or, $T_+ \leq T_0$] | $T \leq T_0$ |

where $T_0$ is found in Table 16 of Appendix B.

*Assumptions:*
1. A random sample of observations has been selected from the population.
2. The absolute differences $y_i - \tau_0$ can be ranked. [No assumptions must be made about the form of the population probability distribution.]
3. Differences equal to 0 are eliminated and $n$ is reduced accordingly. Tied differences are assigned ranks equal to the average of the ranks of the tied observations.

## Applied Exercises

15.23 *Twinned drill holes.* Refer to the *Exploration and Mining Geology* (Vol. 18, 2009) study of drill twinned holes, Exercise 8.47 (p. 405). Recall that the drilling of a new hole, or "twin", next to an earlier drill hole is a traditional method of verifying mineralization grades. The data in the table (p. 857) represent total amount of heavy minerals (THM) percentages for a sample of 15 twinned holes drilled at a diamond mine in Africa. In Exercise 8.47 you used a Student's $T$-test to check for a difference in the true THM means of all original holes and their twin holes drilled at the mine.

a. Explain why the results of the $t$-test may be invalid.
b. What is the appropriate nonparametric test to apply? State $H_0$ and $H_a$ for the test.
c. Compute the difference between the "1st hole" and "2nd hole" measurements for each drilling location.
d. Rank the differences, part **c**.
e. Compute the rank sums of the positive and negative differences.
f. Use the rank sums, part **e**, to conduct the nonparametric test at $\alpha = .05$. Can the geologists conclude that there is no evidence of a difference in the THM distributions of all original holes and their twin holes drilled at the mine?

**TWINHOLE**

| Location | 1st Hole | 2nd Hole |
|---|---|---|
| 1 | 5.5 | 5.7 |
| 2 | 11.0 | 11.2 |
| 3 | 5.9 | 6.0 |
| 4 | 8.2 | 5.6 |
| 5 | 10.0 | 9.3 |
| 6 | 7.9 | 7.0 |
| 7 | 10.1 | 8.4 |
| 8 | 7.4 | 9.0 |
| 9 | 7.0 | 6.0 |
| 10 | 9.2 | 8.1 |
| 11 | 8.3 | 10.0 |
| 12 | 8.6 | 8.1 |
| 13 | 10.5 | 10.4 |
| 14 | 5.5 | 7.0 |
| 15 | 10.0 | 11.2 |

15.24 *Impact of red light cameras on car crashes.* Refer to the June, 2007 Virginia Department of Transportation (VDOT) study of a newly adopted photo-red-light enforcement program, Exercise 7.56 (p. 328). Recall that the VDOT provided crash data both before and after installation of red light cameras at several intersections. The data (measured as the number of crashes caused by red light running per intersection per year) for 13 intersections in Fairfax County, VA are reproduced in the table. The VDOT wants to determine if the photo-red enforcement program is effective in reducing red-light-running crash incidents at intersections. Use the nonparametric Wilcoxon signed-rank test (and the MINITAB printout below) to analyze the data for the VDOT.

**REDLIGHT**

| Intersection | Before Camera | After Camera |
|---|---|---|
| 1 | 3.60 | 1.36 |
| 2 | 0.27 | 0 |
| 3 | 0.29 | 0 |
| 4 | 4.55 | 1.79 |
| 5 | 2.60 | 2.04 |
| 6 | 2.29 | 3.14 |
| 7 | 2.40 | 2.72 |
| 8 | 0.73 | 0.24 |
| 9 | 3.15 | 1.57 |
| 10 | 3.21 | 0.43 |
| 11 | 0.88 | 0.28 |
| 12 | 1.35 | 1.09 |
| 13 | 7.35 | 4.92 |

*Source:* Virginia Transportation Research Council, "Research Report: The Impact of Red Light Cameras (Photo-Red Enforcement) on Crashes in Virginia" June 2007.

**MINITAB Output for Exercise 15.24**

**Wilcoxon Signed Rank Test: Difference**

```
Test of median = 0.000000 versus median > 0.000000

                 N for   Wilcoxon             Estimated
             N   Test   Statistic       P        Median
Difference  13     13        79.0   0.011        0.9650
```

15.25 *NHTSA new car crash tests.* Refer to the National Highway Traffic Safety Administration (NHTSA) crash test data for new cars saved in the CRASH file. In Exercise 7.54 (p. 328) you compared the chest injury ratings of drivers and front-seat passengers using the Student's $T$ procedure for matched pairs. Suppose you want to make the comparison for only those cars that have a driver's star rating of five stars (the highest rating). The data for these 18 cars are listed in the table on p. 858. Now consider analyzing these data using the Wilcoxon signed ranks test.

a. State the null and alternative hypothesis.
b. Use a statistical software package to find the signed ranks test statistic.
c. Give the rejection region for the test using $\alpha = .01$.
d. State the conclusion in practical terms. Report the $p$-value of the test.

### Data for Exercise 15.25

**CRASH5**

| Car | Chest Injury Rating Driver | Chest Injury Rating Passenger | Car | Chest Injury Rating Driver | Chest Injury Rating Passenger |
|---|---|---|---|---|---|
| 1 | 42 | 35 | 10 | 36 | 37 |
| 2 | 42 | 35 | 11 | 36 | 37 |
| 3 | 34 | 45 | 12 | 43 | 58 |
| 4 | 34 | 45 | 13 | 40 | 42 |
| 5 | 45 | 45 | 14 | 43 | 58 |
| 6 | 40 | 42 | 15 | 37 | 41 |
| 7 | 42 | 46 | 16 | 37 | 41 |
| 8 | 43 | 58 | 17 | 44 | 57 |
| 9 | 45 | 43 | 18 | 42 | 42 |

15.26 *Testing electronic circuits.* Refer to the *IEICE Transactions on Information & Systems* (Jan. 2005) comparison of two methods of testing electronic circuits, Exercise 8.50 (p. 406). Recall that each of 11 circuits was tested using the standard compression/depression method and the new Huffman-based coding method and the compression ratio recorded. The data are reproduced in the accompanying table.

a. In Exercise 8.50, you tested the theory that the Huffman coding method will yield a smaller mean compression ratio than the standard method using a $t$ test. Perform the alternative nonparametric test, using $\alpha = .05$.

b. Do the conclusions of the two tests agree?

**CIRCUITS**

| Circuit | Standard Method | Huffman Coding Method | Circuit | Standard Method | Huffman Coding Method |
|---|---|---|---|---|---|
| 1 | .80 | .78 | 7 | .99 | .82 |
| 2 | .80 | .80 | 8 | .98 | .72 |
| 3 | .83 | .86 | 9 | .81 | .45 |
| 4 | .53 | .53 | 10 | .95 | .79 |
| 5 | .50 | .51 | 11 | .99 | .77 |
| 6 | .96 | .68 | | | |

*Source:* Ichihara, H., Shintani, M., and Inoue, T. "Huffman-based test response coding." *IEICE Transactions on Information & Systems*, Vol. E88-D, No. 1, Jan. 2005 (Table 3).

15.27 *Concrete pavement response to temperature.* Refer to *The International Journal of Pavement Engineering* (Sept. 2004) field study of concrete stress at a newly constructed highway, Exercise 8.51 (p. 406). The variable of interest was slab top transverse strain (i.e., change in length per unit length per unit time) at a distance of 1 meter from the longitudinal joint. The 5-hour changes (8:20 P.M. to 1:20 A.M.) in slab top transverse strain for six days are reproduced in the next table. Analyze the data using a nonparametric test. Is there a shift in the change in transverse strain distributions between field measurements and the 3D model? Test using $\alpha = .05$.

**SLABSTRAIN**

| Day | Change in Temperature (°C) | Change in Transverse Strain: Field Measurement | Change in Transverse Strain: 3D Model |
|---|---|---|---|
| Oct. 24 | −6.3 | −58 | −52 |
| Dec. 3 | 13.2 | 69 | 59 |
| Dec. 15 | 3.3 | 35 | 32 |
| Feb. 2 | −14.8 | −32 | −24 |
| Mar. 25 | 1.7 | −40 | −39 |
| May 24 | −.2 | −83 | −71 |

*Source*: Shoukry, S., William, G., and Riad, M. "Validation of 3DFE model of jointed concrete pavement response to temperature variations." *The International Journal of Pavement Engineering*, Vol. 5, No. 3, Sept. 2004 (Table IV).

15.28 *Modeling transport of gases.* Refer to the *AIChE Journal* (Jan. 2005) study of a new method for modeling multicomponent transport of gases, Exercise 8.53 (p. 407). Recall that 12 gas mixtures were prepared and the viscosity of each mixture ($10^{-5}$Pa · s) was measured both experimentally and with the new model. The results are reproduced in the table. Use a nonparametric method to test the chemical engineers' claim that there is "an excellent agreement between our new calculation and experiments."

**VISCOSITY**

| Mixture | Viscosity Measurements: Experimental | Viscosity Measurements: New Method |
|---|---|---|
| 1 | 2.740 | 2.736 |
| 2 | 2.569 | 2.575 |
| 3 | 2.411 | 2.432 |
| 4 | 2.504 | 2.512 |
| 5 | 3.237 | 3.233 |
| 6 | 3.044 | 3.050 |
| 7 | 2.886 | 2.910 |
| 8 | 2.957 | 2.965 |
| 9 | 3.790 | 3.792 |
| 10 | 3.574 | 3.582 |
| 11 | 3.415 | 3.439 |
| 12 | 3.470 | 3.476 |

*Source*: Kerkhof, P., and Geboers, M. "Toward a unified theory of isotropic molecular transport phenomena." *AIChE Journal*, Vol. 51, No. 1, January 2005 (Table 2).

15.29 *Sea turtles and beach nourishment.* According to the National Oceanic and Atmospheric Administration's Office of Protected Species, sea turtle nesting rates have declined in all parts of the southeastern United States over the past ten years. Environmentalists theorize that beach nourishment may improve the nesting rates of these turtles. (Beach nourishment involves replacing the sand on the beach in order to extend the high water line seaward.) A study was undertaken to investigate the effect of beach nourishment on sea turtle nesting rates in Florida. (Aubry Hershorin, unpublished doctoral dissertation, University of Florida, 2010.) For one part of the study, eight beach zones were sampled in Jacksonville, FL. Each beach zone was nourished by the Florida Fish and Wildlife Conservation Commission. Nesting densities (measured as nests per linear meter) were recorded both before and after nourishing at each of the eight beach zones. The data are listed in the following table. Conduct a Wilcoxon signed-ranks test to compare the sea turtle nesting densities before and after beach nourishing. Use $\alpha = .05$.

**NESTDEN**

| Beach Zone | Before Nourishing | After Nourishing |
|---|---|---|
| 401 | 0 | 0.003595 |
| 402 | 0.001456 | 0.007278 |
| 403 | 0 | 0.003297 |
| 404 | 0.002868 | 0.003824 |
| 405 | 0 | 0.002198 |
| 406 | 0 | 0.000898 |
| 407 | 0.000626 | 0 |
| 408 | 0 | 0 |

15.30 *Settlement of shallow foundations.* Refer to the *Environmental & Engineering Geoscience* (Nov., 2012) study of methods for predicting settlement of shallow foundations on cohesive soil, Exercise 8.48 (p. 405). Actual settlement values for a sample of 13 structures built on a shallow foundation were determined, and these values compared to settlement predictions made using a formula that accounts for dimension, rigidity, and embedment depth of the foundation. The data (in millimeters) are reproduced in the table at the top of the next column. Conduct a nonparametric test to determine if the distribution of predicted settlement values is shifted to the right or left of the distribution of actual settlement values. Test using $\alpha = .05$.

**SHALLOW**

| Structure | Actual | Predicted |
|---|---|---|
| 1 | 11 | 11 |
| 2 | 11 | 11 |
| 3 | 10 | 12 |
| 4 | 8 | 6 |
| 5 | 11 | 9 |
| 6 | 9 | 10 |
| 7 | 9 | 9 |
| 8 | 39 | 51 |
| 9 | 23 | 24 |
| 10 | 269 | 252 |
| 11 | 4 | 3 |
| 12 | 82 | 68 |
| 13 | 250 | 264 |

*Source:* Ozur, M. "Comparing Methods for Predicting Immediate Settlement of Shallow Foundations on Cohesive Soils Based on Hypothetical and Real Cases", *Environmental & Engineering Geoscience*, Vol. 18, No. 4, November 2012 (from Table 4).

## Theoretical Exercises

15.31 For the Wilcoxon signed ranks test, show that

$$T_+ + T_- = \frac{n(n+1)}{2}$$

where $n$ is the number of nonzero differences that are ranked.

15.32 For the special case $n = 2$, with no ties in the data (that is, no differences of zero), list the eight different ways in which the two absolute differences can be ranked. (*Note*: The number of arrangements, 8, results from the general formula $2^n \cdot n!$.)

15.33 For the special case described in Exercise 15.32, show that

$$E(T_+) = \frac{n(n+1)}{4}$$

(*Hint:* Find $T_+$ for each of the eight arrangements of the ranks listed in Exercise 15.31; use the fact that any particular arrangement will occur with probability $\frac{1}{8}$.)

## 15.5 Comparing Three or More Populations: Completely Randomized Design

In Section 14.3 we compare the means of $k$ populations based on data collected according to a completely randomized design. The analysis of variance $F$ test, used to test the null hypothesis of equality of means, is based on the assumption that the populations are normally distributed with common variance $\sigma^2$.

The **Kruskal–Wallis *H* test** is the nonparametric equivalent of the analysis of variance $F$ test. It tests the null hypothesis that all $k$ populations possess the same probability distribution against the alternative hypothesis that the distributions differ in location—that is, one or more of the distributions are shifted to the right or left of each other. The advantage of the Kruskal–Wallis $H$ test over the $F$ test is that we need make no assumptions about the nature of the sampled populations.

A completely randomized design specifies that we select independent random samples of $n_1, n_2, \ldots, n_k$ observations from the $k$ populations. To conduct the test, we first rank all $n = n_1 + n_2 + \cdots + n_k$ observations and compute the rank sums, $T_1, T_2, \ldots, T_k$, for the $k$ samples. The ranks of tied observations are averaged in the same manner as for the Wilcoxon rank sum test. Then, if $H_0$ is true, and if the sample sizes, $n_1, n_2, \ldots, n_k$, each equal 5 or more, then the test statistic

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

will have a sampling distribution that can be approximated by a chi-square distribution with $(k-1)$ degrees of freedom. Large values of $H$ imply rejection of $H_0$. Therefore, the rejection region for the test is $H > \chi^2_\alpha$, where $\chi^2_\alpha$ is the value that locates $\alpha$ in the upper tail of the chi-square distribution.

The test is summarized in the box and its use is illustrated in Example 15.6.

### Kruskal–Wallis *H* Test for Comparing *k* Population Probability Distributions: Completely Randomized Design

$H_0$: The $k$ population probability distributions are identical

$H_a$: At least two of the $k$ population probability distributions differ in location

*Test statistic*: $$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

where

$n_i$ = Number of measurements in sample $i$

$T_i$ = Rank sum for sample $i$, where the rank of each measurement is computed according to its relative magnitude in the totality of data for the $k$ samples

$n$ = Total sample size $= n_1 + n_2 + \cdots + n_k$

*Rejection region*: $H > \chi^2_\alpha$ with $(k-1)$ degrees of freedom

$p$-value: $P(\chi^2 > H_c)$

*Assumptions*:
1. The $k$ samples are random and independent.
2. There are 5 or more measurements in each sample.
3. The observations can be ranked.

(*Note:* No assumptions have to be made about the shape of the population probability distributions.)

**Example 15.6**

Kruskal-Wallis Test Application

Independent random samples of three different brands of magnetron tubes (the key components in microwave ovens) were subjected to stress testing, and the number of hours each operated without repair was recorded. Although these times do not represent typical lifelengths, they do indicate how well the tubes can withstand extreme stress. The data are shown in Table 15.7. Experience has shown that the distributions of lifelengths for manufactured products are often nonnormal, thus violating the assumptions required for the proper use of an analysis of variance $F$ test. Use the Kruskal–Wallis $H$ test to determine whether evidence exists to conclude that the brands of magnetron tubes tend to differ in length of life under stress. Test using $\alpha = .05$.

Solution

The first step in performing the Kruskal–Wallis $H$ test is to rank the $n = 15$ observations in the complete data set. The ranks and rank sums for the three samples are shown in Table 15.8.

**MAGTUBES**

**TABLE 15.7 Lengths of Life for Magnetron Tubes in Example 15.6**

| | *Brand* | |
|---|---|---|
| A | B | C |
| 36 | 49 | 71 |
| 48 | 33 | 31 |
| 5 | 60 | 140 |
| 67 | 2 | 59 |
| 53 | 55 | 42 |

**TABLE 15.8 Ranks and Rank Sums for Example 15.6**

| A | Rank | B | Rank | C | Rank |
|---|---|---|---|---|---|
| 36 | 5 | 49 | 8 | 71 | 14 |
| 48 | 7 | 33 | 4 | 31 | 3 |
| 5 | 2 | 60 | 12 | 140 | 15 |
| 67 | 13 | 2 | 1 | 59 | 11 |
| 53 | 9 | 55 | 10 | 42 | 6 |
| | $T_1 = 36$ | | $T_2 = 35$ | | $T_3 = 49$ |

We want to test the null hypothesis

$H_0$: The population probability distributions of length of life under stress are identical for the three brands of magnetron tubes

against the alternative hypothesis

$H_a$: At least two of the population probability distributions differ in location using the test statistic

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{(15)(16)} \left[ \frac{(36)^2}{5} + \frac{(35)^2}{5} + \frac{(49)^2}{5} \right] - 3(16) = 1.22$$

The rejection region for the $H$ test is $H > \chi^2_\alpha$, where $\chi^2_\alpha$ is based on $(k-2)$ degrees of freedom and the tabulated values of $\chi^2_\alpha$ are given in Table 8 of Appendix B. For $\alpha = .05$ and $(k-1) = 2$ degrees of freedom, $\chi^2_{.05} = 5.99147$. The rejection region for the test is $H > 5.99147$, as shown in Figure 15.6. Since the computed value of $H$, $H = 1.22$, is less than $\chi^2_{.05}$, we cannot reject $H_0$. There is insufficient evidence to indicate a difference in location among the distributions of lifelengths for the three brands of magnetron tubes.

A SAS printout of the analysis is shown in Figure 15.7. The test statistic is shaded on the printout, as is the $p$-value of the test. Note that the $p$-value (.5434) exceeds $\alpha = .05$, resulting in a conclusion of "do not reject $H_0$."

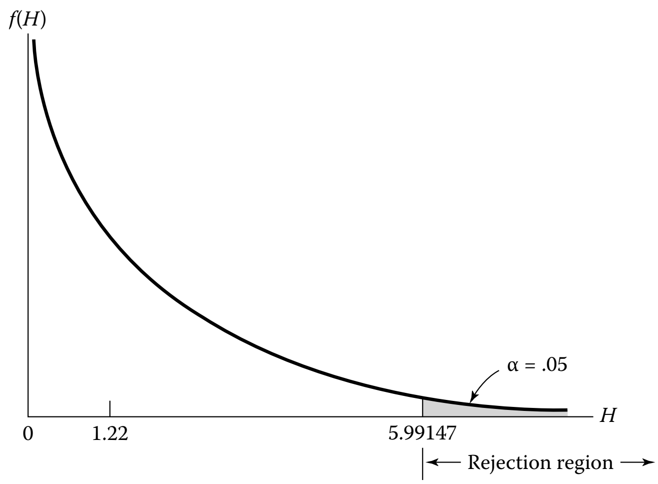



**FIGURE 15.6**
Rejection region for the comparison of three probability distributions

```
                         The NPAR1WAY Procedure

              Wilcoxon Scores (Rank Sums) for Variable LIFE
                      Classified by Variable BRAND

                       Sum of      Expected       Std Dev          Mean
BRAND        N         Scores      Under H0       Under H0         Score

A            5          36.0          40.0        8.164966          7.20
B            5          35.0          40.0        8.164966          7.00
C            5          49.0          40.0        8.164966          9.80

                          Kruskal-Wallis Test

                      Chi-Square          1.2200
                      DF                       2
                      Pr > Chi-Square     0.5434
```

**FIGURE 15.7**
SAS Kruskal–Wallis test for Example 15.6

## Applied Exercises

15.34 *Containing wildfires.* The *International Journal of Wildland Fire* (Dec. 2011) published a study of the time it takes to contain wildfires, both with and without aerial support. Containment time (in hours) was estimated by fire management personnel for a particular wildfire scenario that had no aerial support. Fire management personnel were classified according to one of three primary roles — ground, office, or air support. Data for 21 fire managers (simulated, based on information provided in the article) are listed in the table. One objective is to compare the estimated containment times of the three groups of fire managers.

**WILDFIRE**

| GROUND | OFFICE | AIR |
|---|---|---|
| 7.6 | 5.4 | 2.5 |
| 10.8 | 2.8 | 3.4 |
| 20.9 | 3.9 | 2.7 |
| 15.5 | 5.9 | 2.8 |
| 9.7 | 4.3 | 3.6 |
| 5.9 | 4.6 | |
| | 2.6 | |
| | 3.3 | |
| | 3.2 | |
| | 7.7 | |

a. Why is the ANOVA $F$ test of Chapter 9 inappropriate for analyzing this data? Use graphs to support your answer.
b. What is the appropriate nonparametric test to apply? State $H_0$ and $H_a$ for the test.
c. Rank the 21 estimated containment times, then find the rank sums for the three groups of fire managers.
d. Compute the nonparametric test statistic.
e. Find the rejection region for the test using $\alpha = .10$.
f. State the appropriate conclusion.

15.35 *Effect of scopolamine on memory.* Refer to the *Behavioral Neuroscience* (Feb. 2004) study of the drug scopolamine's effects on memory for word-pair associates, Exercise 14.13 (p. 757). Recall that a completely randomized design with three groups was used—group 1 subjects were injected with scopolamine, group 2 subjects were injected with a placebo, and group 3 subjects were not given any drug. The response variable was number of word pairs recalled. The data for all 28 subjects are reproduced in the table.

**SCOPOLAMINE**

| | |
|---|---|
| *Group 1 (scopolamine)*: | 5 8 8 6 6 6 6 8 6 4 5 6 |
| *Group 2 (placebo)*: | 8 10 12 10 9 7 9 10 |
| *Group 3 (no drug)*: | 8 9 11 12 11 10 12 12 |

a. Rank the data for all 28 observations from smallest to largest.
b. Sum the ranks of the observations from group 1.
c. Sum the ranks of the observations from group 2.
d. Sum the ranks of the observations from group 3.
e. Use the rank sums, parts **b–d**, to compute the Kruskal–Wallis $H$ statistic.
f. Carry out the Kruskal–Wallis nonparametric test (at $\alpha = .05$) to compare the distributions of number of word pairs recalled for the three groups.
g. Recall from Exercise 14.13 that the researchers theorized that group 1 subjects will tend to recall the fewest number of words. Use the Wilcoxon rank sum test to compare the word recall distributions of group 1 and group 2. (Use $\alpha = .05$.)

**GASTURBINE**

15.36 *Cooling method for gas turbines.* Refer to the *Journal of Engineering for Gas Turbines and Power* (Jan. 2005) study of gas turbines augmented with high-pressure inlet

**MINITAB Output for Exercise 15.36**

```
Kruskal-Wallis Test on HEATRATE

ENGINE         N  Median  Ave Rank      Z
Advanced      21    9669      15.0  -5.39
Aeroderiv      7   12414      44.1   1.46
Traditional   39   11183      42.4   4.17
Overall       67              34.0

H = 29.13  DF = 2  P = 0.000
H = 29.13  DF = 2  P = 0.000  (adjusted for ties)
```

fogging, Exercise 15.14 (p. 851). The Kruskal–Wallis test was used to compare the heat rate distributions for traditional, advanced, and aeroderivative gas turbine engines. A MINITAB printout of the analysis is shown above.

a. Give the null and alternative hypotheses tested.
b. Find the critical value of $X^2$ used in the rejection region of the test at $\alpha = .01$.
c. Locate the test statistic on the MINITAB printout and make the appropriate conclusion.
d. Interpret the $p$-value of the test shown at the bottom of the printout.

15.37 *Road safety of neighborhoods.* The *Canadian Journal of Civil Engineering* (Jan., 2013) published a study of the optimal layout (design) of streets in a neighborhood to maximize road safety. Five different community road patterns were compared: (1) traditional grid, (2) fused grid, (3) cul-de-sac, (4) Dutch sustainable road safety (SRS), and (5) 3-way offset. A sample of 30 neighborhoods were selected for each type of road pattern, resulting in a total sample size of 150. Road safety was measured as the number of collisions over a 3-year period. The researchers found that this variable follows a negative binomial distribution rather than a normal distribution; consequently, they analyzed the data using the nonparametric Kruskal-Wallis $H$ test. The rank sums for the 5 road patterns are provided here.

| | Traditional Grid | Fused Grid | Cul-de-sac | Dutch SRS | 3-way offset |
|---|---|---|---|---|---|
| Rank Sum: | 3398 | 2249.5 | 3144 | 1288.5 | 1245 |

a. Specify the null and alternative hypotheses for this nonparametric test.
b. The researchers reported the test statistic as $H = 71.53$. Verify this calculation.
c. State the appropriate conclusion at $\alpha = .05$.

15.38 *Soil scouring and overturned trees.* Refer to the *Landscape Ecology Engineering* (January 2013) investigation of the impact of soil scouring on the characteristics of overturned and uprooted trees, Exercise 14.12 (p. 756). Recall that a completely randomized design was employed with three treatments (scouring conditions): no scouring (NS), shallow scouring (SS), and deep scouring (DS). Five medium-sized trees were uprooted in each condition and the maximum resistive bending moment at the trunk base (kiloNewton-meters) was measured for each tree. The data are reproduced in the following table.

a. Verify that one or more of the assumptions of the ANOVA $F$ test conducted in Exercise 14.12 may be violated.
b. Conduct the appropriate nonparametric test at $\alpha = .05$. Does soil scouring impact the location of the distribution of maximum resistive bending moment at the tree trunk base?

**SCOURING**

| None | Shallow | Deep |
|---|---|---|
| 23.68 | 11.13 | 4.27 |
| 8.88 | 29.19 | 2.36 |
| 7.52 | 13.66 | 8.48 |
| 25.89 | 20.47 | 12.09 |
| 22.58 | 23.24 | 3.46 |

15.39 *Commercial eggs produced from different housing systems.* *Food Chemistry* (Vol. 106, 2008) published a study of the quality of commercial eggs produced from different housing systems for chickens. Four housing systems were investigated: (1) cage, (2) barn, (3) free range, and (4) organic. Twenty-eight commercial grade A eggs were randomly selected from supermarkets — 10 of which were produced in cages, 6 in barns, 6 with free range, and 6 organic. A number of quantitative characteristics were measured for each egg, including penetration strength (Newtons). The data (simulated from summary statistics provided in the journal article) are given in the accompanying table. Use a nonparametric test to make an inference about the strength distributions of the four housing systems.

**EGGS**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CAGE: | 36.9 | 39.2 | 40.2 | 33.0 | 39.0 | 36.6 | 37.5 | 38.1 | 37.8 | 34.9 |
| FREE: | 31.5 | 39.7 | 37.8 | 33.5 | 39.9 | 40.6 | | | | |
| BARN: | 40.0 | 37.6 | 39.6 | 40.3 | 38.3 | 40.2 | | | | |
| ORGANIC: | 34.5 | 36.8 | 32.6 | 38.5 | 40.2 | 33.2 | | | | |

**DDT**

15.40 *DDT contamination of fish.* Refer to the data on DDT levels of contaminated fish saved in the **DDT** file. Suppose you want to compare the DDT levels of the three species of fish: (1) channel catfish, (2) smallmouth buffalofish, and (3) largemouth bass.

a. Use a graphical method to determine whether a parametric or nonparametric procedure is appropriate for analyzing the data. Explain.
b. State the null and alternative hypothesis appropriate for analyzing the data nonparametrically.
c. Analyze the data using the appropriate nonparametric test. Interpret the results.

15.41 *Estimating the age of glacial drifts.* Refer to the *American Journal of Science* (Jan. 2005) study of the chemical makeup of buried tills (glacial drifts) in Wisconsin, Exercise 14.11 (p. 756). Recall that till specimens were obtained from five different boreholes (labeled UMRB-1, UMRB-2, UMRB-3, SWRA, and SD), and the ratio of aluminum to beryllium measured for each specimen. The data are reproduced in the table. Conduct a nonparametric analysis of variance of the data using $\alpha = .10$. Interpret the results.

**TILLRATIO**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *UMRB-1* | 3.75 | 4.05 | 3.81 | 3.23 | 3.13 | 3.30 | 3.21 |
| *UMRB-2* | 3.32 | 4.09 | 3.90 | 5.06 | 3.85 | 3.88 | |
| *UMBR-3* | 4.06 | 4.56 | 3.60 | 3.27 | 4.09 | 3.38 | 3.37 |
| *SWRA* | 2.73 | 2.95 | 2.25 | | | | |
| *SD* | 2.73 | 2.55 | 3.06 | | | | |

*Source:* Adapted from *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

15.42 *Reactivity of phosphate rock.* Phosphoric acid is chemically produced by reacting phosphate rock with sulfuric acid. An important consideration in the chemical process is the length of time required for the chemical reaction to reach a specified temperature. The shorter the length of time, the higher the reactivity of the phosphate rock. An experiment was conducted to compare the reactivity of phosphate rock mined in north, central, and south Florida. Rock samples were collected from each location and placed in vacuum bottles with a 56% strength sulfuric acid solution. The time (in seconds) for the chemical reaction to reach 200°F was recorded for each sample. Do the data (shown below) provide sufficient evidence to indicate a difference in the reactivity of phosphoric rock mined at the three locations? Test using $\alpha = .05$.

**PHOSPHATE**

| South | | Central | | North | | |
|---|---|---|---|---|---|---|
| 40.6 | 38.1 | 41.1 | 33.5 | 25.6 | 31.3 | 27.5 |
| 42.0 | 41.9 | 38.3 | 35.7 | 36.4 | 29.5 | |
| 37.5 | | 40.2 | | 28.2 | 22.8 | |

## Theoretical Exercise

15.43 Use the sum of an arithmetic progression to show that for the Kruskal–Wallis $H$ test,

$$T_1 + T_2 + \cdots + T_k = \frac{n(n+1)}{2}$$

where $k$ is the number of probability distributions being compared and $n$ is the total sample size.

# 15.6 Comparing Three or More Populations: Randomized Block Design

In this section, we present the nonparametric equivalent of the analysis of variance $F$ test for a randomized block design given in Section 14.5. The test, proposed by Milton Friedman (a Nobel prize winner in economics), is particularly appropriate for comparing the relative locations of $k$ or more population probability distributions when the normality and common variance assumptions required for an analysis of variance are not (or may not be) satisfied.

To conduct the $F_r$ test, we first rank the observations within each block and then compute the rank sums, $T_1, T_2, \ldots, T_k$, for the $k$ treatments. If $H_0$ is true—that is, if the population probability distributions are identical—and if the number $n$ of observations is large, then the $F_r$ statistic

$$F_r = \frac{12}{bk(b+1)} \sum_{i=1}^{k} T_i^2 - 3b(k+1)$$

will possess a sampling distribution that can be approximated by a chi-square distribution with $(k-1)$ degrees of freedom. For the approximation to be reasonably good, we require that either $b$, the number of blocks, or $k$, the number of treatments, exceed 5. The rejection region for the test consists of large values of $F_r$. Therefore, we reject $F_r > \chi^2_\alpha$.

The **Friedman $F_r$ test** is summarized in the box and its use is illustrated in Example 15.7.

### The Friedman $F_r$ Test for a Randomized Block Design

$H_0$: The relative frequency distributions for the $k$ populations are identical

$H_a$: At least two of the $k$ populations differ in location (shifted either to the left or to the right of one another)

*Test statistic*: Rank each of the $k$ observations within each block from the smallest (rank 1) to the largest (rank $k$). Calculate the treatment rank sums, $T_1, T_2, \ldots, T_k$. Then the test statistic is

$$F_r = \frac{12}{bk(b+1)} \sum T_i^2 - 3b(k+1)$$

where

$b$ = Number of blocks employed in the experiment

$k$ = Number of treatments

$T_i$ = Sum of the ranks for the $i$th treatment

*Rejection region*: $F_r > \chi^2_\alpha$

$p$-value: $P(\chi^2 > F_r)$

where $\chi^2_\alpha$ is based on $(k-1)$ degrees of freedom

*Assumptions*:
1. The $k$ treatments were randomly assigned to the $k$ experimental units within each block.
2. For the chi-square approximation to be adequate, either the number $b$ of blocks or the number $k$ of treatments should exceed 5.
3. Tied observations are assigned ranks equal to the average of the ranks that would have been assigned to the observations had they not been tied.

**Example 15.7**

Friedman Test Application

The corrosion of different metals is a problem in many mechanical devices. Three sealers used to help retard corrosion were tested to determine whether there were any differences among them. Samples of 10 different metal compositions were treated with each of the three sealers, and the amount of corrosion was measured after exposure to the same environmental conditions for 1 month. The data and their associated ranks are shown in Table 15.9. Is there any evidence of a difference in the probability distributions of the amounts of corrosion among the three types of sealers? Test using $\alpha = .05$.

Solution

We want to test the null hypothesis

$H_0$: The probability distributions of the amounts of corrosion are identical for the three sealers

against the alternative hypothesis

$H_a$: At least two of the probability distributions differ in location

The ranks of the three treatments within each block and the treatment rank sums are shown in Table 15.9. Therefore, the calculated value of the $F_r$ statistic is

$$\begin{aligned} F_r &= \frac{12}{bk(k+1)} \sum_{i=1}^{k} T_i^2 - 3b(k+1) \\ &= \frac{12}{10(3)(4)}[(19.5)^2 + (26.5)^2 + (14)^2] - 3(10)(4) \\ &= 7.85 \end{aligned}$$

CORRODE

**TABLE 15.9 Data and Ranks for the Randomized Block Design of Example 15.7**

| | Sealer | | | | | |
|---|---|---|---|---|---|---|
| *Metal* | 1 | Rank | 2 | Rank | 3 | Rank |
| 1 | 21 | 2 | 23 | 3 | 15 | 1 |
| 2 | 29 | 2 | 30 | 3 | 21 | 1 |
| 3 | 16 | 1 | 19 | 3 | 18 | 2 |
| 4 | 20 | 3 | 19 | 2 | 18 | 1 |
| 5 | 13 | 2 | 10 | 1 | 14 | 3 |
| 6 | 5 | 1 | 12 | 3 | 6 | 2 |
| 7 | 18 | 2.5 | 18 | 2.5 | 12 | 1 |
| 8 | 26 | 2 | 32 | 3 | 21 | 1 |
| 9 | 17 | 2 | 20 | 3 | 9 | 1 |
| 10 | 4 | 2 | 10 | 3 | 2 | 1 |
| | | $T_1 = 19.5$ | | $T_2 = 26.5$ | | $T_3 = 14$ |

Note that this value agrees with the test statistic shaded on the MINITAB printout, Figure 15.8. The rejection region for the test is $F_r > \chi^2_{.05}$, where the tabulated value (given in Table 8 of Appendix B) of $\chi^2_{.05}$, based on $k - 1 = 2$ degrees of freedom, is 5.99147. Thus, we will reject $H_0$ if $F_r > 5.99147$.

Since the computed value of the test statistic, $F_r = 7.85$, exceeds $\chi^2_{.05} = 5.99147$, there is sufficient evidence to reject $H_0$ and conclude that differences exist in the locations of two or more of the corrosion probability distributions. The $p$-value of the test (shaded in Figure 15.8) supports this result. The practical conclusion is that there is evidence to indicate a difference among the sealing abilities of the three sealers.

**Friedman Test: CORROSION versus SEALER blocked by METAL**

```
S = 7.85  DF = 2  P = 0.020
S = 8.05  DF = 2  P = 0.018 (adjusted for ties)

                           Sum
                            of
SEALER    N  Est Median  Ranks
1        10      16.583   19.5
2        10      19.417   26.5
3        10      12.750   14.0

Grand median = 16.250
```

**FIGURE 15.8**
MINITAB Friedman test for Example 15.7

## Applied Exercises

15.44 *Stress in cows prior to slaughter.* Refer to the *Applied Animal Behaviour Science* (June 2010) study of stress in cows prior to slaughter, Exercise 14.22 (p. 770). In the experiment, recall that the heart rate (beats per minute) of a cow was measured at four different pre-slaughter phases — (1) first phase of visual contact with pen mates, (2) initial isolation from pen mates for prepping, (3) restoration of visual contact with pen mates, and (4) first contact with human prior to slaughter. Thus, a randomized block design was employed. The simulated data for eight cows are reproduced in the accompanying table. Consider applying the nonparametric Friedman test to determine whether the heart rate distributions differ for cows in the four preslaughter phases. A MINITAB printout of the analysis follows.

a. Locate the rank sums on the printout.
b. Use the rank sums to calculate the $F_r$ test statistic. Does the result agree with the value shown on the MINITAB printout?
c. Locate the $p$-value of the test on the printout.
d. Provide the appropriate conclusion in the words of the problem if $\alpha = .05$.

**COWSTRESS**

| | PHASE | | | |
|---|---|---|---|---|
| COW | 1 | 2 | 3 | 4 |
| 1 | 124 | 124 | 109 | 107 |
| 2 | 100 | 98 | 98 | 99 |
| 3 | 103 | 98 | 100 | 106 |
| 4 | 94 | 91 | 98 | 95 |
| 5 | 122 | 109 | 114 | 115 |
| 6 | 103 | 92 | 100 | 106 |
| 7 | 98 | 80 | 99 | 103 |
| 8 | 120 | 84 | 107 | 110 |

15.45 *Impact study of distractions while driving.* The consequences of performing verbal and spatial-imagery tasks on visual search while driving were studied and the results published in the *Journal of Experimental Psychology: Applied* (Mar. 2000). Twelve drivers were recruited to drive on a highway in Madrid, Spain. During the drive, each subject was asked to perform three different tasks—a verbal task (repeating words that begin with a certain letter), a spatial-imagery task (imagining letters rotated a certain way), and no mental task. Since each driver performed all three tasks, the design is a randomized block with 12 blocks (drivers) and 3 treatments (tasks). Using a computerized, head-free, eye-tracking system, the researchers kept track of the eye fixations of each driver on three different objects—the interior mirror, the off-side mirror, and the speedometer—and determined the proportion of eye fixations on the object. The researchers used the Friedman nonparametric test to compare the distributions of the eye fixation proportions for the three tasks.

a. Using $\alpha = .01$, find the rejection region for the Friedman test.
b. For the response variable, proportion of eye fixations on the interior mirror, the researchers determined the Friedman test statistic to be $\chi^2 = 19.16$. Give the appropriate conclusion.
c. For the response variable, proportion of eye fixations on the off-side mirror, the researchers determined the Friedman test statistic to be $\chi^2 = 19.16$. Give the appropriate conclusion.
d. For the response variable, proportion of eye fixations on the speedometer, the researchers determined the Friedman test statistic to be $\chi^2 = 20.67$. Give the appropriate conclusion.

15.46 *Containers designed to cool citrus fruit.* Refer to the *Journal of Food Engineering* (September, 2013) study of the cooling performances fruit container designs, Exercise 14.23 (p. 771). Recall that three container types were investigated—Standard, Supervent, and Ecopack. The containers arranged fruit in either two or three rows; thus, row was used as a blocking factor in a randomized block design with container design representing the treatments. The response variable of interest was the half-cooling time, measured as the time (in minutes) required to reduce the temperature difference between the fruit and cooling air by half. Half-cooling times were measured for each row of fruit for each design. The data

**MINITAB Output for Exercise 15.44**

**Friedman Test: BPM versus PHASE blocked by COW**

```
S = 10.39  DF = 3  P = 0.016
S = 10.65  DF = 3  P = 0.014 (adjusted for ties)

                         Sum of
PHASE  N  Est Median     Ranks
1      8      103.63      25.5
2      8       95.63      11.0
3      8      101.13      18.5
4      8      103.13      25.0

Grand median = 100.88
```

is reproduced in the accompanying table. A nonparametric analysis of variance of the data is shown in the accompanying SPSS printout. Interpret the results using $\alpha = .10$.

**COOLING**

| | Standard | Supervent | Ecopack |
|---|---|---|---|
| *Row 1* | 116 | 93 | 115 |
| *Row 2* | 181 | 139 | 164 |
| *Row 3* | 247 | 176 | |

**SPSS Output for Exercise 15.46**

**Friedman Test**

**Ranks**

| | Mean Rank |
|---|---|
| STD | 3.00 |
| SUP | 1.00 |
| ECO | 2.00 |

**Test Statistics[a]**

| | |
|---|---|
| N | 2 |
| Chi-Square | 4.000 |
| df | 2 |
| Asymp. Sig. | .135 |

a. Friedman Test

15.47 *Evaluating lead-free solders.* Refer to the *Soldering & Surface Mount Technology* (Vol. 13, 2001) study to compare four soldering methods, Exercise 14.24 (p. 771). Recall that a measure of plastic hardening (Nm/m$^2$) was obtained for each solder type at each of six different temperatures. The data are reproduced in the table. Analyze the data using the appropriate nonparametric method. Interpret the results at $\alpha = .10$.

**LEADSOLDER**

| Temperature | Tin–Lead | Tin–Silver | Tin–Copper | Tin–Silver–Copper |
|---|---|---|---|---|
| 23°C | 50.1 | 33.0 | 14.9 | 41.0 |
| 50°C | 24.6 | 27.7 | 10.5 | 20.7 |
| 75°C | 23.1 | 10.7 | 9.3 | 17.1 |
| 100°C | 1.8 | 9.0 | 8.8 | 8.7 |
| 125°C | 1.1 | 4.9 | 5.4 | 7.1 |
| 150°C | 0.3 | 3.2 | 5.0 | 4.9 |

*Source:* Harrison, M. R., Vincent, J. H., and Steen, H. A. H. "Lead-free reflow soldering for electronics assembly." *Soldering & Surface Mount Technology*, Vol. 13, No. 3, 2001 (Table X).

15.48 *Testing an optical mark reader.* An optical mark reader (OMR) is a machine that is able to "read" pencil marks that have been entered on a special form. A manufacturer of OMRs believes its product can operate equally well in a variety of temperature and humidity environments. To determine whether operating data contradict this belief, the manufacturer asks a well-known industrial testing laboratory to test its product. Five recently produced OMRs were randomly selected and each was operated in six different environments. The number of forms each was able to process in an hour was recorded and used as a measure of the OMR's operating efficiency. These data appear in the table. Use the Friedman $F_r$ test to determine whether evidence exists to indicate that the probability distributions for the number of forms processed per hour differ in location for at least two of the environments. Test using $\alpha = .10$.

**OMR**

| Machine Number | Environment 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 8,001 | 8,025 | 8,100 | 8,055 | 7,991 | 8,007 |
| 2 | 7,910 | 7,932 | 7,900 | 7,990 | 7,892 | 7,922 |
| 3 | 8,111 | 8,101 | 8,201 | 8,175 | 8,102 | 8,235 |
| 4 | 7,802 | 7,820 | 7,904 | 7,850 | 7,819 | 8,100 |
| 5 | 7,500 | 7,601 | 7,702 | 7,633 | 7,600 | 7,561 |

15.49 *Absentee rates at a jeans plant.* Refer to Exercise 14.26 (p. 772) and the *New Technology, Work, and Employment* (July 2001) study of daily worker absentee rates at a jeans plant. Nine weeks were randomly selected and the absentee rate (percentage of workers absent) determined for each day (Monday through Friday) of the work week. The data are reproduced in the table. Conduct a nonparametric analysis of the data to compare the distributions of absentee rates for the five days of the work week.

**JEANS**

| Week | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| 1 | 5.3 | 0.6 | 1.9 | 1.3 | 1.6 |
| 2 | 12.9 | 9.4 | 2.6 | 0.4 | 0.5 |
| 3 | 0.8 | 0.8 | 5.7 | 0.4 | 1.4 |
| 4 | 2.6 | 0.0 | 4.5 | 10.2 | 4.5 |
| 5 | 23.5 | 9.6 | 11.3 | 13.6 | 14.1 |
| 6 | 9.1 | 4.5 | 7.5 | 2.1 | 9.3 |
| 7 | 11.1 | 4.2 | 4.1 | 4.2 | 4.1 |
| 8 | 9.5 | 7.1 | 4.5 | 9.1 | 12.9 |
| 9 | 4.8 | 5.2 | 10.0 | 6.9 | 9.0 |

*Source:* Boggis, J. J. "The eradication of leisure." *New Technology, Work, and Employment*, Vol. 16, No. 2, July 2001 (Table 3).

15.50 *Nearest neighbor-based imputation algorithms.* For data sets that contain many missing values, methods for estimating the missing values — called *imputation* algorithms — may be applied. In the journal, *Data & Knowledge Engineering* (March 2013), researchers compared several imputation algorithms based on using nearest neighbors to estimate missing values. The five methods studied are named KMI, EACI, IKNNI, KNNI, and SKNN. Each of the methods was applied to each of four different data sets, one data set with 10% missing values, one with 30% missing, one with 50% missing, and one with 70% missing. After each imputation algorithm was applied, the normalized root mean square error (NRMSE) — a measure of the accuracy of the missing value predictions — was determined. These NRMSE values (based on information provided in the journal article) are given in the following table. Conduct a nonparametric analysis of the data. Is there evidence to indicate that the NRMSE distributions differ for the five imputation algorithms? Test using $\alpha = .01$.

**IMPUTE**

| Missing % | KMI | EACI | IKNNI | KNNI | SKNN |
|---|---|---|---|---|---|
| 10 | .42 | .29 | .23 | .23 | .23 |
| 30 | .40 | .30 | .24 | .25 | .24 |
| 50 | .39 | .30 | .25 | .26 | .26 |
| 70 | .40 | .31 | .26 | .27 | .26 |

## 15.7 Nonparametric Regression

We learned in Section 11.13 how to modify the regression analysis when the assumptions about the random error term $\varepsilon$ are violated. For example, if the variance $\sigma^2$ of $\varepsilon$ is not constant, we transform the dependent variable $y$ using one of the variance-stabilizing transformations discussed in Section 11.13. An alternative procedure is to conduct a **nonparametric regression analysis** of the data.

In nonparametric regression, tests of model adequacy do not require any assumptions about the probability distribution of $\varepsilon$; thus, they are distribution-free. Although the tests are intuitively appealing, they can become quite difficult to apply in practice, especially when the number of observations is large. For this reason, and the fact that residual diagnostics are readily available via the computer, most analysts prefer to use the techniques of Section 11.13 when the standard regression assumptions are violated.

For those who are interested, we provide brief descriptions of the nonparametric alternatives to the parametric simple linear regression tests of Chapter 10. Specifically, we discuss a nonparametric test for (1) rank correlation and (2) the slope parameter of the straight-line model.

### Spearman's Rank Correlation

As an alternative to the Pearson product moment correlation coefficient $r$ (Section 10.7), we can compute a correlation coefficient based on ranks. Spearman's rank correlation coefficient, denoted $r_s$, can then be used to test for rank correlation between two variables, $y$ and $x$.

To illustrate, suppose a large manufacturing firm wants to determine whether the number $y$ of work-hours an employee misses per year is correlated with the employee's annual wages $x$ (in thousands of dollars). A sample of 15 employees produced the data shown in Table 15.10.

Spearman's rank correlation coefficient is found by first ranking the values of each variable separately. (Ties are treated by averaging the tied ranks.) Then $r_s$ is computed in exactly the same way as the Pearson correlation coefficient $r$—the only difference is that the values of $x$ and $y$ that appear in the formula for $r$ are replaced by their ranks. That is, the *ranks* of the raw data are used to compute $r_s$ rather than the raw data themselves. When there are no (or few) ties in the ranks, this formula reduces to the simple expression

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where $d_i$ is the difference between the rank of $y$ and $x$ for the $i$th observation.

MISSWORK

**TABLE 15.10 Work-Hours Missed, Annual Wages, and Ranks for 15 Employees**

| Employee | Hours Missed, $y$ | Annual Wages, $x$ | $y$-Rank | $x$-Rank | Difference $d_i$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| 1 | 49 | 15.8 | 6 | 11 | −5 | 25 |
| 2 | 36 | 17.5 | 4 | 12 | −8 | 64 |
| 3 | 127 | 11.3 | 13 | 2 | 11 | 121 |
| 4 | 91 | 13.2 | 12 | 6 | 6 | 36 |
| 5 | 72 | 13.0 | 9 | 5 | 4 | 16 |
| 6 | 34 | 14.5 | 3 | 9 | −6 | 36 |
| 7 | 155 | 11.8 | 14 | 3 | 11 | 121 |
| 8 | 11 | 20.2 | 2 | 14 | −12 | 144 |
| 9 | 191 | 10.8 | 15 | 1 | 14 | 196 |
| 10 | 6 | 18.8 | 1 | 13 | −12 | 144 |
| 11 | 63 | 13.8 | 8 | 7 | 1 | 1 |
| 12 | 79 | 12.7 | 10 | 4 | 6 | 36 |
| 13 | 43 | 15.1 | 5 | 10 | −5 | 25 |
| 14 | 57 | 24.2 | 7 | 15 | −8 | 64 |
| 15 | 82 | 13.9 | 11 | 8 | 3 | 9 |
| | | | | | | $\sum d_i^2 = 1{,}038$ |

The ranks of $y$ and $x$, the differences between the ranks, and the squared differences for each of the 15 employees are also shown in Table 15.10. Note that the sum of the squared differences is $\sum d_i^2 = 1{,}038$. Substituting this value into the formula for $r_s$, we obtain

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6(1{,}038)}{15(224)} = -.854$$

The value of $r_s$ can also be obtained using statistical software. An SPSS printout of the analysis is shown in Figure 15.9. The value of $r_s$, highlighted on the printout, agrees with our calculated value of $-.854$. This large negative value of $r_s$ implies that a fairly strong negative correlation exists between work-hours missed $y$ and annual wages $x$ in the sample.

To determine whether a negative rank correlation exists in the population, we would test $H_0$: $\rho = 0$ against $H_a$: $\rho < 0$ using $r_s$ as a test statistic. As you would expect, we reject $H_0$ for small values of $r_s$. Upper-tailed critical values of Spearman's $r_s$ are provided in Table 17 of Appendix B. This table is partially reproduced in Table 15.11. Since the

**FIGURE 15.9**
SPSS Spearman correlation for data of Table 15.10

**Correlations**

| | | | HOURS | WAGES |
|---|---|---|---|---|
| Spearman's rho | HOURS | Correlation Coefficient | 1.000 | -.854** |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 15 | 15 |
| | WAGES | Correlation Coefficient | -.854** | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 15 | 15 |

**. Correlation is significant at the 0.01 level (2-tailed).

**TABLE 15.11 A Portion of the Spearman's $r_s$ Table, Table 17 of Appendix B**

*The $\alpha$ values correspond to a one-tailed test of $H_0$: $\rho_s = 0$. The tabulated value of $\alpha$ should be doubled for two-tailed tests.*

| $n$ | $\alpha = .05$ | $\alpha = .025$ | $\alpha = .01$ | $\alpha = .005$ |
|---|---|---|---|---|
| 5 | .900 | — | — | — |
| 6 | .829 | .886 | .943 | — |
| 7 | .714 | .786 | .893 | — |
| 8 | .643 | .738 | .833 | .881 |
| 9 | .600 | .683 | .783 | .833 |
| 10 | .564 | .648 | .745 | .794 |
| 11 | .523 | .623 | .736 | .818 |
| 12 | .497 | .591 | .703 | .780 |
| 13 | .475 | .566 | .673 | .745 |
| 14 | .457 | .545 | .646 | .716 |
| 15 | .441 | .525 | .623 | .689 |
| 16 | .425 | .507 | .601 | .666 |

distribution of $r_s$ is symmetric around 0, the lower-tailed critical value is the negative of the corresponding upper-tailed critical value. For, say, $\alpha = .01$ and $n = 15$, the critical value (shaded in Table 15.11) is $r_{.01} = .623$. Thus, the rejection region for the test is

Reject $H_0$ if $r_s < -.623$

Since the test statistic, $r_s = -.854$, falls in the rejection region, there is sufficient evidence (at $\alpha = .01$) of negative correlation between work-hours missed $y$ and annual wages $x$ in the population. (*Note*: The $p$-value of the test is highlighted on Figure 15.9.)

Spearman's nonparametric test for rank correlation in the population is summarized in the box.

### Spearman's Nonparametric Test for Rank Correlation

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\rho = 0$ | $H_0$: $\rho = 0$ |
| $H_a$: $\rho > 0$ (or $H_a$: $\rho < 0$) | $H_a$: $\rho \neq 0$ |

*Test statistic*: $$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i$ is the difference between the $y$ rank and $x$ rank for the $i$th observation.

(*Note:* In the case of ties, calculate $r_s$ by substituting the ranks of the $y$'s and the ranks of the $x$'s for the actual $y$ values and $x$ values in the formula for $r$ given in Section 10.7.)

| *Rejection region*: | *Rejection region*: |
|---|---|
| $r_s > r_\alpha$ (or $r_s < -r_\alpha$) | $\lvert r_s \rvert > r_{\alpha/2}$ |

where the values of $r_\alpha$ and $r_{\alpha/2}$ are given in Table 17 of Appendix B.

*Assumptions*: None

**Theil Test for Zero Slope** Alternatively, we could test for linear correlation in the population by testing the slope parameter $\beta_1$ in the simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

That is, we could test $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 \neq 0$. A distribution-free test for the slope is the **Theil *C* test**.

To conduct this nonparametric test, we first rank the $x$ values in increasing order and list the ordered $(x, y)$ pairs, as shown in Table 15.12. Next, we calculate all possible differences $y_j - y_i$, $i < j$ (where $i$ and $j$ represent the $i$th and $j$th ranked observations), and note the sign (positive or negative) of each difference.

For example, the $y$ value for the employee ranked #2, $y_2 = 127$, is compared to the $y$ value for each employee with a lower rank. In this case, the only employee ranked lower is employee #1, with $y_1 = 191$ (see Table 15.12). The difference

$$y_2 - y_1 = 127 - 191 = -64$$

is negative and is noted as such in Table 15.12.

Similarly, we compare the $y$ value of the employee ranked #3, $y_3 = 155$, to the $y$ value of employees of lower rank, $y_2 = 127$ and $y_1 = 191$, by the differences

$$y_3 - y_2 = 155 - 127 = 28$$

and

$$y_3 - y_1 = 155 - 191 = -36$$

This results in one positive and one negative difference. Continuing in this manner, we obtain a total of 17 positive differences and 88 negative differences, as shown in Table 15.12.

**TABLE 15.12 Data of Table 15.10 Ranked by Annual Wages *x***

| Employee Ranking | Hours Missed, $y$ | Annual Wages, $x$ | Differences, $y_j - y_i (i < j)$ # Negatives | # Positives |
|---|---|---|---|---|
| 1 | 191 | 10.8 | — | — |
| 2 | 127 | 11.3 | 1 | 0 |
| 3 | 155 | 11.8 | 1 | 1 |
| 4 | 79 | 12.7 | 3 | 0 |
| 5 | 72 | 13.0 | 4 | 0 |
| 6 | 91 | 13.2 | 3 | 2 |
| 7 | 63 | 13.8 | 6 | 0 |
| 8 | 82 | 13.9 | 4 | 3 |
| 9 | 34 | 14.5 | 8 | 0 |
| 10 | 43 | 15.1 | 8 | 1 |
| 11 | 49 | 15.8 | 8 | 2 |
| 12 | 36 | 17.5 | 10 | 1 |
| 13 | 6 | 18.8 | 12 | 0 |
| 14 | 11 | 20.2 | 12 | 1 |
| 15 | 57 | 24.2 | 8 | 6 |
| | | | Totals: 88 | 17 |

The test statistic $C$ is obtained by scoring each positive difference as a $+1$ and each negative difference as a $-1$ (differences of 0 are assigned a score of 0) and summing the scores. Therefore, for the data of Table 15.12, we obtain the test statistic

$$C = (+1)(17) + (-1)(88) = -71$$

The observed significance level ($p$-value) of the test is obtained from Table 18 of Appendix B. For this lower-tailed test, i.e., a test for a negative slope, the $p$-value is $P(C \leq -71)$. Searching the $n = 15$ column and the $x = 71$ row of Table 18 of Appendix B, we obtain the $p$-value $\approx 0$. Thus, there is strong evidence to reject $H_0$ and conclude that work-hours missed $y$ is negatively linearly related to annual wages $x$ at this firm.

Theil's test for the slope of a straight-line model is described, in general, in the next box. A nonparametric confidence interval for the slope $\beta_1$ based on the Theil test can also be formed. Consult the references if you want to learn how to construct this interval.

### Theil's Test for Zero Slope in the Straight-Line Model $y = \beta_0 + \beta_1 x + \varepsilon$

| *One-Tailed Test* | *Two-Tailed Test* |
|---|---|
| $H_0$: $\beta_1 = 0$ | $H_0$: $\beta_1 = 0$ |
| $H_a$: $\beta_1 > 0$ (or $H_a$: $\beta_1 < 0$) | $H_a$: $\beta_1 \neq 0$ |

*Test statistic*: $C = (-1)(\text{Number of negative } y_j - y_i \text{ differences}) + (1)(\text{Number of positive } y_j - y_i \text{ differences})$

where $y_i$ and $y_j$ are the $i$th and $j$th observations ranked in increasing order of the $x$ values, $i < j$.

| *Observed significance level*: | *Observed significance level*: |
|---|---|
| $p\text{-value} = \begin{cases} P(x \geq C) \text{ for } H_a\text{: } \beta_1 > 0 \\ P(x \leq C) \text{ for } H_a\text{: } \beta_1 < 0 \end{cases}$ | $p\text{-value} = 2\min(p_1, p_2)$ where $p_1 = P(x \geq C)$, $p_2 = P(x \leq C)$ |

where the values of $P(x \geq C) = P(x \leq -C)$ are given in Table 18 of Appendix B.

*Assumptions*: The random error $\varepsilon$ is independent.

Nonparametric tests are also available for multiple regression models. These tests are very sophisticated, however, and require the use of specialized statistical computer software not yet available on a commercial basis. Consult the references if you want to learn more about these nonparametric techniques.

## Applied Exercises

15.51 *New method for blood typing.* Refer to the *Analytical Chemistry* (May 2010) evaluation of a new method of typing blood, Exercise 10.6 (p. 496). Recall that blood drops were applied to the paper and the rate of absorption (called *blood wicking*) was measured. The next table (p. 874) gives the wicking lengths (millimeters) for six blood drops, each at a different antibody concentration. Let $y$ = wicking length and $x$ = antibody concentration.

a. Rank the wicking length values from 1 to 6.
b. Rank the antibody concentration values from 1 to 6.
c. Use the ranks, parts **a** and **b**, to compute Spearman's rank correlation coefficient.
d. Based on the result, part **c** is there sufficient evidence to indicate that wicking length is negatively rank correlated with antibody concentration? Test using $\alpha = .05$.

### Data for Exercise 15.51

**BLOODTYPE**

| Droplet | Length (mm) | Concentration |
|---|---|---|
| 1 | 22.50 | 0.0 |
| 2 | 16.00 | 0.2 |
| 3 | 13.50 | 0.4 |
| 4 | 14.00 | 0.6 |
| 5 | 13.75 | 0.8 |
| 6 | 12.50 | 1.0 |

*Source:* Khan, M.S., et al. "Paper diagnostic for instant blood typing", *Analytical Chemistry*, Vol. 82, No. 10, May 2010 (adapted from Figure 4b).

15.52 *Extending the life of an aluminum smelter pot.* Refer to *The American Ceramic Society Bulletin* (Feb. 2005) study of the lifelength of an aluminum smelter pot, Exercise 10.9 (p. 497). Since the life of a smelter pot depends on the porosity of the brick lining, the researchers measured the apparent porosity and the mean pore diameter of each of six bricks. The data are reproduced in the accompanying table.

**SMELTPOT**

| Brick | Apparent Porosity (%) | Mean Pore Diameter (micrometers) |
|---|---|---|
| A | 18.8 | 12.0 |
| B | 18.3 | 9.7 |
| C | 16.3 | 7.3 |
| D | 6.9 | 5.3 |
| E | 17.1 | 10.9 |
| F | 20.4 | 16.8 |

*Source:* Bonadia, P., et al. "Aluminosilicate refractories for aluminum cell linings." *The American Ceramic Society Bulletin*, Vol. 84, No. 2, Feb. 2005 (Table II).

a. Rank the apparent porosity values for the six bricks. Then rank the six pore diameter values.
b. Use the ranks, part **a**, to find the rank correlation between apparent porosity ($y$) and mean pore diameter ($x$). Interpret the result.
c. Conduct a test for positive rank correlation. Use $\alpha = .01$.

15.53 *Organizational use of the Internet.* Researchers from the United Kingdom and Germany attempted to develop a theoretically grounded measure of organizational Internet use (OIU) and published their results in *Internet Research* (Vol. 15, 2005). Using data collected from a sample of 77 websites, they investigated the link between OIU level (measured on a 7-point scale) and several observation-based indicators. Spearman's rank correlation coefficient (and associated $p$-values) for several indicators are shown in the table (next column).

a. Interpret each of the values of $r_s$ given in the table.
b. Interpret each of the $p$-values given in the table. (Use $\alpha = .10$ to conduct each test.)

| Indicator | Correlation with OIU Level $r_s$ | $p$-value |
|---|---|---|
| Navigability | .179 | .148 |
| Transactions | .334 | .023 |
| Locatability | .590 | .000 |
| Information Richness | −.115 | .252 |
| Number of files | .114 | .255 |

*Source:* Brock, J. K., and Zhou, Y. "Organizational use of the internet." *Internet Research*, Vol. 15, No. 1, 2005 (Table IV).

15.54 *Assessment of biometric recognition methods.* Biometric technologies have been developed to detect or verify an individual's identity. These methods are based on physiological characteristics (called biometric signatures) such as facial features, eye irises, fingerprints, voice, hand shape, and gait. In *Chance* (Winter 2004), four biometric recognition algorithms were compared. All four methods were applied to 1,196 biometric signatures and "match" scores were obtained. The Spearman correlation between match scores for each possible algorithm pair was determined. The rank correlation matrix is shown here. Interpret the results.

| Method | I | II | III | IV |
|---|---|---|---|---|
| *I* | 1 | .189 | .592 | .340 |
| *II* | | 1 | .205 | .324 |
| *III* | | | 1 | .314 |
| *IV* | | | | 1 |

15.55 *Single machine batch scheduling.* Refer to the *Asian Journal of Industrial Engineering* (Vol. 4, 2012) evaluation of a computerized mathematical model used in single machine batch scheduling, Exercise 10.27 (p. 506). Recall that the performance of the model was graded using a variable called Value of Object Function (VOF).

**SWRUN**

| Software Run | Number of Batches | VOF | Run Time (seconds) |
|---|---|---|---|
| 1 | 3 | 86.68 | 27 |
| 2 | 4 | 232.87 | 14 |
| 3 | 5 | 372.36 | 12 |
| 4 | 6 | 496.51 | 18 |
| 5 | 7 | 838.82 | 42 |
| 6 | 8 | 1183.00 | 33 |

*Source:* Karimi-Nasab, M., Haddad, H., & Ghanbari, P. "A simulated annealing for the single machine batch scheduling deterioration and precedence constraints", *Asian Journal of Industrial Engineering*, Vol. 4, No. 1, 2012 (Table 2).

Data on VOF, run time (in seconds), and number of batches scheduled for six software runs are reproduced in the table (p. 874). Consider a straight-line regression model for $y$ = VOF as a function of either run time or number of batches.

a. Conduct a nonparametric test to determine if the slope of the line relating $y$ = VOF and $x$ = number of batches is positive. Use $\alpha = .05$.

b. Conduct a nonparametric test to determine if the slope of the line relating $y$ = VOF and $x$ = run time is negtive. Use $\alpha = .05$.

15.56 *Removing nitrogen from toxic wastewater.* Refer to the *Chemical Engineering Journal* (April, 2013) study of a method for removing nitrogen from toxic wastewater, Exercise 10.53 (p. 526). Recall that the researchers related $y$ = the amount of nitrogen removed (measured in milligrams per liter) from a wastewater specimen to $x$ = the amount of ammonium (milligrams per liter) used in the removal process using data collected for 120 specimens. The data for the first 5 specimens are shown below.

a. Using only the data for the first 5 specimens, find Spearman's rank correlation between $y$ and $x$.

b. Based on the result, part **a** is the amount of nitrogen removed significantly positively correlated with amount of ammonium? Test using $\alpha = .01$.

c. Repeat parts **a** and **b**, using the full sample of 120 wastewater specimens.

**NITRO (First 5 observations of 120 shown)**

| Nitrogen | Ammonium |
|---|---|
| 18.87 | 67.40 |
| 17.01 | 12.49 |
| 23.88 | 61.96 |
| 10.45 | 15.63 |
| 36.03 | 83.66 |

15.57 *Pressure stabilization of fresh concrete.* Refer to the *Engineering Structures* (July 2013) study of the characteristics of fresh concrete, Exercise 10.38 (p. 512). Recall that the researchers investigated the linear effect of time needed for pressure stabilization ($x$) on each of three different dependent variables: $y_1$ = initial setting time (hours), $Y_2$ = final setting time (hours), and $y_3$ = maturity index (°C - hours). The data on these variables for $n = 8$ fresh concrete lateral pressure tests are reproduced in the table (next column).

a. Apply Spearman's rank correlation test to determine which of the three dependent variables is most strongly positively associated with pressure stabilization time.

b. Consider the linear model, $E(y_i) = \beta_0 + \beta_1 x$, $i = 1, 2, 3$. Apply Theil's nonparametric procedure to determine which of the three slopes are significantly greater than 0. (Test using $\alpha = .05$.)

**CONCRETE2**

| Test | $Y_1$ | $Y_2$ | $Y_3$ | X |
|---|---|---|---|---|
| A1 | 4.63 | 7.17 | 385.81 | 12.03 |
| A2 | 4.32 | 6.52 | 358.44 | 11.32 |
| A3 | 4.54 | 6.31 | 292.71 | 9.51 |
| A4 | 4.09 | 6.19 | 253.16 | 8.25 |
| A5 | 4.56 | 6.81 | 279.82 | 9.02 |
| A6 | 4.48 | 6.98 | 318.74 | 9.97 |
| A7 | 4.35 | 6.45 | 262.14 | 8.42 |
| A8 | 4.23 | 6.69 | 244.97 | 7.53 |

*Source:* Santilli, A., Puente, I., & Tanco, M. "Fresh concrete lateral pressure decay: Kinetics and factorial design to determine significant parameters", *Engineering Structures*, Vol. 52, July 2013 (Table 4).

15.58 *New iron-making process.* Refer to the *Mining Engineering* (Oct. 2004) study of a new iron-making technology, Exercise 10.25 (p. 505). Recall that the carbon content produced in a pilot plant test was compared to that from laboratory furnace tests. The data for 25 pilot tests are reproduced in the accompanying table. Conduct a nonparametric test to determine if the carbon content values from the pilot plant are positively correlated with the values from the lab furnace. Test using $\alpha = .01$.

**CARBON**

| Carbon Content (%) | | Carbon Content (%) | |
|---|---|---|---|
| Pilot Plant | Lab Furnace | Pilot Plant | Lab Furnace |
| 1.7 | 1.6 | 3.4 | 4.3 |
| 3.1 | 2.4 | 3.2 | 3.6 |
| 3.3 | 2.8 | 3.3 | 3.4 |
| 3.6 | 2.9 | 3.1 | 3.3 |
| 3.4 | 3.0 | 3.0 | 3.2 |
| 3.5 | 3.1 | 2.9 | 3.2 |
| 3.8 | 3.2 | 2.6 | 3.4 |
| 3.7 | 3.2 | 2.5 | 3.3 |
| 3.5 | 3.3 | 2.6 | 3.2 |
| 3.4 | 3.3 | 2.6 | 3.1 |
| 3.6 | 3.4 | 2.4 | 3.0 |
| 3.5 | 3.4 | 2.6 | 2.7 |
| 3.9 | 3.8 | | |

*Source:* Hoffman, G., and Tsuge, O. "ITmk3—Application of a new ironmaking technology for the iron ore mining industry." *Mining Engineering*, Vol. 56, No. 9, Oct. 2004 (Figure 8).

15.59 *Thermal performance of copper tubes.* Refer to Exercise 10.14 (p. 499) and the model of the thermal performance of integral-fin tubes used in the refrigeration and process industries (*Journal of Heat Transfer*, Aug. 1990). The data in the table (p. 876) are the unflooded area ratio ($x$) and heat transfer enhancement ($y$) values recorded for the 24 integral-fin tubes. Conduct a nonparametric test for

**FINTUBES**

| Ratio, $x$ | Enhancement, $y$ | Ratio, $x$ | Enhancement, $y$ |
|---|---|---|---|
| 1.93 | 4.4 | 2.00 | 5.2 |
| 1.95 | 5.3 | 1.77 | 4.7 |
| 1.78 | 4.5 | 1.62 | 4.2 |
| 1.64 | 4.5 | 2.77 | 6.0 |
| 1.54 | 3.7 | 2.47 | 5.8 |
| 1.32 | 2.8 | 2.24 | 5.2 |
| 2.12 | 6.1 | 1.32 | 3.5 |
| 1.88 | 4.9 | 1.26 | 3.2 |
| 1.70 | 4.9 | 1.21 | 2.9 |
| 1.58 | 4.1 | 2.26 | 5.3 |
| 2.47 | 7.0 | 2.04 | 5.1 |
| 2.37 | 6.7 | 1.88 | 4.6 |

a positive linear relationship between heat transfer enhancement ($y$) and unflooded area ratio ($x$). Use $\alpha = .10$.

## Theoretical Exercises

15.60 Show that for the special case where $n = 3$, $-1 \le r_s \le 1$. (*Hint:* List each of the $3! \times 3! = 36$ different arrangements of the $x$ and $y$ rankings and compute $r_s$ for each.)

15.61 Show that for the special case where $n = 3$, $E(r_s) = 0$. (This fact is also true in general.) (*Hint*: Use the results of Theoretical Exercise 15.60 and the fact that any of the arrangements has a probability of $\frac{1}{36}$ of occurring.)

## *STATISTICS IN ACTION REVISITED*

### How Vulnerable are New Hampshire Wells to Groundwater Contamination?

We return to the study of MTBE contamination of New Hampshire groundwater wells (p. 838). There are several questions of interest about the level of well contamination in the state that the environmental researchers wanted to answer.

*Research Question 1*: Do fewer than half the New Hampshire wells have MTBE levels that exceed the state-set standard of .5 microgram per liter?

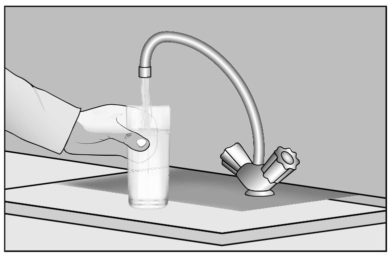
*Research Question 2*: Does the distribution of MTBE levels differ for public and private wells? Also, does the MTBE distribution differ for bedrock and unconsolidated aquifers?

*Research Question 3*: Does the combination of well type (public or private) and aquifer type (bedrock or unconsolidated) impact MTBE levels?

*Research Question 4*: Which of the variables in Table SIA15.1 — pH level, dissolved oxygen, industry percentage, well depth, and distance to tank — are most strongly associated with MTBE level?

Since the researchers discovered that the data on MTBE levels were not normally distributed, they applied nonparametric procedures to answer these questions. A discussion of the analyses follows.

**Research Question 1:**



The Environmental Protection Agency (EPA) has not set a federal standard for MTBE in public water supplies; however, several states have developed their own standards. New Hampshire has a standard of 13 micrograms per liter; that is, no groundwater well should have an MTBE level that exceeds 13 micrograms per liter. Also, only half the wells in the state should have MTBE levels that exceed .5 microgram per liter. This implies that the median MTBE level should be less than .5. Do the data collected by the researchers provide evidence to indicate that the median level of MTBE in New Hampshire groundwater wells is less than .5 microgram per liter? To answer this question, the researchers applied the sign test to the data saved in the MTBE file. The MINITAB printout is shown in Figure SIA15.1

**FIGURE SIA15.1**
MINITAB sign tests for MTBE data

**Sign Test for Median: MTBE**

```
Sign test of median =  0.5000 versus < 0.5000

         N  Below  Equal  Above       P  Median
MTBE   223    180      0     43  0.0000  0.2000
```

**FIGURE SIA 15.2**
SAS rank sum test for comparing public and private wells

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable MTBE
Classified by Variable WellClass

| Well Class | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| Private | 22 | 654.50 | 781.0 | 79.027002 | 29.750000 |
| Public | 48 | 1830.50 | 1704.0 | 79.027002 | 38.135417 |

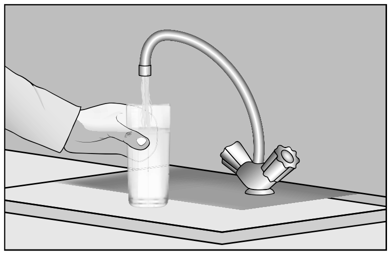
Average scores were used for ties.

Wilcoxon Two-Sample Test

| | |
|---|---|
| Statistic | 654.5000 |
| Normal Approximation | |
| Z | -1.5944 |
| One-Sided Pr < Z | 0.0554 |
| Two-Sided Pr > \|Z\| | 0.1108 |

We want to test $H_0$: $\eta = .5$ versus $H_a$: $\eta < .5$. According to the printout, 180 of the 223 sampled groundwater wells had MTBE levels below .5. Consequently, the test statistic value is $S = 180$. The one-tailed $p$-value for the test (highlighted on the printout) is .0000. Thus, the sign test is significant at $\alpha = .01$. Therefore, the data do provide sufficient evidence to indicate that the median MTBE level of New Hampshire groundwater wells is less than .5 microgram per liter.

**Research Question 2:**
One of the objectives of the study was to determine whether the level of MTBE contamination is different for private and public wells and for bedrock and unconsolidated aquifers. For this objective, the researchers focused on only the 70 sampled wells that had detectable levels of MTBE. They wanted to determine whether the distribution of MTBE levels in public wells is shifted above or below the distribution of MTBE levels in private wells and whether the distribution of MTBE levels in bedrock aquifers is shifted above or below the distribution of MTBE levels in unconsolidated aquifers.



To answer these questions, the researchers applied the Wilcoxon rank sum test for two independent samples. In the first analysis, public and private wells were compared; in the second analysis, bedrock and unconsolidated aquifers were compared. The SAS printouts for these analyses are shown in Figures SIA15.2 and SIA15.3, respectively. Both the test statistics and the two-tailed $p$-values are highlighted on the printouts.

**FIGURE SIA 15.3**
SAS rank sum test for comparing bedrock and unconsolidated aquifers

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable MTBE
Classified by Variable Aquifer

| Aquifer | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| Bedrock | 63 | 2345.50 | 2236.50 | 51.069646 | 37.230159 |
| Unconsoli | 7 | 139.50 | 248.50 | 51.069646 | 19.928571 |

Average scores were used for ties.

Wilcoxon Two-Sample Test

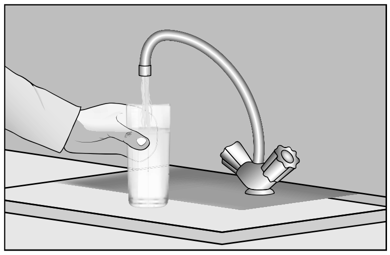
| | |
|---|---|
| Statistic | 139.5000 |
| Normal Approximation | |
| Z | -2.1245 |
| One-Sided Pr < Z | 0.0168 |
| Two-Sided Pr > \|Z\| | 0.0336 |

For the comparison of public and private wells in Figure SIA15.2, $p$-value $= .0118$. Thus, at $\alpha = .05$, there is insufficient evidence to conclude that the distribution of MTBE levels differs for public and private New Hampshire groundwater wells. Although public wells tend to have higher MTBE values than private wells (note the rank sums in Figure SIA15.2 ), the difference is not statistically significant.

For the comparison of bedrock and unconsolidated aquifers in Figure SIA15.3, $p$-value = .0336. At $\alpha = .05$, there is sufficient evidence to conclude that the distribution of MTBE levels differs for bedrock and unconsolidated aquifers. Furthermore, the rank sums shown in Figure SIA15.3 indicate that bedrock aquifers have the higher MTBE levels.

**Research Question 3:**

The environmental researchers also investigated how the combination of well class and aquifer affected the MTBE levels of the 70 wells that had detectable levels of MTBE. Although there are four possible combinations of well class and aquifer, data were available for only three: Private/bedrock, Public/bedrock, and Public/unconsolidated.



The distributions of MTBE levels for these three groups of wells were compared with the use of the Kruskal-Wallis nonparametric test for independent samples. The SAS printout for the analysis is shown in Figure SIA15.4. The test statistic is $H = 9.12$ and the $p$-value is .0104 (highlighted). At $\alpha = .05$, there is sufficient evidence to indicate differences in the distributions of MTBE levels of the three class-aquifer types. (However, at $\alpha = .01$, no significant differences are found.) On the basis of the mean rank sum scores shown on the printout, it appears that public wells with bedrock aquifers have the highest levels of MTBE contamination.

**FIGURE SIA 15.4**
SAS Kruskal-Wallis test for comparing MTBE levels of wells

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable MTBE
Classified by Variable wellaq

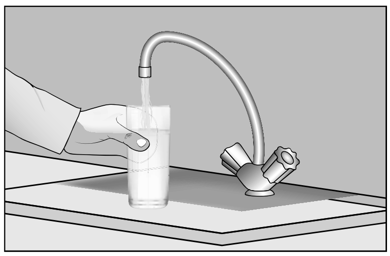
| wellaq | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| Private/Bedro | 22 | 654.50 | 781.00 | 79.027002 | 29.750000 |
| Public/Uncon | 7 | 139.50 | 248.50 | 51.069646 | 19.928571 |
| Public/Bedroc | 41 | 1691.00 | 1455.50 | 83.856064 | 41.243902 |

Average scores were used for ties.

Kruskal-Wallis Test

| | |
|---|---|
| Chi-Square | 9.1244 |
| DF | 2 |
| Pr > Chi-Square | 0.0104 |

**Research Question 4:**



The environmental researchers also wanted an estimate of the correlation between the MTBE level of a groundwater well and each of the other environmental variables listed in Table SIA15.1. Since the MTBE level is not normally distributed, they employed Spearman's rank correlation method. Also, because earlier analyses indicated that public and private wells have different MTBE distributions, the rank correlations were computed separately for each well class. SPSS printouts for this analysis are shown in Figures SIA15.5a–e. The values of $r_S$ (and associated $p$-values) are highlighted on the printouts. Our interpretations follow:

*MTBE vs. pH level* (Figure SIA15.5a). For private wells, $r_S = -.026$ (p-value $= .908$). Thus, there is a low negative association between MTBE level and pH level for private wells—an association that is not significantly different from 0 (at $\alpha = .10$). For public wells, $r_S = -.028$ ($p$-value $= .076$). Consequently, there is a low positive association (significant difference from 0 at $\alpha = .10$) for public wells between MTBE level and pH level.

**FIGURE SIA 15.5a**
SAS Spearman rank correlation test: MTBE and pH level

**Correlations**

| CLASS | | | | MTBE | PH |
|---|---|---|---|---|---|
| Private | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | -.026 |
| | | | Sig. (2-tailed) | . | .908 |
| | | | N | 22 | 22 |
| | | PH | Correlation Coefficient | -.026 | 1.000 |
| | | | Sig. (2-tailed) | .908 | . |
| | | | N | 22 | 22 |
| Public | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | .258 |
| | | | Sig. (2-tailed) | . | .076 |
| | | | N | 48 | 48 |
| | | PH | Correlation Coefficient | .258 | 1.000 |
| | | | Sig. (2-tailed) | .076 | . |
| | | | N | 48 | 48 |

*MTBE vs. Dissolved oxygen* (*Figure SIA15.5b*). For private wells, $r_s = -.086$ ($p$-value $= .702$). For public wells, $r_s = -.119$ ($p$-value $= .422$). Thus, there is a low positive association between MTBE level and dissolved oxygen for private wells, but a low negative association between MTBE level and dissolved oxygen for public wells. However, neither rank correlation is significantly different from 0 (at $\alpha = .10$).

*MTBE vs. Industry percentage* (*Figure SIA15.5c*). For private wells, $r_s = -.123$ ($p$-value $= .586$). This low negative association between MTBE level and industry percentage for private wells is not significantly different

**FIGURE SIA 15.5b**
SAS Spearman rank correlation test: MTBE and dissolved oxygen

**Correlations**

| CLASS | | | | MTBE | DISSOXY |
|---|---|---|---|---|---|
| Private | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | .086 |
| | | | Sig. (2-tailed) | . | .702 |
| | | | N | 22 | 22 |
| | | DISSOXY | Correlation Coefficient | .086 | 1.000 |
| | | | Sig. (2-tailed) | .702 | . |
| | | | N | 22 | 22 |
| Public | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | -.119 |
| | | | Sig. (2-tailed) | . | .422 |
| | | | N | 48 | 48 |
| | | DISSOXY | Correlation Coefficient | -.119 | 1.000 |
| | | | Sig. (2-tailed) | .422 | . |
| | | | N | 48 | 48 |

**FIGURE SIA 15.5c**
SAS Spearman rank correlation test: MTBE and dissolved oxygen

**Correlations**

| CLASS | | | | MTBE | INDUSTRY |
|---|---|---|---|---|---|
| Private | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | -.123 |
| | | | Sig. (2-tailed) | . | .586 |
| | | | N | 22 | 22 |
| | | INDUSTRY | Correlation Coefficient | -.123 | 1.000 |
| | | | Sig. (2-tailed) | .586 | . |
| | | | N | 22 | 22 |
| Public | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | .330* |
| | | | Sig. (2-tailed) | . | .022 |
| | | | N | 48 | 48 |
| | | INDUSTRY | Correlation Coefficient | .330* | 1.000 |
| | | | Sig. (2-tailed) | .022 | . |
| | | | N | 48 | 48 |

*. Correlation is significant at the 0.05 level (2-tailed).

**FIGURE SIA 15.5d**
SAS Spearman rank correlation test: MTBE and depth

**Correlations**

| CLASS | | | | MTBE | DEPTH |
|---|---|---|---|---|---|
| Private | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | -.410 |
| | | | Sig. (2-tailed) | . | .103 |
| | | | N | 22 | 17 |
| | | DEPTH | Correlation Coefficient | -.410 | 1.000 |
| | | | Sig. (2-tailed) | .103 | . |
| | | | N | 17 | 17 |
| Public | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | .444** |
| | | | Sig. (2-tailed) | . | .002 |
| | | | N | 48 | 46 |
| | | DEPTH | Correlation Coefficient | .444** | 1.000 |
| | | | Sig. (2-tailed) | .002 | . |
| | | | N | 46 | 46 |

**. Correlation is significant at the 0.01 level (2-tailed).

from 0 (at $\alpha = .10$). For public wells, $r_S = .330$ ($p$-value $= .022$). Consequently, there is a low positive association (significantly different from 0 at $\alpha = .10$) for public wells between MTBE level and industry percentage.

*MTBE vs. Depth of well* (*Figure SIA15.5d*). For private wells, $r_S = -.410$ ($p$-value $= .103$). This low negative association between MTBE level and depth for private wells is not significantly different from 0 (at $\alpha = .10$). For public wells, $r_S = .444$ ($p$-value $= .002$). Consequently, there is a low positive association (significantly different from 0 at $\alpha = .10$) for public wells between MTBE level and depth.

*MTBE vs. Distance from underground tank* (*Figure SIA15.5e*). For private wells, $r_S = -.136$ ($p$-value $= .547$). For public wells, $r_S = -.093$ ($p$-value $= .527$). Thus, there is a low positive association between MTBE level and distance for private wells, but a low negative association between MTBE level and distance for public wells. However, neither rank correlation is significantly different from 0 (at $\alpha = .10$).

In sum, the only significant rank correlations were for public wells, where the researchers discovered low positive associations of MTBE level with pH level, industry percentage, and depth of the well.

**FIGURE SIA 15.5e**
SAS Spearman rank correlation test: MTBE and distance

**Correlations**

| CLASS | | | | MTBE | DISTANCE |
|---|---|---|---|---|---|
| Private | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | .136 |
| | | | Sig. (2-tailed) | . | .547 |
| | | | N | 22 | 22 |
| | | DISTANCE | Correlation Coefficient | .136 | 1.000 |
| | | | Sig. (2-tailed) | .547 | . |
| | | | N | 22 | 22 |
| Public | Spearman's rho | MTBE | Correlation Coefficient | 1.000 | -.093 |
| | | | Sig. (2-tailed) | . | .527 |
| | | | N | 48 | 48 |
| | | DISTANCE | Correlation Coefficient | -.093 | 1.000 |
| | | | Sig. (2-tailed) | .527 | . |
| | | | N | 48 | 48 |

# Quick Review

## Key Terms

Distribution-free tests 839
Friedman $F_r$ statistic 865
Kruskal-Wallis $H$-test 859
Matched-pairs design 853
Nonparametrics 839
Nonparametric regression 869
Parametric statistical tests 839
Rank statistics 839
Rank sum 845
Rank tests 839
Sign test 840
Spearman's rank correlation coefficient 869
Test for location 840
Theil $C$ test 872
Wilcoxon rank sum test 845
Wilcoxon signed ranks test 853

## Key Formulas

| Test | Test Statistic | Large-Sample Approximation | |
|---|---|---|---|
| Sign | $S =$ number of sample measurements greater than (or less than) hypothesized median, $\tau_0$ | $Z = \dfrac{S - .5n}{.5\sqrt{n}}$ | 842 |
| Wilcoxon rank sum | $T_1 =$ rank sum of sample 1 or<br>$T_2 =$ rank sum of sample 2 | $Z = \dfrac{T_1 - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$ | 845, 850 |
| Wilcoxon signed ranks | $T_- =$ negative rank sum or<br>$T_+ =$ positive rank sum | $Z = \dfrac{T_+ - \dfrac{n(n + 1)}{4}}{\sqrt{\dfrac{n(n + 1)(2n + 1)}{24}}}$ | 854, 855 |
| Kruskal–Wallis | $H = \dfrac{12}{n(n + 1)} \sum \dfrac{T_j^2}{n_j} - 3(n + 1)$ | | 860 |
| Friedman | $F_r = \dfrac{12}{bk(k + 1)} \sum T_j^2 - 3b(k + 1)$ | | 865 |
| Spearman rank correlation (shortcut formula) | $r_s = 1 - \dfrac{6 \sum d_i^2}{n(n^2 - 1)}$<br>where $d_i =$ difference in ranks of $i$th observations for samples 1 and 2 | | 871 |
| Theil's Zero slope | $C = (-1)$(number of negative $y_i - y_j$ differences) $+ (1)$ (Number of positive $y_i - y_j$ differences) | | 873 |

## LANGUAGE LAB

| Symbol | Description |
|---|---|
| $\tau$ (tau) | Population median |
| $S$ | Test statistic for sign test (see Key Formulas) |
| $T_i$ | Sum of ranks of observations in sample $i$ |
| $T_L$ | Critical lower Wilcoxon rank sum value |
| $T_U$ | Critical upper Wilcoxon rank sum value |
| $T_+$ | Sum of ranks of positive differences of paired observations |
| $T_-$ | Sum of ranks of negative differences of paired observations |
| $T_0$ | Critical value of Wilcoxon signed ranks test |
| $H$ | Test statistic for Kruskal–Wallis test (see Key Formulas) |
| $F_r$ | Test statistic for Friedman test (see Key Formulas) |
| $r_s$ | Spearman's rank correlation coefficient (see Key Formulas) |
| $\rho$ (rho) | Population correlation coefficient |
| $C$ | Test statistic for Theil's zero slope test (see Key Formulas) |

## Chapter Summary Notes

- **Distribution-free tests**—do not rely on assumptions about the probability distribution of the sampled population
- **Nonparametrics**—distribution-free tests that are based on **rank statistics**
- *One-sample* nonparametric test for the population median—**sign test**
- Nonparametric test for *matched pairs*—**Wilcoxon rank test**
- Nonparametric test for a *completely randomized design*—**Kruskal–Wallis test**
- Nonparametric test for a *randomized block design*—**Friedman test**
- Nonparametric test for *rank correlation*—**Spearman's test**
- Nonparametric test for *zero slope*—**Theil's *C* test**

## Applied Supplementary Exercises

15.62 *Oil drill bit comparison.* Refer to Exercise 14.81 (p. 832) and the study to compare the speeds of three drill bits. Recall that five drilling sites were randomly assigned to each bit, and the rate of penetration (RoP) in feet per hour was recorded after drilling 3,000 feet at each site. Based on the information given in the table, can you conclude that the RoP probability distributions differ for at least two of the three drill bits? Test at the $\alpha = .05$ level of significance.

**DRILLBIT**

| PD-1 | IADC 1-2-6 | IADC 5-1-7 |
|---|---|---|
| 35.2 | 25.8 | 14.7 |
| 30.1 | 29.7 | 28.9 |
| 37.6 | 26.6 | 23.3 |
| 34.3 | 30.1 | 16.2 |
| 31.5 | 28.8 | 20.1 |

15.63 *Biting rates of flies.* The biting rate of a particular species of fly was investigated in a study reported in the *Journal of the American Mosquito Control Association* (Mar. 1995). Biting rate was defined as the number of flies biting a volunteer during 15 minutes of exposure. This species of fly is known to have a median biting rate of 5 bites per 15 minutes on Stanbury Island, Utah. However, it is theorized that the median biting rate is higher in bright, sunny weather. To test this theory, 122 volunteers were exposed to the flies during a sunny day on Stanbury Island. Of these volunteers, 95 experienced biting rates greater than 5.

a. Set up the null and alternative hypotheses for the test.
b. Calculate the approximate $p$-value of the test. (*Hint*: Use the normal approximation for a binomial probability.)
c. Make the appropriate conclusion at $\alpha = .01$.

15.64 *Biting rates of flies (continued).* Refer to Exercise 15.63. The effect of wind speeds in kilometers per hour (kph) on the biting rate of the flies on Stanbury Island, Utah, was investigated by exposing samples of volunteers to one of six wind speed conditions. The distributions of the biting rates for the six wind speeds were compared using the Kruskal–Wallis test. The rank sums of the biting rates for the six conditions are shown in the next table.

| Wind Speed (kph) | Number of Volunteers ($n_j$) | Rank Sum of Biting Rates ($R_j$) |
|---|---|---|
| $< 1$ | 11 | 1,804 |
| 1–2.9 | 49 | 6,398 |
| 3–4.9 | 62 | 7,328 |
| 5–6.9 | 39 | 4,075 |
| 7–8.9 | 35 | 2,660 |
| 9–20 | 21 | 1,388 |
| **Totals** | 217 | 23,653 |

*Source*: Strickman, D., et al. "Meteorological effects on the biting activity of *Leptoconops americanus* (Diptera: Ceratopogonidae)." *Journal of the American Mosquito Control Association*, Vol. II, No. 1, Mar. 1995, p. 17 (Table 1).

a. The researchers reported the test statistic as $H = 35.2$. Verify this value.
b. Find the rejection region for the test using $\alpha = .01$.
c. Make the proper conclusions.
d. The researchers reported that the $p$-value of the test is less than .01. Does this value support your inference in part **c**? Explain.

15.65 *Real-time scheduling with robots.* Refer to Exercise 8.104 (p. 439) and the study to compare human real-time

**THRUPUT**

| Task | Human Scheduler | Automated Method |
|---|---|---|
| 1 | 185.4 | 180.4 |
| 2 | 146.3 | 248.5 |
| 3 | 174.4 | 185.5 |
| 4 | 184.9 | 216.4 |
| 5 | 240.0 | 269.3 |
| 6 | 253.8 | 249.6 |
| 7 | 238.8 | 282.0 |
| 8 | 263.5 | 315.9 |

*Source:* Yih, Y., Liang, T., and Moskowitz, H. "Robot scheduling in a circuit board production line: A hybrid OR/ANN approach." *IEEE Transactions*, Vol. 25, No. 2, Mar. 1993, p. 31 (Table 1).

scheduling to an automated approach that utilizes computerized robots and sensing devices (*IEEE Transactions*, Mar. 1993). Recall that eight simulated scheduling tasks were performed by a human scheduler and by the automated system. The resulting throughput rates are shown in the table on p. 882. Compare the throughput rates of tasks scheduled by a human and the automated method with a nonparametric test. Use $\alpha = .01$.

15.66 *Breaking strength of sewer pipe.* The building specifications in a certain city require that the sewer pipe used in residential areas have a median breaking strength of more than 2,500 pounds per lineal foot. A manufacturer who would like to supply the city with sewer pipe has submitted a bid and provided the following additional information. An independent contractor randomly selected seven sections of the manufacturer's pipe and tested each for breaking strength. The results (pounds per lineal foot) are shown below. Is there sufficient evidence to conclude that the manufacturer's sewer pipe meets the required specifications? Use a significance level of $\alpha = .10$.

**SEWER**

| | | | | | | |
|---|---|---|---|---|---|---|
| 2,610 | 2,750 | 2,420 | 2,510 | 2,540 | 2,490 | 2,680 |

**MAINELAKE**

15.67 *Mercury poisoning in lakes.* Refer to the EPA study of mercury poisoning in Maine lakes, Exercise 10.77 (p. 552). Lakes can be classified into three trophic states: Oligotrophic lakes have a balance between decaying vegetation and living organisms; eutrophic lakes have a high decay rate in the top layer of water; and, mesotrophic lakes have a moderate amount of nutrients in the water. One goal of the study was to compare the mercury level distributions for the three types of Maine lakes. Data on mercury level (parts per million) and type of 118 Maine lakes are saved in the **MAINELAKE** file.

a. For each lake type, determine if the mercury levels are approximately normally distributed.
b. Given the result, part **a**, explain why a nonparametric analysis is appropriate.
c. Conduct the Kruskal–Wallis test to compare the mercury level distributions for the three types of Maine lakes. Use $\alpha = .05$.

15.68 *Acid rain study.* Refer to Exercise 14.77 (p. 883) and the study to determine the effects of acid rain on the acidity of soils in a natural ecosystem. Recall that experimental plots were irrigated with acid rain at two pH levels: 3.7 and 4.5. The acidity of the soil was then measured at three different depths: 0–15, 15–30, and 30–46 centimeters. Tests were conducted during three different time periods. The resulting soil pH values are reproduced in the next table. The main objective of the experiment was to compare the acidity of soil irrigated with pH 4.5 acid rain to the acidity of soil irrigated with pH 3.7 acid rain.

**ACIDRAIN**

| | | April 3 Acid Rain pH | | June 16 Acid Rain pH | | June 30 Acid Rain pH | |
|---|---|---|---|---|---|---|---|
| | | 3.7 | 4.5 | 3.7 | 4.5 | 3.7 | 4.5 |
| | *0–15 cm* | 5.33 | 5.33 | 5.47 | 5.47 | 5.20 | 5.13 |
| *Soil Depth* | *15–30 cm* | 5.27 | 5.03 | 5.50 | 5.53 | 5.33 | 5.20 |
| | *30–46 cm* | 5.37 | 5.40 | 5.80 | 5.60 | 5.33 | 5.17 |

*Source*: "Acid rain linked to growth of coal-fired power." *Florida Agricultural Research* 83, Vol. 2, No. 1, Winter 1983.

a. Use a nonparametric test to compare the soil pH values of the two treatments on April 3.
b. Use a nonparametric test to compare the soil pH values of the two treatments on June 16.
c. Use a nonparametric test to compare the soil pH values of the two treatments on June 30.
d. Comment on the validity of the tests in parts **a–c**.

15.69 *Mold contamination of corn.* A serious, drought-related problem for farmers is the spread of aflatoxin, a highly toxic substance caused by mold, which contaminates field corn. In higher levels of contamination, aflatoxin is potentially hazardous to animal and possibly human health. (Officials of the FDA have set a maximum limit of 20 parts per billion aflatoxin as safe for interstate marketing.) Three sprays, A, B, and C, have been developed to control aflatoxin in field corn. To determine whether differences exist among the sprays, 10 ears of corn are randomly chosen from a contaminated corn field and each is divided into three pieces of equal size. The sprays are then randomly assigned to the pieces for each ear of corn, thus setting up a randomized block design. The table gives the amount (in parts per billion) of aflatoxin present in the corn samples after spraying. Use the Friedman test to determine whether there are differences among the probability distributions of the amounts of aflatoxin present for the three sprays. Test at the $\alpha = .05$ level of significance.

**AFLATOXIN**

| | Spray | | | | Spray | | |
|---|---|---|---|---|---|---|---|
| Ear | A | B | C | Ear | A | B | C |
| 1 | 21 | 23 | 15 | 6 | 5 | 12 | 6 |
| 2 | 29 | 30 | 21 | 7 | 18 | 18 | 12 |
| 3 | 16 | 19 | 18 | 8 | 26 | 32 | 21 |
| 4 | 20 | 19 | 18 | 9 | 17 | 20 | 9 |
| 5 | 13 | 10 | 14 | 10 | 4 | 10 | 2 |

15.70 *Vehicle congestion study.* Refer to the *Journal of Engineering for Industry* (Aug. 1993) study of an automated warehouse, Exercise 10.73 (p. 550). Recall that the number of vehicles was varied and the congestion time (total time one vehicle blocked another) was recorded for a simulated automated warehouse. The data are reproduced in the accompanying table. Use Spearman's method to test for a correlation between congestion time ($y$) and number of vehicles ($x$) at $\alpha = .05$.

**WAREHOUSE**

| Number of Vehicles | Congestion Time, minutes | Number of Vehicles | Congestion Time, minutes |
|---|---|---|---|
| 1 | 0 | 9 | .02 |
| 2 | 0 | 10 | .04 |
| 3 | .02 | 11 | .04 |
| 4 | .01 | 12 | .04 |
| 5 | .01 | 13 | .03 |
| 6 | .01 | 14 | .04 |
| 7 | .03 | 15 | .05 |
| 8 | .03 | | |

*Source:* Pandit, R., and Palekar, U. S., "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry*, Transactions of the ASME, Vol. 115, Aug. 1993, p. 326 (Table 2).

15.71 *Drift-ratio of a building.* Refer to the *Microcomputers in Civil Engineering* study of lateral drift in a building, Exercise 14.76 (p. 830). The data shown in the table are the lateral displacements (in inches) estimated by three different computer programs at each of five different building levels. Compare the distributions of lateral displacement estimated by the three computer programs with the appropriate nonparametric test. Use $\alpha = .05$.

**STAAD**

| Level | STAAD-III (1) | STAAD-III (2) | Drift |
|---|---|---|---|
| 1 | .17 | .16 | .16 |
| 2 | 1.35 | 1.26 | 1.27 |
| 3 | 3.04 | 2.76 | 2.77 |
| 4 | 4.54 | 3.98 | 3.99 |
| 5 | 5.94 | 4.99 | 5.00 |

*Source:* Valles, R. E., et al. "Simplified drift evaluation of wall-frame structures." *Microcomputers in Civil Engineering*, Vol. 8, 1993, p. 242 (Table 2).

15.72 *Solving a mathematical program.* A hybrid algorithm for solving a polynomial zero–one mathematical program was presented in *IIE Transactions* (June 1990). The algorithm incorporates a mixture of pseudo-Boolean concepts and time-proven implicit enumeration procedures. Twenty-five random problems were solved using the hybrid algorithm; the times to solution (CPU time in seconds) are listed in the next table. Conduct a test to determine if more than half of random polynomial zero–one mathematical problems will require a solution time of 1 CPU second or less. Use $\alpha = .01$.

**MATHCPU**

| | | | | |
|---|---|---|---|---|
| .045 | 1.055 | .136 | 1.894 | .379 |
| .136 | .336 | .258 | 1.070 | .506 |
| .088 | .242 | 1.639 | .912 | .412 |
| .361 | 8.788 | .579 | 1.267 | .567 |
| .182 | .036 | .394 | .209 | .445 |

*Source:* Snyder, W. S., and Chrissis, J. W. "A hybrid algorithm for solving zero–one mathematical programming problems." *IIE Transaction*, Vol. 22, No. 2, June 1990, p. 166 (Table 1).

15.73 *PCBs in soil.* A preliminary study was conducted to obtain information on the background levels of the toxic substance polychlorinated biphenyl (PCB) in soil samples in the United Kingdom (*Chemosphere*, Feb. 1986). Such information could then be used as a benchmark against which PCB levels at waste disposal facilities in the United Kingdom can be compared. The accompanying table contains the measured PCB levels of soil samples taken at 14 rural and 15 urban locations in the United Kingdom (PCB concentration is measured in .0001 gram per kilogram of soil). From these preliminary results, the researchers reported "a significant difference between (the PCB levels) for rural areas . . . and for urban areas." Do the data support the researchers' conclusions? Test using $\alpha = .05$.

**PCB2**

| Rural | | Urban | |
|---|---|---|---|
| 3.5 | 5.3 | 24.0 | 11.0 |
| 8.1 | 9.8 | 29.0 | 49.0 |
| 1.8 | 15.0 | 16.0 | 22.0 |
| 9.0 | 12.0 | 21.0 | 13.0 |
| 1.6 | 8.2 | 107.0 | 18.0 |
| 23.0 | 9.7 | 94.0 | 12.0 |
| 1.5 | 1.0 | 141.0 | 18.0 |
| | | 11.0 | |

*Source:* Badsha, K., and Eduljee, G. "PCB in the U. K. environment—A preliminary survey." *Chemosphere*, Vol. 15, No. 2, Feb. 1986, p. 213 (Table 1). Copyright 1986, Pergamon Press, Ltd. Reprinted with permission.

15.74 *Synthetic fiber study.* Synthetic fibers (such as rayon, nylon, and polyester) account for approximately 70% of all fibers used by American mills in their production of textile products. An experiment was conducted to compare the breaking tenacity of synthetic fibers produced using two methods of spinning: wet spinning and dry spinning. Specimens of 10 different synthetic fibers were selected, and each was split into two filaments. One filament was processed using the wet spinning method, and the other using the dry spinning method; the breaking tenacity (grams per denier) of each filament was then measured. Do the data shown in the table provide sufficient evidence to indicate a difference in the breaking tenacity of synthetic fibers produced by the two methods? Test using $\alpha = .05$.

**SYNFIBER**

| Fiber | Dry Spinning | Wet Spinning |
|---|---|---|
| Acetate | 1.3 | 1.0 |
| Acrylic | 2.7 | 2.5 |
| Aramid | 4.8 | 4.7 |
| Modacrylic | 2.6 | 2.8 |
| Nylon | 4.5 | 4.2 |
| Olefin | 5.9 | 5.8 |
| Polyester | 4.5 | 4.3 |
| Rayon | 1.6 | 1.1 |
| Spandex | .7 | .9 |
| Triacetate | 1.3 | .9 |

15.75 *Resistivity of an alloy.* Refer to the *Corrosion Science* (Sept. 1993) study on the resistivity of an amorphous iron–boron–silicon alloy after crystallization, Exercise 10.74 (p. 551). Five alloy specimens were annealed at 700°C, each for a different length of time. The passivation potential—a measure of resistivity of the crystallized alloy—was then measured for each specimen. The experimental data are reproduced here.

**ALLOY**

| Annealing Time *x*, minutes | Passivation Potential *y*, mV |
|---|---|
| 10 | −408 |
| 20 | −400 |
| 45 | −392 |
| 90 | −379 |
| 120 | −385 |

*Source:* Chattoraj, I., et al. "Polarization and resistivity measurements of post-crystallization changes in amorphous Fe-B-Si alloys." *Corrosion Science*, Vol. 49, No. 9, Sept. 1993, p. 712 (Table 1).

a. Calculate Spearman's correlation coefficient between annealing time ($x$) and passivation potential ($y$). Interpret the result.

b. Use the result, part **a**, to test for a significant correlation between annealing time and passivation potential. Use $\alpha = .10$.

15.76 *Strength of masonry joints.* Refer to Exercise 10.79 (p. 553) and the straight-line model relating the mean shear strength $E(y)$ of masonry joints to precompression stress, $x$. To test this model, a series of stress tests were performed on solid bricks arranged in triplets and joined with mortar (*Proceedings of the Institute of Civil Engineers*, Mar. 1990). The precompression stress was varied for each triplet, and the ultimate shear load just before failure (called the shear strength) was recorded. The stress results for seven triplets (measured in N/mm$^2$) are shown in the next table. Conduct a nonparametric test of $H_0$: $\beta_1 = 0$ against the alternative $H_0$: $\beta_1 > 0$. Test using $\alpha = .05$.

**TRIPLETS**

| *Triplet Test* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| *Shear strength, y* | 1.00 | 2.18 | 2.24 | 2.41 | 2.59 | 2.82 | 3.06 |
| *Precompression stress, x* | 0 | .60 | 1.20 | 1.33 | 1.43 | 1.75 | 1.75 |

*Source:* Riddington, J. R., and Ghazali, M. Z. "Hypothesis for shear failure in masonry joints." *Proceedings of the Institute of Civil Engineers, Part 2*, Mar. 1990, Vol. 89, p. 96 (Figure 7).

15.77 *Impact of water temperature on fish.* The EPA wants to determine whether temperature changes in the ocean's water caused by a nuclear power plant will have a significant effect on the animal life in the region. Recently hatched specimens of a certain species of fish are randomly divided into four groups. The groups are placed in separate simulated ocean environments that are identical in every way except for water temperature. Six months later, the specimens are weighed. The results (in ounces) are given in the table. Do the data provide sufficient evidence to indicate that one (or more) of the temperatures tend(s) to produce larger weight increases than the other temperatures? Test using $\alpha = .10$.

**OCEANTEMP**

| *Water Temperature* | | | |
|---|---|---|---|
| 38°F | 42°F | 46°F | 50°F |
| 22 | 15 | 14 | 17 |
| 24 | 21 | 28 | 18 |
| 16 | 26 | 21 | 13 |
| 18 | 16 | 19 | 20 |
| 19 | 25 | 24 | 21 |
| | 17 | 23 | |

15.78 *Nickel alloy study.* Oil producers are interested in finding high-strength nickel alloys that are corrosion-resistant. Nickel alloys are especially susceptible to hydrogen embrittlement, a process that results when the alloy is cathodically charged in a sulfuric acid solution. To rate the performance of two incoloy alloys, 800 and 902, hydrogen charged tensile of each alloy were measured for the amount of ductility loss (recorded as a percentage reduction of area). The measurements for eight tensile specimens of each type are given in the table. Conduct a test to determine whether the probability distributions of ductility losses differ for the two nickel alloys. Use $\alpha = .05$.

**NICKEL2**

| *Alloy 800* | | *Alloy 902* | |
|---|---|---|---|
| 59.2 | 66.3 | 67.2 | 61.3 |
| 78.8 | 69.8 | 46.8 | 58.7 |
| 79.2 | 66.2 | 50.2 | 40.9 |
| 75.0 | 70.7 | 44.5 | 55.4 |

15.79 *Agent Orange and Vietnam Vets.* Agent Orange, the code name for a herbicide developed for the U.S. armed forces in the 1960s, was found to be extremely contaminated with TCDD, or dioxin. During the Vietnam War, an estimated 19 million gallons of Agent Orange was used to destroy the dense plant and tree cover of the Asian jungle. As a result of this exposure, many Vietnam veterans have dangerously high levels of TCDD in their blood and adipose (fatty) tissue. A study published in *Chemosphere* (Vol. 20, 1990) reported on the TCDD levels of 20 Massachusetts Vietnam vets who were possibly exposed to Agent Orange. The TCDD amounts (measured in parts per trillion) in both plasma and fat tissue of the 20 vets are listed in the accompanying table (left column).

**TCDD**

| Vet | Fat | Plasma |
|---|---|---|
| 1 | 4.9 | 2.5 |
| 2 | 6.9 | 3.5 |
| 3 | 10.0 | 6.8 |
| 4 | 4.4 | 4.7 |
| 5 | 4.6 | 4.6 |
| 6 | 1.1 | 1.8 |
| 7 | 2.3 | 2.5 |
| 8 | 5.9 | 3.1 |
| 9 | 7.0 | 3.1 |
| 10 | 5.5 | 3.0 |
| 11 | 7.0 | 6.9 |
| 12 | 1.4 | 1.6 |
| 13 | 11.0 | 20.0 |
| 14 | 2.5 | 4.1 |
| 15 | 4.4 | 2.1 |
| 16 | 4.2 | 1.8 |
| 17 | 41.0 | 36.0 |
| 18 | 2.9 | 3.3 |
| 19 | 7.7 | 7.2 |
| 20 | 2.5 | 2.0 |

*Source:* Schecter, A., et al. "Partitioning of 2,3,7,8-chlorinated dibenzo-*p*-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans," *Chemosphere*, Vol. 20, Nos. 7–9, 1990, pp. 954–955 (Tables I and II).

a. Medical researchers consider a TCDD level of 3 parts per trillion (ppt) to be dangerously high. Do the data provide evidence (at $\alpha = .05$) to indicate that the median level of TCDD in the fat tissue of Vietnam vets exceeds 3 ppt?
b. Repeat part **a** for plasma.
c. Medical researchers also are interested in comparing the TCDD levels in fat tissue and plasma for Vietnam veterans. Specifically, they want to determine whether the distribution of TCDD levels in fat is shifted above or below the distribution of TCDD levels in plasma. Conduct this analysis (at $\alpha = .05$) and make the appropriate inference.
d. Find the rank correlation between the TCDD level in fat tissue and the TCDD level in plasma. Is there sufficient evidence (at $\alpha = .05$) of a positive association between the two TCDD measures?

15.80 *Study of guppy migration.* In zoology, the phenomenon of fish moving excessively from one confined area to another is known as excessive transitory migration (ETM). To investigate the ETM of guppy populations, 40 adult female guppies were placed into the left compartment of an experimental aquarium tank that was divided in half by a glass plate. After the plate was removed, the numbers of fish passing through the slit from the left compartment to the right one, and vice versa, was monitored every minute for 30 minutes (*Zoological Science*, Vol. 6, 1989). If an equilibrium is reached, the researchers would expect the median number of fish remaining in the left compartment to be 20. The data for the 30 observations (i.e., numbers of fish in the left compartment at the end of the minute interval) are shown below. Use the large-sample sign test to determine whether the median is less than 20. Test using $\alpha = .05$.

**GUPPY**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 11 | 12 | 15 | 14 | 16 | 18 | 15 | 13 | 15 |
| 14 | 14 | 16 | 13 | 17 | 17 | 14 | 22 | 18 | 19 |
| 17 | 17 | 20 | 23 | 18 | 19 | 21 | 17 | 21 | 17 |

*Source*: Terami, H., and Watanabe, M. "Excessive transitory migration of guppy populations, III. Analysis of perception of swimming space and a mirror effect." Zoological Science, Vol. 6, 1989, p. 977, (Figure 2).

## Theoretical Supplementary Exercises

(*Note*: These exercises require the use of a computer and computer simulation techniques.)

15.81 Throughout this chapter we have omitted the theoretical derivations of the null distributions of the various nonparametric test statistics. However, we can use computer simulation to derive approximate rejection regions for the tests. Consider the problem of finding the approximate sampling distribution of the Wilcoxon rank sum statistic for the case $n_1 = n_2 = 10$.

a. Write a computer program that will randomly order the $n = n_1 + n_2 = 20$ ranks and compute the corresponding Wilcoxon rank sum $T_1$. This can be accomplished using a random number generator.

b. Write a computer program that will repeat the instructions of part **a** $N = 1{,}000$ times.

c. Construct a relative frequency distribution for the $N = 1{,}000$ computer-generated values of $T_1$ (refer to Chapter 2). This simulated distribution represents an approximation to the sampling distribution of $T_1$.

d. Use the simulated sampling distribution to determine the value $T_{.05}$, such that $P(T_1 \leq T_{.05}) = .05$. This value represents the one-tailed critical value of the Wilcoxon rank sum test for $\alpha = .05$.

15.82 Follow the steps outlined in Exercise 15.81 to find the approximate critical value (at $\alpha = .05$) of Spearman's test for rank correlation for the case $n = 10$. (*Hint*: In part **a** you will need to randomly order the $n = 10$ $y$ ranks and $n = 10$ $x$ ranks.)

CHAPTER 16

# Statistical Process and Quality Control

## OBJECTIVE

*To present some statistical procedures for monitoring the quality of a manufactured product and for controlling the quality of products shipped to consumers*

## CONTENTS



*STATISTICS IN ACTION*

Testing Jet Fuel Additive for Safety

## STATISTICS IN ACTION

### Testing Jet Fuel Additive for Safety

The American Society of Testing and Materials (ASTM) International provides standards and guidelines for materials, products, systems, and services. The Federal Aviation Administration (FAA) has a huge conglomerate of testing requirements for jet fuel safety that are spelled out in ASTM methods. This Statistics in Action involves an engineering firm that is developing a new method of surfactant detection in jet fuel.

Surfactants (surface active agents) are basically soaps that can form due to acids in the fuel but are more commonly caused by contamination from other products, such as engine cleaning additives. Although the surfactants do not directly cause problems, they reduce the ability of coalescing filters to remove water. Water in jet fuel carries bacteria that are deposited in tanks and engine components, causing major corrosion and engine damage.

The standard test for surfactants (described in ASTM Rule D-3948) is to use a miniature filter (Filter-A) with a pumping mechanism (Pump-A). A water/fuel mixture is pumped through the filter at a specific rate, and the amount of water that passes through the filter is detected with an optical transmittance test. Test measurements will typically yield a result between 80 and 85.

In an attempt to improve the precision of the surfactant test, the engineering firm compared the standard test (Pump-A with Filter-A) to three other pumping mechanism and filter option combinations—Pump-A with Filter-B, Pump-B with Filter-A, and Pump-B with Filter-B. Each day, a routine batch of jet fuel was created by adding 0.4 ppm of a surfactant solution. Twelve samples of the fuel were randomly selected and randomly divided into four groups of three samples each. The three samples in a group were tested for surfactants using one of the four pump/ filter combinations. Consequently, each day there were three test results for each pump/filter method. This pattern of sampling continued for over 100 days. The test measurements are saved in four JET files. (Data for the first 5 days of the sampling experiment are listed in Table SIA16.1).

The firm wants to monitor the results of the surfactant tests and determine if one of the test methods yields the most stable process. In the Statistics in Action Revisited section of this chapter, we show how to analyze the data using methods for quality and process control.

***Data Sets:***
**JETA-A, JETA-B, JETB-A, JETB-B**

**TABLE SIA16.1 Selected Data in the JET Files**

| Weekday | Month | Day | Sample | Pump-B Filter-A | Pump-A Filter-A | Pump-B Filter-B | Pump-A Filter-B |
|---|---|---|---|---|---|---|---|
| Tue | May | 9 | 1 | 76 | 84 | 85 | 85 |
| | | | 2 | 81 | 91 | 84 | 84 |
| | | | 3 | 81 | 86 | 84 | 88 |
| Wed | May | 10 | 1 | 84 | 92 | 87 | 92 |
| | | | 2 | 81 | 93 | 82 | 95 |
| | | | 3 | 86 | 94 | 85 | 90 |
| Thu | May | 11 | 1 | 83 | 94 | 82 | 90 |
| | | | 2 | 82 | 96 | 85 | 87 |
| | | | 3 | 79 | 92 | 84 | 81 |
| Fri | May | 12 | 1 | 81 | 96 | 81 | 90 |
| | | | 2 | 84 | 91 | 82 | 91 |
| | | | 3 | 83 | 96 | 88 | 92 |
| Mon | May | 15 | 1 | 80 | 90 | 87 | 94 |
| | | | 2 | 88 | 92 | 85 | 94 |
| | | | 3 | 87 | 91 | 86 | 84 |

## 16.1 Total Quality Management

When we think of product or service quality, we think of a set of characteristics that we expect a product to possess. We want lightbulbs to have a long life, paper towels to be strong and absorbent, service waiting time to be reasonably short, and a quarter-pound hamburger to weigh at least one-quarter pound. But producing a quality product is not an easy job. Variations in the characteristics of raw materials and workmanship tend to produce variations in product quality. The length of life of a lightbulb produced in an automated production line may differ markedly from the length of life of a bulb produced seconds later. Similarly, the strength of paper produced by a paper machine may vary from one point in time to another because of variations in the characteristics of the pulp fed into the machine. Consequently, it is vital that manufacturers monitor the quality of the product they produce.

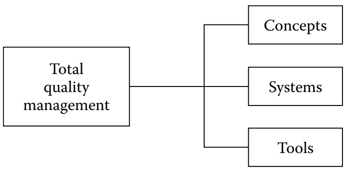



**FIGURE 16.1**
TQM components

Today, U.S. business leaders are promoting the concept of **total quality management (TQM)**. As shown in Figure 16.1, TQM has three key components: (1) concepts, (2) systems, and (3) tools. The concepts component of TQM includes a number of ideas that surround the total quality movement. These include *customer satisfaction, all work is a process, speak with data*, and *upstream management*. (Speaking with data is a particularly relevant concept for this text, since it involves measuring and monitoring process variables.)

The second component, systems, involves the notion of systems management. Systems such as *general management, market creation, product creation*, and *product supply* must be responsibly managed by the company's owners.

Finally, several tools are avaiable to implement a TQM program. These include *flowcharts, cause-and-effect diagrams*, and *statistical process control charts*.

All three components of Figure 16.1 are necessary to successfully implement a TQM methodology at a company. In this chapter, we focus on the statistical process control element of TQM. **Statistical process control (SPC)** allows engineers to understand and monitor process variation through **control charts**.

## 16.2 Variable Control Charts

Although TQM in U.S. business is a recent trend, the idea of a control chart to monitor process data was developed in 1924 by W. A. Shewhart. Control charts are constructed by plotting a product's quality variable over time in a sequence plot, as shown in Figure 16.2. The variable plotted can be either a quantitative characteristic (e.g., diameter of an eyescrew) or a qualitative attribute (e.g., defective or nondefective lightbulb) of a manufactured product. The power of this simple chart lies in its ability to separate two types of variation in a product quality characteristic: (1) variation due to **assignable causes** and (2) **random variation**.

**Definition 16.1**

A **control chart for a quality variable** is obtained by plotting the variable's measurements periodically over time.

Variations due to assignable causes are produced by such things as the wear in a metallic cutting machine, the wear in an abrasive wheel, changes in the humidity and temperature in the production area, worker fatigue, and so on. The effects of wear in cutting edges, abrasive surfaces, or changes in the environment are usually evidenced by gradual trends in a characteristic over time (see Figure 16.2a). In contrast, the raw material will often produce an abrupt change in the level of a quality characteristic (see Figure 16.2b). Quality control and production engineers attempt to identify trends or abrupt changes in a quality characteristic when they occur and to modify the process to reduce or eliminate this type of variation.

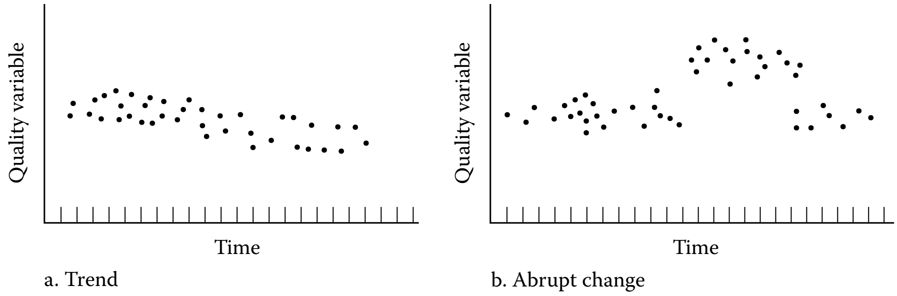



**FIGURE 16.2**
Plots of a quality characteristic that suggest variation due to assignable causes

Even when variation due to assignable causes is accounted for, measurements taken on a product quality characteristic tend to vary in a random manner from one point in time to another. This second category of variation—random (or chance) variation—is caused by minute and random changes in raw materials, worker behavior, and so on. Since some stable system of chance causes is inherent in any production process, this type of variation is accepted as the normal variation of the process. When the quality characteristics of a product are subject only to random variation, the process is said to be **in statistical control**.

**Definition 16.2**

The variation in a product characteristic that measures quality is due to either an **assignable cause** or **random (chance) variation**.

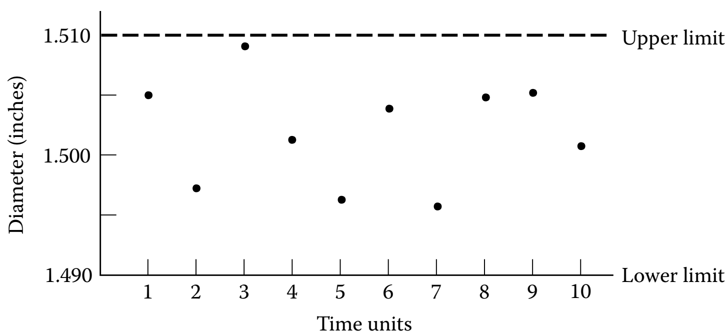
**Definition 16.3**

A production process is said to be **in control** when the quality characteristics of a product are subject only to random variation. Otherwise, the process is **out of control**.

To illustrate these ideas, consider a manufacturing process that produces shafts for an electrical motor. A quality control inspector might select one shaft every 10 minutes and measure its diameter. These measurements, plotted against time, provide visual evidence of the ability of the process to produce shafts with diameters that are subject only to random variation. For example, the diameters of 10 shafts might appear as shown in Figure 16.3. Although the diameters of these 10 shafts vary from one point in time to another, all fall within a set of *control limits* established by the manufacturer. The process appears to be "in control."

How are these **control limits** established? A widely used (and successful) technique is to monitor the process during a period when it is known to be in control and

**FIGURE 16.3**
A plot of the diameters of 10 motor shafts



calculate the mean and standard deviation of the sample quality measurements. Then, for future measurements, apply the $z$-score rule for detecting outliers (Section 2.6). We know that it is highly unlikely that a sample measurement will fall more than 3 standard deviations away from the mean. Consequently, if a quality measurement falls below $\bar{x} - 3s$ or above $\bar{x} + 3s$, we say the process is "out of control" and modifications in the production process may be necessary.

## Example 16.1

Variable Control Chart for Firing Pins

A corporation that manufactures field rifles for the Department of Defense operates a production line that turns out finished firing pins. To monitor the process, an inspector randomly selects a firing pin from the production line every 30 minutes and measures its length (in inches). The lengths for a sample of 20 firing pins obtained in this manner are provided in Table 16.1. Construct a control chart for the length of the firing pins. Is the process out of control?

FIREPIN

**TABLE 16.1 Lengths of Firing Pins, Example 16.1**

| Pin | Length | Pin | Length |
|---|---|---|---|
| 1 | 1.00 | 11 | 1.01 |
| 2 | .99 | 12 | .99 |
| 3 | .98 | 13 | .98 |
| 4 | 1.01 | 14 | .99 |
| 5 | 1.01 | 15 | .87 |
| 6 | .99 | 16 | 1.01 |
| 7 | 1.06 | 17 | .99 |
| 8 | .99 | 18 | .99 |
| 9 | .99 | 19 | .97 |
| 10 | 1.03 | 20 | .99 |

Solution

The first step in constructing the control chart is to calculate the mean and standard deviation of the sample firing pin lengths. These values, obtained using a computer, are shown in the SAS printout, Figure 16.4. You can see that the sample mean is $\bar{x} = .992$ and the sample standard deviation is $s = .035$.

Next, we plot the 20 sample measurements in time order, as shown in the MINITAB printout, Figure 16.5. Typically, three horizontal lines are drawn on the control chart. For variable control charts, the **center line** is the sample mean, $\bar{x}$. The center line estimates the mean value $\mu$ of the process. For this example, we estimate that the mean length of firing pins is .992 inch.

The two lines located above and below the center line on Figure 16.5 establish the **upper control limit (UCL)** and the **lower control limit (LCL)**, between which we expect the measurements to fall if the process is in control. For variable control charts, $\text{LCL} = \bar{x} - 3s$ and $\text{UCL} = \bar{x} + 3s$. Consequently, we have

$$\text{LCL} = .992 - 3(.035) = .887 \quad \text{and} \quad \text{UCL} = .992 + 3(.035) = 1.097$$

Note that the length measurement of firing pin #15 falls below the LCL. Thus, the process is "out of control," indicating possible trouble in the production line. In this situation, process engineers are usually assigned to determine the cause of the unusually small (or large) measurement.

The MEANS Procedure

Analysis Variable : LENGTH

| N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|
| 20 | 0.9920000 | 0.0348833 | 0.8700000 | 1.0600000 |

**FIGURE 16.4** SAS descriptive statistics for firing pin lengths

**Definition 16.4**

The **center line** for a variable control chart is the mean of the sample measurements, $\bar{x}$.

**Definition 16.5**

The **lower control limit (LCL)** and **upper control limit (UCL)** for a variable control chart are calculated as follows:

$$\text{LCL} = \bar{x} - 3s \qquad \text{UCL} = \bar{x} + 3s$$

where $s$ is the standard deviation of the sample measurements.

In concluding our discussion of an individual variable control chart (or, as it is often called, an **individuals control chart**), it is important to note that the chart describes the process as it is, not the way we want it to be. The process mean and control limits may differ markedly from the specifications set by the manufacturer of the product. For example, although a manufacturer may want to produce electrical motor

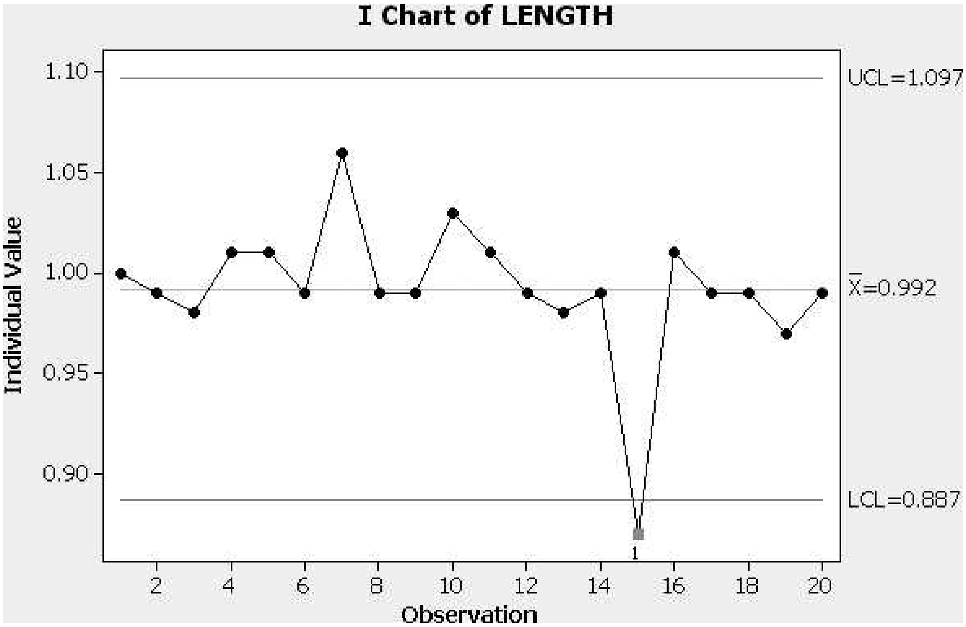



**FIGURE 16.5**
MINITAB variable control chart for firing pin lengths, Example 16.1

shafts with a diameter of 1.5 inches, the actual process mean will usually differ from 1.5, at least by some small amount. Also, the control limits obtained from the control chart are appropriate only for analyzing past data—that is, the data that were used in their calculation. Thus, they may require modification before they are applied to future production data. For example, in cases where the process is found to be out of control (Example 16.1), the control limits and center line would be modified by recalculating their values using only the sample measurements that fell within the original control limits. If the cause of the problem has been corrected, these new values serve as control limits for future data.

### Warning

The control limits obtained from a control chart are appropriate only for analyzing past data—that is, the data that were used in their calculation. They can be applied to future data only when the process is in control and/or the control limits are modified.

In this section, we presented control charts for a single quality variable (e.g., firing pin length). Control charts can also be constructed for process means, process variation, percent defectives, and number of defects per item. We present these types of control charts in the sections that follow.

## Applied Exercises

16.1 *Software file updates.* Refer to the *Software Quality Professional* (Nov. 2004) evaluation of the performance of a software engineering team's performance at Motorola, Inc., Exercise 5.42 (p. 209). Recall that the variable of interest was the number of updates to a file changed because of a problem report. The monthly number of updates reported by a particular team was recorded for 12 consecutive months. The data are shown in the accompanying table.

**SWUPDATE**

| Month | Number of Updates | Month | Number of Updates |
|---|---|---|---|
| 1 | 323 | 7 | 249 |
| 2 | 268 | 8 | 181 |
| 3 | 290 | 9 | 92 |
| 4 | 405 | 10 | 80 |
| 5 | 383 | 11 | 30 |
| 6 | 368 | 12 | 75 |

*Source*: Holmes, J. S. "Software measurement using SCM." *Software Quality Professional*, Vol. 7, No. 1, Nov. 2004 (Figure 5).

a. Locate the center line for a variable control chart of the data.
b. Locate the upper and lower control limits.
c. Is the process in control?

16.2 *New iron-making process.* Refer to the *Mining Engineering* (Oct. 2004) study of a new technology for producing high-quality iron nuggets directly from raw iron ore and coal, Exercise 10.25 (p. 505). For one phase of the study, the percentage change in the carbon content of the produced nuggets was measured at 4-hour intervals for 33 consecutive intervals. The data for the 33 time intervals are listed in the table. Construct and interpret a variable control chart for the data. Is the process in control?

**CARBON2**

| Interval | Carbon Change (%) | Interval | Carbon Change (%) |
|---|---|---|---|
| 1 | 3.25 | 18 | 3.55 |
| 2 | 3.30 | 19 | 3.48 |
| 3 | 3.23 | 20 | 3.42 |
| 4 | 3.00 | 21 | 3.40 |
| 5 | 3.51 | 22 | 3.50 |
| 6 | 3.60 | 23 | 3.45 |
| 7 | 3.65 | 24 | 3.75 |
| 8 | 3.50 | 25 | 3.52 |
| 9 | 3.40 | 26 | 3.10 |
| 10 | 3.35 | 27 | 3.25 |
| 11 | 3.48 | 28 | 3.78 |
| 12 | 3.50 | 29 | 3.70 |
| 13 | 3.25 | 30 | 3.50 |
| 14 | 3.60 | 31 | 3.40 |
| 15 | 3.55 | 32 | 3.45 |
| 16 | 3.60 | 33 | 3.30 |
| 17 | 2.90 | | |

*Source*: Hoffman, G., and Tsuge, O. "ITmk3—Application of a new ironmaking technology for the iron ore mining industry." *Mining Engineering*, Vol. 56, No. 9, October 2004 (Figure 5).

16.3 *Drug content assessment.* Refer to the *Analytical Chemistry* (Dec. 15, 2009) study of a method—called high-performance liquid chromatography—of determining the amount of drug in a tablet, Exercise 5.45 (p. 210). Assume the data in the accompanying table represent the drug concentrations (measured as a percentage) in tablets selected from two different production sites, one tablet selected each hour for 25 consecutive hours at each site. Use control chart methodology to monitor the drug content assessment process at each site. Is the process in statistical control at both sites?

**DRUGCON**

*Site 1*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 91.28 | 92.83 | 89.35 | 91.90 | 82.85 | 94.83 | 89.83 | 89.00 | 84.62 |
| 86.96 | 88.32 | 91.17 | 83.86 | 89.74 | 92.24 | 92.59 | 84.21 | 89.36 |
| 90.96 | 92.85 | 89.39 | 89.82 | 89.91 | 92.16 | 88.67 | | |

*Site 2*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 89.35 | 86.51 | 89.04 | 91.82 | 93.02 | 88.32 | 88.76 | 89.26 | 90.36 |
| 87.16 | 91.74 | 86.12 | 92.10 | 83.33 | 87.61 | 88.20 | 92.78 | 86.35 |
| 93.84 | 91.20 | 93.44 | 86.77 | 83.77 | 93.19 | 81.79 | | |

Note: Read across rows for consecutive drug concentration measurements.

16.4 *Monitoring urinary tract infections.* In *Quality Engineering* (Vol. 25, 2013), statisticians at Minitab, Inc., applied control chart methodology to monitor the time between discharges of male patients with hospital-acquired urinary tract infections (URI). The data (days between discharges) for $n = 54$ consecutive URI discharges at a large hospital system are shown in the table, p. 895.

a. Construct a control chart for the variable, time between discharges (in days). Does the process appear to be in control?
b. The statisticians demonstrated that time between discharges follows an exponential distribution rather than a normal distribution. Consequently, the control limits on the control chart, part **a**, will not yield the rare event probabilities that form the basis of control chart methodology. They derived the control limits for an exponential random variable with mean $\theta$ as follows:

$$\text{LCL} = .001351(\theta),\ \text{Center line} = \theta\ln(2),\ \text{UCL} = 6.60773(\theta)$$

Show that for an exponential random variable $Y$, $P(Y < \text{LCL}) = P(Y > \text{UCL}) \approx .001$ and $P(Y > \text{Center line}) = .5$.

c. Use the control limits given in part **b** and the fact that the mean time between discharges is .21 days to construct a control chart for the time between urinary tract infections. Does the process appear to be in control now?

**Data for Exercise 16.4**

**URI**

| DISCHARGE ORDER | DAYS | DISCHARGE ORDER | DAYS |
|---|---|---|---|
| 1 | .57014 | 28 | .01389 |
| 2 | .07431 | 29 | .03819 |
| 3 | .15278 | 30 | .46806 |
| 4 | .14583 | 31 | .22222 |
| 5 | .13889 | 32 | .29514 |
| 6 | .14931 | 33 | .53472 |
| 7 | .03333 | 34 | .15139 |
| 8 | .08681 | 35 | .52569 |
| 9 | .33681 | 36 | .07986 |
| 10 | .03819 | 37 | .27083 |
| 11 | .24653 | 38 | .04514 |
| 12 | .29514 | 39 | .13542 |
| 13 | .11944 | 40 | .08681 |
| 14 | .05208 | 41 | .40347 |
| 15 | .12500 | 42 | .12639 |
| 16 | .25000 | 43 | .18403 |
| 17 | .40069 | 44 | .70833 |
| 18 | .02500 | 45 | .15625 |
| 19 | .12014 | 46 | .24653 |
| 20 | .11458 | 47 | .04514 |
| 21 | .00347 | 48 | .01736 |
| 22 | .12014 | 49 | .08889 |
| 23 | .04861 | 50 | .05208 |
| 24 | .02778 | 51 | .02778 |
| 25 | .32639 | 52 | .03472 |
| 26 | .64931 | 53 | .23611 |
| 27 | .14931 | 54 | .35972 |

*Source:* Santiago, E. & Smith, J. "Control Charts Based on the Exponential Distribution: Adapting Runs Rules for the *t* Chart", *Quality Engineering,* Vol. 25, 2013 (Table B1).

16.5 *Bottle weights.* Each month, the quality control engineer at a bottle manufacturing company randomly selects one finished bottle from the production process at 20 points in time (days) and records the weight of each bottle (in ounces). The data for last month's inspection are provided in the next table.

a. Construct a variable control chart for the weights of the finished bottles.
b. Does the process appear to be in control for this particular month?

**BOTTLE**

| Day | Weight | Day | Weight |
|---|---|---|---|
| 1 | 5.6 | 11 | 6.2 |
| 2 | 5.7 | 12 | 5.9 |
| 3 | 6.1 | 13 | 5.2 |
| 4 | 6.3 | 14 | 6.0 |
| 5 | 5.2 | 15 | 6.3 |
| 6 | 6.0 | 16 | 5.8 |
| 7 | 5.8 | 17 | 6.1 |
| 8 | 5.8 | 18 | 6.2 |
| 9 | 6.4 | 19 | 5.3 |
| 10 | 6.0 | 20 | 6.0 |

16.6 *Molded-rubber expansion joints.* Molded-rubber expansion joints, used in heating and air-conditioning systems, are designed to have internal diameters of 5 inches. To monitor the manufacturing process, one joint was randomly selected each hour from the production line and its diameter (in inches) measured, for a period of 12 hours, as shown in the table. The data will be used to construct a variable control chart.

**RUBBERJNT**

| Hour | Diameter | Hour | Diameter |
|---|---|---|---|
| 1 | 5.08 | 7 | 5.02 |
| 2 | 4.88 | 8 | 4.91 |
| 3 | 4.99 | 9 | 5.06 |
| 4 | 5.04 | 10 | 4.92 |
| 5 | 5.00 | 11 | 5.01 |
| 6 | 4.83 | 12 | 4.92 |

a. Locate the center line for the variable control chart.
b. Locate upper and lower control limits.
c. Does the process appear to be in control?

16.7 *Rheostat knob insert.* A rheostat knob, produced by plastic molding, contains a metal insert. The fit of this knob into its assembly is determined by the distance from the back of the knob to the far side of a pin hole. To monitor the molding operation, one knob from each hour's production was randomly sampled and the dimension measured on each. The next table (p. 896) gives the distance measurements (in inches) for the first 27 hours the process was in operation.

a. Construct a variable control chart for the process.
b. Locate the center line, upper control limit, and lower control limit on the chart.
c. Does the process appear to be in control?

**Data for Exercise 16.7**

**KNOB**

| *Hour* | *Distance* | *Hour* | *Distance* | *Hour* | *Distance* | *Hour* | *Distance* |
|---|---|---|---|---|---|---|---|
| 1 | .140 | 8 | .143 | 15 | .144 | 22 | .139 |
| 2 | .138 | 9 | .141 | 16 | .140 | 23 | .140 |
| 3 | .139 | 10 | .142 | 17 | .137 | 24 | .134 |
| 4 | .143 | 11 | .137 | 18 | .137 | 25 | .138 |
| 5 | .142 | 12 | .137 | 19 | .142 | 26 | .140 |
| 6 | .136 | 13 | .142 | 20 | .136 | 27 | .145 |
| 7 | .142 | 14 | .137 | 21 | .142 | | |

## 16.3 Control Chart for Means: $\bar{x}$-Chart

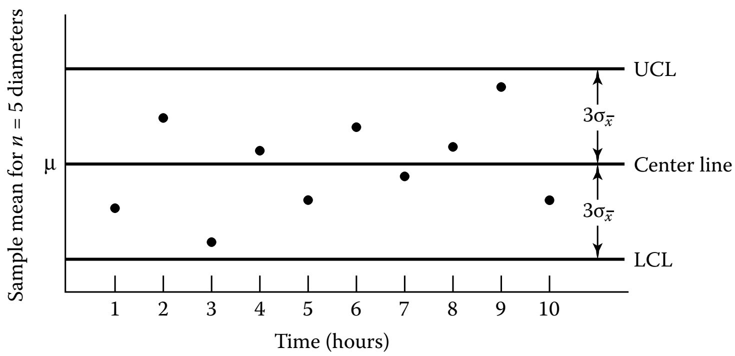
A control chart constructed to monitor a quantitative quality characteristic is usually based on random samples of several units of the product rather than on the characteristics of individual industrial units as shown in Figure 16.3. For example, the manufacturer of electrical shafts in Section 16.2 might select a sample of five shafts at the end of each hour. A plot showing the mean diameters of the samples, one mean corresponding to each point in time, is called a **control chart for means** or an **$\bar{x}$-chart**.*

In practice, control charts are constructed after a process has been adjusted to correct for assignable causes of variation and the process is deemed to be in control. When the process is in control, an $\bar{x}$-chart would show only random variation in the sample mean over time. Theoretically, $\bar{x}$ should vary about the process mean, $E(x) = \mu$, and fall within the limits $\mu \pm 3\sigma_{\bar{x}}$ or $\mu \pm 3\sigma/\sqrt{n}$ with a high probability. A control chart constructed for the means of samples of $n = 5$ motor shafts taken each hour might appear as shown in Figure 16.6.

An $\bar{x}$-chart, such as that shown in Figure 16.6, contains three horizontal lines. The **center line** establishes the mean value $\mu$ of the process. Although this value is usually unknown, it can be estimated by averaging a large number (for example, 20) of sample means obtained when the process is in control. For example, if we average the values of $k$ sample means, then

$$\text{Center line} = \bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

**FIGURE 16.6**
$\bar{x}$-chart for samples of $n = 5$ shaft diameters




*To be consistent with the symbols used in quality control literature, we will use $x$ (rather than $y$) to denote a quantitative quality characteristic variable.

The two lines located above and below the center line establish the **upper control limit** (**UCL**) and the **lower control limit** (**LCL**), between which we would expect the sample means to fall if the process is in control. They are located a distance of $3\sigma_{\bar{x}} = 3\sigma/\sqrt{n}$ above and below the center line.

The process standard deviation $\sigma$ is usually unknown, but it can be estimated from a large sample of data collected while the process is in control. Prior to the advent of statistical software, it was common to estimate $\sigma$ by first computing the sample range $R$, the difference between the largest and smallest sample measurements. The process standard deviation $\sigma$ was then estimated by dividing the average $\bar{R}$ of $k$ sample ranges by a constant $d_2$, the value of which depended on the sample size $n$:

$$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{\sum_{i=1}^{k} R_i/k}{d_2}$$

Since the control limits are located a distance of $3\sigma_{\bar{x}} = 3\sigma/\sqrt{n}$ above and below the center line, this distance was estimated to be

$$3\hat{\sigma}_{\bar{x}} = \frac{3(\bar{R}/d_2)}{\sqrt{n}} = \frac{3}{d_2\sqrt{n}}\bar{R} = A_2\bar{R}$$

where

$$A_2 = \frac{3}{d_2\sqrt{n}}$$

Values of $A_2$ and $d_2$ for sample sizes $n = 2$ to $n = 25$ are given in Table 19 of Appendix B.

## Location of Center Line and Control Limits for an $\bar{x}$-Chart

*Center line*: $\bar{\bar{x}} = \dfrac{\sum_{i=1}^{k} \bar{x}_i}{k}$

UCL: $\bar{\bar{x}} + A_2\bar{R}$

LCL: $\bar{\bar{x}} - A_2\bar{R}$

where

$k$ = Number of samples, each of size $n$

$\bar{x}_i$ = Sample mean for the $i$th sample

$R_i$ = Range of the $i$th sample

$$\bar{R} = \frac{\sum_{i=1}^{k} R_i}{k}$$

and $A_2$ is given in Table 19 of Appendix B.

(*Note*: For large samples (say, $n > 15$) collected from a process with no time trend, the upper and lower control limits may be computed as follows:

UCL: $\bar{\bar{x}} + 3s/\sqrt{n}$  LCL: $\bar{\bar{x}} - 3s/\sqrt{n}$

where $s$ is the standard deviation of all $nk$ sample measurements.)

Today, statistical software technology allows quality control inspectors to compute the means and standard deviations of the individual samples, as well as the means and standard deviations of the data contained in any set of $k$ samples. When no time trend exists (see Section 16.5) and the samples are large, the best estimate of $\sigma$ is then the standard deviation $s$ of the data contained in the $k$ sets of data.* Software programs calculate $\bar{\bar{x}}$ and $s$ and provide a printout of the control chart. However, the simplicity of calculating a sample range is not to be overlooked. Time, energy, and money often can be saved by reporting the sample ranges rather than $s$. Thus, in practice, $\bar{x}$-charts based on either $\bar{R}$ or $s$ are employed.

**Example 16.2**

$\bar{x}$-Chart for Shaft Diameters

Suppose the process for manufacturing electrical shafts is in control. At the end of each hour, for a period of 20 hours, the manufacturer selected a random sample of four shafts and measured the diameter of each. The measurements (in inches) for the 20 samples are recorded in Table 16.2. Construct a control chart for the sample means and interpret the results.

Solution

The first step in constructing an $\bar{x}$-chart is to compute the sample mean, $\bar{x}$, and range, $R$, for each of the 20 samples. These values are shown in the last two columns of Table 16.2.

Next, we calculate $\bar{\bar{x}}$, the average of the 20 sample means, and $\bar{R}$, the average of the 20 sample ranges, using SAS. These values, shown on the printout, Figure 16.7, are $\bar{\bar{x}} = 1.500425$ and $\bar{R} = .01985$.

SHAFTS

**TABLE 16.2 Samples of $n = 4$ Shaft Diameters, Example 16.2**

| Sample Number Number | Sample Measurements, inches | | | | Sample Mean $\bar{x}$ | Range $R$ |
|---|---|---|---|---|---|---|
| 1 | 1.505 | 1.499 | 1.501 | 1.488 | 1.4983 | .017 |
| 2 | 1.496 | 1.513 | 1.512 | 1.501 | 1.5055 | .017 |
| 3 | 1.516 | 1.485 | 1.492 | 1.503 | 1.4990 | .031 |
| 4 | 1.507 | 1.492 | 1.511 | 1.491 | 1.5003 | .020 |
| 5 | 1.502 | 1.491 | 1.501 | 1.502 | 1.4990 | .011 |
| 6 | 1.502 | 1.488 | 1.506 | 1.483 | 1.4948 | .023 |
| 7 | 1.489 | 1.512 | 1.496 | 1.501 | 1.4995 | .023 |
| 8 | 1.485 | 1.518 | 1.494 | 1.513 | 1.5025 | .033 |
| 9 | 1.503 | 1.495 | 1.503 | 1.496 | 1.4993 | .008 |
| 10 | 1.485 | 1.519 | 1.503 | 1.507 | 1.5035 | .034 |
| 11 | 1.491 | 1.516 | 1.497 | 1.493 | 1.4993 | .025 |
| 12 | 1.486 | 1.505 | 1.487 | 1.492 | 1.4925 | .019 |
| 13 | 1.510 | 1.502 | 1.515 | 1.499 | 1.5065 | .016 |
| 14 | 1.495 | 1.485 | 1.493 | 1.503 | 1.4940 | .018 |
| 15 | 1.504 | 1.499 | 1.504 | 1.500 | 1.5018 | .005 |
| 16 | 1.499 | 1.503 | 1.508 | 1.497 | 1.5018 | .011 |
| 17 | 1.501 | 1.493 | 1.509 | 1.491 | 1.4985 | .018 |
| 18 | 1.497 | 1.510 | 1.496 | 1.500 | 1.5008 | .014 |
| 19 | 1.503 | 1.526 | 1.497 | 1.500 | 1.5065 | .029 |
| 20 | 1.494 | 1.501 | 1.508 | 1.519 | 1.5055 | .025 |

*Grant and Leavenworth (1988) suggest using $s$ to estimate $\sigma$ when the sample size $n$ is greater than 15. For smaller samples, $\bar{R}/d_2$ will usually provide a better estimate.

The MEANS Procedure

| Variable | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|
| XBAR | 20 | 1.5004250 | 0.0039413 | 1.4925000 | 1.5065000 |
| RANGE | 20 | 0.0198500 | 0.0080934 | 0.0050000 | 0.0340000 |

**FIGURE 16.7**
SAS descriptive statistics for $\bar{x}$ and range of shaft diameters

The value of $\bar{\bar{x}} = 1.500425$ locates the center line on the control chart. To find upper and lower control limits we need the value of the control limit factor $A_2$, found in Table 19 of Appendix B. For $n = 4$ measurements in each sample, $A_2 = .729$. Then

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 1.500425 + (.729)(.01985) = 1.51489$$

$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 1.500425 - (.729)(.01985) = 1.48596$$

Using these limits, we construct the control chart for the sample means shown in the MINITAB printout, Figure 16.7. Note that all 20 sample means fall within the control limits.

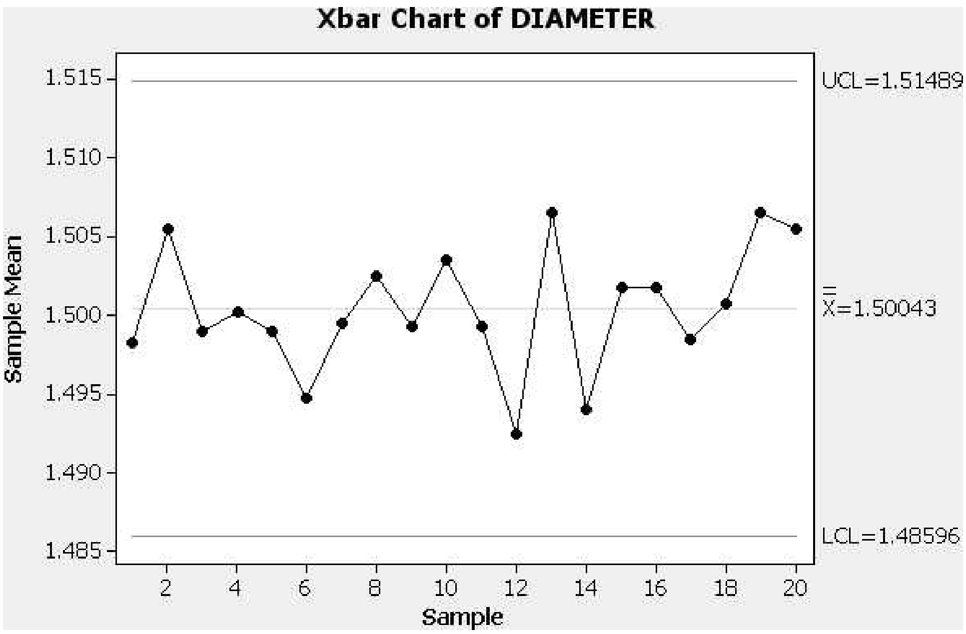



**FIGURE 16.8**
MINITAB $\bar{x}$-chart for shaft diameters, Example 16.2

The purpose of the $\bar{x}$-chart is to detect departures from process control. If the process is in control, the probability that a sample mean will fall within the control limits is very high. This result is due to the central limit theorem, which guarantees that the sampling distribution of $\bar{x}$ will be approximately normal for large samples. Consequently, the probability that $\bar{x}$ will fall within the control limits, i.e., $\pm\ 3\hat{\sigma}_{\bar{x}}$, is approximately .997. Therefore, a sample mean falling outside the control limits is taken as an indication of possible trouble in the production process. When this occurs, we say with a high degree of confidence that the process is out of control, and process engineers are usually assigned to determine the cause of the unusually large (or small) value of $\bar{x}$.

On the other hand, when all the sample means fall within the control limits (as in Figure 16.8), we say that the process is in control. However, we do not have the same degree of confidence in this statement as with the "out of control" conclusion above. In one sense, we are using the control chart to test the null hypothesis $H_0$: Process in control (i.e., no assignable causes of variation are present). As you recall from Chapter 9, we must be careful not to accept $H_0$ since the probability of a Type II error is unknown. In practice, when quality control engineers say "the process is in control," they really mean that "it pays to act as if no assignable causes of variation are present." In this situation, it is better to leave the process alone than to spend a great deal of time and money looking for trouble that may not exist.

Before concluding our discussion of the $\bar{x}$-chart, two important points must be made. First, *in practice, the $\bar{x}$-chart is typically used in conjunction with a chart that monitors the variation of the process, called an R-chart*. In fact, since the sample range (or standard deviation) is used to construct the $\bar{x}$-chart, *it is essential to examine an R-chart first to be sure that the process variation is stable*. The *R*-chart is the topic of the next section.

### Interpreting an $\bar{x}$-Chart

*Process "out of control"*: One or more of the sample means fall outside the control limits.* This indicates possible trouble in the production process and efforts should be made to determine the cause of the unusually large (or small) values of $\bar{x}$.

*Process "in control"*: All sample means fall within the control limits. Although assignable causes of variation may be present, it is better to leave the process alone than to look for trouble that may not exist.

The second point to be made about $\bar{x}$-charts, and control charts in general, is the importance of the sampling plan. Ideally, we want to choose samples of items over time so that we maximize the chance of detecting process change, if it exists. To do this, we choose **rational subgroups** (samples) of items so that the change in the process mean (if it exists) occurs *between* samples, *not within* samples (i.e., not during the period that the sample is drawn). The next example illustrates this point.

**Definition 16.6**

**Rational subgroups** are samples of items collected that maximize the chance that (1) quality measurements within each sample are similar and (2) quality measurements between samples differ.

### Example 16.3

Rational Subgrouping Strategy

Refer to the discussion of the process for manufacturing electrical motor shafts in Example 16.2. Suppose that the operations manager suspects that workers on the night shift are producing shafts with larger mean diameters than workers on the morning and afternoon shifts. The manager wants to use an $\bar{x}$-chart to determine whether the process mean has changed. Suggest a sampling plan for the manager that follows rational subgrouping strategy. That is, how should the samples of four shafts be selected so that the chance of detecting the shift in means is maximized?

Solution

Obviously, the control chart should be constructed using samples of shafts that are drawn within each shift. For example, the manager could sample four shafts each hour for 24 consecutive hours. Then the first eight samples will come from the morning

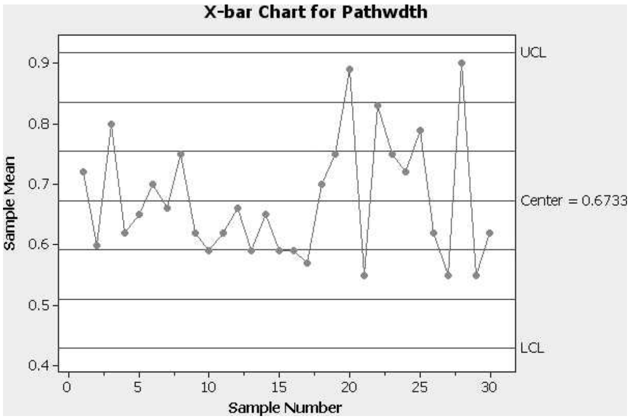
*In addition to this "one or more points beyond the control limits" rule, there are several other "pattern analysis" rules that help the analyst determine whether the process is out of control. For example, the process is also out of control if four of five consecutive points are beyond $\mu + 2\sigma_{\bar{x}}$ or $\mu - 2\sigma_{\bar{x}}$. Consult the references for a detailed discussion of these other rules of thumb.

shift, the next eight from the afternoon shift, and the last eight from the night shift. In this way, none of the samples would span shifts. (This is in contrast to a sample of, say, two shafts from the afternoon shift, and two from the night shift.) These 24 samples represent rational subgroups of shafts designed to maximize the chance of detecting the change in mean shaft diameters attributable to the night shift workers.

## Applied Exercises

16.8 *CPU of a computer chip.* The central processing unit (CPU) of a microcomputer is a computer chip containing millions of transistors. Connecting the transistors are slender circuit paths only .5 to .85 micron wide. A manufacturer of CPU chips knows that if the circuit paths are not .5–.85 micron wide, a variety of problems will arise in the chips' performance. The manufacturer sampled four CPU chips six times a day (every 90 minutes from 8:00 A.M. until 4:30 P.M.) for five consecutive days and measured the circuit path widths. These data and MINITAB were used to construct the $\bar{x}$-chart shown below.

a. Assuming that $\bar{R} = .335$, calculate the chart's upper and lower control limits, the upper and lower A–B boundaries, and the upper and lower B–C boundaries.
b. What does the chart suggest about the stability of the process used to put circuit paths on the CPU chip? Justify your answer.
c. Should the control limits be used to monitor future process output? Explain.




16.9 *Pain levels of ICU patients.* Various interventions are available for nurses to help relieve patients' pain (e.g., heat/cold applications, breathing exercises, massage). The journal *Research in Nursing & Health* (Vol. 35, 2012) demonstrated the utility of statistical process control in determining the effectiveness of a pain intervention. The researchers presented the following illustration. Pain levels (measured on a 100-point scale) were recorded for a sample of 10 intensive care unit (ICU) patients 24-hours post-surgery each week for 20 consecutive weeks. The accompanying table provides the means and ranges for each of the 20 weeks. To establish that the pain management process is "in statistical control", an $\bar{x}$-chart is constructed.

a. Compute the value of the center line for the $\bar{x}$-chart.
b. Compute the value of $\bar{R}$.
c. Compute the UCL and LCL for the $\bar{x}$-chart.
d. Plot the means for the 20 weeks on the $\bar{x}$-chart. Is the pain management process "in control"?
e. After the 20th week, a pain intervention occurred in the ICU. The goal of the intervention was to reduce the average pain level of ICU patients. To determine if the intervention was effective, the sampling of ICU patients was continued for eight more consecutive weeks. The mean pain levels of these patients were (in order): 71, 72, 69, 67, 66, 65, 64, and 62. Plot these means on the $\bar{x}$-chart. Do you detect a shift in the mean pain level of the patients following the intervention? Explain.

**ICUPAIN**

| Week | X-Bar | Range |
|---|---|---|
| 1 | 65 | 28 |
| 2 | 75 | 41 |
| 3 | 72 | 31 |
| 4 | 69 | 35 |
| 5 | 73 | 35 |
| 6 | 63 | 33 |
| 7 | 77 | 34 |
| 8 | 75 | 29 |
| 9 | 69 | 30 |
| 10 | 64 | 39 |
| 11 | 70 | 34 |
| 12 | 74 | 37 |
| 13 | 73 | 25 |
| 14 | 62 | 33 |
| 15 | 68 | 28 |
| 16 | 75 | 35 |
| 17 | 72 | 29 |
| 18 | 70 | 32 |
| 19 | 62 | 33 |
| 20 | 72 | 29 |

*Source:* Polit, D.F. & Chaboyer, W. "Statistical process control in nursing research", *Research in Nursing & Health,* Vol. 35, No. 1, 2012 (adapted from Figure 1).

16.10 *Quality control for irrigation data.* Most farmers budget water by using an irrigation schedule. The success of the schedule hinges on collecting accurate data on *evapotranspiration* (ETo), a term that describes the sum of evaporation and plant transpiration. The California Irrigation Management Information System (CIMIS) collects daily weather data (e.g., air temperature, wind speed, and vapor pressure) used to estimate ETo and supplies this information to farmers. Researchers at CIMIS demonstrated the use of quality-control charts to monitor daily ETo measurements (*IV International Symposium on Irrigation of Horticultural Crops,* Dec. 31, 2004). Daily minimum air temperatures (°C) collected hourly during the month of May at the Davis CIMIS station yielded the following summary statistics (where five measurements are collected each hour): $\bar{\bar{x}} = 10.16$ and $\bar{R} = 14.87°$.

a. Use the information provided to find the lower and upper control limits for an $\bar{x}$-chart.

b. Suppose that one day in May the mean air temperature at the Davis CIMIS station was recorded as $\bar{x} = 20.3°$. How should the manager of the station respond to this observation?

16.11 *Lengths of firing pins.* A corporation that manufactures field rifles for the Department of Defense operates a production line that turns out finished firing pins. To monitor the process, an inspector randomly selects five firing pins from the production line, measures their lengths (in inches), and repeats this process at 30-minute intervals over a 5-hour period.

a. Use the data for the 10 time periods listed in the table below to calculate the center line for an $\bar{x}$-chart.

b. Compute upper and lower control limits.

c. Calculate and plot the 10 sample means to form an $\bar{x}$-chart for the firing pin lengths.

d. Suppose the Defense Department's specification for the firing pins is that they be 1.00 inch plus or minus .08 inch in length. Does the manufacturing process appear to be in control?

**FIREPINS**

| 30-Minute Interval | Firing Pin Lengths | | | | |
|---|---|---|---|---|---|
| 1 | 1.05 | 1.03 | .99 | 1.00 | 1.03 |
| 2 | .93 | .96 | 1.01 | .98 | .97 |
| 3 | 1.02 | .99 | .99 | 1.00 | .98 |
| 4 | .98 | 1.01 | 1.02 | .99 | .97 |
| 5 | 1.02 | .99 | 1.04 | 1.07 | .98 |
| 6 | 1.05 | .98 | .96 | .91 | 1.02 |
| 7 | .92 | .95 | 1.00 | .99 | 1.01 |
| 8 | 1.06 | .98 | .98 | 1.04 | 1.00 |
| 9 | .97 | .99 | .99 | .98 | 1.01 |
| 10 | 1.00 | .96 | 1.02 | 1.03 | .99 |

16.12 *Molded-rubber expansion joints.* Refer to the production of molded-rubber expansion joints used in heating and air-conditioning systems, Exercise 16.6. (p. 895). To monitor the mean of the manufacturing process, eight joints (rather than one joint) were randomly selected from the production line and their diameters (in inches) measured each hour, for a period of 12 hours, as shown in the table. The data for the 12 samples will be used to construct an $\bar{x}$-chart.

a. Locate the center line for the $\bar{x}$-chart.

b. Locate upper and lower control limits.

c. Calculate and plot the 12 sample means to produce an $\bar{x}$-chart for the joint diameters. Does the process appear to be in control?

**RUBBERJNT2**

| Hour | Molded-Rubber Expansion Joint Diameters | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.08 | 5.01 | 4.99 | 4.93 | 4.98 | 5.00 | 5.04 | 4.97 |
| 2 | 4.88 | 5.10 | 4.93 | 5.02 | 5.06 | 4.99 | 4.92 | 4.91 |
| 3 | 4.99 | 5.00 | 5.02 | 5.01 | 5.03 | 4.92 | 4.97 | 5.01 |
| 4 | 5.04 | 4.96 | 5.01 | 5.00 | 5.00 | 4.98 | 4.91 | 4.96 |
| 5 | 5.00 | 4.93 | 4.94 | 5.02 | 5.01 | 4.97 | 5.08 | 5.11 |
| 6 | 4.83 | 4.92 | 4.96 | 4.91 | 5.01 | 5.03 | 4.93 | 5.00 |
| 7 | 5.02 | 5.01 | 4.96 | 4.98 | 5.00 | 5.07 | 4.94 | 5.01 |
| 8 | 4.91 | 5.00 | 4.97 | 5.03 | 5.02 | 4.99 | 4.98 | 4.99 |
| 9 | 5.06 | 5.04 | 4.99 | 5.02 | 4.97 | 5.00 | 5.01 | 5.01 |
| 10 | 4.92 | 4.98 | 5.01 | 5.01 | 4.97 | 5.00 | 5.02 | 4.93 |
| 11 | 5.01 | 5.00 | 5.02 | 4.98 | 4.99 | 5.00 | 5.01 | 5.01 |
| 12 | 4.92 | 5.12 | 5.06 | 4.93 | 4.98 | 5.02 | 5.04 | 4.97 |

16.13 *Selecting the best wafer slicing machine.* Silicon wafer slicing is a critical step in the production of semiconductor devices (e.g., diodes, solar cells, transistors). Yuanpei (China) University researchers used control charts to aid in selecting the best silicon wafer slicing machine (*Computers & Industrial Engineering,* Vol. 52, 2007). Samples of $n = 2$ wafers were sliced each hour for 67 consecutive hours and bow measurements (a measure of precision) were recorded. The resulting $\bar{x}$-chart for one of the machines tested revealed that the cutting process was out of control on the 19th, 40th, and 59th hour. For each of these three hours, the mean bow measurement fell above the upper control limit. Assume the mean bow measurements are normally distributed.

a. If the process is in control, what is the probability that a mean bow measurement for a randomly selected hour will fall above the upper control limit?

b. If the process is in control, what is the probability that 3 of 67 mean bow measurements fall above the upper control limit?

16.14 *Detecting under-reported emissions.* The Environmental Protection Agency (EPA) regulates the level of carbon dioxide ($CO_2$) emissions. Periodically these emissions measurements are under-reported due to leakage or faulty equipment. Such problems are often detected only by an

## Data for Exercise 16.14

**CO2**

Daily Average $CO_2$ Measurements for 30 Consecutive Days

| *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12.9 | 13.2 | 13.4 | 13.3 | 13.1 | 13.2 | 13.1 | 13.0 | 12.5 | 12.5 | 12.7 | 12.8 | 12.7 | 12.9 | 12.0 |
| *16* | *17* | *18* | *19* | *20* | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* |
| 12.9 | 12.8 | 12.7 | 13.2 | 13.2 | 13.3 | 13.0 | 13.0 | 13.2 | 13.2 | 13.4 | 13.1 | 13.3 | 13.4 | 13.4 |

expensive test (RATA) that is typically conducted only once per year. Just recently, the EPA began applying an automated control chart methodology to detect undermeasurement of emissions data (*EPRI CEM Users Group Conference,* Nashville, TN, May 13, 2008). Each day, the EPA collects emissions data by measuring $CO_2$ concentration for each of 6 randomly selected hours. The daily average $CO_2$ levels for each of 30 days are shown in the above table. The EPA considers these values to truly represent emissions levels because the RATA test was recently performed and showed no problems with under-reporting. The lower and upper control limits for the averages were established as LCL $= 12.26$ and UCL $= 13.76$.

a. Construct a control chart for the daily average $CO_2$ levels.
b. Based on the control chart, describe the behavior of the measurement process.
c. The following average $CO_2$ levels were determined for a later 10-day period: 12.7, 12.1, 12.0, 12.0, 11.8, 11.7, 11.6, 11.7, 11.8, 11.7. Make an inference about the potential under-reporting of the emissions data for this 10-day period.

16.15 *Rheostat knob insert.* Refer to the manufacture of a rheostat knob, Exercise 16.7 (p. 895). To monitor the process mean, five knobs from each hour's production were randomly sampled and the distance from the back of the knob to the far side of a pin hole was measured on each. The measurements (in inches) for the first 27 hours the process was in operation are shown in the table below.

a. Construct an $\bar{x}$-chart for the process.
b. Locate the center line, upper control limit, and lower control limit on the $\bar{x}$-chart.
c. Does the process appear to be in control?

**CHUNKY**

16.16 *Chunky data.* BPI Consulting, a leading provider of statistical process control software and training in the United States, recently alerted its clients to problems with "chunky" data. In an April 2007 report, BPI Consulting identified "chunky" data as data that result when the range between possible values of the variable of interest becomes too large. This typically occurs when the data are rounded. For example, a company monitoring the time it takes shipments to arrive from a given supplier rounded

**KNOB2**

| *Hour* | *Distance Measurements* | | | | | *Hour* | *Distance Measurements* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .140, | .143, | .137, | .134, | .135 | 15 | .144, | .142, | .143, | .135, | .145 |
| 2 | .138, | .143, | .143, | .145, | .146 | 16 | .140, | .132, | .144, | .145, | .141 |
| 3 | .139, | .133, | .147, | .148, | .139 | 17 | .137, | .137, | .142, | .143, | .141 |
| 4 | .143, | .141, | .137, | .138, | .140 | 18 | .137, | .142, | .142, | .145, | .143 |
| 5 | .142, | .142, | .145, | .135, | .136 | 19 | .142, | .142, | .143, | .140, | .135 |
| 6 | .136, | .144, | .143, | .136, | .137 | 20 | .136, | .142, | .140, | .139, | .137 |
| 7 | .142, | .147, | .137, | .142, | .138 | 21 | .142, | .144, | .140, | .138, | .143 |
| 8 | .143, | .137, | .145, | .137, | .138 | 22 | .139, | .146, | .143, | .140, | .139 |
| 9 | .141, | .142, | .147, | .140, | .140 | 23 | .140, | .145, | .142, | .139, | .137 |
| 10 | .142, | .137, | .145, | .140, | .132 | 24 | .134, | .147, | .143, | .141, | .142 |
| 11 | .137, | .147, | .142, | .137, | .135 | 25 | .138, | .145, | .141, | .137, | .141 |
| 12 | .137, | .146, | .142, | .142, | .140 | 26 | .140, | .145, | .143, | .144, | .138 |
| 13 | .142, | .142, | .139, | .141, | .142 | 27 | .145, | .145, | .137, | .138, | .140 |
| 14 | .137, | .145, | .144, | .137, | .140 | | | | | | |

*Source*: Grant, E. L., and Leavenworth, R. S., *Statistical Quality Control*, 5th ed. New York McGraw-Hill, 1950 (Table 1-2). Reprinted with permission.

off the data to the nearest day. To show the effect of chunky data on a control chart, BPI Consulting considered a process with a quality characteristic that averages about 100. Data on the quality characteristic for a random sample of three observations collected each hour for 40 consecutive hours are given in the accompanying table. (*Note:* BPI Consulting cautions its clients that out-of-control data points in this example were actually due to the measurement process and not to an "out-of-control" process.)

a. Show that the process is "in control," according to Rule 1, by constructing an $\bar{x}$-chart for the data.
b. Round each measurement in the data set to a whole number and then form an $\bar{x}$-chart for the rounded data. What do you observe?

| *Sample* | *Quality Levels* | | | *Sample* | *Quality Levels* | | |
|---|---|---|---|---|---|---|---|
| 1 | 99.69 | 99.73 | 99.81 | 21 | 99.43 | 99.63 | 100.08 |
| 2 | 98.67 | 99.47 | 100.20 | 22 | 100.04 | 99.71 | 100.40 |
| 3 | 99.93 | 99.97 | 100.22 | 23 | 101.08 | 99.84 | 99.93 |
| 4 | 100.58 | 99.40 | 101.08 | 24 | 99.98 | 99.50 | 100.25 |
| 5 | 99.28 | 99.48 | 99.10 | 25 | 101.18 | 100.79 | 99.56 |
| 6 | 99.06 | 99.61 | 99.85 | 26 | 99.24 | 99.90 | 100.03 |
| 7 | 99.81 | 99.78 | 99.53 | 27 | 99.41 | 99.18 | 99.39 |
| 8 | 99.78 | 100.10 | 99.27 | 28 | 100.84 | 100.47 | 100.48 |
| 9 | 99.76 | 100.83 | 101.02 | 29 | 99.31 | 100.15 | 101.08 |
| 10 | 100.20 | 100.24 | 99.85 | 30 | 99.65 | 100.05 | 100.12 |
| 11 | 99.12 | 99.74 | 100.04 | 31 | 100.24 | 101.01 | 100.71 |
| 12 | 101.58 | 100.54 | 100.53 | 32 | 99.08 | 99.73 | 99.61 |
| 13 | 101.51 | 100.52 | 100.50 | 33 | 100.30 | 100.02 | 99.31 |
| 14 | 100.27 | 100.77 | 100.48 | 34 | 100.38 | 100.76 | 100.37 |
| 15 | 100.43 | 100.67 | 100.53 | 35 | 100.48 | 99.96 | 99.72 |
| 16 | 101.08 | 100.54 | 99.89 | 36 | 99.98 | 100.30 | 99.07 |
| 17 | 99.63 | 100.77 | 99.86 | 37 | 100.25 | 99.58 | 101.27 |
| 18 | 99.29 | 99.49 | 99.37 | 38 | 100.49 | 100.16 | 100.86 |
| 19 | 99.89 | 100.75 | 100.73 | 39 | 100.44 | 100.53 | 99.84 |
| 20 | 100.54 | 101.51 | 100.54 | 40 | 99.45 | 99.41 | 99.27 |

## 16.4 Control Chart for Process Variation: *R*-Chart

In quality control, we want to control not only the mean value of some quality characteristic but also its variability. An increase in the process standard deviation $\sigma$ means that the quality characteristic variable will vary over a wider range, thereby increasing the probability of producing an inferior product. Consequently, a process that is in control generates data with a relatively constant process mean $\mu$ and standard deviation $\sigma$.

The variation in a quality characteristic is monitored using a **range chart** or ***R*-chart**. Thus, in addition to calculating the mean $\bar{x}$ for each sample, we also calculate and plot the sample range $R$. As with an $\bar{x}$-chart, an $R$-chart also contains a center line and lines corresponding to the upper and lower control limits.*

The expected value and standard deviation of the sample range are

$$E(R) = d_2\sigma \quad \text{and} \quad \sigma_R = d_3\sigma$$

*We could also monitor process variation by plotting sample standard deviations in an ***S*-chart**. However, in this chapter we focus on just the $R$-chart because (1) when using samples of size 9 or less, the $S$-chart and the $R$-chart reflect about the same information, and (2) the $R$-chart is used much more widely by engineers than the $S$-chart (primarily because the sample range is easier to calculate than the sample standard deviation). For more information on $S$-charts, consult the references for this chapter.

where $d_2$ and $d_3$ are constants (see Table 19 of Appendix B) that depend on the sample size $n$. Therefore, we would locate the center line of the $R$-chart at $d_2\sigma$, where, if $\sigma$ is unknown, $E(R)$ is estimated by the mean $\overline{R}$ of the ranges of $k$ samples.*

### Location of Center Line and Control Limits for an *R*-Chart

*Center line:* $\overline{R}$

UCL: $D_4\overline{R}$

LCL: $D_3\overline{R}$

where

$$k = \text{Number of samples, each of size } n$$

$$R_i = \text{Range of the } i\text{th sample}$$

$$\overline{R} = \frac{\sum_{i=1}^{k} R_i}{k}$$

and $D_3$ and $D_4$ are given in Table 19 of Appendix B for $n = 2$ to $n = 25$.

The upper and lower control limits are located a distance $3\sigma_R = 3d_3\sigma$ above and below the center line. Using $\overline{R}/d_2$ to estimate $\sigma$, we locate the upper and lower control limits as follows:

$$\text{UCL:} \quad \overline{R} + 3\frac{d_3}{d_2}\overline{R} = \left(1 + 3\frac{d_3}{d_2}\right)\overline{R} = D_4\overline{R}$$

where

$$D_4 = 1 + 3\frac{d_3}{d_2}$$

and

$$\text{LCL: } \overline{R} - 3\frac{d_3}{d_2}\overline{R} = \left(1 - 3\frac{d_3}{d_2}\right)\overline{R} = D_3\overline{R}$$

where

$$D_3 = 1 - 3\frac{d_3}{d_2}$$

Values of $D_3$ and $D_4$ have been computed for sample sizes of $n = 2$ to $n = 25$, and appear in Table 19 of Appendix B.

*As an alternative procedure, we could estimate $\sigma$ using the standard deviation of all the data contained in the $k$ samples.

**Example 16.4**

R-chart for Shaft Diameters

Refer to the problem of monitoring the manufacturing of electrical shafts, Example 16.2. Recall that the manufacturer selected a sample of four shafts each hour, for a period of 20 hours, and measured the diameter of each. Assuming the process is in control, construct and interpret an *R*-chart for process variation.

Solution

In Example 16.2 we calculated the mean of the 20 sample ranges to be $\overline{R} = .01985$. This value is the center line. For $n = 4$, the values of $D_3$ and $D_4$ given in Table 19 of Appendix B are $D_3 = 0$ and $D_4 = 2.282$. Then the upper and lower control limits for the *R*-chart are:

$$\text{UCL} = D_4\overline{R} = (2.282)(.01985) = .0453$$

$$\text{LCL} = D_3\overline{R} = (0)(.01985) = 0$$

An $\overline{R}$-chart for the 20 sample ranges of Table 16.2 is shown in the MINITAB printout, Figure 16.9.

To monitor the variation in shaft diameters produced by the manufacturing process, a quality control engineer would check to determine that the sample range does not exceed the UCL of .0453 inch. (Since the LCL is 0, no diameter can fall below this value.)

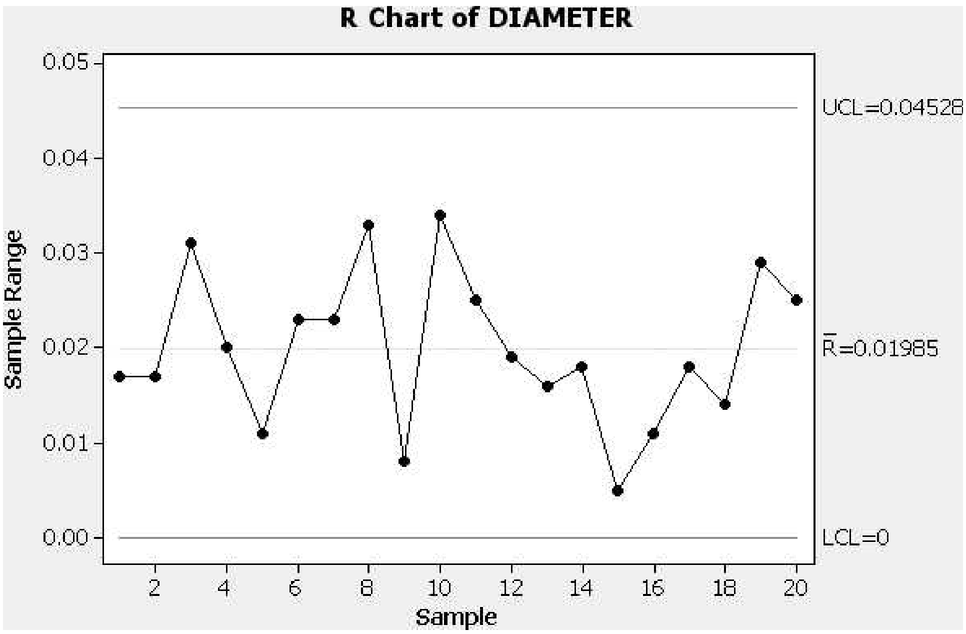



**FIGURE 16.9**
MINITAB *R*-chart for shaft diameters, Example 16.4

The practical implications to be derived from an *R*-chart are similar to those associated with an $\bar{x}$-chart. Values of *R* that fall outside of the control limits are suspect and suggest a possible change in the process. Trends in the sample range may also indicate problems, such as wear within a machine. (We investigate this type of problem in the next section.) As in the case of the $\bar{x}$-chart, the *R*-chart can provide an indication of possible trouble in a process. A process engineer then attempts to locate the difficulty, if in fact it exists.

### Interpreting an $\overline{R}$-Chart

*Process "out of control"*: One or more of the sample means fall outside the control limits. As with the $\bar{x}$-chart, this indicates a possible change in the production process and efforts should be made to locate the trouble.

*Process "in control"*: All sample means fall within the control limits. In this case, it is better to leave the process alone than to look for trouble that may not exist.

In practice, the $\bar{x}$-chart and the $R$-chart are not used in isolation, as our presentation so far might suggest. Rather, they are used together to monitor the mean (i.e., the location) of the process and the variation of the process simultaneously. In fact, many practitioners plot them on the same piece of paper.

One important reason for dealing with them as a unit is that the control limits of the $\bar{x}$-chart are a function of $R$—that is, the control limits depend on the variation of the process. (Recall that the control limits are $\bar{x} \pm A_2R$.) Thus, if the process variation is out of control, the control limits of the $\bar{x}$-chart have little meaning. This is because when the process variation is changing any single estimate of the variation (such as $R$ or $s$) is not representative of the process. Accordingly, **the appropriate procedure is to first construct and then interpret the $R$-chart. If it indicates that the process variation is in control, then it makes sense to construct and interpret the $\bar{x}$-chart.**

## Applied Exercises

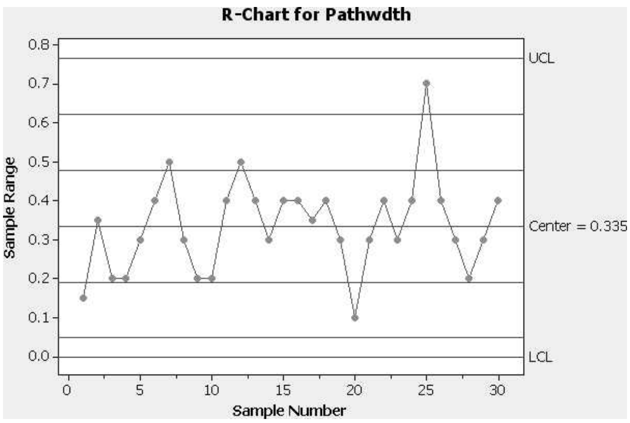
16.17 *CPU of a computer chip.* Refer to Exercise 16.8 (p. 901), where the desired circuit path widths were .5 to .85 micron. The manufacturer sampled four CPU chips six times a day (every 90 minutes from 8:00 A.M. until 4:30 P.M.) for five consecutive days. The path widths were measured and used to construct the MINITAB $R$-chart shown below.

a. Calculate the chart's upper and lower control limits.

b. What does the $R$-chart suggest about the presence of special causes of variation during the time when the data were collected?

c. Should the control limit(s) be used to monitor future process output? Explain.

d. How many different $R$ values are plotted on the control chart? Notice how most of the $R$ values fall along three horizontal lines. What could cause such a pattern?




16.18 *Pain levels of ICU patients.* Refer to the *Research in Nursing & Health* (Vol. 35, 2012) study of the effectiveness of a pain intervention, Exercise 16.9 (p. 901). Recall that pain levels (measured on a 100-point scale) were recorded for a sample of 10 ICU patients 24-hours post-surgery each week for 20 consecutive weeks. The data are repeated in the accompanying table. Now, you want to check process variation using an $R$-chart

a. Compute the centerline for the chart.

b. Compute the UCL and LCL for the $R$-chart.

c. Plot the ranges for the 20 weeks on the $R$-chart. Does the variation of the pain management process appear to be "in control"?

d. Recall that after the 20th week, a pain intervention occurred in the ICU. The ranges of the pain levels for the samples of patients over the next eight consecutive weeks were (in order): 22, 29, 16, 15, 23, 19, 30, and 32. Plot these ranges on the $R$-chart.

**ICUPAIN**

| Week | X-Bar | Range | Week | X-Bar | Range |
|---|---|---|---|---|---|
| 1 | 65 | 28 | 11 | 70 | 34 |
| 2 | 75 | 41 | 12 | 74 | 37 |
| 3 | 72 | 31 | 13 | 73 | 25 |
| 4 | 69 | 35 | 14 | 62 | 33 |
| 5 | 73 | 35 | 15 | 68 | 28 |
| 6 | 63 | 33 | 16 | 75 | 35 |
| 7 | 77 | 34 | 17 | 72 | 29 |
| 8 | 75 | 29 | 18 | 70 | 32 |
| 9 | 69 | 30 | 19 | 62 | 33 |
| 10 | 64 | 39 | 20 | 72 | 29 |

*Source:* Polit, D.F. & Chaboyer, W. "Statistical process control in nursing research", *Research in Nursing & Health,* Vol. 35, No. 1, 2012 (adapted from Figure 1).

16.19 *Quality control for irrigation data.* Refer to Exercise 16.10 (p. 902) and the monitoring of irrigation data by the CIMIS. Recall that daily minimum air temperatures (°C) collected hourly during the month of May at the Davis CIMIS station yielded the following summary statistics (where five measurements are collected each hour): $\bar{\bar{x}} = 10{:}16^\circ$ and $\bar{R} = 14.87^\circ$.

a. Use the information provided to find the lower and upper control limits for an $R$-chart.

b. Suppose that one day in May the air temperature at the Davis CIMIS station had a high of 24.7° and a low of 2.2°. How should the manager of the station respond to this observation?

**FIREPINS**

16.20 *Lengths of firing pins.* Refer to Exercise 16.11 (p. 902). Suppose the inspector wants to monitor the variation in firing pin lengths with an *R*-chart.

a. Locate the center line for the *R*-chart.
b. Locate upper and lower control limits for the *R*-chart.
c. Calculate and plot the 10 sample ranges in an *R*-chart. Does the process variation appear to be in control?

**RUBBERJNT2**

16.21 *Molded-rubber expansion joints.* Construct an *R*-chart for the data of Exercise 16.12 (p. 902) to monitor the variation in the diameters of the molded-rubber expansion joints produced by the manufacturing process. Does the process appear to be in control?

**COLA**

16.22 *Cola bottle filling process.* A soft-drink bottling company is interested in monitoring the amount of cola injected into 16-ounce bottles by a particular filling head. The process is entirely automated and operates 24 hours a day. At 6:00 a.m. and 6:00 p.m. each day, a new dispenser of carbon dioxide capable of producing 20,000 gallons of cola is hooked up to the filling machine. To monitor the process using control charts, the company decided to sample five consecutive bottles of cola each hour beginning at 6:15 A.M. (i.e., 6:15 A.M., 7:15 A.M., 8:15 A.M., etc.). The data for the first day are saved in the file. An SPSS descriptive statistics printout for the data is shown below.

**Descriptive Statistics for 24 Cola Samples**

colaamt

| sample | N | Mean | Minimum | Maximum | Range |
|---|---|---|---|---|---|
| 1 | 5 | 16.0060 | 15.98 | 16.03 | .05 |
| 2 | 5 | 16.0000 | 15.97 | 16.03 | .06 |
| 3 | 5 | 16.0080 | 15.98 | 16.04 | .06 |
| 4 | 5 | 16.0020 | 15.98 | 16.03 | .05 |
| 5 | 5 | 16.0080 | 15.97 | 16.04 | .07 |
| 6 | 5 | 16.0080 | 15.97 | 16.04 | .07 |
| 7 | 5 | 16.0040 | 15.96 | 16.05 | .09 |
| 8 | 5 | 16.0100 | 15.97 | 16.05 | .08 |
| 9 | 5 | 15.9920 | 15.95 | 16.03 | .08 |
| 10 | 5 | 16.0120 | 15.95 | 16.06 | .11 |
| 11 | 5 | 16.0040 | 15.93 | 16.07 | .14 |
| 12 | 5 | 16.0180 | 15.94 | 16.08 | .14 |
| 13 | 5 | 15.9900 | 15.96 | 16.01 | .05 |
| 14 | 5 | 15.9980 | 15.98 | 16.02 | .04 |
| 15 | 5 | 16.0020 | 15.98 | 16.03 | .05 |
| 16 | 5 | 16.0040 | 15.97 | 16.02 | .05 |
| 17 | 5 | 16.0140 | 15.99 | 16.05 | .06 |
| 18 | 5 | 16.0080 | 15.98 | 16.04 | .06 |
| 19 | 5 | 15.9840 | 15.96 | 16.01 | .05 |
| 20 | 5 | 16.0060 | 15.96 | 16.04 | .08 |
| 21 | 5 | 16.0140 | 15.97 | 16.05 | .08 |
| 22 | 5 | 16.0100 | 15.95 | 16.07 | .12 |
| 23 | 5 | 16.0200 | 15.95 | 16.07 | .12 |
| 24 | 5 | 15.9920 | 15.93 | 16.08 | .15 |
| Total | 120 | 16.0047 | 15.93 | 16.08 | .15 |

a. Will the rational subgrouping strategy that was used enable the company to detect variation in fill caused by differences in the carbon dioxide dispensers?
b. Construct an *R*-chart from the data.
c. What does the *R*-chart indicate about the stability of the filling process during the time when the data were collected? Justify your answer.
d. Should the control limit(s) be used to monitor future process output? Explain.
e. Given your answer to part **c**, should an $\bar{x}$-chart be constructed from the given data? Explain.

16.23 *Lowering the thickness of an expensive blow-molded container.* *Quality* (Mar. 2009) presented a problem that actually occurred at a plant that produces a high-volume, blow-molded container with multiple layers. One of the layers is very expensive to manufacture. The quality manager at the plant desires to lower the average thickness for the expensive layer of material and still meet specifications. To estimate the actual thickness for this layer, the manager measured the thickness for one container from each of two cavities every 2 hours for 2 consecutive days. The data (in millimeters) are shown in the following tables.

a. Construct an *R*-chart for the data.
b. Construct an $\bar{x}$-chart for the data.
c. Based on the control charts in parts **a** and **b**, comment on the current behavior of the manufacturing process. As part of your answer, give an estimate of the true average thickness of the expensive layer.

**BLWMLD**

Day 1

| Time | 7A.M. | 9A.M. | 11A.M. | 1P.M. | 3P.M. | 5P.M. | 7P.M. | 9 P.M. |
|---|---|---|---|---|---|---|---|---|
| Thickness | .167 | .241 | .204 | .221 | .255 | .224 | .216 | .235 |
| (mm) | .232 | .203 | .214 | .190 | .207 | .238 | .210 | .210 |
| Average | .1995 | .2220 | .2090 | .2055 | .2310 | .2310 | .2310 | .2225 |
| Range | .065 | .038 | .010 | .031 | .048 | .014 | .006 | .025 |

Day 2

| Time | 7A.M. | 9A.M. | 11A.M. | 1P.M. | 3P.M. | 5P.M. | 7P.M. | 9 P.M. |
|---|---|---|---|---|---|---|---|---|
| Thickness | .223 | .202 | .258 | .243 | .248 | .192 | .208 | .223 |
| (mm) | .216 | .215 | .228 | .221 | .252 | .221 | .245 | .224 |
| Average | .2195 | .2085 | .2430 | .2320 | .2500 | .2065 | .2265 | .2235 |
| Range | .007 | .013 | .030 | .022 | .004 | .029 | .037 | .001 |

**KNOB2**

16.24 *Rheostat knob insert.* Construct an *R*-chart for the data of Exercise 16.15 (p. 903). Does the process variation appear to be in control?

**CHUNKY**

16.25 *Chunky data.* Refer to Exercise 16.16 (p. 903) and the hourly data collected by BPI consulting.

a. Construct an *R*-chart for the data. Is the process variation in control?
b. Round each measurement in the data set to a whole number, like in Exercise 16.16b. Form an *R*-chart for the rounded data. What do you observe?

16.26 *Precision of an R-chart.* The *Journal of Quality Technology* (July 1998) published an article examining the effects of the precision of measurement on the *R*-chart. The authors presented data from a British nutrition company that fills containers labeled "500 grams" with a powdered dietary supplement. Once every 15 minutes, five containers are sampled from the filling process and the fill weight is measured. The first table (**FILLWT1** file) lists the measurements for 25 consecutive samples made with a scale that is accurate to .5 gram, followed by a second table (**FILLWT2** file) that gives measurements for the same samples made with a scale that is accurate to only 2.5 grams. Throughout the time period over which the samples were drawn, it is known that the filling process was in statistical control, with mean 500 grams and standard deviation 1 gram.

a. Construct an *R*-chart for the data that is accurate to .5 gram. Is the process under statistical control? Explain.
b. Given your answer to part **a**, is it appropriate to construct an $\bar{x}$-chart for the data? Explain.
c. Construct an *R*-chart for the data that is accurate to only 2.5 grams. What does it suggest about the stability of the filling process?
d. Based on your answers to parts **a** and **c**, discuss the importance of the accuracy of measurement instruments in evaluating the stability of production processes.

**FILLWT1**

| Sample | Fill Weights Accurate to .5 Gram | | | | | Range |
|---|---|---|---|---|---|---|
| 1 | 500.5 | 499.5 | 502.0 | 501.0 | 500.5 | 2.5 |
| 2 | 500.5 | 499.5 | 500.0 | 499.0 | 500.0 | 1.5 |
| 3 | 498.5 | 499.0 | 500.0 | 499.5 | 500.0 | 1.5 |
| 4 | 500.5 | 499.5 | 499.0 | 499.0 | 500.5 | 1.5 |
| 5 | 500.0 | 501.0 | 500.5 | 500.5 | 500.0 | 1.0 |
| 6 | 501.0 | 498.5 | 500.0 | 501.5 | 500.5 | 3.0 |
| 7 | 499.5 | 500.0 | 499.0 | 501.0 | 499.5 | 2.0 |
| 8 | 498.5 | 498.0 | 500.0 | 500.5 | 500.5 | 2.5 |
| 9 | 498.0 | 499.0 | 502.0 | 501.0 | 501.5 | 4.0 |
| 10 | 499.0 | 499.5 | 499.5 | 500.0 | 499.5 | 1.0 |
| 11 | 502.5 | 499.5 | 501.0 | 501.5 | 502.0 | 3.0 |
| 12 | 501.5 | 501.5 | 500.0 | 500.0 | 501.0 | 1.5 |
| 13 | 498.5 | 499.5 | 501.0 | 500.5 | 498.5 | 2.5 |
| 14 | 499.5 | 498.0 | 500.0 | 499.5 | 498.5 | 2.0 |
| 15 | 501.0 | 500.0 | 498.0 | 500.5 | 500.0 | 3.0 |
| 16 | 502.5 | 501.5 | 502.0 | 500.5 | 500.5 | 2.0 |
| 17 | 499.5 | 500.5 | 500.0 | 499.5 | 499.5 | 1.0 |
| 18 | 499.0 | 498.5 | 498.0 | 500.0 | 498.0 | 2.0 |
| 19 | 499.0 | 498.0 | 500.5 | 501.0 | 501.0 | 3.0 |
| 20 | 501.5 | 499.5 | 500.0 | 500.5 | 502.0 | 2.5 |
| 21 | 501.0 | 500.5 | 502.0 | 502.5 | 502.5 | 2.0 |
| 22 | 501.5 | 502.5 | 502.5 | 501.5 | 502.0 | 1.0 |
| 23 | 499.5 | 502.0 | 500.0 | 500.5 | 502.0 | 2.5 |
| 24 | 498.5 | 499.0 | 499.0 | 500.5 | 500.0 | 2.0 |
| 25 | 500.0 | 499.5 | 498.5 | 500.0 | 500.5 | 2.0 |

**FILLWT2**

| Sample | Fill Weights Accurate to 2.5 Grams | | | | | Range |
|---|---|---|---|---|---|---|
| 1 | 500.0 | 500.0 | 502.5 | 500.0 | 500.0 | 2.5 |
| 2 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 3 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 4 | 497.5 | 500.0 | 497.5 | 497.5 | 500.0 | 2.5 |
| 5 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 6 | 502.5 | 500.0 | 497.5 | 500.0 | 500.0 | 5.0 |
| 7 | 500.0 | 500.0 | 502.5 | 502.5 | 500.0 | 2.5 |
| 8 | 497.5 | 500.0 | 500.0 | 497.5 | 500.0 | 2.5 |
| 9 | 500.0 | 500.0 | 497.5 | 500.0 | 502.5 | 5.0 |
| 10 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 11 | 500.0 | 505.0 | 502.5 | 500.0 | 500.0 | 5.0 |
| 12 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 13 | 500.0 | 500.0 | 497.5 | 500.0 | 500.0 | 2.5 |
| 14 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 15 | 502.5 | 502.5 | 502.5 | 500.0 | 502.5 | 2.5 |
| 16 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 17 | 497.5 | 497.5 | 497.5 | 497.5 | 497.5 | 0.0 |
| 18 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 19 | 495.0 | 497.5 | 500.0 | 500.0 | 500.0 | 5.0 |
| 20 | 500.0 | 502.5 | 500.0 | 500.0 | 502.5 | 2.5 |
| 21 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 22 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 23 | 500.0 | 500.0 | 500.0 | 500.0 | 500.0 | 0.0 |
| 24 | 497.5 | 497.5 | 500.0 | 497.5 | 497.5 | 2.5 |
| 25 | 500.0 | 500.0 | 497.5 | 500.0 | 500.0 | 2.5 |

*Source*: Adapted from Tricker, A., Coates, E. and Okell, E. "The Effects on the *R*-chart of Precision of Measurement." *Journal of Quality Technology*, Vol. 30, No. 3, July 1998, pp. 232–239.

## 16.5 Detecting Trends in a Control Chart: Runs Analysis

As mentioned in the previous two sections, control charts are also examined for trends in the values of $\bar{x}$ or $R$ collected over time. Even when the sample values fall within the control limits, such a trend may indicate the presence of one or more assignable causes of variation. For example, the true process mean may have shifted slightly as a result of wear in the machine.

Trends in the process can be detected by observing runs of points above or below the center line of a control chart. In quality control, a **run** is defined as a sequence of one or more consecutive points that all fall above (or all fall below) the center line.

**Definition 16.7**

A **run** is a sequence of one or more consecutive points that fall on the same side of the center line in a control chart.

The runs (indicated in brackets) for the $R$-chart of Figure 16.9 are shown in Figure 16.10. Sample ranges that fall above the center line are denoted by a "+" symbol, and ranges that fall below the center line by a "–" symbol.

Note that the sequence of 20 points consists of a total of eight runs, starting with a run of two "–", followed by a run of two "+", and so on. Considerable work has been done by researchers on the development of statistical tests based on the **theory of runs**. Many of these techniques are useful for testing whether the sample observations have been drawn at random from the target population. These tests require that the total number of runs, long and short alike, be determined. In quality control, however, a few simple rules have been developed for detecting trends that are based on only the *extreme* (*or longest*) *runs*, in the control chart.

To illustrate, consider the sequence of runs in Figure 16.10. The extreme run in the sequence is composed of seven "–" symbols. These represent the seven consecutive sample ranges that all fell below the center line during hours 12, 13, . . . , 18. How likely is it to observe seven consecutive points on the control chart, all on the same side of the center line, if in fact no assignable causes of variation are present? To answer this question, we use the laws of probability learned in Chapter 3.

First, note that the probability of any one point falling above (or below) the center line is $\frac{1}{2}$ when the process is in control. Then, from the Multiplicative Law of Probability for independent events (see Chapter 3), the probability of seven consecutive points falling, say, *above* the center line is

$$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^7 = \frac{1}{128}$$

Likewise, the probability of seven consecutive points falling *below* the center line is $\left(\frac{1}{2}\right)^7 = \frac{1}{128}$. Therefore, the probability of seven consecutive points falling on the same side of the center line is, by the Additive Law of Probability,

$$\begin{aligned}
&P(\text{7 consecutive points on the same side of the center line}) \\
&= P(\text{7 consecutive points above the center line}) \\
&\quad + P(\text{7 consecutive points below the center line}) \\
&= \frac{1}{128} + \frac{1}{128} = \frac{2}{128} = \frac{1}{64}
\end{aligned}$$

**FIGURE 16.10**
Runs for the $k = 20$ sample ranges in the $R$-chart, Figure 16.9

$$\text{Run:}\quad \underbrace{-\ -}_{1}\quad \underbrace{+\ +}_{2}\quad \underbrace{-}_{3}\quad \underbrace{+\ +\ +}_{4}\quad \underbrace{-}_{5}\quad \underbrace{+\ +}_{6}\quad \underbrace{-\ -\ -\ -\ -\ -\ -}_{7}\quad \underbrace{+\ +}_{8}$$

or .0156. Since it is very unlikely (probability of .0156) to observe such a pattern if the process is in control, the trend in the control chart is taken as a signal of possible trouble in the production process.

A probability such as the one above can be calculated for any run in the control chart, and, based on its value, a decision made about whether to look for trouble in the process. Grant and Leavenworth (1988) recommend looking for assignable causes of variation if any one of the following sequences of points occurs in the control chart:

### Detecting Trend in a Control Chart: Runs Analysis

If any one of the following sequence of runs occurs in a control chart, assignable causes of variation (e.g., trend) are likely to be present:

- Seven or more consecutive points on the same side of the center line
- At least 10 out of 11 consecutive points on the same side of the center line
- At least 12 out of 14 consecutive points on the same side of the center line
- At least 14 out of 17 consecutive points on the same side of the center line

The rules in the box are easy to apply in practice since they simply require one to count consecutive points in the control chart. In each case, the probability of observing that sequence of points when the process is in control is approximately .01. (We leave proof of this result to you as an exercise.) Consequently, if one of these sequences occurs, we are highly confident that some problem in the production process, possibly a shift in the process mean, exists.

More formal statistical tests of runs are available. Consult the references at the end of this chapter if you want to learn more about these techniques.

## Applied Exercises

16.27 *Detecting trends.* Examine the sequences of points in parts **a**–**f** for any trends.

a. + + − − − − − + + + +
b. − + − − − + + − + + + +
c. − − − − − + + + + + − −
d. − +++++ − ++++++++
e. + − + + + − − + + −
f. − + + + + + + + + + −

16.28 *CPU of a computer chip.* Refer to the $\bar{x}$- and $R$-charts, Exercises 16.8 and 16.17 (p. 901, 907). Conduct a runs analysis to detect any trend in the process.

**ICUPAIN**

16.29 *Pain levels of ICU patients.* Refer to the *Research in Nursing & Health* (Vol. 35, 2012) study of the effectiveness of a pain intervention, Exercises 16.9 and 16.18 (p. 901, 907). Conduct a runs analysis on both the $\bar{x}$-chart and the $R$-chart. Interpret the results.

**FIREPINS**

16.30 *Lengths of firing pins.* Refer to the $\bar{x}$- and $R$-charts, Exercises 16.11 and 16.20 (p. 902, 908). Conduct a runs analysis to detect any trend in the process.

**RUBBERJNT2**

16.31 *Molded-rubber expansion joints.* Refer to the $\bar{x}$- and $R$-charts, Exercises 16.12 and 16.21 (p. 902, 908). Conduct a runs analysis to detect any trend in the process.

**KNOB2**

16.32 *Rheostat knob insert.* Refer to the $\bar{x}$- and $R$-charts, Exercises 16.15 and 16.24 (p. 903, 908). Conduct a runs analysis to detect any trend in the process.

**CHUNKY**

16.33 *Chunky data.* Refer to Exercises 16.16 and 16.25 (p. 903, 908) and the hourly data collected by BPI consulting. Conduct a runs analysis on both the $\bar{x}$-chart and the $R$-chart. Interpret the results.

## 16.6 Control Chart for Percent Defectives: *p*-Chart

In addition to measuring quantitative quality characteristics, we are also interested in monitoring the binomial proportion $p$ of the items produced that are defective. As in the case of the $\bar{x}$-chart, random samples of $n$ items are selected from the production line at the end of some specified interval of time. For each sample, we compute the sample proportion

$$\hat{p} = \frac{y}{n}$$

where $y$ is the number of defective items in the sample. The sample proportions are then plotted against time and displayed in a ***p*-chart**.

The center line for a $p$-chart is determined by combining the data contained in a large number $k$ of samples. The estimate of the process proportion defective $p$ is

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number inspected}} = \frac{n\sum_{i=1}^{k}\hat{p}_i}{nk} = \frac{\sum_{i=1}^{k}\hat{p}_i}{k}$$

The upper and lower control limits are located a distance of

$$3\sigma_{\hat{p}} = 3\sqrt{\frac{p(1-p)}{n}}$$

### Location of Center Line and Control Limits for *p*-Chart

*Center line*: $$\bar{p} = \frac{\text{Total number of defectives in } k \text{ samples}}{\text{Total number of items inspected}} = \frac{\sum_{i=1}^{k}\hat{p}_i}{k}$$

UCL: $$\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

LCL: $$\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where

$k$ = Number of samples, each of size $n$

$y_i$ = Number of defectives in the $i$th sample

$\hat{p} = y_i/n$ is the proportion of defectives in the $i$th sample

above and below the center line. Using $\bar{p}$ to estimate the process proportion defective $p$, we find

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

The interpretation of a $p$-chart is similar to the interpretations of $\bar{x}$- and $R$-charts. We expect the sample proportions defective to fall within the control limits. Failure to do so suggests difficulties with the production process and should be investigated.

**Example 16.5**

*p*-chart for Defective Bearings

BEARINGS

To monitor the manufacturing process of rubber support bearings used between the superstructure and foundation pads of nuclear power plants, a quality control engineer randomly samples 100 bearings from the production line each day over a 15-day period. The bearings were inspected for defects and the number of defectives found each day are recorded in Table 16.3. Construct a *p*-chart for the fraction of defective bearings.

**TABLE 16.3 Defective Bearings in 15 Samples of $n = 100$, Example 16.5**

| Day | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Number of Defectives* | 2 | 12 | 3 | 4 | 4 | 1 | 3 | 5 |
| *Proportion of Defectives* | .02 | .12 | .03 | .04 | .04 | .01 | .03 | .05 |

| Day | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Totals |
|---|---|---|---|---|---|---|---|---|
| *Number of Defectives* | 3 | 2 | 10 | 3 | 3 | 2 | 3 | 60 |
| *Proportion of Defectives* | .03 | .02 | .10 | .03 | .03 | .02 | .03 | .04 |

Solution

The center line for the *p*-chart is the proportion of defective bearings in the combined sample of $nk = 1{,}500$ bearings:

$$\bar{p} = \frac{\text{Total number of defective bearings}}{\text{Total number inspected}} = \frac{60}{1{,}500} = .04$$

Upper and lower control limits are then computed as follows:

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = .04 + 3\sqrt{\frac{(.04)(.96)}{100}}$$

$$= .04 + .0588 = .0988$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = .04 - 3\sqrt{\frac{(.04)(.96)}{100}}$$

$$= .04 - .0588 = -.0188$$

Thus, if the process is in control, we expect the sample proportion of defective rubber bearings to fall between 0 (since no sample proportion can be negative) and .099 with a high probability.

A control chart for the percentage of defective bearings is shown in the MINITAB printout, Figure 16.11. Note that on days 2 and 11, the sample proportion fell outside the control limits. This suggests possible problems with the manufacturing process and warrants further investigation.

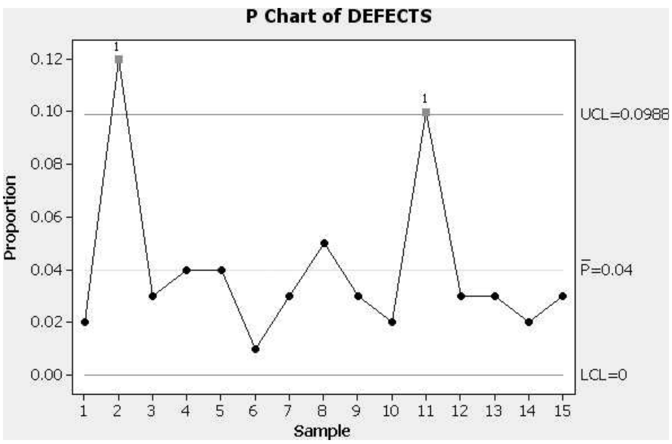



**FIGURE 16.11**
MINITAB *p*-chart for percentage of defective bearings, Example 16.5

### Interpreting a $p$-Chart

*Process "out of control"*: One or more of the sample proportions fall outside the control limits. This indicates possible trouble in the production process and warrants further investigation.

*Process "in control"*: All sample proportions fall within the control limits. In this case it is better to leave the process alone than to look for trouble that may not exist.

Once the problem that caused the two unusually large percentages of defectives in Example 16.5 has been identified and corrected, the control limits should be modified so that they can be applied to future data. As mentioned in Section 16.2, one method of adjusting is to recalculate their values based on only the sample points that fall within the control limits of Figure 16.11. Omitting the data for days 2 and 11, we obtain the modified values

$$\bar{p} = \frac{\text{Total number of defective bearings (excluding days 2 and 11)}}{\text{Total number inspected (excluding days 2 and 11)}}$$

$$= \frac{38}{1{,}300} = .029$$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = .029 + 3\sqrt{\frac{(.029)(.971)}{100}}$$

$$= .029 + .050 = .079$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = .029 - 3\sqrt{\frac{(.029)(.971)}{100}}$$

$$= .029 - .050 = -.021$$

Now a control chart with center line $\bar{p} = .029$, UCL $= .079$, and LCL $= 0$ can be used to monitor the percentage defective produced in future days of the process.

## Applied Exercises

16.34 *Rental-car call center study.* A world-wide rental-car company receives about 10,000 calls per month at its European call center. These calls typically involve customer issues with the level of service or the billing/invoice process. In an effort to reduce the proportion of issues that are not resolved on the customer's first call, management conducted a thorough study of the call center's procedures. The results were published in the *International Journal of Productivity and Performance Management* (Vol. 59, 2010). After making major changes at the call center, management constructed a $p$-chart to monitor the process improvements. Assume that 18 calls to the center were sampled each day for 60 consecutive days. The article reported that the proportion of all calls in the sample that had unresolved issues at the end of the call was .107. (This was a major improvement over the previous unresolved first-call rate of .845.)

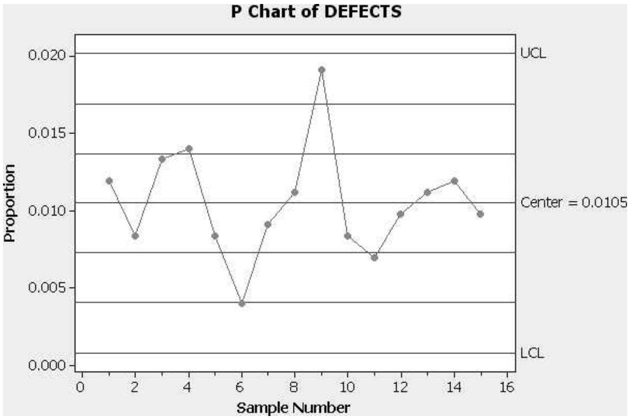
a. What is the centerline for the $p$-chart?

b. Compute the lower and upper control limits for the $p$-chart.

c. When the proportions of daily calls that resulted in unresolved issues are plotted on the $p$-chart, all fall within the LCL and UCL boundaries. What does this imply about the process?

16.35 *Defective micron chips.* A manufacturer produces micron chips for personal computers. From past experience, the production manager believes that 1% of the chips are defective. The company collected a sample of the first 1,000 chips manufactured after 4:00 P.M. every other day for a month. The chips were analyzed for defects, then these data and MINITAB were used to construct the $p$-chart shown below.

a. Calculate the chart's upper and lower control limits.
b. What does the $p$-chart suggest about the presence of special causes during the time when the data were collected?
c. Critique the rational subgrouping strategy used by the disk manufacturer.




16.36 *Monitoring surgery complications.* An article on the use of control charts for monitoring the proportion of post-operative complications at a large hospital was published in the *International Journal for Quality in Health Care* (Oct. 2010). A random sample of surgical procedures was selected each month for 30 consecutive months, and the number of procedures with post-operative complications was recorded. The data are listed in the accompanying table.

a. Identify the attribute of interest to the hospital.
b. What are the rational subgroups for this study?
c. Find the value of $\bar{p}$ for use in a $p$-chart.
d. Compute the proportion of post-op complications in each month.
e. Compute the critical boundaries for the $p$-chart (i.e., UCL, LCL, upper AB boundary, etc.).
f. Construct a $p$-chart for the data.
g. Interpret the chart. Does the process appear to be in control? Explain.

**POSTOP**

| Month | Complications | Procedures Sampled |
|---|---|---|
| 1 | 14 | 105 |
| 2 | 12 | 97 |
| 3 | 10 | 115 |
| 4 | 12 | 100 |
| 5 | 9 | 95 |
| 6 | 7 | 111 |
| 7 | 9 | 68 |
| 8 | 11 | 47 |
| 9 | 9 | 83 |
| 10 | 12 | 108 |
| 11 | 10 | 115 |
| 12 | 7 | 94 |
| 13 | 12 | 107 |
| 14 | 9 | 99 |
| 15 | 15 | 105 |
| 16 | 13 | 110 |
| 17 | 7 | 97 |
| 18 | 10 | 105 |
| 19 | 8 | 71 |
| 20 | 5 | 48 |
| 21 | 12 | 95 |
| 22 | 9 | 110 |
| 23 | 7 | 103 |
| 24 | 9 | 95 |
| 25 | 15 | 105 |
| 26 | 12 | 100 |
| 27 | 8 | 116 |
| 28 | 2 | 110 |
| 29 | 9 | 105 |
| 30 | 10 | 120 |
| Totals | 294 | 2939 |

*Source:* Duclois, A. & Voirin, N. "The $p$-Conrol Chart: A Tool for Care Improvement", *International Journal for Quality in Health Care,* Vol. 22, No. 5, Oct. 2010 (Table 1).

16.37 *Leaky process pumps. Quality* (Feb. 2008) presented a problem that actually occurred at a company that produces process pumps for a variety of industries. The company recently introduced a new pump model and immediately began receiving customer complaints about "leaky pumps." There were no complaints about the old pump model. For each of the first 13 weeks of production of the new pump, quality-control inspectors tested 500 randomly selected pumps for leaks. The results of the leak tests are summarized by week in the accompanying table. Construct an appropriate control chart for the data. What does the chart indicate about the stability of the process?

**PUMPS**

| Week | Number Tested | Number with Leaks |
|---|---|---|
| 1 | 500 | 36 |
| 2 | 500 | 28 |
| 3 | 500 | 24 |
| 4 | 500 | 26 |
| 5 | 500 | 20 |
| 6 | 500 | 56 |
| 7 | 500 | 26 |
| 8 | 500 | 28 |
| 9 | 500 | 31 |
| 10 | 500 | 26 |
| 11 | 500 | 34 |
| 12 | 500 | 26 |
| 13 | 500 | 32 |

16.38 *Testing tires.* Goodstone Tire & Rubber Company is interested in monitoring the proportion of defective tires generated by the production process at its Akron, Ohio, production plant. The company's chief engineer believes that the proportion is about 7%. Because the tires are destroyed during the testing process, the company would like to keep the number of tires tested to a minimum. The chief engineer recommended that the company randomly sample and test 120 tires from each day's production. To date, 20 samples have been taken. The data are presented in the table below.

a. Construct a $p$-chart for the tire production process.
b. What does the chart indicate about the stability of the process? Explain.
c. Is it appropriate to use the control limits to monitor future process output? Explain.
d. Is the $p$-chart you constructed in part **b** capable of signaling hour-to-hour changes in $p$? Explain.

**DEFTIRES**

| Sample | Sample Size | Defectives | Sample | Sample Size | Defectives |
|---|---|---|---|---|---|
| 1 | 120 | 11 | 11 | 120 | 10 |
| 2 | 120 | 5 | 12 | 120 | 12 |
| 3 | 120 | 4 | 13 | 120 | 8 |
| 4 | 120 | 8 | 14 | 120 | 6 |
| 5 | 120 | 10 | 15 | 120 | 10 |
| 6 | 120 | 13 | 16 | 120 | 5 |
| 7 | 120 | 9 | 17 | 120 | 10 |
| 8 | 120 | 8 | 18 | 120 | 10 |
| 9 | 120 | 10 | 19 | 120 | 3 |
| 10 | 120 | 11 | 20 | 120 | 8 |

16.39 *Stress cracks in PCCP.* Prestressed concrete cylinder pipe (PCCP) is a rigid pipe designed to take optimum advantage of the tensile strength of steel and the compressive strength and corrosive-inhibiting properties of concrete. PCCP, produced in laying lengths of 24 feet, is susceptible to major stress cracks during the manufacturing process. To monitor the process, 20 sections of PCCP were sampled each week for a 6-week period. The number of defective sections (i.e., sections with major stress cracks) in each sample is recorded in the table.

**PCCP**

| *Week* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *Number of Defectives* | 1 | 0 | 2 | 2 | 3 | 1 |

a. Construct a $p$-chart for the sample percentage of defective PCCP sections manufactured.
b. Locate the center line on the $p$-chart.
c. Locate upper and lower control limits on the $p$-chart. Does the process appear to be in control?

16.40 *Defective fuses.* A manufacturer of computer terminal fuses wants to establish a control chart to monitor the production process. Each hour, for a period of 25 hours, during a time when the process is known to be in control, a quality control engineer randomly selected and tested 100 fuses from the production line. The number of defective fuses found each hour is recorded in the table, p. 917.

a. Construct a $p$-chart for the sample percentage of defective terminal fuses.
b. Locate the center line on the $p$-chart.
c. Locate upper and lower control limits on the $p$-chart. Does the process appear to be in control?

### Data for Exercise 16.40

**CTFUSE**

| *Hour* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Number Defective* | 6 | 4 | 9 | 3 | 0 | 6 | 4 | 2 | 1 | 2 |
| *Hour* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Number Defective* | 1 | 3 | 4 | 5 | 5 | 2 | 1 | 1 | 0 | 3 |
| *Hour* | 21 | 22 | 23 | 24 | 25 | | | | | |
| *Number Defective* | 7 | 9 | 2 | 10 | 3 | | | | | |

d. Conduct a runs analysis on the points on the $p$-chart. What does this imply?
e. Suppose the next sample of 100 terminal fuses selected from the production line contains 11 defectives. Is the process now out of control? Explain.

16.41 *Cathode-ray tubes.* An electronics company manufactures several types of cathode-ray tubes on a mass production basis. To monitor the process, 50 tubes of a certain type were randomly sampled from the production line and inspected each day over a 1-month period. The number of defectives found each day is provided in the accompanying table.

a. Construct a $p$-chart for the sample fraction of defective cathode-ray tubes.
b. Locate the center line on the $p$-chart.
c. Locate the upper and lower control limits on the $p$-chart.
d. Does the process appear to be in control? If not, modify the control limits for future data.
e. Conduct a runs analysis to detect a trend in the production process.

**CRTUBE**

| *Day* | *Number Defective* | *Day* | *Number Defective* |
|---|---|---|---|
| 1 | 11 | 12 | 23 |
| 2 | 15 | 13 | 15 |
| 3 | 12 | 14 | 12 |
| 4 | 10 | 15 | 11 |
| 5 | 9 | 16 | 11 |
| 6 | 12 | 17 | 16 |
| 7 | 12 | 18 | 15 |
| 8 | 14 | 19 | 10 |
| 9 | 9 | 20 | 13 |
| 10 | 13 | 21 | 12 |
| 11 | 15 | | |

## 16.7 Control Chart for the Number of Defects per Item: *c*-Chart

In addition to various other quality characteristics, we may be interested in the number of defects or blemishes contained in each single item of the product. For example, a manufacturer of office furniture might randomly sample one piece of furniture from the production line every 15 minutes and record the number of blemishes on the finish. Similarly, a textile manufacturer might inspect a randomly selected 1-square-foot piece of material each hour and count the number of minor defects that it contains. The objective of this procedure is to monitor the number of defects per item and to detect situations where this variable is out of control. In the notation used in quality control, the number of defects per item is denoted by the symbol $c$ and a control chart used to monitor this variable over time is called a ***c*-chart**.

The Poisson probability distribution (Section 4.10) provides a good model for the probability distribution for the number $c$ of defects contained in some manufactured product. From Section 4.10, we recall that if $c$ possesses a Poisson probability distribution with parameter $\lambda$, then

$$E(c) = \lambda$$

and

$$\sigma_c = \sqrt{\lambda}$$

To construct a $c$-chart, we observe $c$ over a reasonably large number, $k$, of equally spaced points in time and use the average value of $c$, $\bar{c}$, to estimate $\lambda$. Then since $E(c) = \lambda$, we would locate the center line of the $c$-chart at

$$\textit{Center line:} \quad \bar{c} = \frac{\sum_{i=1}^{k} c_i}{k}$$

The upper and lower control limits are located a distance of $3\sigma_c$ (estimated to be $3\sqrt{\bar{c}}$) above and below the center line. Thus, the upper and lower control limits are located at

$$\text{UCL:} \quad \bar{c} + 3\sqrt{\bar{c}}$$
$$\text{LCL:} \quad \bar{c} - 3\sqrt{\bar{c}}$$

### Location of Center Line and Control Limits for a *c*-Chart

$$\textit{Center line}: \quad \bar{c} \qquad \text{UCL:} \quad \bar{c} + 3\sqrt{\bar{c}} \qquad \text{LCL:} \quad \bar{c} - 3\sqrt{\bar{c}}$$

where

$k$ = Number of time periods sampled

$c_i$ = Number of defects per item observed at time $i$

$$\bar{c} = \frac{\sum_{i=1}^{k} c_i}{k} = \text{Average number of defects per item observed over all time periods}$$

**Example 16.6**

*c*-Chart for Defects in Woolen Fabric

The number of noticeable defects found by quality control inspectors in a randomly selected 1-square-meter specimen of woolen fabric from a certain loom is recorded each hour for a period of 20 hours. The results are shown in Table 16.4. Assuming that the number of defects per square meter has an approximate Poisson probability distribution, construct a *c*-chart to monitor the textile production process.

**WOOLFAB**

**TABLE 16.4 Number of Defects Observed in Specimens of Woolen Fabric over 20 Consecutive Hours, Example 16.6**

| *Hour* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Number of Defects* | 11 | 14 | 10 | 8 | 3 | 9 | 10 | 2 | 5 | 6 |
| *Hour* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Number of Defects* | 12 | 3 | 4 | 5 | 6 | 8 | 11 | 8 | 7 | 9 |

Solution

The first step is to estimate $\lambda$, the mean number of defects per square meter of woolen fabric. This value, $\bar{c}$, also represents the center line for the control chart:

$$\bar{c} = \frac{\sum c_i}{n} = \frac{151}{20} = 7.55$$

Upper and lower control limits are then calculated as follows:

$$\text{UCL} = \bar{c} + 3\sqrt{\bar{c}} = 7.55 + 3\sqrt{7.55} = 15.79$$
$$\text{LCL} = \bar{c} - 3\sqrt{\bar{c}} = 7.55 - 3\sqrt{7.55} = -.69$$

Since a negative number of defects cannot be observed, the LCL is adjusted up to 0.

The control chart for the data appears in the MINITAB printout, Figure 16.12. According to current standards, the textile process produces an allowable number of defects in woolen fabric if the number of defects per square meter does not exceed 15. At no time during the 20-hour period did the process appear to be out of control.

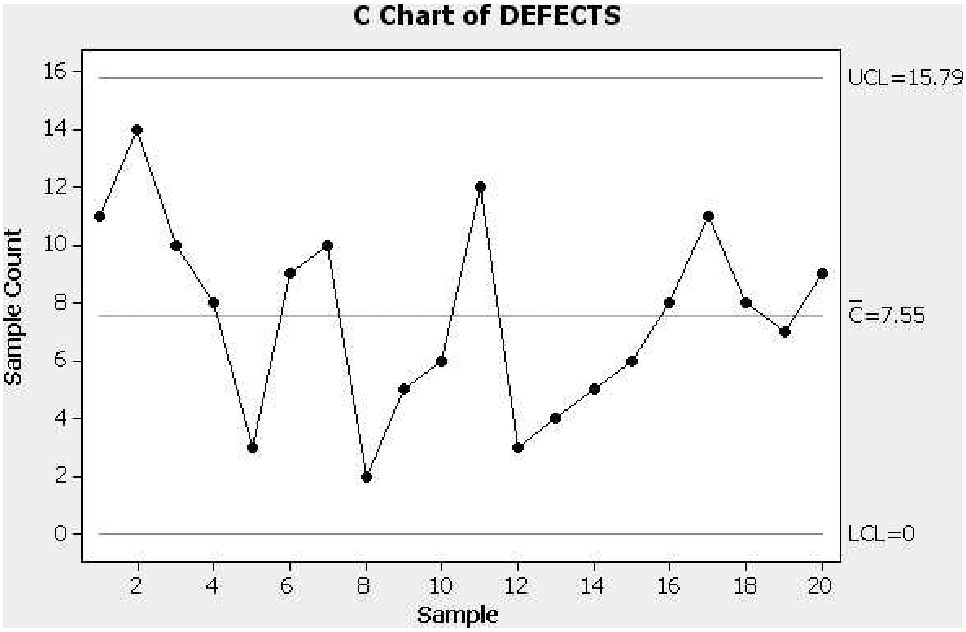



**FIGURE 16.12**
MINITAB *c*-chart for number of defects per square meter, Example 16.6

However, before we conclude that the process is in control, we should check for trends on the number of defects over time, i.e., we should perform a runs analysis as described in Section 16.5. Using the symbols "+" and "–" to denote points above and below the center line, respectively, we obtain the sequence of runs shown in Figure 16.13. Note that the extreme runs in the sequence (runs 1 and 6) include only four points. Also, none of the other unlikely sequences given in the box in Section 16.5 occurs. Therefore, it does not appear that any trend exists in the data. At this point in time, the process appears to be in control.

**FIGURE 16.13**
Runs for the $k = 20$ numbers of defects in the *c*-chart, Figure 16.12

$$\text{Run:}\quad \underbrace{+\ +\ +\ +}_{1}\quad \underbrace{-}_{2}\quad \underbrace{+\ +}_{3}\quad \underbrace{-\ -\ -}_{4}\quad \underbrace{+}_{5}\quad \underbrace{-\ -\ -\ -}_{6}\quad \underbrace{+\ +\ +}_{7}\quad \underbrace{-}_{8}\quad \underbrace{+}_{9}$$

### Interpreting a *c*-Chart

*Process "out of control"*: One or more of the sample numbers of defects fall outside the control limits. This indicates possible trouble in the production process and warrants further investigation.

*Process "in control"*: All of the sample numbers of defects fall within the control limits. In this case it is better to leave the process alone than to look for trouble that may not exist.

## Applied Exercises

16.42 *Imperfections in wood panels.* The number of imperfections (scratches, chips, cracks, and blisters) in manufactured custom wood cabinet panels is important both to the customer and to the custom builder. To monitor the manufacturing process, each hour for 15 consecutive hours a finished panel 4 feet by 8 feet was selected and inspected for imperfections. The number of imperfections per panel is recorded in the table.

**WOODPANEL**

| *Panel* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Defects* | 4 | 2 | 3 | 3 | 9 | 4 | 5 | 3 | 8 | 7 | 3 | 6 | 5 | 7 | 3 |

a. Plot the number of defects per panel in a *c*-chart.
b. Locate the center line for the *c*-chart.
c. Locate upper and lower control limits for the *c*-chart. Is the process in control?
d. Conduct a runs analysis for the *c*-chart. What does this imply?

16.43 *Computer module coding errors.* A quality control study was undertaken by the supervisor of a group of computer programmers. For each of the last 20 days, 25 program modules were randomly selected and inspected for coding errors. The numbers of errors observed per day are recorded in the table.

**CODING**

| *Day* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Coding Errors* | 7 | 3 | 9 | 8 | 2 | 5 | 10 | 5 | 7 | 6 |
| *Day* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Coding Errors* | 1 | 4 | 11 | 8 | 6 | 6 | 9 | 2 | 12 | 9 |

a. Plot the number of coding errors per day in a $c$-chart.
b. Locate the center line for the $c$-chart.
c. Locate upper and lower control limits for the $c$-chart. Is the process in control?
d. Conduct a runs analysis for the $c$-chart. What does this imply?

16.44 *Aircraft alignment.* A certain airplane model is susceptible to alignment errors in the manufacturing process. To monitor this process, the total number of alignment errors observed at final inspection for each of the first 25 aircraft produced were recorded, as shown in the table, right column.

a. Construct a $c$-chart for the number of alignment errors per aircraft.
b. Locate the center and upper and lower control limits on the $c$-chart.
c. Does the process appear to be in control? Would you recommend using these control limits for future data?

**AIRALIGN**

| *Airplane* | *Number of Alignment Errors* | *Airplane* | *Number of Alignment Errors* |
|---|---|---|---|
| 1 | 7 | 14 | 9 |
| 2 | 6 | 15 | 8 |
| 3 | 6 | 16 | 15 |
| 4 | 7 | 17 | 6 |
| 5 | 4 | 18 | 4 |
| 6 | 7 | 19 | 13 |
| 7 | 8 | 20 | 7 |
| 8 | 12 | 21 | 8 |
| 9 | 9 | 22 | 15 |
| 10 | 9 | 23 | 6 |
| 11 | 8 | 24 | 6 |
| 12 | 5 | 25 | 10 |
| 13 | 5 | | |

*Source*: Grant, E. L., and Leavenworth, R. S., *Statistical Quality Control*, 5th ed. New York: McGraw-Hill, 1980 (Table 8–1). Reprinted with permission.

16.45 *Aircraft alignment (continued).* Refer to Exercise 16.44. The numbers of alignment errors observed for each of the next 25 aircraft produced are shown in the table below.

a. Add these 25 points to the $c$-chart of Exercise 16.44. Does the process still appear to be in control?
b. Conduct a runs analysis for the revised $c$-chart. What do you detect?

**AIRALIGN2**

| *Airplane* | *Number of Alignment Errors* | *Airplane* | *Number of Alignment Errors* | *Airplane* | *Number of Alignment Errors* | *Airplane* | *Number of Alignment Errors* |
|---|---|---|---|---|---|---|---|
| 26 | 7 | 33 | 6 | 40 | 8 | 47 | 9 |
| 27 | 13 | 34 | 7 | 41 | 10 | 48 | 11 |
| 28 | 4 | 35 | 14 | 42 | 8 | 49 | 11 |
| 29 | 5 | 36 | 18 | 43 | 7 | 50 | 8 |
| 30 | 9 | 37 | 11 | 44 | 16 | | |
| 31 | 3 | 38 | 11 | 45 | 13 | | |
| 32 | 4 | 39 | 11 | 46 | 12 | | |

*Source*: Grant, E. L., and Leavenworth, R. S., *Statistical Quality Control*, 5th ed. New York: McGraw-Hill, 1980 (Table 8–1). Reprinted with permission.

16.46 *Defects per million opportunities.* Electronics products (e.g., backplanes, complex motherboards for server systems, etc.) can have as many as thousands of opportunities for defects per printed circuit board (PCB). The defects can be traced to improper solder joints (potentially thousands on a PCB), missing components, improperly placed components, and others. (*International Journal of Industrial Engineering,* Vol. 16, 2009). The data in the table represent the total number of defects per day in a PCB assembly process where 100 PCB assemblies are inspected each day, for 24 consecutive days.

a. Construct a $c$-chart for the data. Is the process in control?

b. Each PCB assembly has 3,000 opportunities for defects. Because there are so many opportunities for defects, quality control engineers are interested in the *average number of defects per million opportunities* (or, *dpmo*). The author of the journal article demonstrated how to convert a standard defects $c$-chart into a dpmo-chart. The points plotted in the chart are calculated as follows:

$$\text{dpmo} = (1{,}000{,}000)(c/n)/(\text{number of defect opportunities per unit})$$

where $c$ is the total number of defects per day and $n$ is the number of units inspected per day. The center line is the average dpmo, i.e., $\overline{\text{dpmo}} = \Sigma\, \text{dpmo}/k$, where $k$ = the number of subgroups (days). The lower and upper control limits are:

$$(\text{LCL, UCL}) = \text{dpmo} \pm 3\sqrt{\overline{\text{dpmo}}(1{,}000{,}000)/n(\text{number of defect opportunities})}$$

For this application, $n = 100$ PCB assemblies per day, $k = 24$ days, and number of defect opportunities $= 3{,}000$. Use this information to construct a dpmo-chart for the data. Interpret the resulting graph.

**DPMO**

| Day | Defects |
|---|---|
| 1 | 19 |
| 2 | 19 |
| 3 | 22 |
| 4 | 19 |
| 5 | 21 |
| 6 | 17 |
| 7 | 29 |
| 8 | 13 |
| 9 | 15 |
| 10 | 17 |
| 11 | 16 |
| 12 | 17 |
| 13 | 17 |
| 14 | 15 |
| 15 | 23 |
| 16 | 22 |
| 17 | 27 |
| 18 | 17 |
| 19 | 20 |
| 20 | 22 |
| 21 | 20 |
| 22 | 23 |
| 23 | 30 |
| 24 | 24 |

## 16.8 Tolerance Limits

The Shewhart control charts described in the previous sections provide valuable information on the quality of the production process as a whole. Even if the process is deemed to be in control, however, an individual manufactured item may not always meet specifications. Therefore, in addition to process control, it is often important to know that a large proportion of the individual quality measurements fall within certain limits with a high degree of confidence. An interval that includes a certain percentage of measurements with a specified probability is called a **tolerance interval** and the endpoints of the interval are called **tolerance limits**.

Tolerance intervals are identical to the confidence intervals of Chapter 8, except that we are attempting to capture a proportion $\gamma$ of measurements in a population rather than a population parameter (e.g., the population mean $\mu$). For example, a production supervisor may want to establish tolerance limits for 99% of the length measurements of eyescrews manufactured on the production line, using a 95% tolerance interval. Here, the confidence coefficient is $1 - \alpha = .95$ and the proportion of measurements the supervisor wants to capture is $\gamma = .99$. The confidence coefficient, .95, has the same meaning as in Chapter 8. That is, approximately 95 out of every 100 similarly constructed tolerance intervals will contain 99% of the length measurements in the population.

**Definition 16.8**

A $100(1-\alpha)\%$ **tolerance interval** for $100(\gamma)\%$ of the quality measurements of a product is an interval that includes $100(\gamma)\%$ of the measurements with confidence coefficient $(1-\alpha)$.

**Definition 16.9**

The endpoints of a tolerance interval are called **tolerance limits**.

When the population of measurements that characterize the product is normally distributed with known mean $\mu$ and known standard deviation $\sigma$, tolerance limits are easily constructed. In fact, such an interval is a 100% tolerance interval, i.e., the confidence coefficient is 1.0. For example, suppose the lengths of the eyescrews above have a normal distribution with $\mu = .50$ inch and $\sigma = .01$ inch. From our knowledge of the standard normal ($z$) distribution, we know with certainty (i.e., with probability $1 - \alpha = 1.0$) that 99% of the measurements will fall within $z = 2.58$ standard deviations of the mean (see Figure 16.14). Thus, a 100% tolerance interval for 99% of the length measurements is

$$\begin{aligned} \mu \pm 2.58\sigma &= .50 \pm 2.58(.01) \\ &= .50 \pm .0258 \end{aligned}$$

or (.4742, .5258).

In practice, quality control engineers will rarely know the true values of $\mu$ and $\sigma$. Fortunately, tolerance intervals can be constructed by substituting the sample estimates $\bar{x}$ and $s$ for $\mu$ and $\sigma$, respectively. Due to the errors introduced by the sample estimators, however, the confidence coefficient for the tolerance interval will no longer equal 1.0. The procedure for constructing tolerance limits for a normal population of measurements is described in the following box.

## A Tolerance Interval for the Measurements in a Normal Population

A $100(1 - \alpha)\%$ tolerance interval for $100\gamma\%$ of the measurements in a normal population is given by

$$\bar{x} \pm Ks$$

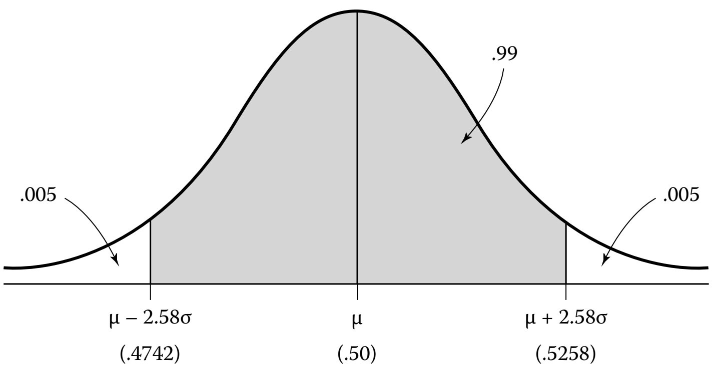
where

$$\begin{aligned} \bar{x} &= \text{Mean of a sample of } n \text{ measurements} \\ s &= \text{Sample standard deviation} \end{aligned}$$

and $K$ is found from Table 20 of Appendix B, based on the values of the confidence coefficient $(1 - \alpha)$, $\gamma$, and the sample size $n$.

*Assumption*: The population of measurements is approximately normal.

**FIGURE 16.14**
Normal distribution of eyescrew lengths



**Example 16.7**

A 95% Tolerance Interval for Shaft Diameter

Refer to Example 16.2. Use the sample information provided in Table 16.2 to find a 95% tolerance interval for 99% of the shaft diameters produced by the manufacturing process. Assume that the distribution of shaft diameters is approximately normal.

Solution

Table 16.2 (p. 898) contains diameters for 20 samples of four shafts each, or a total of $n = 80$ shaft diameters. Descriptive statistics for the 80 diameters are shown on the SPSS printout, Figure 16.15. The mean diameter of the entire sample, highlighted on the printout, is $\bar{x} = 1.50043$. (Note that this is the same value as the center line, $\bar{\bar{x}}$, in Example 16.2.) The sample standard deviation (also highlighted on Figure 16.15), is $s = .009246$.

**FIGURE 16.15**
SPSS descriptive statistics for shaft diameters, Example 16.7

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| DIAMETER | 80 | 1.483 | 1.526 | 1.50043 | .009246 |
| Valid N (listwise) | 80 | | | | |

Since we desire a tolerance interval for 99% of the shaft diameters, $\gamma = .99$. Also, the confidence coefficient is $1 - \alpha = .95$. Table 20 of Appendix B gives the values of $K$ for several values of $\gamma$ and $1 - \alpha$. For $\gamma = .99$, $1 - \alpha = .95$, and $n = 80$, Table 20 gives $K = 2.986$. Then, the 95% tolerance interval is

$$\begin{aligned} \bar{x} \pm 2.986s &= 1.50043 \pm (2.986)(.009246) \\ &= 1.50043 \pm .02761 \end{aligned}$$

or (1.47282, 1.52804). Thus, the lower and upper 95% tolerance limits for 99% of the shaft diameters are 1.47282 inches and 1.52804 inches, respectively. Our confidence in the procedure is based on the premise that approximately 95 out of every 100 similarly constructed tolerance intervals will contain 99% of the shaft diameters in the population.

The technique applied in Example 16.7 gives tolerance limits for a normal distribution of measurements. If we are unwilling or unable to make the normality assumption, we must resort to a nonparametric method. Nonparametric tolerance limits are based on only the smallest and largest measurements in the sample data, as shown in the box. These tolerance intervals can be applied to any distribution of measurements.

### A Nonparametric Tolerance Interval

Let $x_{\text{min}}$ and $x_{\text{max}}$ be the smallest and largest observations, respectively, in a sample of size $n$ from any distribution of measurements. Then we can select $n$ so that $(x_{\text{min}}, x_{\text{max}})$ forms a $100(1 - \alpha)\%$ tolerance interval for at least $100\gamma\%$ of the population. Values of $n$ for several values of the confidence coefficient $(1 - \alpha)$ and $\gamma$ are given in Table 21 of Appendix B.

**Example 16.8**

Finding the Sample Size for a Tolerance Interval

Refer to Example 16.7. Find the sample size required so that the interval $(x_{\text{min}}, x_{\text{max}})$ forms a 95% tolerance interval for at least 90% of the shaft diameters produced by the manufacturing process.

Solution

Here, the confidence coefficient is $1 - \alpha = .95$ and the proportion of measurements we want to capture is $\gamma = .90$. From Table 21 of Appendix B, the sample size corresponding to $1 - \alpha = .95$ and $\gamma = .90$ is $n = 46$. Therefore, if we randomly sample $n = 46$ shafts, the smallest and largest diameters in the sample will represent the lower and upper tolerance limits, respectively, for at least 90% of the shaft diameters with confidence coefficient .95.

The information provided by tolerance intervals is often used to determine whether product specifications are being satisfied. **Specification limits**, unlike tolerance or control limits, are not determined by sampling the process. Rather, they define acceptable values of the quality variable that are set by customers, management, and/or product designers. To determine whether the specifications are realistic, the specification limits are compared to the "natural" tolerance limits of the process, that is, the tolerance limits obtained from sampling. If the tolerance limits do not fall within the specification limits, a review of the production process is strongly recommended. An investigation may reveal that the specifications are tighter than necessary for the functioning of the product, and, consequently, should be widened. Or, if the specifications cannot be changed, a fundamental change in the production process may be necessary to reduce product variability.

**Definition 16.10**

**Specification limits** are boundary points that define the acceptable values for an output variable (i.e., for a quality characteristic) of a particular product or service. They are determined by customers, management, and product designers. Specification limits may be two-sided, with upper and lower limits, or one-sided, with either an upper or a lower limit.

## Applied Exercises

16.47 *Robotics clamp gap width.* University of Waterloo (Canada) statistician S. H. Steiner applied control chart methodology to the manufacturing of a horseshoe-shaped metal fastener called a robotics clamp (*Applied Statistics*, Vol. 47, 1998). Users of the clamp were concerned with the width of the gap between the two ends of the fastener. Their preferred target width is .054 inches. An optical measuring device was used to measure the gap width of the fastener during the manufacturing process. The manufacturer sampled five finished clamps every 15 minutes throughout its 16-hour daily production schedule and optically measured the gap. Data for 4 consecutive hours of production are presented in the table.

a. Use all the sample information to find a 95% tolerance interval for 99% of all the gap widths. Assume the distribution of gap widths is approximately normal.
b. Specifications require the gap width of a clamp to fall within $54 \pm 4$ thousandths of an inch. Based on the "natural" tolerance limits of the process (i.e., the tolerance limits of part **a**), does it appear that the specifications are being met?
c. How large a sample is required to construct a nonparametric 95% tolerance interval for at least 95% of the gap widths? If $n$ is large enough for this case, give the nonparametric tolerance limits.

**CLAMPGAP**

| Time | Gap Width (thousandths of an inch) | | | | |
|---|---|---|---|---|---|
| 00:15 | 54.2 | 54.1 | 53.9 | 54.0 | 53.8 |
| 00:30 | 53.9 | 53.7 | 54.1 | 54.4 | 55.1 |
| 00:45 | 54.0 | 55.2 | 53.1 | 55.9 | 54.5 |
| 01:00 | 52.1 | 53.4 | 52.9 | 53.0 | 52.7 |
| 01:15 | 53.0 | 51.9 | 52.6 | 53.4 | 51.7 |
| 01:30 | 54.2 | 55.0 | 54.0 | 53.8 | 53.6 |
| 01:45 | 55.2 | 56.6 | 53.1 | 52.9 | 54.0 |
| 02:00 | 53.3 | 57.2 | 54.5 | 51.6 | 54.3 |
| 02:15 | 54.9 | 56.3 | 55.2 | 56.1 | 54.0 |
| 02:30 | 55.7 | 53.1 | 52.9 | 56.3 | 55.4 |
| 02:45 | 55.2 | 51.0 | 56.3 | 55.6 | 54.2 |
| 03:00 | 54.2 | 54.2 | 55.8 | 53.8 | 52.1 |
| 03:15 | 55.7 | 57.5 | 55.4 | 54.0 | 53.1 |
| 03:30 | 53.7 | 56.9 | 54.0 | 55.1 | 54.2 |
| 03:45 | 54.1 | 53.9 | 54.0 | 54.6 | 54.8 |
| 04:00 | 53.5 | 56.1 | 55.1 | 55.0 | 54.0 |

*Source*: Adapted from Steiner, Stefan, H. "Grouped Data Exponentially Weighted Moving Average Control Charts." *Applied Statistics—Journal of the Royal Statistical Society*, Vol. 47, Part 2, 1998, pp. 203–216.

**FIREPINS**

16.48 *Lengths of firing pins.* Refer to Exercise 16.11 (p. 902). Use all the sample information to find a 95% tolerance interval for 90% of the firing pin lengths. Assume the distribution of pin lengths is approximately normal.

16.49 *Customer complaint study.* J. Namias used the techniques of statistical quality control to determine when to conduct a search for specific causes of consumer complaints at a beverage company (*Journal of Marketing Research*, Aug. 1964). Namias discovered that when the process was in control, the biweekly complaint rate of a bottled product (i.e., the number of customer complaints per 10,000 bottles sold in a 2-week period) had an approximately normal distribution with $\mu = 26$ and $\sigma = 11.3$. Customer complaints primarily concerned chipped bottles that looked dangerous.

a. Find a tolerance interval for 99% of the complaint rates when the bottling process is assumed to be in control. What is the confidence coefficient for the interval? Explain.

b. In one 2-week period, the observed complaint rate was 93.12 complaints per 10,000 bottles sold. Based on your knowledge of statistical quality control, do you think the observed rate is due to chance or some specific cause? (In actuality, a search for a possible problem in the bottling process led to a discovery of rough handling of the bottled beverage in the warehouse by newly hired workers. As a result, a training program for new workers was instituted.)

**KNOB2**

16.50 *Rheostat knob insert.* Refer to Exercise 16.15 (p. 903). Find a 99% tolerance interval for at least 95% of the distance measurements assuming each of the following:

a. A normal distribution
b. A nonnormal distribution

16.51 *Mechanical hand tools.* Many hand tools used by mechanics involve attachments that fit into sockets (e.g., a socket wrench). In the manufacturing of the tools, specifications require that the inside diameter of the socket be larger than the outside diameter of the extension. That is, there must be enough clearance so that the extensions actually fit in the sockets. To establish tolerances for the tools, independent random samples of 50 sockets and 50 attachments were selected from the production process and the diameters (inside for sockets and outside for extensions) were measured. An analysis revealed that the distributions for both dimensions were approximately normal. The means and standard deviations (in inches) for the two samples are given in the accompanying table.

| | Sockets (1) | Attachments (2) |
|---|---|---|
| *Sample Mean* | .5120 | .5005 |
| *Standard Deviation* | .0010 | .0015 |

a. Find a 95% tolerance interval for 99% of the socket diameters.

b. Find a 95% tolerance interval for 99% of the attachment diameters.

c. Specifications require that the clearance between attachment and socket (i.e., the difference between the inside socket diameter and outside attachment diameter) be at least .004 inch. Based on the tolerance limits from parts **a** and **b**, is it likely to find an extension and socket with less than the desired minimum clearance of .004 inch?

d. Specifications also require a maximum of .015-inch clearance between attachment and socket, to prevent fits that are too loose. Based on the tolerance intervals from parts **a** and **b**, would you expect to find some attachment and socket pairs that fit too loosely?

e. Refer to part **d**. Calculate the approximate probability of observing a loose fit. [*Hint*: Use the fact that the difference between the inside socket diameter and outside attachment diameter is approximately normal (since the two distributions are normal) with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$ (from Theorem 6.6)].

## 16.9 Capability Analysis (*Optional*)

As we have seen in the previous sections, the achievement of process stability is vitally important to process improvement efforts. But it is not an end in itself. A process may be in statistical control, but due to a high level of variation may not be capable of producing output that is acceptable to customers.

To see this, consider Figure 16.16. The figure displays six different in-control processes. Recall that if a process is under statistical control, its output distribution does not change over time and the process can be characterized by a single probability distribution, as in each of the panels of the figure. The upper and lower specification limits for the output of each of the six processes are also indicated on each panel, as

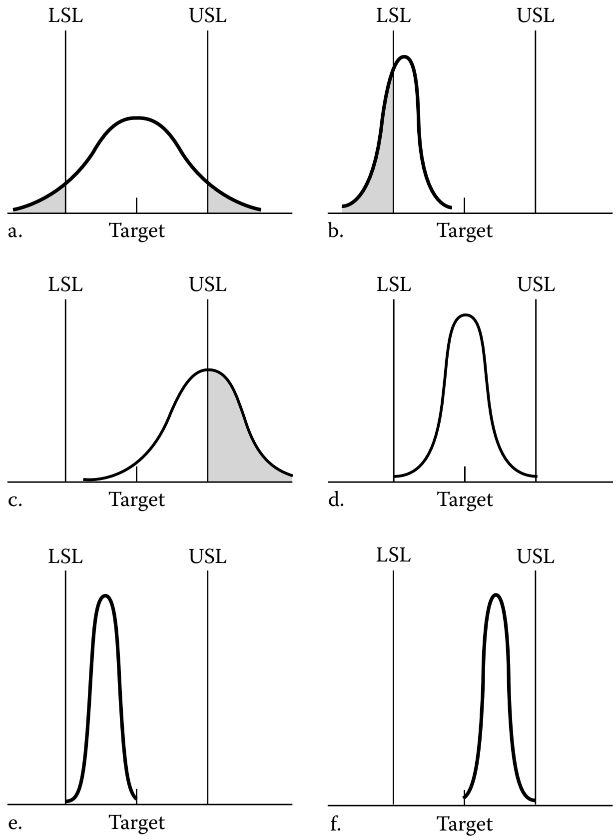



**FIGURE 16.16**
Output distributions of six different in-control processes, where LSL = lower specification limit and USL = upper specification limit

is the target value for the output variable. Recall from Definition 16.10 that the specification limits are boundary points that define the acceptable values for an output variable.

The processes of panels (a), (b), and (c) produce a high percentage of items that are outside the specification limits. None of these processes is *capable* of satisfying its customers. In panel (a), the process is centered on the target value, but the variation due to common causes is too high. In panel (b), the variation is low relative to the width of the specification limits, but the process is off-center. In panel (c), both problems exist: The variation is too high and the process is off-center. Thus, bringing a process into statistical control is not sufficient to guarantee the capability of the process.

All three processes in panels (d), (e), and (f) are capable. In each case, the process distribution fits comfortably between the specification limits. Virtually all of the individual items produced by these processes would be acceptable. However, any significant tightening of the specification limits—whether by customers or internal managers or engineers—would result in the production of unacceptable output and necessitate the initiation of process improvement activities to restore the process' capability. Further, even though a process is capable, continuous improvement of a process requires constant improvement of its capability.

In this optional section, we present a methodology—called **capability analysis**—designed to assess process capability. When a process is known to be in control, the most direct way to assess its capability is to construct a frequency distribution (e.g., dot plot, histogram, or stem-and-leaf display) for a large sample of individual measurements (usually 50 or more) from the process. Then, add the specification limits and the target value for the output variable on the graph. This is called a **capability analysis diagram**. It is a simple visual tool for assessing process capability.

**Example 16.9**

Capability Analysis Diagram

In a paint manufacturing process, 1-gallon cans of paint are consecutively filled by the same filling nozzle. To monitor the process, it was decided to sample five consecutive cans once each hour for the next 25 hours and measure the weight (in pounds) of each can. The sample data are presented in Table 16.5. Specifications are that the weight be between 9.995 pounds and 10.005 pounds, with a target weight of 10 pounds. Construct a capability analysis diagram for the data and interpret the results.

Solution

The MINITAB histogram shown in Figure 16.17 is a capability analysis diagram for the total sample of 125 weights. You can see that the process is roughly centered on the target of 10 pounds of paint, but that a large number of paint cans fall outside the specification limits (12.8% below 9.995 and 11.2% above 10.005, as shown at the bottom left of Figure 16.17). This tells us that the process is not capable of satisfying customer requirements.

**PAINT125**

TABLE 16.5 **Twenty-Five Samples of Size 5 from the Paint-Filling Process**

| Sample | Measurements | | | | |
|---|---|---|---|---|---|
| 1 | 10.0042 | 9.9981 | 10.0010 | 9.9964 | 10.0001 |
| 2 | 9.9950 | 9.9986 | 9.9948 | 10.0030 | 9.9938 |
| 3 | 10.0028 | 9.9998 | 10.0086 | 9.9949 | 9.9980 |
| 4 | 9.9952 | 9.9923 | 10.0034 | 9.9965 | 10.0026 |
| 5 | 9.9997 | 9.9983 | 9.9975 | 10.0078 | 9.9891 |
| 6 | 9.9987 | 10.0027 | 10.0001 | 10.0027 | 10.0029 |
| 7 | 10.0004 | 10.0023 | 10.0024 | 9.9992 | 10.0135 |
| 8 | 10.0013 | 9.9938 | 10.0017 | 10.0089 | 10.0001 |
| 9 | 10.0103 | 10.0009 | 9.9969 | 10.0103 | 9.9986 |
| 10 | 9.9980 | 9.9954 | 9.9941 | 9.9958 | 9.9963 |
| 11 | 10.0013 | 10.0033 | 9.9943 | 9.9949 | 9.9999 |
| 12 | 9.9986 | 9.9990 | 10.0009 | 9.9947 | 10.0008 |
| 13 | 10.0089 | 10.0056 | 9.9976 | 9.9997 | 9.9922 |
| 14 | 9.9971 | 10.0015 | 9.9962 | 10.0038 | 10.0022 |
| 15 | 9.9949 | 10.0011 | 10.0043 | 9.9988 | 9.9919 |
| 16 | 9.9951 | 9.9957 | 10.0094 | 10.0040 | 9.9974 |
| 17 | 10.0015 | 10.0026 | 10.0032 | 9.9971 | 10.0019 |
| 18 | 9.9983 | 10.0019 | 9.9978 | 9.9997 | 10.0029 |
| 19 | 9.9977 | 9.9963 | 9.9981 | 9.9968 | 10.0009 |
| 20 | 10.0078 | 10.0004 | 9.9966 | 10.0051 | 10.0007 |
| 21 | 9.9963 | 9.9990 | 10.0037 | 9.9936 | 9.9962 |
| 22 | 9.9999 | 10.0022 | 10.0057 | 10.0026 | 10.0032 |
| 23 | 9.9998 | 10.0002 | 9.9978 | 9.9966 | 10.0060 |
| 24 | 10.0031 | 10.0078 | 9.9988 | 10.0032 | 9.9944 |
| 25 | 9.9993 | 9.9978 | 9.9964 | 10.0032 | 10.0041 |

Most quality-management professionals and statisticians agree that the capability analysis diagram is the best way to describe the performance of an in-control process. However, many quality engineers have found it useful to have a numerical measure of capability. There are several different approaches to quantifying capability. We will

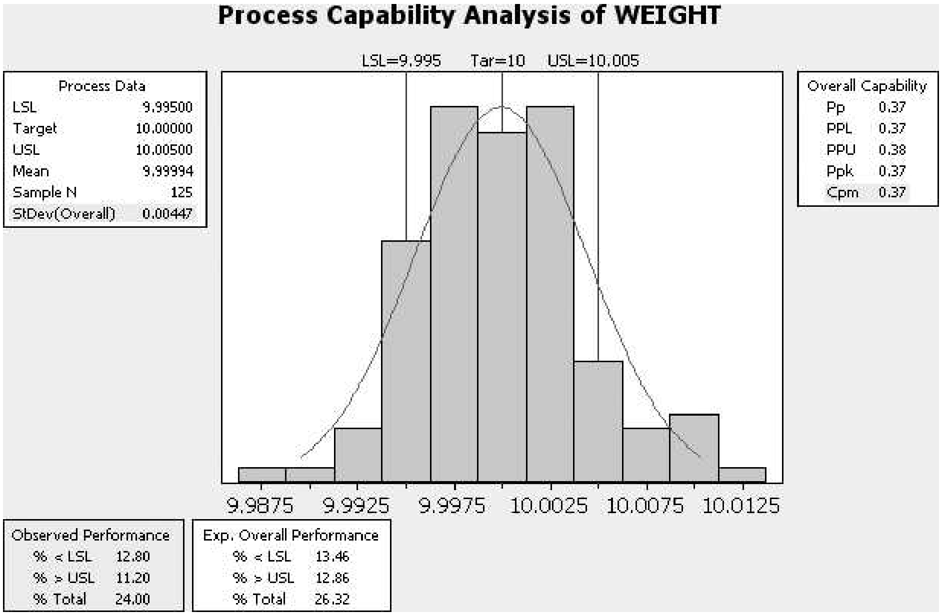



**FIGURE 16.17**
MINITAB capability analysis diagram for the paint-filling process, Example 16.9

briefly describe two of them. The first (and most direct) consists of counting the number of items that fall outside the specification limits in the capability analysis diagram and reporting the percentage of such items in the sample. As shown in Figure 16.17, 24% of the 125 paint cans sampled in Example 16.9 fall outside the specification limits (12.8% below 9.995 and 11.2% above 10.005). Thus, 24% of the 125 cans in the sample, (i.e., 30 cans) are unacceptable.

When this percentage is used to characterize the capability of the process, the implication is that over time, if this process remains in control, roughly 24% of the paint cans will be unacceptable. Remember, however, that this percentage is only an estimate, a sample statistic, not a known parameter. It is based on a sample of size 125 and is subject to both sampling error and measurement error. We discussed such percentages and proportions in detail in Chapter 7.

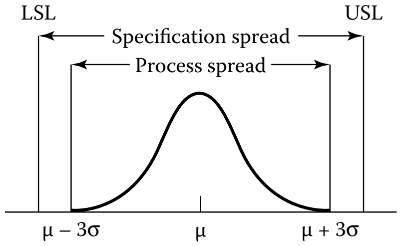
If it is known that the process follows approximately a normal distribution, as is often the case, a similar approach to quantifying process capability can be used. In this case, the mean and standard deviation of the sample of measurements used to construct the capability analysis diagram can be taken as estimates of the mean and standard deviation of the process. Then, the fraction of items that would fall outside the specification limits can be found by solving for the associated area under the normal curve, as we did in Chapter 5. As stated above, if you use this percentage to characterize process capability, remember that it is only an estimate and is subject to sampling error.

The second approach to measuring capability is to construct a **capability index**. Several such indexes have been developed. We will describe one used for stable processes that are centered on the target value. It is known as the $\boldsymbol{C_p}$ **index**.*

When the capability analysis diagram indicates that the process is centered, capability can be measured through a comparison of the distance between the upper specification limit (USL) and the lower specification limit (LSL), called the **specification**

*For off-center processes, its sister index, $C_{pk}$, is used. Consult the chapter references for a description of $C_{pk}$.

**FIGURE 16.18**
Process spread versus specification spread




**spread**, and the spread of the output distribution. The spread of the output distribution—called the **process spread**—is defined as $6\sigma$ and is estimated by $6s$, where $s$ is the standard deviation of the sample of measurements used to construct the capability analysis diagram. These two distances are illustrated in Figure 16.18. The ratio of these distances is the capability index known as $C_p$.

**Definition 16.11**

The **capability index** for a process *centered on the desired mean* is

$$C_p = \frac{(\text{Specification spread})}{(\text{Process spread})} = \frac{\text{USL} - \text{LSL}}{6\sigma}$$

where $\sigma$ is estimated by $s$, the standard deviation of the sample of measurements used to construct the capability analysis diagram.

### Interpretation of Capability Index, $C_p$

$C_p$ summarizes the performance of a stable, centered process relative to the specification limits. It indicates the extent to which the output of the process falls within the specification limits.

1. If $C_p = 1$ (specification spread = process spread), process is capable
2. If $C_p > 1$ (specification spread > process spread), process is capable
3. If $C_p < 1$ (specification spread < process spread), process is not capable

If the process follows a normal distribution,

$C_p = 1.00$ means about 2.7 units per 1,000 will be unacceptable
$C_p = 1.33$ means about 63 units per million will be unacceptable
$C_p = 1.67$ means about .6 units per million will be unacceptable
$C_p = 2.00$ means about 2 units per billion will be unacceptable

In applications where the process follows a normal distribution (approximately), quality engineers typically require a $C_p$ of at least 1.33. With a $C_p$ of 1.33 the process spread takes up only 75% of the specification spread, leaving a little wiggle room in case the process moves off center.

**Example 16.10**

Finding $C_p$, the Capability Index

Let's return to the paint-filling process analyzed in Example 16.9. Recall that 25 samples of size 5 (125 weight measurements), were collected (see Table 16.5). The specification limits for the acceptable amount of paint fill per can are LSL = 9.995 and USL = 10.005 pounds.

a. Is it appropriate to construct a capability index for this process?
b. Find $C_p$ for this process and interpret its value.

Solution

a. First, we must demonstrate that the process is in a state of statistical control. The data of Table 16.5 were entered into MINITAB, and both an $R$-chart and $\bar{x}$-chart were created. Both control charts, shown in Figure 16.19, indicate that the paint-filling process is "in control." Since the process is stable, its output distribution can be characterized by the same probability distribution at any point in time. Accordingly, it is appropriate to assess the performance of the process using that distribution and related performance measures such as $C_p$.

b. From Definition 16.11,

$$C_p = \frac{(\text{USL} - \text{LSL})}{6\sigma}$$

Now, USL $= 10.005$ and LSL $= 9.995$. But what is $\sigma$? Since the output distribution will never be known exactly, neither will $\sigma$, the standard deviation of the output distribution. It must be estimated with $s$, the standard deviation of a large sample drawn from the process. In this case, we use the standard deviation of the 125 measurements used to construct the capability analysis diagram. This value, $s = .00447$, is highlighted in the upper left of the MINITAB printout, Figure 16.17 (p. 928). Then,

$$C_p = \frac{(10.005 - 9.995)}{6(.00447)} = \frac{.01}{.02682} = .373$$

(This value of $C_p$ is also highlighted in the upper right corner of Figure 16.17.) Since $C_p$ is less than 1.0, the process is not capable. The process spread is wider than the specification spread. Thus, the $C_p$ statistic confirms the results shown on the capability analysis diagram (Figure 16.17), where 24% of the sampled cans were found to be unacceptable.

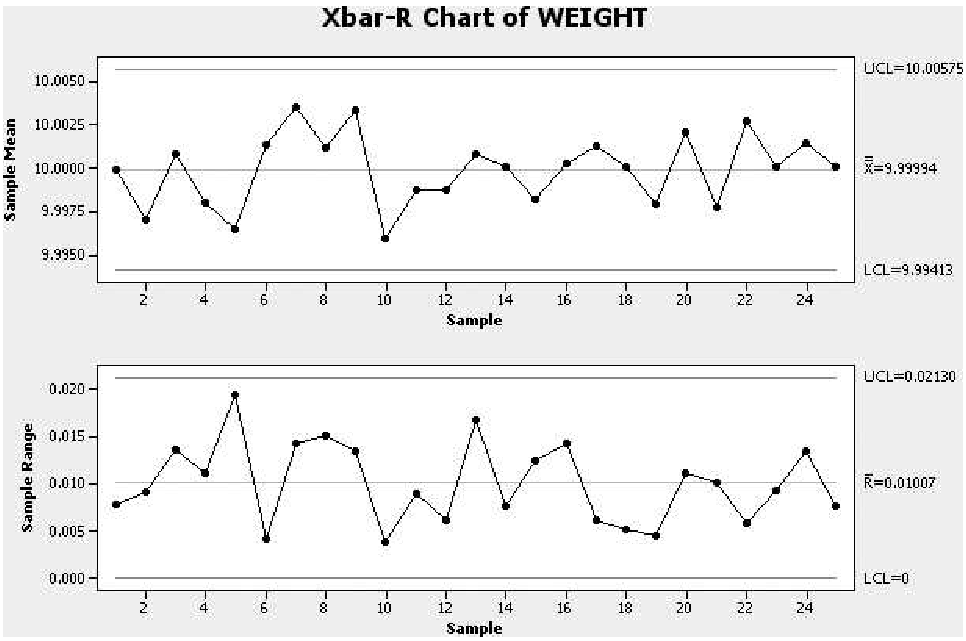



**FIGURE 16.19**
MINITAB control charts for paint can weights, Example 16.10

For two reasons, great care should be exercised in using and interpreting $C_p$. First, like the sample standard deviation, $s$, used in its computation, $C_p$ is a statistic and is subject to sampling error. That is, the value of $C_p$ will change from sample to sample. Thus, unless you understand the magnitude of the sampling error, you should be cautious in comparing the $C_p$'s of different processes. Second, $C_p$ does not reflect the shape of the output distribution. Distributions with different shapes can have the same $C_p$ value. Accordingly, $C_p$ should not be used in isolation, but in conjunction with the capability analysis diagram.

If a capability analysis study indicates that an in-control process is not capable, as in the paint-filling example, it is usually variation, rather than off-centeredness, that is the culprit. Thus, capability is typically achieved or restored by seeking out and eliminating common causes of variation.

## Applied Exercises

16.52 *Determining specification limits.* An in-control, centered process that follows a normal distribution has a $C_p = 2.0$. How many standard deviations away from the process mean is the upper specification limit?

16.53 *Finding $C_p$.* A process is in control with a normally distributed output distribution with mean 1,000 and standard deviation 100. The upper and lower specification limits for the process are 1,020 and 980, respectively.

a. Assuming no changes in the behavior of the process, what percentage of the output will be unacceptable?
b. Find and interpret the $C_p$ value of the process.

16.54 *Water use at a thermal power plant.* Thermal power plants use de-mineralized (DM) water for steam generation. Since it is costly to replace, power plants must conserve the use of DM water. DM water consumption was monitored at a thermal power plant in India, and the results published in *Total Quality Management* (Feb. 2009). Plant management set the target for DM water consumption at .5%, the upper specification limit at .7% and the lower specification limit at .1%. Based on data collected for a sample of 182 flow meter measurements, the overall standard deviation of the process was .265%. Use this information to find the capability index for this process. Interpret the result.

**CARBON2**

16.55 *New iron-making process.* Refer to the *Mining Engineering* (Oct. 2004) study of a new technology for producing high-quality iron nuggets, Exercise 16.2 (p. 894). The data on percent carbon change in produced nuggets for 33 time intervals is saved in the **CARBON2** file. Specifications state that the carbon content should be within $3.42 \pm 0.3\%$.

a. Construct a capability analysis diagram for the iron-making process.
b. Determine the proportion of carbon measurements that fall outside specifications.
c. Find the capability index for the process and interpret its value.

16.56 *Cereal box filing process.* A machine fills boxes with bran flakes. The target weight for the filled boxes is 24 ounces. To monitor the process, five boxes are randomly sampled from each day's production and weighed. The data for 20 consecutive days is given in the table below. Assume the specification limits for the weights are USL = 24.2 ounces and LSL = 23.8 ounces.

a. Assuming the process is under control, construct a capability analysis diagram for the process.
b. Is the process capable? Support your answer with a numerical measure of capability.

**CEREAL**

| Day | Weight of Cereal Boxes (ounces) | | | | |
|---|---|---|---|---|---|
| 1 | 24.02 | 23.91 | 24.12 | 24.06 | 24.13 |
| 2 | 23.89 | 23.98 | 24.01 | 24.00 | 23.91 |
| 3 | 24.11 | 24.02 | 23.99 | 23.79 | 24.04 |
| 4 | 24.06 | 23.98 | 23.95 | 24.01 | 24.11 |
| 5 | 23.81 | 23.90 | 23.99 | 24.07 | 23.96 |
| 6 | 23.87 | 24.12 | 24.07 | 24.01 | 23.99 |
| 7 | 23.88 | 24.00 | 24.05 | 23.97 | 23.97 |
| 8 | 24.01 | 24.03 | 23.99 | 23.91 | 23.98 |
| 9 | 24.06 | 24.02 | 23.80 | 23.79 | 24.07 |
| 10 | 23.96 | 23.99 | 24.03 | 23.99 | 24.01 |
| 11 | 24.10 | 23.90 | 24.11 | 23.98 | 23.95 |
| 12 | 24.01 | 24.07 | 23.93 | 24.09 | 23.98 |
| 13 | 24.14 | 24.07 | 24.08 | 23.98 | 24.02 |
| 14 | 23.91 | 24.04 | 23.89 | 24.01 | 23.95 |
| 15 | 24.03 | 24.04 | 24.01 | 23.98 | 24.10 |
| 16 | 23.94 | 24.07 | 24.12 | 24.00 | 24.02 |
| 17 | 23.88 | 23.94 | 23.91 | 24.06 | 24.07 |
| 18 | 24.11 | 23.99 | 23.90 | 24.01 | 23.98 |
| 19 | 24.05 | 24.04 | 23.97 | 24.08 | 23.95 |
| 20 | 24.02 | 23.96 | 23.95 | 23.89 | 24.04 |

16.57 *Military aircraft bolts.* A precision parts manufacturer produces bolts for use in military aircraft. The company sampled four consecutively produced bolts each hour on the hour for 25 consecutive hours and measured the length of each bolt. The data on lengths of bolts are shown in the table at the bottom of the page. Management has specified upper and lower specification limits of 37 cm and 35 cm, respectively.

a. Assuming the process is in control, construct a capability analysis diagram for the process.
b. Find the percentage of bolts that fall outside the specification limits.
c. Find the capability index, $C_p$.
d. Is the process capable? Explain.

16.58 *Bioreactor production of antibodies.* Bench-top bioreactors are used to produce antibodies for anti-cancer drugs. Engineers calibrate bioreactors in order to maximize production. The *African Journal of Biotechnology* (Dec. 2011) published a study designed to achieve a high percentage of antibody production from a bioreactor. The variable of interest was the natural logarithm of the number of viable cells produced in a bioreactor run. Data were collected for a sample of four bioreactor runs every six hours for 20 consecutive time periods. These data (simulated from information provided in the article) are listed in the table (right column). Engineers have specified the following for the bioreactor runs: target mean = 6.3, LSL = 5.9, and USL = 6.5. Run a complete capability analysis on the data. How would you categorize the performance of the process?

**BIOREACTOR**

| Time Period | Hour | Run 1 | Run 2 | Run 3 | Run 4 |
|---|---|---|---|---|---|
| 1 | 0 | 5.83 | 5.90 | 5.91 | 5.93 |
| 2 | 6 | 5.98 | 5.94 | 5.97 | 5.84 |
| 3 | 12 | 5.99 | 5.98 | 5.99 | 5.98 |
| 4 | 18 | 6.09 | 6.04 | 5.93 | 6.02 |
| 5 | 24 | 6.20 | 6.30 | 6.30 | 6.20 |
| 6 | 30 | 6.04 | 6.08 | 6.23 | 6.15 |
| 7 | 36 | 6.19 | 6.13 | 6.13 | 6.29 |
| 8 | 42 | 6.37 | 6.27 | 6.27 | 6.27 |
| 9 | 48 | 6.56 | 6.46 | 6.36 | 6.26 |
| 10 | 54 | 6.36 | 6.36 | 6.16 | 6.16 |
| 11 | 60 | 6.36 | 6.37 | 6.37 | 6.27 |
| 12 | 66 | 6.27 | 6.27 | 6.27 | 6.17 |
| 13 | 72 | 6.26 | 6.26 | 6.26 | 6.16 |
| 14 | 78 | 6.29 | 6.46 | 6.16 | 6.26 |
| 15 | 84 | 6.26 | 6.16 | 6.25 | 6.15 |
| 16 | 90 | 6.35 | 6.45 | 6.25 | 6.53 |
| 17 | 96 | 6.16 | 6.16 | 6.55 | 6.56 |
| 18 | 102 | 6.24 | 6.23 | 6.24 | 6.24 |
| 19 | 108 | 6.15 | 6.16 | 6.15 | 6.15 |
| 20 | 114 | 6.30 | 6.52 | 6.13 | 6.48 |

**BOLTS**

| Hour | Bolt Lengths (centimeters) | | | | Hour | Bolt Lengths (centimeters) | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 37.03 | 37.08 | 36.90 | 36.88 | 14 | 37.08 | 37.07 | 37.10 | 37.04 |
| 2 | 36.96 | 37.04 | 36.85 | 36.98 | 15 | 37.03 | 37.04 | 36.89 | 37.01 |
| 3 | 37.16 | 37.11 | 36.99 | 37.01 | 16 | 36.95 | 36.98 | 36.90 | 36.99 |
| 4 | 37.20 | 37.06 | 37.02 | 36.98 | 17 | 36.97 | 36.94 | 37.14 | 37.10 |
| 5 | 36.81 | 36.97 | 36.91 | 37.10 | 18 | 37.11 | 37.04 | 36.98 | 36.91 |
| 6 | 37.13 | 36.96 | 37.01 | 36.89 | 19 | 36.88 | 36.99 | 37.01 | 36.94 |
| 7 | 37.07 | 36.94 | 36.99 | 37.00 | 20 | 36.90 | 37.15 | 37.09 | 37.00 |
| 8 | 37.01 | 36.91 | 36.98 | 37.12 | 21 | 37.01 | 36.96 | 37.05 | 36.96 |
| 9 | 37.17 | 37.03 | 36.90 | 37.01 | 22 | 37.09 | 36.95 | 36.93 | 37.12 |
| 10 | 36.91 | 36.99 | 36.87 | 37.11 | 23 | 37.00 | 37.02 | 36.95 | 37.04 |
| 11 | 36.88 | 37.10 | 37.07 | 37.03 | 24 | 36.99 | 37.07 | 36.90 | 37.02 |
| 12 | 37.06 | 36.98 | 36.90 | 36.99 | 25 | 37.10 | 37.03 | 37.01 | 36.90 |
| 13 | 36.91 | 37.22 | 37.12 | 37.03 | | | | | |

## 16.10 Acceptance Sampling for Defectives

In the preceding sections, we have learned how control charts can be used during the manufacturing process to monitor and improve the quality of a product. After manufacturing, items of the product are stored (and packaged) in *lots* containing anywhere from two to many thousands of items per lot, the *lot size* depending on the nature of the product. At this point, just prior to shipment, a second statistical tool—an **acceptance sampling plan**—is often employed to reduce the proportion of defective items shipped to customers.

An acceptance sampling plan works in the following way. A fixed number $n$ of items is sampled from each lot, carefully inspected, and each item is judged to be either defective or nondefective. If the number $y$ of defectives in the sample is less than or equal to prespecified **acceptance number** $\boldsymbol{a}$, the lot is accepted. If the number of defectives exceeds $a$, the lot is rejected and withheld for either a second sampling, a complete inspection, or some other procedure (see Figure 16.20). The objectives of the sampling plan are to accept and ship lots containing a small fraction $p$ of defectives, to reject and withhold lots containing a high fraction of defectives, and to do both with a high probability.

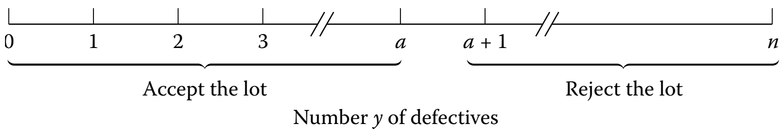



**FIGURE 16.20**
Accepting or rejecting lots based on the number of defectives in a sample of $n$ items

At this point you may wonder why quality control engineers resort to sampling rather than an inspection of all items in the lot. That is, why not 100% inspection? First, 100% inspection often turns out to be impractical or uneconomical. Second, studies have shown that the quality of the product shipped is often better with acceptance sampling than with 100% inspection, especially when there are a great many similar items of a product to be inspected. With 100% inspection, inspectors' fatigue on repetitive operations is always a danger. Also, psychologically, laborers have more of a tendency to make a quality product when only a few items are inspected.

Upon reflection, you can see that the decision procedure for accepting or rejecting a lot with acceptance sampling is simply a test of a hypothesis about the lot fraction defective $p$. The manufacturer (or customer) has in mind some lot fraction defective, say, $p_0$, called the **acceptable quality level** (**AQL**). If the lot fraction $p$ is below $p_0 = \text{AQL}$, the lot is deemed acceptable. The probability $\alpha$ of rejecting

$$H_0: \quad p = p_0$$

if in fact $p = p_0$ (that is, if the lot is actually acceptable) is called the producer's risk. In other words, even if $p = p_0$, the manufacturer (the producer) will withhold $100\alpha\%$ of the acceptable lots from shipment and be subjected to the cost of resampling, and so on.

**Definition 16.12**

The **acceptable quality level (AQL)** is an upper limit, $p_0$, on the fraction defective that a producer is willing to tolerate.

**Definition 16.13**

The **producer's risk** is the probability $\alpha$ of rejecting lots if in fact the lot fraction defective is equal to $p_0$, the acceptable quality level. In the terminology of hypothesis testing, the producer's risk is the probability of a Type I error.

The consumer, the purchaser of the product, is also subject to a risk—namely, the risk of accepting lots containing a high fraction defective $p$. The consumer will usually have in mind a lot fraction defective $p_1$, which is the largest lot fraction defective that he or she will tolerate. The probability $\beta$ of accepting lots containing fraction defective $p_1$ is called the **consumer's risk**.

**Definition 16.14**

The **consumer's risk** is the probability $\beta$ of accepting lots containing fraction defective $p_1$, where $p_1$ is the upper limit in lot fraction defective acceptable to the consumer. In the terminology of hypothesis testing, the consumer's risk is the probability of a Type II error.

An **operating characteristic curve** is a graph of the probability of lot acceptance $P(A)$ versus lot fraction defective $p$. A typical operating characteristic curve, shown in Figure 16.21, completely characterizes a sampling plan and shows the probability of lot acceptance equal to 1 when $p = 0$ and equal to 0 when $p = 1$. As the lot fraction defective $p$ increases, the probability $P(A)$ of lot acceptance decreases until it reaches 0. The producer's risk $\alpha$ is equal to $1 - P(A)$ when $p = p_0$. The consumer's risk $\beta$ is equal to $P(A)$ when $p = p_1$.

**Definition 16.15**

The **operating characteristic (OC) curve** for a sampling plan is a graph of the probability of lot acceptance, $P(A)$, versus the lot fraction defective, $p$.

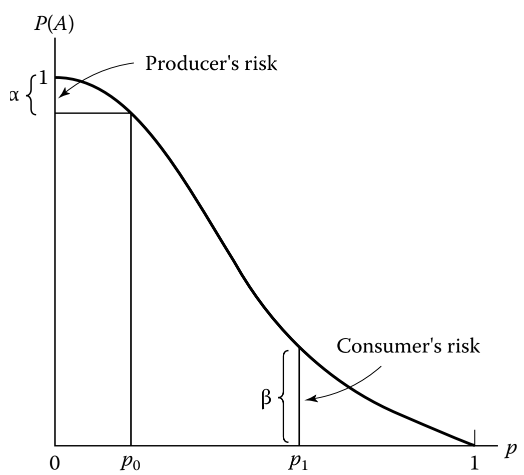
The operating characteristic curve for a sampling plan can be constructed by calculating $P(A)$ for various values of the lot fraction defective $p$. As explained in Sections 4.6 and 4.9, the probability distribution for the number $y$ of defectives in a sample of $n$ items from a lot will depend on the lot size $N$. If $N$ is large and $n$ is small relative to $N$, then the probability distribution for $y$ can be approximated by a *binomial probability distribution* (Section 4.6):

$$p(y) = \binom{n}{y} p^y q^{n-y} \qquad y = 0, 1, 2, \ldots, n$$

where

$$q = 1 - p$$

**FIGURE 16.21**
A typical operating characteristic curve



If $N$ is small or $n$ is large relative to $N$, then $y$ will have a *hypergeometric probability distribution* (Section 4.9):

$$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

where

$$\begin{aligned} N &= \text{Lot size} \\ r &= \text{Number of defectives in the lot} \\ p &= \frac{r}{N} = \text{Lot fraction defective} \\ n &= \text{Sample size} \\ y &= \text{Number of defectives in the sample} \end{aligned}$$

Using the appropriate probability distribution for a sampling plan with sample size $n$ and acceptance number $\alpha$, we can compute the probability of accepting a lot with lot fraction defective $p$:

$$P(A) = P(y \leq a) = p(0) + p(1) + \cdots + p(a)$$

We will illustrate the procedure with the next example.

## Example 16.11

Producer's and Consumer's Risk

A manufacturer of metal gaskets ships a particular gasket in lots of 500 each. The acceptance sampling plan used prior to shipment is based on a sample size $n = 10$ and acceptance number $\alpha = 1$.

a. Find the producer's risk if the AQL is .05.
b. Find the consumer's risk if the lot fraction defective is $p_1 = .20$.
c. Draw a rough sketch of the operating characteristic curve for the sampling plan.

Solution

a. The producer's risk is $\alpha = 1 - P(A)$ when $p = p_0 = .05$. For $N = 500$ and $n = 10$, $y$ will possess approximately a binomial probability distribution. Then, if in fact $p = .05$,

$$\begin{aligned} P(A) &= p(0) + p(1) \\ &= \binom{10}{0}(.05)^0(.95)^{10} + \binom{10}{1}(.05)^1(.95)^9 = .914 \end{aligned}$$

and the producer's risk is

$$\alpha = 1 - P(A) = 1 - .914 = .086$$

This means that the producer will reject 8.6% of the lots, even if the lot fraction defective is as small as .05.

b. The consumer's risk is $\beta = P(A)$ when $p = .20$:

$$\begin{aligned} \beta = P(A) &= p(0) + p(1) \\ &= \binom{10}{0}(.2)^0(.8)^{10} + \binom{10}{1}(.2)^1(.8)^9 = .376 \end{aligned}$$

Thus, the consumer risks accepting lots containing a lot fraction defective equal to $p_1 = .20$ approximately 37.6% of the time. The fact that $\beta$ is so large for $p_1 = .20$ indicates that this sampling plan would be of little value in practice. The plan needs to be based on a larger sample size.

c. A rough sketch of the operating characteristic curve for the sampling plan can be obtained using the two points calculated in parts **a** and **b** and the fact that $P(A) = 1$ when $p = 0$ and $P(A) = 0$ when $p = 1$. The sketch is shown in Figure 16.22.

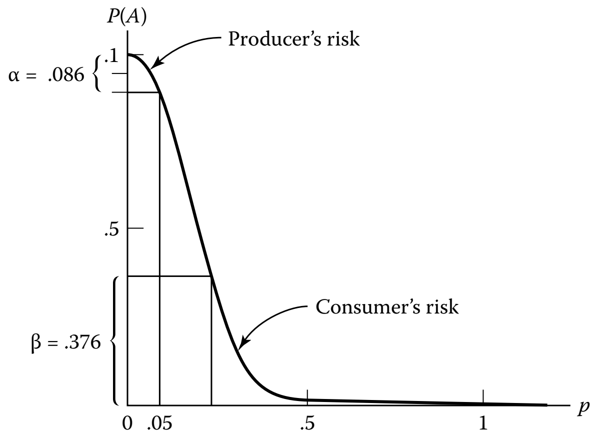



**FIGURE 16.22**
A rough sketch of the operating characteristic curve of $n = 10$ and $\alpha = 1$

In practice, engineers do not construct sampling plans for specific lot sizes and AQLs because they have been constructed and have been in use for years. One of the most widely used collections of sampling plans is known as the **Military Standard 105D (MIL-STD-105D)**. The sampling plans contained in MIL-STD-105D employ a sample size $n$ that varies with the lot size $N$. The sample sizes specified in the plans were chosen to give reasonable values of consumer risk. In addition, the plans have been constructed so that each falls into one of three levels of inspection categories: reduced (I), normal (II), or tightened (III). Lower consumer risks are associated with tighter plans.

Two of the MIL-STD-105D tables are reproduced in Tables 22 and 23 of Appendix B. The following example illustrates their use.

## Example 16.12

An Inspection Sampling Plan

Find the appropriate MIL-STD-105D normal (level) general inspection sampling plan for a lot size of 500 items and an acceptable quality level of .065.

Solution

The first step in selecting the sampling plan is to identify the MIL-STD-105D code corresponding to a lot size of 500 and a normal inspection level—that is, level II. This code letter, H, is found in Table 22 of Appendix B in the row corresponding to lot size 281–500 and in the column labeled II.

The second step in selecting the plan is to determine the sample size and acceptance number from Table 23 of Appendix B. The sample size code letters appear in the first column of the table. The recommended sample sizes are shown in the second column. Moving down column 1 to code letter H, we see that the recommended sample size (column 2) is $n = 50$. To find the acceptance number, move across the top row to 6.5%, or, equivalently, AQL $= .065$. The acceptance (Ac) number, $a = 7$, is shown at the intersection of the 6.5 column and the H row. The number 8 that also appears at this intersection is the rejection number for the sampling plan—that is, we reject a lot if $y$ is greater than or equal to 8.

You can see that this MIL-STD-105D sampling plan uses a much larger sample ($n = 50$) than the plan of Example 16.11. Because of this larger sample size, the probability of lot acceptance, $P(A)$, calculated for a given lot fraction defective $p$, would be much smaller than for the plan of Example 16.11. We would say that the MIL-STD-105D plan is *tighter* than the plan of Example 16.11. The consumer risk is less or, equivalently, it allows fewer bad lots to be shipped.

The probability of acceptance $P(A)$ for the MIL-STD-105D sampling plan can be calculated as described earlier in this section. For example, for a lot fraction defective $p = .10$ in Example 16.12, we have

$$P(A) = P(y \leq 7) = \sum_{y=0}^{7} p(y)$$

where $p(y)$ is a hypergeometric probability distribution with $N = 500$, $n = 50$, and the number $r$ of defectives in the lot is $Np = (500)(.1) = 50$. The actual calculation of $P(A)$ is tedious and is best accomplished by using a computer.

## Applied Exercises

16.59 *Sampling plan analysis.* Consider a sampling plan with sample size $n = 15$ and acceptance number $\alpha = 1$.

a. Calculate the probability of lot acceptance for fractions defective $p = .1, .2, .3, .4$, and $.5$. Sketch the operating characteristic curve for the plan.
b. Find the producer's risk if AQL $= .05$.
c. Find the consumer's risk if $p_1 = .20$.

16.60 *Sampling plan analysis.* Consider a sampling plan with sample size $n = 5$ and acceptance number $a = 0$.

a. Calculate the probability of lot acceptance for fractions defective $p = .1, .3$, and $.5$. Sketch the operating characteristic curve for the plan.
b. Find the producer's risk if AQL $= .01$.
c. Find the consumer's risk if $p_1 = .10$.

16.61 *Wire tensile strength.* The tensile strengths of wires in a certain lot of size 400 are specified to exceed 5 kilograms. Consider an acceptance sampling plan based on a sample of $n = 10$ wires and acceptance number $\alpha = 1$.

a. Find the producer's risk if the AQL is 2.5%.
b. Find the consumer's risk if the lot fraction failing to meet specifications is $p_1 = .15$.
c. Draw a rough sketch of the operating characteristic curve for the sampling plan. Do you think the sampling plan is acceptable? Explain.

16.62 *Wire tensile strength (continued).* Refer to Exercise 16.61. Find the appropriate MIL-STD-105D normal (level) general inspection sampling plan for a lot size of 400 wires and an AQL of 2.5%.

16.63 *Finding a sampling plan.* Find the appropriate MIL-STD-105D general inspection sampling plan for a lot size of 5,000 items and an AQL of 4% under each of the following inspection categories:

a. Reduced (I) inspection level
b. Normal (II) inspection level
c. Tightened (III) inspection level

## 16.11 Other Sampling Plans (*Optional*)

In Section 16.10, we presented a sampling plan based on the number of defectives contained in a single sample. A second type of acceptance sampling plan is one based on double or multiple sampling. A **double sampling plan** involves the selection of $n_1$ items from the lot. The lot is accepted if the number $y_1$ of defectives in the sample is $y_1 \leq a_1$ and rejected if $y_1 \geq r_1$ (where $r_1 > a_1$), as shown in Figure 16.23. If $y_1$ falls *between* $\alpha_1$ and $r_1$, then a second sample of $n_2$ items is selected from the lot and the total number $y$ of defectives in the $(n_1 + n_2)$ sampled items is recorded. If $y$ is less than or equal to a second acceptance number $\alpha_2$, the lot is accepted; otherwise, it is rejected.

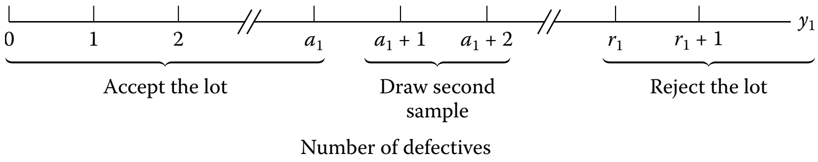



**FIGURE 16.23**
Location of the acceptance number $a_1$ and rejection number $r_1$ for the first sample in a double sampling plan

The ultimate in multiple sampling is **sequential sampling**. In a sequential sampling plan, the items are selected from the lot, one-by-one. As each item is selected, a decision is made to accept the lot, to reject the lot, or to sample the next item from the lot. With this type of sampling, the decision to accept (or to reject) the lot might occur as early as the first, second, or third items sampled. It is also possible that the decision to accept or to reject the lot might require a very large sample. Thus, in sequential sampling, the sample size $n$ is a random variable.

In addition to single, multiple, and sequential sampling plans based on the number $y$ of defects observed, similar plans have been developed to utilize measurements on quantitative variables. Thus, instead of examining each item in a sample and rating it as defective or nondefective, we make our decision to reject or to accept the lot based on a quantitative measurement taken on each of the items. For example, a purchaser of 50-gallon barrels of acetone might be primarily concerned that each barrel contain at least 50 gallons. A typical sampling plan might involve sampling 10 barrels from each lot and measuring the exact number $y$ of gallons in each barrel. We could classify each barrel that contains less than 50 gallons as defective and base our decision to reject or to accept the lot on the number of defective barrels in the sample. Alternatively, we could base our decision on the sample mean, $\bar{y}$, the average amount of acetone in the 10 barrels. A sampling plan based on the mean of a sample of quantitative measurements is called **acceptance sampling by variables**. One of the most widely used collections of such sampling plans is **Military Standard 414 (MIL-STD-414)**.

The literature on acceptance sampling plans is extensive. For collections of sampling plans and for more information on the subject, we refer you to the references at the end of the chapter. Before leaving this discussion, however, we leave you with this thought: *It does not always pay to sample*. There may be certain situations where the cost of sampling is so prohibitive that the only alternatives are either 100% inspection or no inspection at all. Thus, total cost plays an important role in the acceptance sampling plan selection process.

## 16.12 Evolutionary Operations (*Optional*)

An **evolutionary operation** is a technique designed to improve the yield and/or the quality of an industrial product by extracting information from an operating process. To illustrate the procedure, suppose that some quality characteristic of a chemical product—say, viscosity—is dependent on a number of variables, including the temperature of the raw materials and the pressure maintained within the vat in which they are mixed. To investigate the effect of these variables on the viscosity of a batch, we could simulate the process in a laboratory and conduct a multivariable experiment (for example, a factorial experiment) as described in Chapter 13. But this process would be costly and it is possible that the simulation would behave differently from the production process.

A second and less costly procedure is to concentrate on only two or three of the independent variables and to vary the settings of these variables according to a designed experiment. The key is to make the changes in the independent variables so

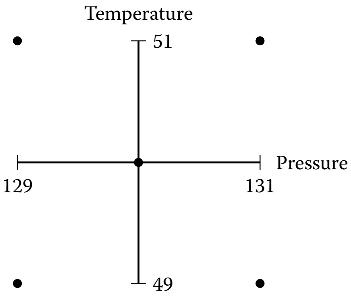



**FIGURE 16.24**
An experimental design for an evolutionary operation

small that there is no *observable* change in the quality of the product. To detect the effect of these small changes, we repeat the experiment over and over again until the sample sizes are so large that even small changes in the mean value of the quality variable are significant when tested statistically.

For example, suppose we know that a number of controllable process variables, including the temperature and pressure of raw materials, affect the viscosity of a batch-produced chemical. We are afraid to make experimental changes in these variables out of fear that we might produce a bad product and an accompanying financial loss. However, we know that very slight changes in temperature and pressure—say, changes of 2°F and 2 pounds per square inch (psi)—would have a negligible effect on product quality.

To investigate the effects of temperature and pressure, we will conduct an experiment in the operating process using the experimental design shown in Figure 16.24. The four temperature–pressure combinations at the corners of the design are the four factor-level combinations of a $2 \times 2$ factorial experiment. The pressure–temperature combination (50°F, 130 psi) was added at the center of the design region to enable us to detect a relatively high (or low) mean viscosity in the center of the experimental region, in case it exists.

To conduct the evolutionary operation, we would assign one of the five pressure–temperature combinations to each batch of chemical and measure the viscosity $y$ for each. If the manufacturer produces 10 batches per day, we would obtain two replications of the five treatments contained in the design shown in Figure 16.24. If we were to conduct statistical tests to detect changes in the mean viscosity based on the data for 1 day, or perhaps even for 100 days, it is conceivable that no changes in mean viscosity would be evident. However, if we continue to collect data over a long period of time, obtaining two replications of the experiment each day, we would eventually detect changes in mean viscosity (if they exist). Thus, the logic of an evolutionary operation is that a production process produces data at the same time that it generates a product. Why not utilize the information that is free (except for the cost of collection)? Although the individual observations contain very little information on the effect that pressure and temperature have on mean viscosity, the weight of huge amounts of data eventually will show us how to change these variables to produce desirable changes in mean viscosity. Thus, repeated experimentation over time enables the process to evolve to a higher level of quality and/or yield.

## *STATISTICS IN ACTION REVISITED*
### Testing Jet Fuel Additive for Safety

We now return to the problem of testing for surfactants (surface active agents) in jet fuel. A standard test for surfactants involves pumping a water/fuel mixture through the filter at a specific rate. Recall that an engineering firm wants to compare the standard test (Pump-A with Filter-A) to three other pumping mechanism and filter option combinations—Pump-A with Filter-B, Pump-B with Filter-A, and Pump-B with Filter-B. For each of over 100 days, the firm obtained three test results for each pump/filter method. The test measurements are saved in four **JET** files.

Does one of the test methods yield the most stable process? To answer this question, we will apply the quality control methods of this chapter to the data. Since a “safe” surfactant additive measurement should range between 80 and 90, this range represents the specification limits of the process.

Treating the three samples collected on the same day as a rational subgroup, four MINITAB $\bar{x}$-charts are produced (one for each pump/filter method) in Figures SIA16.1a–d. As an option, MINITAB will highlight (in gray) any sample means that match any of six pattern-analysis rules for detecting special causes of

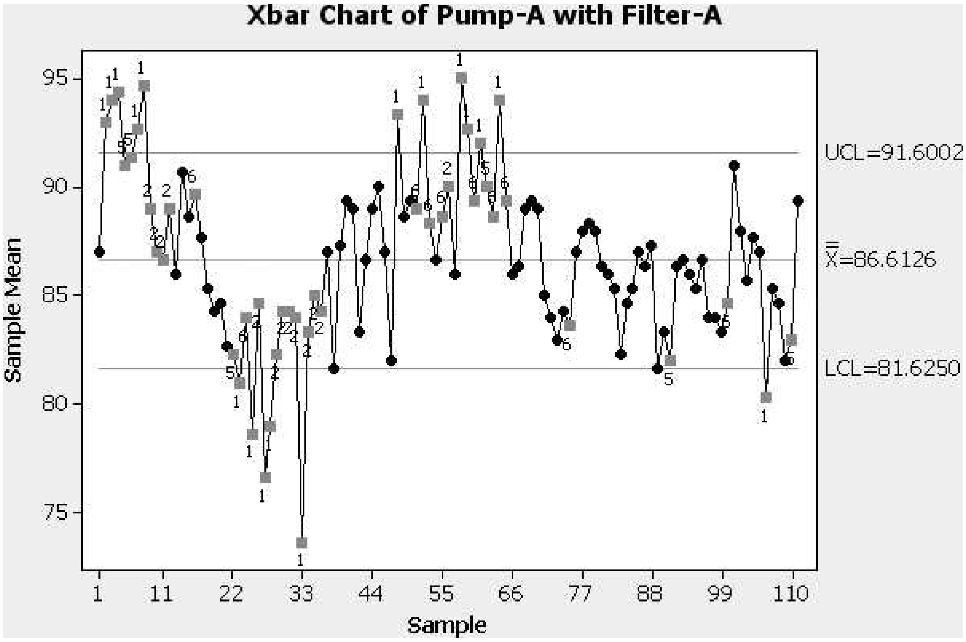



**FIGURE SIA16.1a**
MINITAB $\bar{x}$-chart for pump-A with filter-A (standard) method

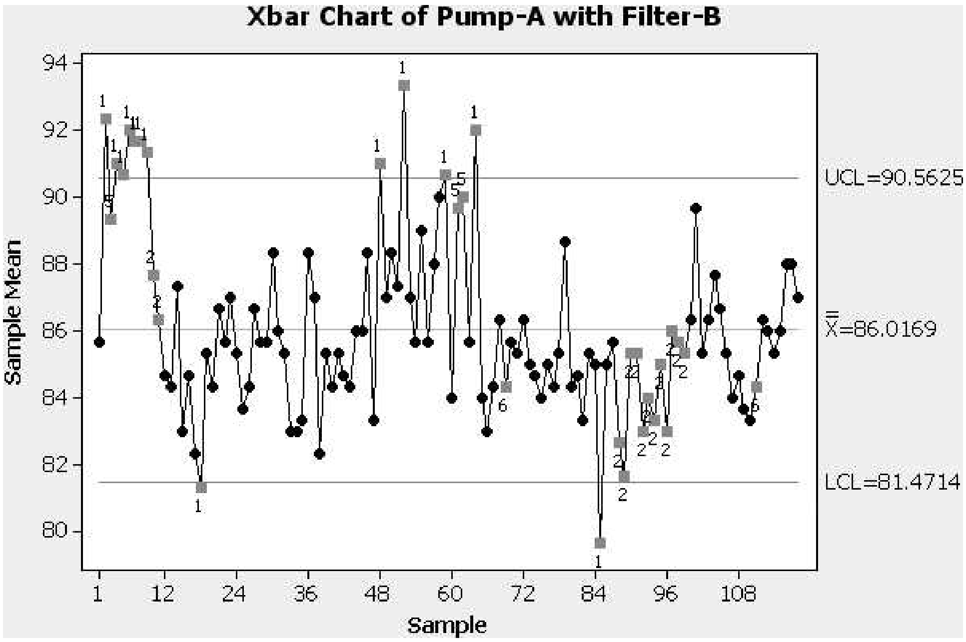



**FIGURE SIA16.1b**
MINITAB $\bar{x}$-chart for pump-A with filter-B method

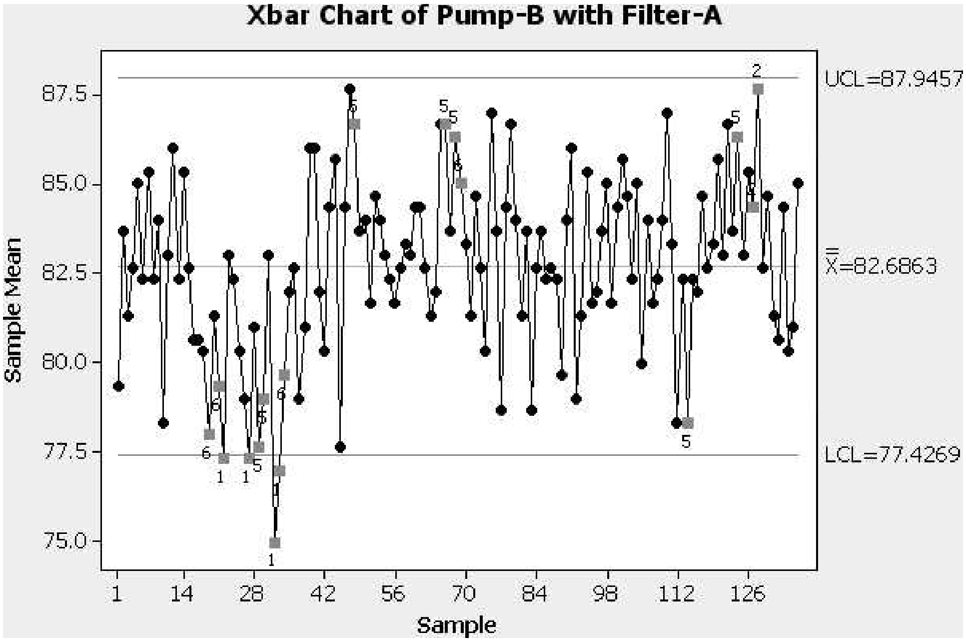



**FIGURE SIA16.1c**
MINITAB $\bar{x}$-chart for pump-B with filter-A method

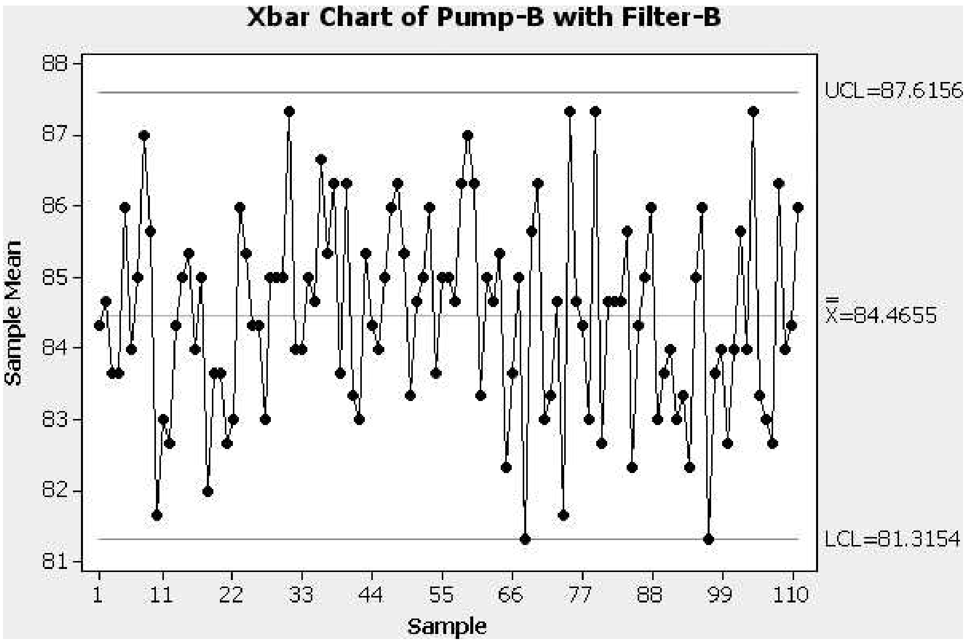



**FIGURE SIA16.1d**
MINITAB $\bar{x}$-chart for pump-B with filter-B method

**TABLE SIA16.2 Pattern Analysis Rules for Detecting Special Causes of Variation in a Control Chart**

| | |
|---|---|
| **Rule 1:** | At least one point falling beyond the 3 standard deviation control limits. |
| **Rule 2:** | Nine or more points in a row falling on the same side of the center line. |
| **Rule 3:** | Six or more points steadily increasing (or decreasing). |
| **Rule 4:** | Fourteen or more points in a row alternating up and down. |
| **Rule 5:** | Two out of three points in a row falling beyond the 2 standard deviation control limit above (or below) the center line. |
| **Rule 6:** | Four out of five points in a row falling beyond the 1 standard deviation control limit above (or below) the center line. |

*Note*: Rules 1–6 are used for $\bar{x}$-charts. Rules 1–4 are used for *R*-charts.

variation shown in Table SIA16.2. (*Note*: In this chapter, we discussed Rule 1—points that fall outside the 3 standard deviation control limits.) The number of the rule that is violated is shown next to the sample mean on the chart.

You can see that only one of the process means is "in control"—the mean for the Pump-B with Filter-B test method—as shown in Figure SIA16.1d. There is at least one pattern-analysis rule violated in each of the other three $\bar{x}$-charts. Also, each sample mean for Pump-B/Filter-B falls within the specification limits (80–90%). In contrast, the other injection methods have several means that fall outside the specification limits of the process. Of the three nonstandard surfactant test methods, the Pump-B/Filter-B method appears to have the most promise. However, as discussed in this chapter, the variation of the process should be checked first before interpreting the $\bar{x}$-chart.

Figure SIA16.2 is a MINITAB *R*-chart for the test results using Pump-B with Filter-B. As an option, we instructed MINITAB to highlight (in gray) any sample ranges that match any of the first four pattern-analysis

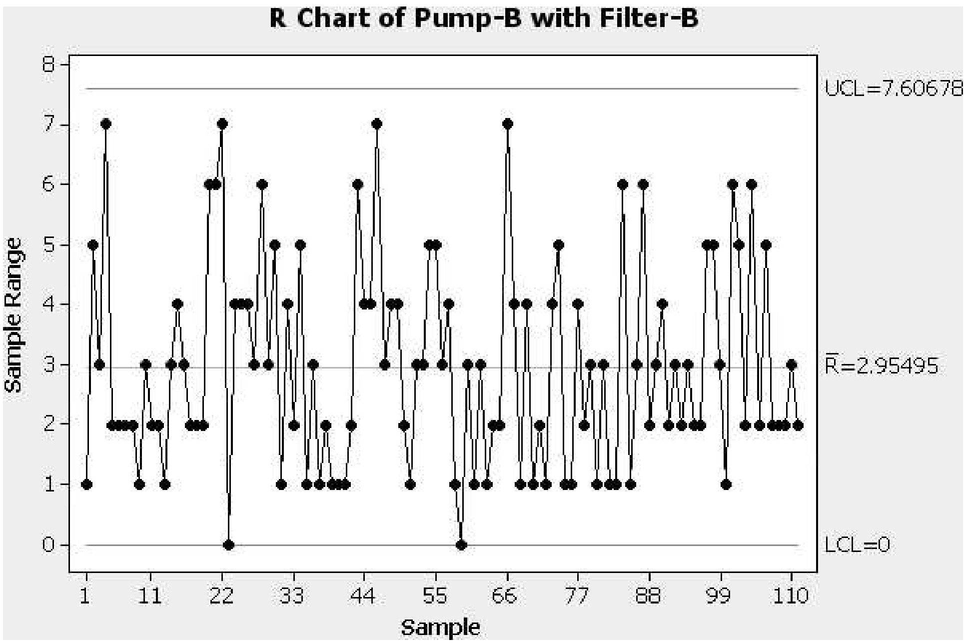



**FIGURE SIA16.2**
*R*-chart for pump-B with filter-B method

rules given in Table SIA16.2. (If a rule is violated, the rule number will be shown next to the sample range on the chart.) Figure SIA16.2 shows that the process variation is "in control"—none of the pattern-analysis rules for ranges are matched. Now that we've established the stability of the process variance, the $\bar{x}$-chart of Figure SIA16.1d can be meaningfully interpreted. Together, the $\bar{x}$-chart and $R$-chart helped the engineering firm establish the Pump-B with Filter-B surfactant test method as a viable alternative to the standard test, one which appears to have no special causes of variation present and with more precision than the standard.

(*Note*: Extensive testing done with the Navy concluded the improved precision of the "new" surfactant test was valid. However, the new test was unable to detect several light surfactants that can still cause problems in jet engines. The original test for surfactants in jet fuel additive remains the industry standard.)

# Quick Review

## Key Terms

*Note*: Starred (*) items are from the optional sections in this chapter.

acceptable number 933
acceptable quality level 933
acceptance sampling plan 933
*acceptance sampling by variables 938
assignable cause variation 890
*capability analysis 926
*capability analysis diagram 926
*capability index 928
$c$-chart 917
center line 918
consumer's risk 934
control chart 890
control limits 891
*double sampling plan 937
*evolutionary operations 938
in control 891
individuals chart 892
lower control limit 892
means control chart 896
operating characteristic curve 934
out of control 891
$p$-chart 912
*process spread 929
process variation 904
producer's risk 933
quality variable 890
$R$-chart 904
random (chance) variation 891
range control chart 904
rational subgroups 900
run 910
*sequential sampling 938
specification limits 924
*specification spread 929
statistical process control 890
theory of runs 910
tolerance interval 921
tolerance limits 921
total quality management 890
upper control limit 892
variable control chart 890
$\bar{x}$-chart 896

## Key Formulas

| Control Chart | Center Line | Control Limits (Lower, Upper) | |
|---|---|---|---|
| Variable chart | $\bar{x} = \dfrac{\sum x_i}{n}$ | $\bar{x} \pm 3s$ | 892 |
| $\bar{x}$-chart | $\bar{\bar{x}} = \dfrac{\sum_{i=1}^{k} \bar{x}_i}{k}$ | $\bar{\bar{x}} \pm A_2\bar{R}$ or $\bar{\bar{x}} \pm 3\dfrac{(\bar{R}/d_2)}{\sqrt{n}}$ | 897 |
| $R$-chart | $\bar{R} = \dfrac{\sum_{i=1}^{k} R_i}{k}$ | $(\bar{R}D_3, \bar{R}D_4)$ | 905 |
| $p$-chart | $\bar{p} = \dfrac{\text{Total number defectives}}{\text{Total number units sampled}}$ | $\bar{p} \pm 3\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}$ | 912 |
| $c$-chart | $\bar{c} =$ Average number of defects per item | $\bar{c} \pm 3\sqrt{\bar{c}}$ | 918 |
| USL − LSL | Specification spread* | | 929 |

## Key Formulas (continued)

| | | |
|---|---|---|
| $6\sigma \approx 6s$ | Process spread* | 929 |
| $(\text{USL} - \text{LSL})/6\sigma$ | $C_p$ index* | 929 |
| $\bar{x} \pm K_s$ where $K$ depends on $\alpha, \gamma, n$ | $(1-\alpha)100\%$ Tolerance interval for $100\gamma$ % of the measurements | 922 |
| $(x_{\text{min}}, x_{\text{max}})$ | Nonparametric tolerance interval | 925 |

## Chapter Summary Notes

- **Total quality management (TQM)**—involves the management of quality in all phases of a business.
- A process **in statistical control** has an output distribution that does not change over time; if it does change, the process is **out of control**.
- **Statistical process control (SPC)**—the process of monitoring and eliminating variation to keep a process in control.
- Two causes of variation—**assignable causes** and **random (chance) variation**.
- **Specification limits**—define acceptable values for an output variable.
- **Rational subgroups**—samples designed to make it more likely that process changes will occur between (rather than within) subgroups.
- A control chart to monitor a variable—the **variable (individuals) chart**.
- A control chart to monitor the **process mean**—the **$\bar{x}$-chart**.
- A control chart to monitor **process variation**—the ***R*-chart**.
- A control chart to detect trends—the **run chart**.
- A control chart to monitor the **proportion noncomforming**—the ***p*-chart**.
- A control chart to monitor number of defects per item—the ***c*-chart**.
- Interpret the $\bar{x}$-chart only after establishing that the process variation is in control with the $R$-chart.
- **Capability analysis**—used to determine if process is capable of satisfying its customers.
- **Capability index ($C_p$)**—summarizes the performance of a process relative to the specification limits.
- **Quality assurance sampling plans**—used to prevent bad lots of product from being shipped.
- An **operating characteristic curve**—a graph of the probability of accepting the lot versus the fraction defective.
- **Evolutionary operations**—experimenting and improving quality during ongoing manufacturing operations.

## LANGUAGE LAB

*Note: Starred (*) terms are from the optional sections in this chapter.*

| Symbol | Pronunciation | Description |
|---|---|---|
| SPC | S-P-C | Statistical process control |
| TQM | T-Q-M | Total quality management |
| LCL | L-C-L | Lower control limit |
| UCL | U-C-L | Upper control limit |
| $\bar{\bar{x}}$ | $x$-bar-bar | Average of the sample means |
| $\bar{R}$ | $R$-bar | Average of the sample ranges |
| $A_2$ | $A$-two | Constant obtained from Table 19, Appendix B |
| $D_3$ | $D$-three | Constant obtained from Table 19, Appendix B |
| $D_4$ | $D$-four | Constant obtained from Table 19, Appendix B |
| $d_2$ | $d$-two | Constant obtained from Table 19, Appendix B |
| $d_3$ | $d$-three | Constant obtained from Table 19, Appendix B |
| $\hat{p}$ | $p$-hat | Estimated number of defectives in sample |

## LANGUAGE LAB (continued)

| | | |
|---|---|---|
| $\bar{p}$ | *p*-bar | Overall proportion of defective units in all *nk* samples |
| $\bar{c}$ | *c*-bar | Average number of defects per item over all *k* time periods |
| *K* | | Constant obtained from Table 20, Appendix B |
| USL* | U-S-L | Upper specification limit |
| LSL* | L-S-L | Lower specification limit |
| $C_p$* | *C-p* | Capability index |
| $\gamma$ | gamma | Proportion of measurements in a population |
| AQL | A-Q-L | Acceptable quality level |

## Applied Supplementary Exercises

*Note*: Starred (*) exercises are from the optional sections in this chapter.

16.64 *Pitch diameters of threads.* One of the operations in a plant consists of thread-grinding a fitting for an aircraft hydraulic system. To monitor the process, a production supervisor randomly sampled five fittings for each hour, for a period of 20 hours, and measured the pitch diameters of the threads. The measurements, expressed in units of .0001 inch in excess of .4000 inch, are shown in the table. (For example, the value 36 represents .4036 inch.)

**THREADS**

| Hour | Pitch Diameters of Threads | | | | |
|---|---|---|---|---|---|
| 1 | 36 | 35 | 34 | 33 | 32 |
| 2 | 31 | 31 | 34 | 32 | 30 |
| 3 | 30 | 30 | 32 | 30 | 32 |
| 4 | 32 | 33 | 33 | 32 | 35 |
| 5 | 32 | 34 | 37 | 37 | 35 |
| 6 | 32 | 32 | 31 | 33 | 33 |
| 7 | 33 | 33 | 36 | 32 | 31 |
| 8 | 23 | 33 | 36 | 35 | 36 |
| 9 | 43 | 36 | 35 | 24 | 31 |
| 10 | 36 | 35 | 36 | 41 | 41 |
| 11 | 34 | 38 | 35 | 34 | 38 |
| 12 | 36 | 38 | 39 | 39 | 40 |
| 13 | 36 | 40 | 35 | 26 | 33 |
| 14 | 36 | 35 | 37 | 34 | 33 |
| 15 | 30 | 37 | 33 | 34 | 35 |
| 16 | 28 | 31 | 33 | 33 | 33 |
| 17 | 33 | 30 | 34 | 33 | 35 |
| 18 | 27 | 28 | 29 | 27 | 30 |
| 19 | 35 | 36 | 29 | 27 | 32 |
| 20 | 33 | 35 | 35 | 39 | 36 |

*Source*: Grant, E. L., and Leavenworth, R. S. *Statistical Quality Control*, 5th ed. New York: McGraw-Hill, 1980 (Table 1-1). Reprinted with permission.

a. Construct an *R*-chart to monitor the variation in pitch diameter. Is the process in control?
b. Modify the control limits on the *R*-chart so that it can be applied to future data.
c. Construct an $\bar{x}$-chart for the process. Does the process mean appear to be in control?
d. Eliminate the points that fall outside the control limits and recalculate their values. Would you recommend using these modified control limits for future data?

16.65 *Diameters of electrical shafts.* Suppose the process for manufacturing electrical shafts is in control. At the end of each hour, for a period of 20 hours, the manufacturer randomly selects one shaft and measures the diameter. The measurements (in inches) for the 20 samples are recorded in the table. Construct and interpret a control chart for shaft diameter.

**ELECSHAFT**

| Sample | Diameter (inches) | Sample | Diameter (inches) |
|---|---|---|---|
| 1 | 1.505 | 11 | 1.491 |
| 2 | 1.496 | 12 | 1.486 |
| 3 | 1.516 | 13 | 1.510 |
| 4 | 1.507 | 14 | 1.495 |
| 5 | 1.502 | 15 | 1.504 |
| 6 | 1.502 | 16 | 1.499 |
| 7 | 1.489 | 17 | 1.501 |
| 8 | 1.485 | 18 | 1.497 |
| 9 | 1.503 | 19 | 1.503 |
| 10 | 1.485 | 20 | 1.494 |

16.66 *Nickel in steel valves.* Specifications require the nickel content of manufactured stainless steel hydraulic valves to be 13% by weight. To monitor the production process, four valves were selected from the production line each hour over an 8-hour period and the percentage nickel content was measured for each, with the results recorded in the next table.

**PCTNICKEL**

| *Hour* | *Nickel Content* | | | |
|---|---|---|---|---|
| 1 | 13.1 | 12.8 | 12.7 | 12.9 |
| 2 | 12.5 | 13.0 | 13.6 | 13.1 |
| 3 | 12.9 | 12.9 | 13.2 | 13.3 |
| 4 | 12.4 | 13.0 | 12.1 | 12.6 |
| 5 | 12.8 | 11.9 | 12.7 | 12.4 |
| 6 | 13.0 | 13.6 | 13.2 | 12.9 |
| 7 | 13.5 | 13.5 | 13.1 | 12.7 |
| 8 | 12.6 | 13.9 | 13.3 | 12.8 |

a. Construct a control chart for the mean nickel content of the hydraulic valves.
b. Establish control limits for the mean using Table 19 of Appendix B.
c. Establish control limits for the mean using the standard deviation of the overall sample. Compare to the limits obtained in part **b**.
d. Do all observed sample means lie within the control limits? What are the consequences of this?
e. Find a 99% tolerance interval for 99% of the nickel contents in the hydraulic valves. Assume that the distribution of nickel contents is approximately normal.
f. Construct a control chart with control limits for the variability in the nickel contents of the hydraulic valves. Interpret your results.

16.67 *Bottle weights.* Refer to the bottle manufacturing process, Exercise 16.5 (p. 895). To monitor the process mean, three finished bottles are sampled from the production process at 20 points in time (days). The data (weight, in ounces) for last month's inspection are provided in the table. Construct both an $R$-chart and $\bar{x}$-chart for the weights of the finished bottles. Interpret the results.

**BOTTLE2**

| *Day* | *Bottle Weights* | | | *Day* | *Bottle Weights* | | |
|---|---|---|---|---|---|---|---|
| 1 | 5.6 | 5.8 | 5.8 | 11 | 6.2 | 5.6 | 5.8 |
| 2 | 5.7 | 6.3 | 6.0 | 12 | 5.9 | 5.7 | 5.9 |
| 3 | 6.1 | 5.3 | 6.0 | 13 | 5.2 | 5.5 | 5.7 |
| 4 | 6.3 | 5.8 | 5.9 | 14 | 6.0 | 6.1 | 6.0 |
| 5 | 5.2 | 5.9 | 6.3 | 15 | 6.3 | 5.7 | 5.9 |
| 6 | 6.0 | 6.7 | 5.2 | 16 | 5.8 | 6.2 | 6.1 |
| 7 | 5.8 | 5.7 | 6.1 | 17 | 6.1 | 6.4 | 6.6 |
| 8 | 5.8 | 6.0 | 6.2 | 18 | 6.2 | 5.7 | 5.7 |
| 9 | 6.4 | 5.6 | 5.9 | 19 | 5.3 | 5.5 | 5.4 |
| 10 | 6.0 | 5.7 | 6.1 | 20 | 6.0 | 6.1 | 6.0 |

16.68 *Mudbag data.* B. Render (Rollins College) and R. M. Stair (Florida State University) presented the case of the Bayfield Mud Company (*Quantitative Analysis of Management*, 1997). Bayfield supplies boxcars of 50-pound bags of mud-treating agents to the Wet-Land Drilling Company. Mud-treating agents are used to control the pH and other chemical properties of the cone during oil drilling operations. Wet-Land has complained to Bayfield that its most recent shipment of bags were underweight by about 5%. (The use of underweight bags may result in poor chemical control during drilling, which may hurt drilling efficiency, resulting in serious economic consequences.) Afraid of losing a long-time customer, Bayfield immediately began investigating their production process. Management suspected that the causes of the problem were the recently added third shift and the fact that all three shifts were under pressure to increase output to meet increasing demand for the product. Their quality control staff began randomly sampling and weighing six bags of output each hour. The average weight of each sample over the last three days is recorded in the table (p. 947) along with the weight of the heaviest and lightest bag in each sample.

a. Construct both an $R$-chart and an $\bar{x}$-chart for these data.
b. Is the process under statistical control?
c. Does it appear that management's suspicion about the third shift is correct? Explain?

*16.69 *Sampling plan analysis.* A quality control inspector is studying the alternative sampling plans $(n = 5, a = 1)$ and $(n = 25, a = 5)$.

a. Sketch the operating characteristic curves for both plans, using lot fractions defective .05, .10, .20, .30, and .40.
b. As a seller producing lots with AQL = .10, which of the two sampling plans would you prefer? Why?
c. As a buyer wanting to protect against accepting lots with fraction defective exceeding $p_1 = .30$, which of the two sampling plans would you prefer? Why?

16.70 *Monitoring rolled steel.* A company manufactures rolled steel for nuclear submarines. To monitor the production process, a quality control inspector sampled finished rolls of steel from the production line, one each hour for 12 consecutive hours. The number of imperfections discovered on each roll is recorded in the table.

**STEELROLL**

| *Hour* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Number of Imperfections* | 14 | 10 | 8 | 7 | 11 | 12 | 6 | 15 | 13 | 4 | 9 | 10 |

a. Construct a control chart for the number of imperfections per finished roll of steel.
b. Locate the center line and upper and lower control limits on the chart.
c. Does the manufacturing process appear to be in control?

**Data for Exercise 16.68**

**MUDBAGS**

| Time | Average Weight (pounds) | Lightest | Heaviest | Time | Average Weight (pounds) | Lightest | Heaviest |
|---|---|---|---|---|---|---|---|
| 6:00 A.M. | 49.6 | 48.7 | 50.7 | 6:00 P.M. | 46.8 | 41.0 | 51.2 |
| 7:00 | 50.2 | 49.1 | 51.2 | 7:00 | 50.0 | 46.2 | 51.7 |
| 8:00 | 50.6 | 49.6 | 51.4 | 8:00 | 47.4 | 44.0 | 48.7 |
| 9:00 | 50.8 | 50.2 | 51.8 | 9:00 | 47.0 | 44.2 | 48.9 |
| 10:00 | 49.9 | 49.2 | 52.3 | 10:00 | 47.2 | 46.6 | 50.2 |
| 11:00 | 50.3 | 48.6 | 51.7 | 11:00 | 48.6 | 47.0 | 50.0 |
| 12 noon | 48.6 | 46.2 | 50.4 | 12 midnight | 49.8 | 48.2 | 50.4 |
| 1:00 P.M. | 49.0 | 46.4 | 50.0 | 1:00 A.M. | 49.6 | 48.4 | 51.7 |
| 2:00 | 49.0 | 46.0 | 50.6 | 2:00 | 50.0 | 49.0 | 52.2 |
| 3:00 | 49.8 | 48.2 | 50.8 | 3:00 | 50.0 | 49.2 | 50.0 |
| 4:00 | 50.3 | 49.2 | 52.7 | 4:00 | 47.2 | 46.3 | 50.5 |
| 5:00 | 51.4 | 50.0 | 55.3 | 5:00 | 47.0 | 44.1 | 49.7 |
| 6:00 | 51.6 | 49.2 | 54.7 | 6:00 | 48.4 | 45.0 | 49.0 |
| 7:00 | 51.8 | 50.0 | 55.6 | 7:00 | 48.8 | 44.8 | 49.7 |
| 8:00 | 51.0 | 48.6 | 53.2 | 8:00 | 49.6 | 48.0 | 51.8 |
| 9:00 | 50.5 | 49.4 | 52.4 | 9:00 | 50.0 | 48.1 | 52.7 |
| 10:00 | 49.2 | 46.1 | 50.7 | 10:00 | 51.0 | 48.1 | 55.2 |
| 11:00 | 49.0 | 46.3 | 50.8 | 11:00 | 50.4 | 49.5 | 54.1 |
| 12 midnight | 48.4 | 45.4 | 50.2 | 12 noon | 50.0 | 48.7 | 50.9 |
| 1:00 A.M. | 47.6 | 44.3 | 49.7 | 1:00 P.M. | 48.9 | 47.6 | 51.2 |
| 2:00 | 47.4 | 44.1 | 49.6 | 2:00 | 49.8 | 48.4 | 51.0 |
| 3:00 | 48.2 | 45.2 | 49.0 | 3:00 | 49.8 | 48.8 | 50.8 |
| 4:00 | 48.0 | 45.5 | 49.1 | 4:00 | 50.0 | 49.1 | 50.6 |
| 5:00 | 48.4 | 47.1 | 49.6 | 5:00 | 47.8 | 45.2 | 51.2 |
| 6:00 | 48.6 | 47.4 | 52.0 | 6:00 | 46.4 | 44.0 | 49.7 |
| 7:00 | 50.0 | 49.2 | 52.2 | 7:00 | 46.4 | 44.4 | 50.0 |
| 8:00 | 49.8 | 49.0 | 52.4 | 8:00 | 47.2 | 46.6 | 48.9 |
| 9:00 | 50.3 | 49.4 | 51.7 | 9:00 | 48.4 | 47.2 | 49.5 |
| 10:00 | 50.2 | 49.6 | 51.8 | 10:00 | 49.2 | 48.1 | 50.7 |
| 11:00 | 50.0 | 49.0 | 52.3 | 11:00 | 48.4 | 47.0 | 50.8 |
| 12 noon | 50.0 | 48.8 | 52.4 | 12 midnight | 47.2 | 46.4 | 49.2 |
| 1:00 P.M. | 50.1 | 49.4 | 53.6 | 1:00 A.M. | 47.4 | 46.8 | 49.0 |
| 2:00 | 49.7 | 48.6 | 51.0 | 2:00 | 48.8 | 47.2 | 51.4 |
| 3:00 | 48.4 | 47.2 | 51.7 | 3:00 | 49.6 | 49.0 | 50.6 |
| 4:00 | 47.2 | 45.3 | 50.9 | 4:00 | 51.0 | 50.5 | 51.5 |
| 5:00 | 46.8 | 44.1 | 49.0 | 5:00 | 50.5 | 50.0 | 51.9 |

*Source*: Kinard, J., Western Carolina University, as reported in Render, B., and Stair, Jr., R., *Quantitative Analysis for Management*, 6th ed. Upper Saddle River, NJ: Prentice Hall, 1997.

16.71 *Sampling electron tubes.* For a lot of 250 electron tubes with an acceptance quality level of 10%, find the appropriate MIL-STD-105D general inspection sampling plan under each of the following inspection categories:

a. Normal inspection level
b. Tightened inspection level

16.72 *Defective robots.* High-level computer technology has developed bit-sized microprocessors for use in operating industrial "robots." To monitor the fraction of defective microprocessors produced by a manufacturing process, 50 microprocessors are sampled each hour. The results for 20 hours of sampling are provided in the table.

**ROBOTS2**

| *Sample* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Defectives* | 5 | 6 | 4 | 7 | 1 | 3 | 6 | 5 | 4 | 5 |
| *Sample* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Defectives* | 8 | 3 | 2 | 1 | 0 | 1 | 1 | 2 | 3 | 3 |

a. Construct a control chart for the proportion of defective microprocessors.
b. Locate the center line and upper and lower control limits on the chart. Does the process appear to be in control?
c. Conduct a runs analysis for the control chart. Interpret the result.

16.73 *Strength of steel cable.* A construction engineer buys steel cable in large rolls to use in supporting equipment and temporary structures during the process of erecting permanent structures. Specifications require the breaking strength of the steel cable to exceed 200 pounds. For a lot size of 1,500 large rolls of steel cable, consider an acceptance sampling plan based on a sample of $n = 20$ rolls and acceptance number $a = 2$.

a. Find the producer's risk if the AQL is .05.
b. Find the consumer's risk if the lot fraction failing to meet breaking strength specifications is $p_1 = .10$.
c. Draw a rough sketch of the operating characteristic curve for the sampling plan. Is the sampling plan reasonable?
d. Find the appropriate MIL-STD-105D normal (level) general inspection sampling plan for a lot size of 1,500 large rolls of steel cable and an AQL of .05.
e. Refer to part **d.** Find the producer's risk under the inspection sampling plan. (*Hint*: Use the normal approximation to the binomial.)
f. Refer to part **d.** Find the consumer's risk if the lot fraction failing to meet breaking strength specifications is $p_1 = .08$. (*Hint*: Use the normal approximation to the binomial.)

16.74 *Epoxy-repaired joints.* Refer to the stress analysis on epoxy-repaired truss joints described in Exercise 7.20 (p. 307). Tests were conducted on epoxy-bonded truss joints made of wood to determine tolerances for actual glue line shear stress (*Journal of Structural Engineering*, Feb. 1986). The mean and standard deviation of the shear strengths (pounds per square inch) for a random sample of 100 Southern pine truss joints are

$$\bar{x} = 1{,}312 \qquad s = 422$$

a. Assuming the distribution of strength measurements is approximately normal, construct a 95% tolerance interval for 99% of the shear strengths.
b. Interpret the interval obtained in part **b**.
c. Explain how you could obtain a tolerance interval when the normality assumption is not satisfied.

16.75 *Defective plastic mold.* A company that manufactures plastic molded parts believes it is producing an unusually large number of defects. To investigate this suspicion, each shift drew seven random samples of 200 parts, visually inspected each part to determine whether it was defective, and tallied the primary type of defect present (Hart, 1992). These data are presented in the table.

a. Construct a $p$-chart for this manufacturing process.
b. Should the control limits be used to monitor future process output? Explain.

**MOLD**

| | | | Type of Defect | | | | |
|---|---|---|---|---|---|---|---|
| Sample | Shift | *# of Defects* | *Crack* | *Burn* | *Dirt* | *Blister* | *Trim* |
| 1 | 1 | 4 | 1 | 1 | 1 | 0 | 1 |
| 2 | 1 | 6 | 2 | 1 | 0 | 2 | 1 |
| 3 | 1 | 11 | 1 | 2 | 3 | 3 | 2 |
| 4 | 1 | 12 | 2 | 2 | 2 | 3 | 3 |
| 5 | 1 | 5 | 0 | 1 | 0 | 2 | 2 |
| 6 | 1 | 10 | 1 | 3 | 2 | 2 | 2 |
| 7 | 1 | 8 | 0 | 3 | 1 | 1 | 3 |
| 8 | 2 | 16 | 2 | 0 | 8 | 2 | 4 |
| 9 | 2 | 17 | 3 | 2 | 8 | 2 | 2 |
| 10 | 2 | 20 | 0 | 3 | 11 | 3 | 3 |
| 11 | 2 | 28 | 3 | 2 | 17 | 2 | 4 |
| 12 | 2 | 20 | 0 | 0 | 16 | 4 | 0 |
| 13 | 2 | 20 | 1 | 1 | 18 | 0 | 0 |
| 14 | 2 | 17 | 2 | 2 | 13 | 0 | 0 |
| 15 | 3 | 13 | 3 | 2 | 5 | 1 | 2 |
| 16 | 3 | 10 | 0 | 3 | 4 | 2 | 1 |
| 17 | 3 | 11 | 2 | 2 | 3 | 2 | 2 |
| 18 | 3 | 7 | 0 | 3 | 2 | 2 | 0 |
| 19 | 3 | 6 | 1 | 2 | 0 | 1 | 2 |
| 20 | 3 | 8 | 1 | 1 | 2 | 3 | 1 |
| 21 | 3 | 9 | 1 | 2 | 2 | 2 | 2 |

16.76 *Waiting times of airline passengers.* Officials at Mountain Airlines are interested in monitoring the length of time customers must wait in line to check in at their airport counter in Reno, Nevada. In order to develop a control chart, five customers were sampled each day for 20 days. The data, in minutes, are presented in the table.

**CHECKIN**

| Sample | Waiting Time (mins.) | | | | |
|---|---|---|---|---|---|
| 1 | 3.2 | 6.7 | 1.3 | 8.4 | 2.2 |
| 2 | 5.0 | 4.1 | 7.9 | 8.1 | .4 |
| 3 | 7.1 | 3.2 | 2.1 | 6.5 | 3.7 |
| 4 | 4.2 | 1.6 | 2.7 | 7.2 | 1.4 |
| 5 | 1.7 | 7.1 | 1.6 | .9 | 1.8 |
| 6 | 4.7 | 5.5 | 1.6 | 3.9 | 4.0 |
| 7 | 6.2 | 2.0 | 1.2 | .9 | 1.4 |
| 8 | 1.4 | 2.7 | 3.8 | 4.6 | 3.8 |
| 9 | 1.1 | 4.3 | 9.1 | 3.1 | 2.7 |
| 10 | 5.3 | 4.1 | 9.8 | 2.9 | 2.7 |
| 11 | 3.2 | 2.9 | 4.1 | 5.6 | .8 |
| 12 | 2.4 | 4.3 | 6.7 | 1.9 | 4.8 |
| 13 | 8.8 | 5.3 | 6.6 | 1.0 | 4.5 |
| 14 | 3.7 | 3.6 | 2.0 | 2.7 | 5.9 |
| 15 | 1.0 | 1.9 | 6.5 | 3.3 | 4.7 |
| 16 | 7.0 | 4.0 | 4.9 | 4.4 | 4.7 |
| 17 | 5.5 | 7.1 | 2.1 | .9 | 2.8 |
| 18 | 1.8 | 5.6 | 2.2 | 1.7 | 2.1 |
| 19 | 2.6 | 3.7 | 4.8 | 1.4 | 5.8 |
| 20 | 3.6 | .8 | 5.1 | 4.7 | 6.3 |

a. Construct an $R$-chart from these data.
b. What does the $R$-chart suggest about the stability of the process? Explain.
c. Explain why the $R$-chart should be interpreted prior to the $\bar{x}$-chart.
d. Construct an $\bar{x}$-chart from these data.
e. What does the $\bar{x}$-chart suggest about the stability of the process? Explain.
f. Should the control limits for the $R$-chart and $\bar{x}$-chart be used to monitor future process output? Explain.

*16.77 *Waiting times of airline passengers.* Consider the airline check-in process described in Exercise 16.76.

a. Assume the process is under control and construct a capability analysis diagram for the process. Management has specified an upper specification limit of 5 minutes.
b. Is the process capable? Justify your answer.
c. If it is appropriate to estimate and interpret $C_p$ for this process, do so. If it is not, explain why.
d. Why didn't management provide a lower specification limit?

16.78 *Defects in graphite shafts.* A manufacturer of golf clubs has received numerous complaints about the performance of its graphite shafts. To monitor the shaft production process, a pultrusion method was used. A fabric is pulled through a thermosetting polymer bath and then through a long, heated steel die. As it moves through the die, the shaft is cured. Finally, it is cut to the desired length. Defects that can occur during the process are internal voids, broken strands, gaps between successive layers, and microcracks caused by improper curing. The quality department sampled 10 consecutive shafts every 30 minutes and nondestructive testing was used to seek out flaws in the shafts. The data from each 8-hour work shift were combined to form a shift sample of 160 shafts. Data on the proportion of defective shafts for 36 shift samples are presented in the table on p. 950. Data on the types of flaws identified are shown below. (*Note*: Each defective shaft may have more than one flaw.)

a. Use the appropriate control chart to determine whether the process proportion remains stable over time.
b. To help diagnose the causes of variation in process output, construct a Pareto diagram for the types of shaft defects observed. Which are the "vital few"? The "trivial many"?

**SHAFT2**

| Type of Defect | Number of Defects |
|---|---|
| Internal voids | 11 |
| Broken strands | 96 |
| Gaps between layer | 72 |
| Microcracks | 150 |

## Data for Exercise 16.78

**SHAFT1**

| Shift Number | Number of Defective Shafts | Proportion of Defective Shafts | Shift Number | Number of Defective Shafts | Proportion of Defective Shafts |
|---|---|---|---|---|---|
| 1 | 9 | .05625 | 19 | 6 | .03750 |
| 2 | 6 | .03750 | 20 | 12 | .07500 |
| 3 | 8 | .05000 | 21 | 8 | .05000 |
| 4 | 14 | .08750 | 22 | 5 | .03125 |
| 5 | 7 | .04375 | 23 | 9 | .05625 |
| 6 | 5 | .03125 | 24 | 15 | .09375 |
| 7 | 7 | .04375 | 25 | 6 | .03750 |
| 8 | 9 | .05625 | 26 | 8 | .05000 |
| 9 | 5 | .03125 | 27 | 4 | .02500 |
| 10 | 9 | .05625 | 28 | 7 | .04375 |
| 11 | 1 | .00625 | 29 | 2 | .01250 |
| 12 | 7 | .04375 | 30 | 6 | .03750 |
| 13 | 9 | .05625 | 31 | 9 | .05625 |
| 14 | 14 | .08750 | 32 | 11 | .06875 |
| 15 | 7 | .04375 | 33 | 8 | .05000 |
| 16 | 8 | .05000 | 34 | 9 | .05625 |
| 17 | 4 | .02500 | 35 | 7 | .04375 |
| 18 | 10 | .06250 | 36 | 8 | .05000 |

*Source*: Kolarik, W. *Creating Quality: Concepts, Systems, Strategies, and Tools*. New York: McGraw-Hill, 1995.

CHAPTER 17

# Product and System Reliability

**OBJECTIVE**

*To present some statistical methods for estimating the probability that a manufactured product or a system will perform satisfactorily for a specified period of time*

**CONTENTS**

## STATISTICS IN ACTION

### Modeling the Hazard Rate of Reinforced Concrete Bridge Deck Deterioration

In this chapter, we will learn about an important conditional probability associated with product failure, called the *hazard rate*. In simple terms, the hazard rate measures the probability of failure in a certain time frame, given the product has not failed prior to that time. We demonstrate (Section 17.3) that knowledge of the hazard rate of a failure time distribution can aid in the selection of the appropriate failure time density function and vice versa. However, some failure time distributions are dynamically changing and may not follow a prespecified probability density function. For example, the deterioration of reinforced concrete bridge decks is a continuous, gradual, and relatively slow process that varies widely with several factors such as traffic loading, current structural condition of the deck, bridge design, environmental factors, and material properties. The failure time distribution of these bridge decks cannot be accurately approximated with a single, known probability distribution.

In the *Journal of Infrastructure Systems* (June, 2001), civil and environmental engineers at the University of California-Berkeley developed a probabilistic model of the distribution of deterioration times for reinforced concrete bridge decks in Indiana. Their goal was to predict the probability that a bridge deck will undergo a significant change in condition-state (deterioration) at a given time. The researchers used data on the characteristics of concrete bridges obtained from the Indiana Bridge Inventory (IBI) data base to fit the model.

We provide details on the researchers modeling approach in the *Statistics in Action Revisited* section at the end of this chapter.

## 17.1 Introduction

Do your high-definition TV and your automobile perform well for a reasonably long period of time? If they do, we would say that these products are *reliable*. The **reliability** of a product is the probability that the product will meet certain specifications for a given period of time. For example, suppose we want a new automobile to perform without malfunction for a period of 2 years or for 20,000 miles. The probability that an automobile will meet these specifications is the *reliability* of the automobile.

**Definition 17.1**

The **reliability** of a product is the probability that the product will meet a set of specifications for a given period of time.

Some products need to function on a one-time basis. Others repeat a function over and over until they eventually fail. For example, a fuse either works or does not work when an electrical circuit is overloaded. The reliability of a fuse is the probability that it will work when subjected to a specific overload. In contrast, an automobile is used over and over again; its reliability is the probability that the automobile will perform without a major malfunction for some specified period of time.

## 17.2 Failure Time Distributions

The *length of life* of a product is the length of time until the product fails to perform according to specifications. When the product fails to perform according to specifications, it is said to have *failed*.

The time at which a single product item fails is called the **failure time** for the item. For example, the length of life of an abrasive grinding wheel is the length of time

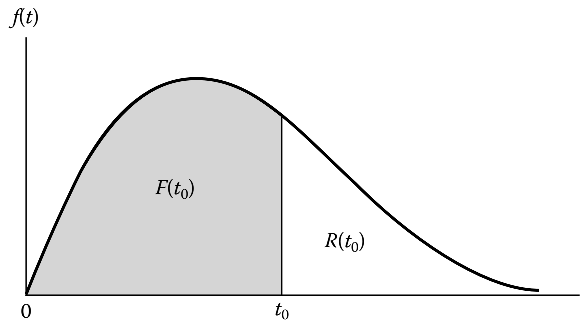



**FIGURE 17.1**
A failure time distribution

until the wheel fails to perform according to specifications. The specifications may have been determined by the manufacturer or the user may have written his or her own specifications. The length of time until failure is called the failure time of the wheel.

**Definition 17.2**

The **failure time *T*** of a product is a random variable that represents the length of time that the product performs according to specifications.

The failure time $T$ for any product varies from one item to another and is, in fact, a random variable. The density function for a product failure time is called a **failure time distribution**. A typical failure time distribution might appear as shown in Figure 17.1.

**Definition 17.3**

The **failure time distribution** for a product is the density function $f(t)$ of the failure time $T$.

If we denote the failure time density function by the symbol $f(t)$, then the probability that the product will fail before time $t_0$ is

$$P(T \leq t_0) = F(t_0) = \int_0^{t_0} f(t)\, dt$$

This probability is the shaded area under the density function shown in Figure 17.1.

Suppose that a product is said to be *reliable* if it survives until time $t_0$. Then the **reliability** of the product—that is, the probability that it will survive until time $t_0$—is

$$R(t_0) = 1 - F(t_0)$$

This probability, $R(t_0)$, is the unshaded area under the density function to the right of $t_0$ in Figure 17.1. The reliability, $R(t_0)$, is also called the **survival function** for the product.

Realistically, the failure time distribution is a conceptual relative frequency distribution of the lengths of life of some group of product items of specific interest—say, those manufactured in a given week, month, or year. Based on an analysis of sample data, we may select one of the density functions described in Chapter 5 to model this distribution. The family of density functions represented by the Weibull distribution (discussed in Section 5.8) is often used for this purpose.

**Definition 17.4**

The **reliability** (or **survival function**) $R(t_0)$ for a product is the probability that it will survive until time $t_0$:

$$R(t_0) = 1 - F(t_0)$$

where $F(t)$ is the cumulative distribution function for the failure time $T$.

## 17.3 Hazard Rates

The failure time distribution for a product enables us to calculate the probability $F(t_0)$ that an item will fail before time $t_0$ and the probability $R(t_0) = 1 - F(t_0)$ that the item will survive until time $t_0$. For some small change in failure time, denoted $\Delta t$, the probability that an item will fail in the interval $(t, t + \Delta t)$ is the shaded area shown in Figure 17.2. The density $f(t)$, the height of the shaded rectangle, is proportional to this probability.

Another way to describe the life characteristics of a product is to use a measure of the probability of failure as the product gets older—that is, the probability that the product will fail in the interval $(t, t + \Delta t)$, given that the item has survived to time $t$.

If we define the events

$A$: Item fails in the interval $(t, t + \Delta t)$

$B$: Item survives until time $t$

then the probability of failure in the time interval $(t, t + dt)$, given that the item has survived to time $t$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

But, the event $A \cap B$ is equivalent to the event $A$—that is, an item must have survived to time $t$ for it to be able to fail in the interval $(t, t + \Delta t)$. Therefore,

$$P(A \cap B) = P(A)$$

This probability is approximately equal to the shaded area in Figure 17.2. Then the probability of failure in the interval $(t, t + \Delta t)$, given that the item has survived to time $t$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \approx \frac{f(t)\,\Delta t}{1 - F(t)} = \frac{f(t)\,\Delta t}{R(t)}$$

The quantity

$$z(t) = \frac{f(t)}{R(t)}$$

is proportional to this conditional probability and is called the **hazard rate** for the product. Knowledge about a product's hazard rate often helps us to select the appropriate failure time density function for the product. The following example illustrates the point.

**Definition 17.5**

The **hazard rate** for a product is defined to be

$$z(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)}$$

where $f(t)$ is the density function of the product's failure time distribution.

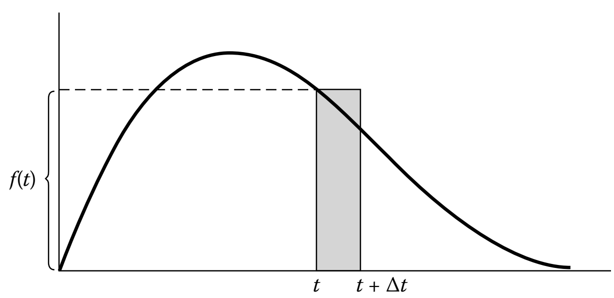



**FIGURE 17.2**
A failure time distribution showing the approximate probability of failure during the interval $(t, t + dt)$

**Example 17.1**

**Hazard Rate for an Exponential Failure Time Distribution**

The exponential distribution (discussed in Section 5.7) is often used in industry to model the failure time distribution of a product. Find the hazard rate for the exponential distribution.

**Solution**

The exponential density function and cumulative distribution function are, respectively,

$$f(t) = \frac{e^{-t/\beta}}{\beta} \qquad 0 \le t < \infty, \quad \beta > 0$$

and

$$F(t) = \int_{-\infty}^{t} f(y)\,dy = \int_{0}^{t} \frac{e^{-y/\beta}}{\beta} dy = 1 - e^{-t/\beta}$$

Then the hazard rate for the exponential distribution is

$$z(t) = \frac{f(t)}{1 - F(t)} = \frac{\dfrac{e^{-t/\beta}}{\beta}}{1 - (1 - e^{-t/\beta})} = \frac{1}{\beta}$$

Since $\beta = E(t)$ is the mean life of the product, it follows that the hazard rate is constant (see Figure 17.3). Therefore, a product that has an exponential failure time distribution never becomes fatigued. It is just as likely to survive any one unit of time as it is any other.

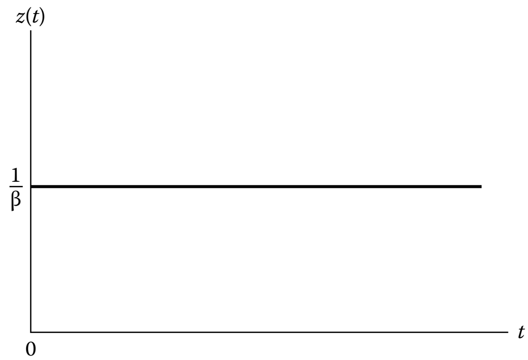



**FIGURE 17.3**
Hazard rate for the exponential failure time distribution

Clearly, the exponential distribution would not provide a good model for the failure time distribution of humans or for industrial products that become fatigued and more prone to failure as they get older. But it does provide a good model for some products, particularly for complex systems whose parts are replaced as they fail. After such systems have been in operation for a while, the probability of failure tends to be as likely in any one unit of time as in any other. Failure time distributions that exhibit this property (i.e., a constant hazard rate) are often called **memoryless distributions**.

The Weibull distribution density function (discussed in Section 5.8) and cumulative distribution function are, respectively,

$$f(t) = \frac{\alpha}{\beta} t^{\alpha - 1} e^{-t^{\alpha}/\beta} \qquad 0 \le t < \infty; \quad \alpha > 0; \quad \beta > 0$$

and

$$F(t) = 1 - e^{-t^{\alpha}/\beta}$$

By changing the shape parameter $\alpha$ and the scale parameter $\beta$, we obtain a variety of density functions useful for modeling failure time distributions for many industrial products. For $\alpha = 1$, we obtain the exponential distribution.

**Example 17.2**

Hazard Rate for a Weibull Failure Time Distribution

Find the hazard rate for the Weibull distribution and graph $z(t)$ versus time for $\alpha = 1, 2$, and 3.

Solution

Using the density function and cumulative distribution functions given above, we determine the hazard rate for the Weibull distribution:

$$z(t) = \frac{f(t)}{1 - F(t)} = \frac{\left(\frac{\alpha}{\beta}\right)t^{\alpha-1}e^{-t^{\alpha}/\beta}}{1 - (1 - e^{-t^{\alpha}/\beta})} = \frac{\alpha}{\beta}t^{\alpha-1}$$

When the shape parameter $\alpha$ is equal to 1, we obtain

$$z(t) = \frac{1}{\beta}$$

which is the constant hazard rate for the exponential distribution. For $\alpha = 2$,

$$z(t) = \frac{2}{\beta}t$$

the equation of a straight line passing through the origin. For $\alpha = 3$,

$$z(t) = \frac{3}{\beta}t^2$$

a second-order function of time $t$. Graphs of these hazard rates are shown in Figure 17.4. Note that the hazard rate increases more rapidly with time for larger values of the shape parameter $\alpha$.

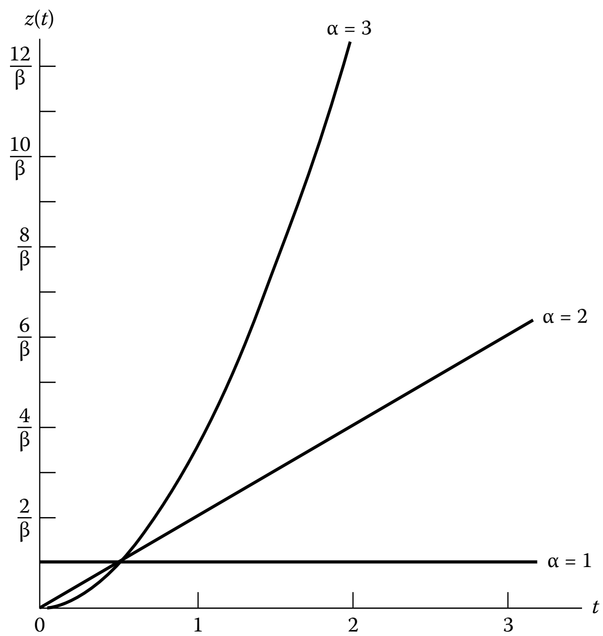



**FIGURE 17.4** Graphs of the hazard rate for Weibull distribution with $\alpha = 1, 2, 3$

## Applied Exercises

17.1 *Preventative maintenance tests.* The optimal scheduling of preventative maintenance tests of some (but not all) of $n$ independently operating components was developed in *Reliability Engineering and System Safety* (Jan. 2006). The number of failures per hour of a component was approximated by a Poisson distribution with mean $\lambda$. Consequently, the time between failures of a component is exponentially distributed with $\beta = 1/\lambda$. Find and graph the hazard rate for the time between failures of a component.

17.2 *Reliability of tension-leg platforms.* Tension-leg platforms (TLPs) are used for oil and gas exploration under deep-sea conditions. The reliability of TLPs under impulsive loading was assessed in *Reliability Engineering and System Safety* (Jan. 2006). The researchers examined several random variables that contribute to the impulsive force under dynamic loading of a TLP. One variable, tide and surge stress until failure (measured in MPa), was assumed to have a uniform distribution over the interval (4, 5). Find and graph the hazard rate for this variable.

17.3 *Normal failure time distribution.* Suppose the failure time distribution for a product can be approximated by a normal distribution with $\mu = 3$ and $\sigma = 1$.

a. Find $f(t)$, $F(t)$, and $z(t)$ for $t = 0, 1, 2, \ldots, 6$.
b. Plot the values of $z(t)$ for corresponding values of $t$ and obtain a graph of the hazard rate for this normal failure time distribution.

17.4 *Reliability of an electronic component.* The lifetime $T$ (in hours) of a certain electronic component is a random variable with density function

$$f(t) = \begin{cases} \dfrac{1}{100}e^{-t/100} & t > 0 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find $F(t)$ and $R(t)$.
b. What is the reliability of the component at $t = 25$ hours?
c. Find $z(t)$ and interpret the result.

17.5 *Failures of gas station fuel dispensers.* An article published in the annual *Journal of Industrial Engineering* (2013) studied preventive maintenance operations for fuel dispensers in a chain of gas stations. The time between failures of fuel dispensers at high failure stations was assumed to have an exponential distribution with a mean of 460 hours. Also, the time between preventive maintenance activities for the fuel dispensers has an exponential distribution with mean 2,880 hours.

a. Find and interpret the hazard rate for the failure time distribution.
b. Find and interpret the hazard rate for the preventive maintenance time distribution.
c. Assuming the failure time and preventive maintenance time distributions are independent, the research showed that the time between maintenance activities (either preventive maintenance or corrective maintenance due to failures) has an exponential distribution with mean 395 hours. Find and interpret the hazard rate for the time between any maintenance activity (either preventive or corrective).

17.6 *Reliability of an electric power system.* The reliability of an electrical power system in Iraq was analyzed in the journal *PLoS ONE* (Aug. 2013). The failure time distribution, $f(t)$, followed a *Dagum* probability distribution with parameters $\beta$, $\alpha$, and $\kappa$. The density function takes the form:

$$f(t) = \kappa\alpha(1 + (t/\beta)^{\alpha})^{-1-\kappa})\beta^{-\kappa\alpha}t^{-1-\kappa\alpha},$$
$$\beta > 0, \alpha > 0, \kappa > 0$$

The researchers showed that the reliability function for this distribution is:

$$R(t) = \beta^{-\kappa\alpha}(1 + (t/\beta)^{\alpha})^{-\kappa}t^{\kappa\alpha}$$

a. Show that the hazard rate for this distribution is

$$z(t) = \frac{\kappa\alpha}{(1 + (t/\beta)^{\alpha})(-1 + (1 + (\beta/t)^{\alpha})^{\kappa})t}$$

b. The best fitting distribution for the time between failures of the electrical power system had parameter values $\beta = 90$, $\alpha = 2.4$, and $\kappa = .34$. Use these values to find and interpret the hazard rate for this failure time distribution.

17.7 *Failure time of a drill bit.* A drill bit has a failure time distribution given by the density

$$f(t) = \begin{cases} \dfrac{2te^{-t^2}/100}{100} & 0 \le t < \infty \\ 0 & \text{elsewhere} \end{cases}$$

a. Find $F(t)$.
b. Find expressions for the reliability $R(t)$ and the hazard rate $z(t)$ of the drill bit at time $t$.
c. Use the results of part **b** to find $R(8)$ and $z(8)$.

17.8 *Failure of a computer disk pack.* The failure of a computer disk pack is considered to be an *initial failure* if it occurs prior to time $t = \alpha$ and a *wear-out failure* if it occurs after time $t = \beta$. Suppose the failure time distribution during the useful life of the disk pack is given by

$$f(t) = \frac{1}{\beta - \alpha} \qquad \alpha \le t \le \beta$$

a. Find $F(t)$ and $R(t)$.
b. Find the hazard rate $z(t)$.
c. Graph the hazard rate of the disk pack for $\alpha = 100$ hours and $\beta = 1{,}500$ hours.
d. For $\alpha = 100$ and $\beta = 1{,}500$, what is the reliability of the disk pack at time $t = 500$ hours? What is the hazard rate?

17.9 *Failure of lead-free solder joints.* Mechanical engineers at Purdue University conducted a reliability analysis of lead-free solder joints used in microscopic electronic packages (*Electronic Components and Technology Conference*, May 2005). The minimum time required by cracks in the solder joints to propagate through the weakest joint (i.e., the failure time) was approximated using a Weibull distribution with parameters $\alpha$ and $\beta$. The researchers estimated the shape parameter $\alpha = 3.5$ and the mean failure time $\mu = 2{,}370$ hours.

a. Use these estimates and your knowledge of the Weibull distribution to find an estimate of the scale parameter $\beta$.
b. Give an expression for the hazard rate of the failure time distribution.
c. Find the hazard rate at $t = 5{,}000$ hours.

17.10 *Solder joint fatigue.* Solder joints are widely used in the electronic packaging industry to provide a connection between wafer level chips and a printed circuit board (PCB). The most common failure of the connection is caused by cracks in the bulk solder joint close to the wafer package-pad. Researchers at National Semiconductor Corporation estimated the failure time (in hours) of the solder joint connections for a 64L bump micro SMD package using the Weibull distribution (*Electronic Components and Technology Conference*, May 2003). The median failure time was estimated at 590 hours.

a. Write an expression for the median failure time as a function of the Weibull distribution parameters $\alpha$ and $\beta$.
b. Find the value of $\beta$ when $\alpha = 1$ and the median is 590 hours.
c. Find the value of $\beta$ when $\alpha = 2$ and the median is 590 hours.
d. Find the hazard rate when $\alpha = 2$ and the median is 590 hours.

## 17.4 Life Testing: Censored Sampling

A **life test** is an experiment conducted to obtain sample values of the lengths of life of some product items. Typically, a random sample of $n$ items is placed on test under specified environmental conditions and left on test until they fail. The recorded times to failure, $t_1, t_2, \ldots, t_n$, provide a random sample of observations on the length of life $T$ of the product. If for convenience we let $t_1$ represent the smallest failure time, $t_2$ the second smallest, . . . , and $t_n$ the largest, then the times might appear as points on a time line, as shown in Figure 17.5.

In many situations, life tests are conducted to determine the quality of a manufactured product prior to sale. Waiting for the last few items in a sample to fail can be time-consuming and expensive. To reduce the cost of waiting for some long-life items, tests are often concluded after a specified length $T = t_c$ of test time. When we do this, we say that the life test is censored at time $t_c$. A second type of **censored sampling** occurs when we conclude the testing after a fixed number $r$ of items have failed.

If a life test is censored at a fixed time $t_c$, the length of testing time is fixed. This makes it easier to schedule the life-testing equipment, but the number $R$ of failures observed prior to time $t_c$ is a random variable. Thus, $R$ could assume any integer value $r$ in the interval $0 \le r \le n$, and it is possible that no failure times would be observed. If the test is censored after a fixed number $R = r$ of failures have been observed, we know that we will always acquire the values of $r$ failure times, but the length of the life test will be variable and equal to the length of time $t_r$ until failure of the $r$th item.

There are many other types of life-testing procedures. In **life testing with replacement**, product items are replaced on the test equipment as soon as an item fails, a procedure that makes maximum use of the test equipment. Other tests are designed to investigate the effect of various stresses on a product by testing the items under varying stresses. Tests of this type are called **accelerated life tests**. Descriptions of these and other life test procedures, as well as methods for using the **censored data** to estimate the parameters of failure time distributions, are described in the references for this chapter.

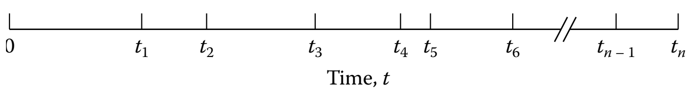



**FIGURE 17.5**
Failure times of $n$ items of some product

## 17.5 Estimating the Parameters of an Exponential Failure Time Distribution

The methods for finding estimators of the parameters of failure time distributions are those described in the preceding chapters. We can use the method of moments (Chapter 7), method of maximum likelihood (Chapter 7), or the method of least squares (Chapter 10). Depending on the failure time distribution and the number of parameters involved, finding the estimator may be easy or difficult. For example, finding the maximum likelihood estimator of the parameter of an exponential distribution based on simple random sampling is easy (see Example 7.5), but solving the maximum likelihood equations obtained for estimating the parameters of a Weibull distribution is relatively difficult.

It is also difficult to obtain estimators and their sampling distributions based on certain types of sampling, especially when the sampling has been censored at a fixed time $t_c$. Consequently, in this and the following sections, we will present estimation procedures for the exponential and the Weibull failure time distributions. Estimation procedures for these and other failure time distributions are discussed in the literature or in texts on product and system reliability.

We will first consider estimators for the mean failure time $\beta$ for the exponential distribution. Estimators for $\beta$ are the same regardless of whether the life test is censored or uncensored; the estimator is always equal to the total observed life divided by the number $r$ of failures observed. For example, if a random sample of $n$ items is selected from the population and the life test is concluded after the $r$th failure is observed $(r > 0)$, then

$$\hat{\beta} = \frac{\sum_{i=1}^{r} t_i + (n - r)t_r}{r} = \frac{\text{Total observed life}}{r}$$

If we wait until all $n$ items fail, then $r = n$ and the estimator is the sample mean failure time:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} t_i}{n} = \bar{t}$$

Note that for both the censored and uncensored sampling situations, the numerator in the above expressions is equal to the total length of life observed for the $n$ items during the length of the life test.

For censored sampling with a fixed time of testing $t_c$,

$$\hat{\beta} = \frac{\sum_{i=1}^{r} t_i + (n - r)t_c}{r} = \frac{\text{Total observed life}}{r} \quad \text{for } r \geq 1$$

Again, note that the numerator in this expression is the total length of life recorded for the $n$ items until the life test is concluded at time $t_c$.

### Point Estimators of the Mean Life $\beta$ for an Exponential Distribution

For uncensored life testing:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} t_i}{n}$$

For censored sampling with $r$ fixed:

$$\hat{\beta} = \frac{\sum_{i=1}^{r} t_i + (n - r)t_r}{r}$$

For censored sampling with test time $t_c$ fixed:

$$\hat{\beta} = \frac{\sum_{i=1}^{r} t_i + (n - r)t_c}{r}$$

The formulas for $(1 - \alpha)100\%$ confidence intervals for $\beta$ are shown in the following boxes. The confidence interval based on sampling censored at a fixed point in time is only approximate.

### A $(1 - \alpha)100\%$ Confidence Interval for $\beta$ Based on Censored Sampling with $r$ Fixed

$$\frac{2(\text{Total life})}{\chi^2_{\alpha/2}} \le \beta \le \frac{2(\text{Total life})}{\chi^2_{(1-\alpha/2)}}$$

where

$$\text{Total life} = \sum_{i=1}^{r} t_i + (n - r)t_r$$

and $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are the tabulated values of a chi-square statistic, based on $2r$ degrees of freedom, that locate $\alpha/2$ in the upper and lower tails, respectively, of the chi-square distribution.

### An Approximate $(1 - \alpha)100\%$ Confidence Interval for $\beta$ Based on Censored Sampling with $t_c$ Fixed

$$\frac{2(\text{Total life})}{\chi^2_{\alpha/2}} \le \beta \le \frac{2(\text{Total life})}{\chi^2_{(1-\alpha/2)}}$$

where

$$\text{Total life} = \sum_{i=1}^{r} t_i + (n - r)t_c$$

and $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are the tabulated upper- and lower-tail values of a chi-square distribution based on $(2r + 2)$ degrees of freedom.

### Example 17.3

**Mean Time Between Aircraft Engine Malfunctions**

Suppose that the length of time between malfunctions for a particular type of aircraft engine has an exponential failure time distribution. Ten of the engines were tested until six of the engines malfunctioned. The times to malfunction were 48, 35, 91, 62, 59, and 77 hours, respectively. Find a 95% confidence interval for the mean time $\beta$ between malfunctions for the engines.

**Solution** Since this life test was concluded after the sixth failure was observed, it represents censored sampling with $r = 6$. The total observed life for the test was

$$\begin{aligned}\text{Total life} &= \sum_{i=1}^{r} t_i + (n - r)t_r \\ &= 372 + 364 = 736 \text{ hours}\end{aligned}$$

The tabulated values of $\chi^2_{.025}$ and $\chi^2_{.975}$, based on $2r = 2(6) = 12$ degrees of freedom, are 23.3367 and 4.40379, respectively. Then the 95% confidence interval for $\beta$ is

$$\frac{2(736)}{23.3367} \leq \beta \leq \frac{2(736)}{4.40379}$$

or $63.08 \leq \beta \leq 334.26$. Our interpretation is that the true mean time $\beta$ between malfunctions of this particular type of aircraft engine falls between 63.08 hours and 334.26 hours, with 95% confidence.

**Example 17.4**

Hazard Rate and Reliability Confidence Intervals

Refer to Example 17.3.

a. Find a 95% confidence interval for the hazard rate of the aircraft engine.
b. Find a 95% confidence interval for the reliability of the system at time 50 hours.

Solution

a. Recall from Section 17.3 that the hazard rate for the exponential distribution is $1/\beta$. We therefore begin with the 95% confidence interval for $\beta$ derived in Example 17.3 and transform it to a confidence interval for $1/\beta$:

$$63.08 \leq \beta \leq 334.26$$
$$\frac{1}{334.26} \leq \frac{1}{\beta} \leq \frac{1}{63.08}$$
$$.003 \leq \frac{1}{\beta} \leq .016$$

Thus, the hazard rate for the aircraft engine at time $t$ (which is proportional to the probability that the engine will fail during a fixed small interval of time, given that the engine has survived to time $t$) falls between .003 and .016 with 95% confidence.

b. From Example 17.1, the cumulative distribution function for the exponential distribution is

$$F(t) = 1 - e^{-t/\beta}$$

By definition, the reliability of the aircraft engine at time $t_0$ is

$$R(t_0) = 1 - F(t_0)$$
$$= 1 - (1 - e^{-t_0/\beta}) = e^{-t_0/\beta}$$

or, for $t_0 = 50$ hours, $R(50) = e^{-50/\beta}$.

Then $63.08 \leq \beta \leq 334.26$ is equivalent to

$$e^{-50/63.08} \leq e^{-50/\beta} \leq e^{-50/334.26}$$
$$.453 \leq e^{-50/\beta} \leq .861$$

Therefore, the probability that the engine survives at least 50 hours may be as low as .453 or as high as .861, with 95% confidence.

## Applied Exercises

17.11 *Reliability of wafer-level chips.* The reliability of wafer-level-chip-scale packages mounted on a printed circuit board (PCB) in handheld devices such as cellular phones, pagers, and PDAs was investigated by researchers at National Semiconductor Corporation (*Electronic Components and Technology Conference*, May 2005). In one experiment, PCBs were mounted 2 millimeters from the point of plunger contact and the number of cycles to failure measured for each. The data (simulated) for a sample of 20 PCBs are listed in the table. Assume the number of cycles to failure can be approximated by an exponential distribution.

**PCB3**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1534 | 333 | 1179 | 679 | 1186 | 197 | 279 | 263 | 682 | 240 |
| 331 | 508 | 361 | 593 | 420 | 2028 | 271 | 176 | 525 | 538 |

a. Find and interpret a 95% confidence interval for the mean number of cycles to failure.
b. Find and interpret a 95% confidence interval for the hazard rate of the distribution.

17.12 *Strength of shotcrete.* A wet-mix, steel-fiber-reinforced microsilica concrete (called *shotcrete*), used extensively in Scandinavia, is now being marketed in the United States. The material is said to have a minimum 28-day breaking strength of 9,000 pounds per square inch (psi) of compression. To investigate the breaking strength of the new product, seven pieces of shotcrete were subjected to 9,000 psi of compression daily until they failed. The times to failure were 33, 35, 61, 38, 21, 41, and 52 days. Assume that the shotcrete has an exponential failure time distribution when subjected to 9,000 psi of compression.

a. Find a 90% confidence interval for the mean time $\beta$ until the shotcrete fails.
b. Find a 90% confidence interval for the probability that the shotcrete will not fail before the 28-day specified minimum.
c. Find a 90% confidence interval for the hazard rate of the shotcrete.

17.13 *Machine failures at a tire factory.* Failure times for two machines at the Babel tire factory in Iraq were obtained and used to estimate mean failure time. (*Iraqi Journal of Statistical Science*, Vol. 14, 2008.) The data (in hours) for the cutting layers machine and the coating machine are listed in the accompanying table. Failure time is known to follow an exponential probability distribution.

**TIREIRAQ**

Cutting Machine:

| | | | | | |
|---|---|---|---|---|---|
| 1.00 | 1.00 | 5.00 | 5.50 | 12.50 | 16.75 |
| 17.75 | 20.75 | 22.50 | 22.75 | 25.00 | 25.00 |
| 27.25 | 30.25 | 43.75 | 45.00 | 48.00 | 48.25 |
| 97.50 | 99.75 | 136.75 | 143.50 | 207.75 | 215.00 |
| 225.50 | 235.00 | 283.50 | 567.00 | 970.50 | |

Coating Machine:

| | | | | | |
|---|---|---|---|---|---|
| 3.5 | 6.5 | 10.5 | 23.25 | 23.5 | 43.5 |
| 69 | 70.5 | 75.5 | 83.25 | 95.5 | 109.5 |
| 111.25 | 144 | 164 | 167.25 | 253 | 383.75 |
| 417.75 | 428.25 | 453 | 1215 | | |

a. Find and interpret a 90% confidence interval for the mean failure time of the cutting layers machine.
b. Find and interpret a 90% confidence interval for the mean failure time of the coating machine.

17.14 *Integrated circuit chips.* Suppose that an integrated circuit chip possesses an exponential failure time distribution. Fifteen chips were put on accelerated life test until five of the chips failed. The first five failures occurred at 18.2, 19.5, 24.8, 31.0, and 45.6 (in thousands of hours).

a. Find a 95% confidence interval for the mean time between failures of the circuit chips.
b. Find a 95% confidence interval for the reliability of the circuit chips at 20,000 hours.

17.15 *High-reliability capacitors.* A sample of 100 high-reliability capacitors was placed on test for 2,000 hours. At the end of this period only three capacitors had malfunctioned, with failure times of 810, 1,422, and 1,816 hours. Assuming an exponential failure time distribution, construct a 99% confidence interval for the mean time between failures of the capacitors. Interpret the interval.

17.16 *High-reliability capacitors (continued).* Refer to Exercise 17.15.

a. Find a 99% confidence interval for the hazard rate of the capacitors.
b. Find a 99% confidence interval for the reliability of the capacitors at 3,000 hours.
c. Find a 99% confidence interval for the probability that a capacitor will fail before 2,000 hours.

17.17 *Lifelengths of locomotives.* A study was conducted to estimate the mean life (in miles) of a certain type of locomotive using censored sampling (*Technometrics*, May 1985). Ninety-six locomotives were operated for either 135 thousand miles or until failure. Of these, 37 failed before the 135-thousand-mile period. The accompanying table contains the miles to failure for these locomotives. Assuming an exponential failure time distribution, construct a 95% confidence interval for the mean miles to failure of the locomotives.

**LOCOMOTIVE**

| | | | | | |
|---|---|---|---|---|---|
| 22.5 | 57.5 | 78.5 | 91.5 | 113.5 | 122.5 |
| 37.5 | 66.5 | 80.0 | 93.5 | 116.0 | 123.0 |
| 46.0 | 68.0 | 81.5 | 102.5 | 117.0 | 127.5 |
| 48.5 | 69.5 | 82.0 | 107.0 | 118.5 | 131.0 |
| 51.5 | 76.5 | 83.0 | 108.5 | 119.0 | 132.5 |
| 53.0 | 77.0 | 84.0 | 112.5 | 120.0 | 134.0 |
| 54.5 | | | | | |

*Source*: Schmee, J., Gladstein, D., and Nelson, W. "Confidence limits for parameters of a normal distribution from singly-censored samples, using maximum likelihood." *Technometrics*, Vol. 27, No. 2, May 1985, p. 119.

## 17.6 Estimating the Parameters of a Weibull Failure Time Distribution

The method of maximum likelihood (discussed in Section 7.3) can be used to obtain estimates of the shape and scale parameters of the Weibull distribution, but the procedure is difficult and beyond the scope of this text. The interested reader should consult the references listed at the end of the chapter. The disadvantage of the method of maximum likelihood is that the estimates of $\alpha$ and $\beta$ are obtained by solving a complicated pair of simultaneous nonlinear equations. The advantage of the method is that when the sample size $n$ is large, maximum likelihood estimators possess sampling distributions that are approximately normal with known means and variances. This fact can be used to form large-sample confidence intervals using the method described in Section 7.3.

Instead of using the method of maximum likelihood to estimate $\alpha$ and $\beta$, we will use the method of least squares. You will recall that the cumulative distribution function for the Weibull distribution is

$$F(t) = 1 - e^{-t^{\alpha}/\beta}$$

Then the probability of survival to time $t$ is

$$\begin{aligned} R(t) &= 1 - F(t) \\ &= e^{-t^{\alpha}/\beta} \end{aligned}$$

and

$$\frac{1}{R(t)} = e^{t^{\alpha}/\beta}$$

Taking the natural logarithms of both sides of this equation, we obtain

$$\begin{aligned} \ln\left[\frac{1}{R(t)}\right] &= \frac{t^{\alpha}}{\beta} \\ -\ln R(t) &= \frac{t^{\alpha}}{\beta} \\ \ln[-\ln R(t)] &= -\ln \beta + \alpha \ln t \end{aligned}$$

To use the method of least squares, we need to estimate the survival function based on life test data. One way to do this is to place a random sample of $n$ items on life test and count the number of survivors at the end of one unit of time (for example, a week, or a month), after two units of time, and, in general, after $i$ units of time, $i = 1, 2, \ldots$. The intervals of time are shown in Figure 17.6. An estimate of the proportion of survivors at time $i$ is

$$\hat{R}(i) = \frac{n_i}{n}$$

where

$n_i$ = Number of survivors at the end of the $i$th time unit

$n$ = Total number of items placed on test

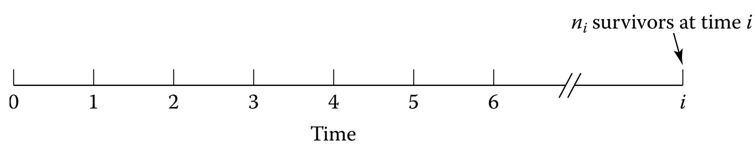
We would calculate $\hat{R}(i)$ for $i = 1, 2, \ldots$, and then fit the least-squares line

$$\underbrace{\ln[-\ln \hat{R}(i)]}_{y} = \underbrace{-\ln \beta}_{\beta_0} + \underbrace{\alpha \ln i}_{\beta_1 \quad x}$$

to the data points $(x_i, y_i)$, $i = 1, 2, \ldots$, where the $i$th data point is

$$y_i = \ln[-\ln \hat{R}(i)] \quad \text{and} \quad x_i = \ln i$$

**FIGURE 17.6**



The procedure is outlined in the box and illustrated with an example.

### Point Estimation of the Weibull Life Parameters, $\alpha$ and $\beta$

Assume a sample of $n$ items are placed on test and the number of items surviving at the end of several time intervals are recorded.

*Step 1* For each time interval, find $\hat{R}(i) = n_i/n$, where $n_i$ = number of survivors at the end of the $i$th interval and $n$ = total number of items placed on test.

*Step 2* For each time interval, compute $y_i = \ln[-\ln(\hat{R}(i))]$.

*Step 3* For each time interval, compute $x_i = \ln(i)$.

*Step 4* Using the variables computed in steps 2–3, fit the simple linear regression model, $E(y) = \beta_0 + \beta_1 x$.

*Step 5* The Weibull parameter estimates are: $\hat{\alpha} = \hat{\beta}_1$ and $\hat{\beta} = e^{-\hat{\beta}_0}$.

**Example 17.5**

Estimating Weibull Parameters

A manufacturer of hydraulic seals conducted a life test during which the seals were subjected to a fluid pressure that was 200% of the pressure normally maintained in hydraulic systems in which the seal is used. One hundred seals were placed on test and the number of survivors was recorded at the end of each day for a period of 7 days, as listed in Table 17.1

a. Use the data to estimate the parameters $\alpha$ and $\beta$ for a Weibull distribution.

b. Find 95% confidence intervals for $\alpha$ and $\beta$.

**HYDSEAL**

**TABLE 17.1 Daily Number of Survivors**

| Day | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Number of Survivors | 69 | 48 | 33 | 21 | 13 | 7 | 4 |

Solution

a. The first step is to calculate $\hat{R}(i)$ and $\ln[-\ln\hat{R}(i)]$ for each of the seven time intervals. These calculations are shown in Table 17.2. The SAS printout for a simple linear regression for the data is shown in Figure 17.7 (p. 965).

From the printout, you can see that the least-squares estimates (shaded) are

$$\hat{\beta}_0 = -1.05098 \quad \text{and} \quad \hat{\beta}_1 = 1.11019$$

Since $\beta_0 = -\ln\beta$ and $\beta_1 = \alpha$, we have

$$\hat{\alpha} = \hat{\beta}_1 = 1.11019$$

and

$$\hat{\beta}_0 = -\ln\hat{\beta} \quad \text{or} \quad \hat{\beta} = e^{-\beta_0} = 2.86045$$

**TABLE 17.2 Calculations for Example 17.5**

| *Time* $i$ | $x_i = \ln i$ | *Number of Survivors* | $\hat{R}(i)$ | $-\ln\hat{R}(i)$ | $y_i = \ln[-\ln\hat{R}(i)]$ |
|---|---|---|---|---|---|
| 1 | 0 | 69 | .69 | .37106 | −.99138 |
| 2 | .69315 | 48 | .48 | .73397 | −.30929 |
| 3 | 1.09861 | 33 | .33 | 1.10866 | .10315 |
| 4 | 1.38629 | 21 | .21 | 1.56065 | .44510 |
| 5 | 1.60944 | 13 | .13 | 2.04022 | .71306 |
| 6 | 1.79176 | 7 | .07 | 2.65926 | .97805 |
| 7 | 1.94591 | 4 | .04 | 3.21888 | 1.16903 |

```
                          The REG Procedure
                            Model: MODEL1
                        Dependent Variable: Y

              Number of Observations Read          7
              Number of Observations Used          7

                        Analysis of Variance

                                   Sum of          Mean
Source                  DF        Squares        Square    F Value    Pr > F

Model                    1        3.46814       3.46814    1074.81    <.0001
Error                    5        0.01613       0.00323
Corrected Total          6        3.48428

          Root MSE              0.05680    R-Square     0.9954
          Dependent Mean        0.30110    Adj R-Sq     0.9944
          Coeff Var            18.86544

                        Parameter Estimates

                    Parameter     Standard
Variable     DF      Estimate        Error   t Value   Pr > |t|     95% Confidence Limits

Intercept     1      -1.05098      0.04650    -22.60     <.0001     -1.17050      -0.93146
X             1       1.11019      0.03386     32.78     <.0001      1.02314       1.19724
```

**FIGURE 17.7**
SAS simple linear regression printout for Example 17.5

Therefore, based on the method of least squares, we would use a Weibull distribution with parameters $\alpha = 1.11019$ and $\beta = 2.86045$ to model the failure time distribution of the hydraulic seals.

b. Confidence intervals for $\alpha$ and $\beta$ can be obtained using the confidence limits for $\beta_0$ and $\beta_1$. The confidence interval for $\alpha$ is the usual regression confidence interval for $\beta_1$ because $\alpha = \beta_1$. The confidence limits for $\beta$ are computed by substituting the upper and lower confidence limits for $\beta_0$ into the relationship $\beta = e^{-\beta_0}$.

The 95% confidence intervals for $\beta_0$ and $\beta_1$ are highlighted on the SAS printout, Figure 17.7. The 95% confidence interval for $\beta_1$ is (1.023, 1.197). This also represents the 95% confidence interval for the Weibull parameter $\alpha$. The 95% confidence interval for $\beta_0$ is $(-1.17050, -.93146)$. Consequently, a 95% confidence interval for the Weibull parameter, $\beta$ is

$$(e^{.93146}, e^{1.17050})$$

or (2.5382, 3.2236).

**Example 17.6**

Estimating a Weibull Probability

Use the estimates of $\alpha$ and $\beta$ derived in Example 17.5 to find the probability that a hydraulic seal placed on test will survive at least 3 days.

Solution

Recall that the probability of survival to time $t_0$ under a Weibull distribution is given by

$$R(t_0) = 1 - F(t_0) = e^{-t_0^{\alpha}/\beta}$$

Substituting the estimates $\hat{\alpha} = 1.11019$ and $\hat{\beta} = 2.86045$ into the equation for $t_0 = 3$ days, we have

$$\hat{R}(3) = e^{-3^{1.11019}/2.86045} = e^{-1.18375} = .30613$$

Therefore, the probability that the hydraulic seal will survive at least 3 days is estimated to be .30613.

**Reliability for Small Samples:**

When $n$ is small, we can still use the method of least squares to estimate the parameters of a Weibull distribution, but the preferred procedure is to estimate the probability of survival to time $t$, $R(t) = 1 - F(t)$, after each failure time has been observed. The data points used for the least-squares methods are $[t_1, \hat{R}(t_1)]$, $[t_2, \hat{R}(t_2)], \ldots, [t_r, \hat{R}(t_r)]$, where $t_1$ is the first observed failure, $t_2$ the second, and so on. When this method of defining the data points is used, the estimator of the survival rates used in Example 17.5 is modified to

$$\hat{R}(t_i) = \frac{n_i + 1}{n + 1}$$

where $n_i$ is the number of survivors when the $i$th failure time $t_i$ is observed and $n$ is the sample size.*

In concluding, note that $\hat{\alpha}$ and $\hat{\beta}$ will not possess the properties of the usual least-squares estimators of $\beta_0$ and $\beta_1$. The response variable

$$y = \ln[-\ln \hat{R}(t)]$$

is not a normally distributed random variable and, in addition, the observed values of $y$ are correlated. This is because the number of survivors at one point in time is dependent on the number observed at some previous point in time. The extent to which these violations of the regression analysis assumptions affect the properties of the estimators is unknown, but it is probably slight when the sample size $n$ and the number $r$ of observed failures are large.

## Applied Exercises

17.18 *Brain cancer survival.* The Weibull probability distribution was used to model the survival time of brain cancer patients at an atomic medicine and radiance hospital (*Journal of Basra Research Science*, Vol. 37, 2011). The data in the table represent the number of 50 brain cancer patients surviving each month, for 10 consecutive months. Use this information to estimate the parameters of the Weibull survival time distribution.

**BRAINPAT**

| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of Survivors | 42 | 33 | 13 | 8 | 5 | 3 | 2 | 1 | 1 | 0 |

17.19 *Computer memory chips.* Suppose the lifelength (in years) of a memory chip in a mainframe computer has a Weibull failure time distribution. To estimate the Weibull parameters, $\alpha$ and $\beta$, 50 chips were placed on test and the number of survivors was recorded at the end of each year, for a period of 8 years. The data are shown in the accompanying table.

**MEMCHIP**

| *Year* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Number of Survivors* | 47 | 39 | 29 | 18 | 11 | 5 | 3 | 1 |

a. Use the method of least squares to derive estimates of $\alpha$ and $\beta$.
b. Construct a 95% confidence interval for $\alpha$.
c. Construct a 95% confidence interval for $\beta$.

17.20 *Computer memory chips (continued).* Refer to Exercise 17.19.

a. Use the estimates of $\alpha$ and $\beta$ to find the probability that a memory chip will fail before 5 years.
b. Estimate the reliability of the memory chips at time $t = 7$ years.

17.21 *Computer memory chips (continued).* Refer to Exercise 17.19.

a. Using the least-squares estimates of $\alpha$ and $\beta$, find and graph the hazard rate, $z(t)$.
b. Compute the hazard rate at time $t = 4$ years and interpret its value.

17.22 *Life lengths of roller bearings.* Engineers often use a Weibull failure time distribution for a "weakest link" product, i.e., a product consisting of multiple parts (e.g., roller bearings) that fails when the first part (or weakest link) fails. Nelson (*Journal of Quality Technology*, July 1985) applied the Weibull distribution to the lifelengths of a sample of $n = 138$ roller bearings. The table at the top of the next page gives the number of bearings still in operation at the end of each 100-hour period until all bearings failed.

a. Use the method of least squares to estimate the Weibull parameters $\alpha$ and $\beta$.
b. Construct a 99% confidence interval for $\alpha$. If you have access to a regression computer package, obtain a 99% confidence interval for $\beta$.

*Some statisticians use $\hat{R}(t_i) = (n_i + 1/2)/n$. See Miller and Freund (1977).

**Data for Exercise 17.22**

**BEARINGS2**

| *Hours (hundreds)* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 12 | 13 | 17 | 19 | 24 | 51 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Number of Bearings* | 138 | 114 | 104 | 64 | 37 | 29 | 20 | 10 | 8 | 6 | 4 | 3 | 2 | 1 |

*Source:* Nelson, W. "Weibull analysis of reliability data with few or no failures." *Journal of Quality Technology*, Vol. 17, No. 3, July 1985, p. 141 (Table 1). © 1985 American Society for Quality Control. Reprinted by permission.

c. Estimate the reliability of the roller bearings at $t = 300$ hours.

d. Estimate the probability that a roller bearing will fail before 200 hours.

17.23 *Washing machine repairs.* A manufacturer of washing machines conducted a life test during which it monitored 12 new machines for a period of 3 years and recorded the time to a major repair for each. At the end of the 3-year testing period, two machines had not yet required a major repair. The failure times (in months) of the remaining 10 washing machines were 14, 28, 9, 13, 6, 20, 10, 17, 30, and 20. Assume the lifelength (in years) of the machines has a Weibull failure time distribution with unknown parameters $\alpha$ and $\beta$.

a. Construct a table for the data listing the number of machines surviving (that is, without major repair) at the end of each year.

b. Apply the method of least squares to the data in the table of part **a** to derive estimates of $\alpha$ and $\beta$.

c. Find a 95% confidence interval for $\alpha$. If you have access to a regression computer package, find a 95% confidence interval for $\beta$.

d. The manufacturer guarantees all machines against a major repair for 2 years. Using the least-squares estimates of $\alpha$ and $\beta$, find the probability that a new washer will have to be repaired under the guarantee.

17.24 *Rebuilt hydraulic pumps.* To evaluate the performance of rebuilt hydraulic pumps at an aircraft rework facility, 20 pumps were placed on test and the number of pumps still running at the end of each week was recorded for a period of 6 weeks, as listed in the accompanying table.

**HYDPUMP**

| *Week* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *Number of Pumps* | 14 | 11 | 9 | 7 | 5 | 4 |

a. Use the data to estimate the parameters, $\alpha$ and $\beta$, for a Weibull failure time distribution.

b. Construct a 90% confidence interval for $\alpha$. If you have access to a regression computer package, find a 90% confidence interval for $\beta$.

c. Find the reliability of the rebuilt hydraulic pumps at time $t = 2$ weeks.

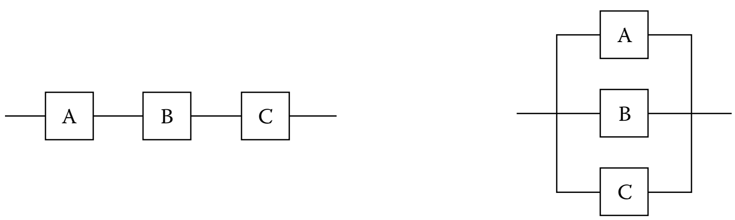
## 17.7 System Reliability

Systems—electronic, mechanical, or a combination of both—are composed of components, some of which are combined to form smaller subsystems. We will identify a component of a system by a capital letter and portray it graphically as a box. Two systems, each composed of three components, A, B, and C, are shown in Figure 17.8.

Suppose that a system is composed of $k$ components. If the system fails when any one of the components fails, it is called a **series system**. A three-component series system is represented graphically in Figure 17.8a. If a system fails only when *all* of its components fail, it is called a **parallel system**. A three-component parallel system is represented graphically in Figure 17.8b.

Figure 17.9a shows a system composed of five components, A, B, C, D, and E. Components D and E form a two-component parallel subsystem. This subsystem is connected in series with components A, B, and C. Figure 17.9b represents a system containing two parallel subsystems connected in series. The first parallel subsystem contains three components, A, B, and C. The second contains two series subsystems—the first composed of components D and E, and the second composed of components F and G.

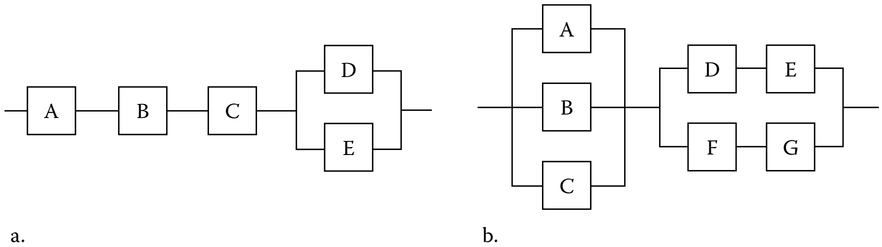
**FIGURE 17.8**
Two systems each composed of three components, A, B, and C




a. Series system

b. Parallel system

**FIGURE 17.9**
Two systems




**Definition 17.6**
A **series system** is one that fails if any one of its components fails.

**Definition 17.7**
A **parallel system** is one that fails only if all of its components fail.

Suppose that the reliability of component $i$—the probability that it will function properly under specified conditions—is $p_i$ and that the $k$ components of a system are mutually independent. That is, we assume that the operation of one component does not affect the operation of any of the others. Then the reliability of a system can be calculated using the multiplicative rule of probability.

Since a series system will function only if all of its components function, the reliability of a series system is

$$P(\text{Series system functions}) = P(\text{All components function})$$

Then, because the components operate independently of each other, we can apply the multiplicative rule of probability:

$$\begin{aligned} &P(\text{Series system functions}) \\ &\quad = P(\text{A functions})\,P(\text{B functions})\cdots P(\text{K functions}) \\ &\quad = p_A p_B p_C \cdots p_K \end{aligned}$$

**THEOREM 17.1**
The **reliability of a series system** consisting of $k$ independently operating components, A, B, . . . , K, is

$$P(\text{Series system functions}) = p_{\text{A}} p_{\text{B}} \ldots p_{\text{K}}$$

where $p_i$ is the probability that the $i$th component functions, $i = \text{A, B}, \ldots, \text{K}$.

The reliability of a parallel system containing $k$ components can be calculated in a similar manner. Since a parallel system will fail only if all components fail,

$$P(\text{Parallel system fails}) = (1 - p_{\text{A}})(1 - p_{\text{B}})\cdots(1 - p_{\text{K}})$$

and

$$\begin{aligned} P(\text{Parallel system functions}) &= 1 - P(\text{Parallel system fails}) \\ &= 1 - (1 - p_{\text{A}})(1 - p_{\text{B}})\cdots(1 - p_{\text{K}}) \end{aligned}$$

**THEOREM 17.2**
The **reliability of a parallel system** consisting of $k$ independently operating components is

$$P(\text{Parallel system functions}) = 1 - (1 - p_{\text{A}})(1 - p_{\text{B}})\cdots(1 - p_{\text{K}})$$

where $p_i$ is the probability that the $i$th component functions, $i = A, \text{B}, \ldots, \text{K}$.

Theorems 17.1 and 17.2 can be used to calculate the reliability of series systems, parallel systems, or any combinations thereof, as long as the systems satisfy the assumption that the components operate independently. The following examples illustrate the procedure.

**Example 17.7**

Reliability of a Series System

Given that $p_A = .90$, $p_B = .95$, and $p_C = .90$, find the reliability of the series system shown in Figure 17.8a.

Solution

Since this is a series system consisting of three components, A, B, and C, it follows from Theorem 17.1 that the reliability of this system is

$$P(\text{System functions}) = p_A p_B p_C = (.90)(.95)(.90) = .7695$$

**Example 17.8**

Reliability of a Parallel System

Suppose that the components in Example 17.7 were connected in parallel, as shown in Figure 17.8b. Find the reliability of the system.

Solution

To find the reliability of this parallel system, we apply Theorem 17.2:

$$\begin{aligned} P(\text{System functions}) &= 1 - (1 - p_A)(1 - p_B)(1 - p_C) \\ &= 1 - (.10)(.05)(.10) \\ &= 1 - .0005 \\ &= .9995 \end{aligned}$$

Examples 17.7 and 17.8 demonstrate that the *reliability of a series system is always less than the reliability of its least reliable component. In contrast, the reliability of a parallel system is always greater than the reliability of its most reliable component.*

To find the reliability of a system containing subsystems, we first find the reliability of the smallest subsystems. Then we find the reliability of the systems in which they are contained.

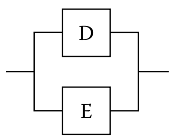
**Example 17.9**

Reliability of a Mixed System

Find the reliability of the system shown in Figure 17.9a, given the following component reliabilities: $p_A = .95$, $p_B = .99$, $p_C = .97$, $p_D = .90$, and $p_E = .90$.

Solution

The complete system is composed of three components, A, B, and C, and a subsystem connected in series. The parallel subsystem, comprised of components D and E, is shown here:




The reliability of this subsystem is

$$\begin{aligned} P(\text{Subsystem D and E functions}) &= p_{DE} \\ &= 1 - (1 - P_D)(1 - p_E) \\ &= 1 - (.1)(.1) = .99 \end{aligned}$$

We now view the complete system as one consisting of four components: components A, B, and C and the subsystem (D, E), connected in series. To find its reliability, we apply Theorem 17.1. The reliability of the complete system is

$$\begin{aligned} P(\text{System functions}) = p_A p_B p_C p_{DE} &= (.95)(.99)(.97)(.99) \\ &= .9031622 \end{aligned}$$

**Example 17.10**

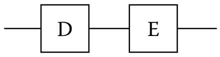
Reliability of Another Mixed System

Find the reliability of the system shown in Figure 17.9b, given that $p_A = .90$, $p_B = .95$, $p_C = .95$, $p_D = .92$, $p_E = .97$, $p_F = .92$, and $p_G = .97$.

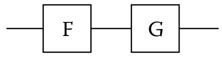
**Solution**

An examination of Figure 17.9b shows that the system is a series of two parallel subsystems. The first parallel subsystem contains components A, B, and C. The second is a parallel subsystem of two series subsystems, the first containing components D and E, and the second containing components F and G.

Since the reliabilities of the pairs of components (D, E) and (F, G) are identical, it follows that the reliabilities of these two series subsystems are equal:



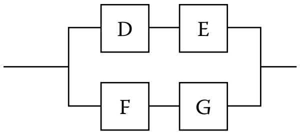

and




By Theorem 17.1, the reliability of these series subsystems is

$$p_{DE} = p_{FG} = p_D p_E = p_F p_G = (.92)(.97) = .8924$$

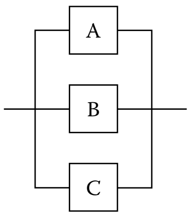
We now consider the reliability of the parallel subsystem containing these two series subsystems:




By Theorem 17.2, we have

$$\begin{aligned} p_{DEFG} &= 1 - (1 - p_{DE})(1 - p_{FG}) \\ &= 1 - (1 - .8924)(1 - .8924) \\ &= 1 - .0115778 \\ &= .9884222 \end{aligned}$$

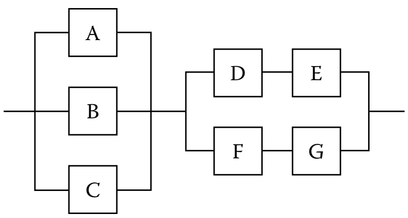
Next, we compute the reliability of the parallel subsystem consisting of components A, B, and C:




By Theorem 17.2,

$$\begin{aligned} p_{ABC} &= 1 - (1 - p_A)(1 - p_B)(1 - p_C) \\ &= 1 - (1 - .90)(1 - .95)(1 - .95) \\ &= 1 - .00025 \\ &= .99975 \end{aligned}$$

We have calculated the reliabilities of the two parallel subsystems. These two subsystems are connected in series, as shown here:



Thus, the reliability of the complete system is

$$\begin{aligned} P(\text{System functions}) &= p_{\text{ABC}}\, p_{\text{DEFG}} \\ &= (.99975)(.9884222) \\ &= .9881751 \end{aligned}$$

## Applied Exercises

17.25 *Series system reliability.* Consider a series system consisting of four components, A, B, C, and D, with probabilities of functioning given by $p_A = .88$, $p_B = .95$ , $p_C = .90$, and $p_D = .80$. Find the reliability of the system.

17.26 *Parallel system reliability.* Consider a parallel system consisting of four components, A, B, C, and D, with probabilities of functioning given by $p_A = .90$, $p_B = .99$, $p_C = .92$, and $p_D = .85$. Find the reliability of the system.

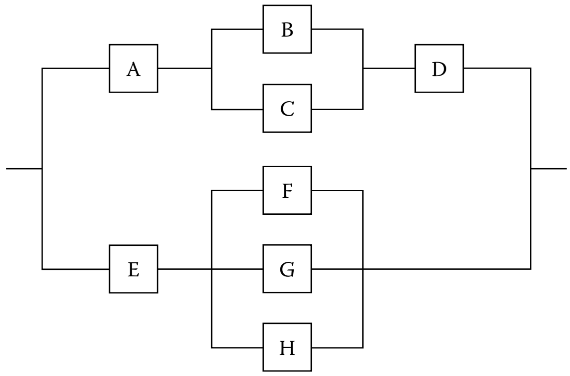
17.27 *Testing safety of system software.* In *Reliability Engineering and System Safety* (Jan. 2006), nuclear and quantum engineers at the Korea Advanced Institute of Science and Technology designed a digital safety system for testing system software. Consider $k$ independent software statements in system software code, each with probability of failure of $p_i$, $i = 1, 2, \ldots, k$. The system software will fail if at least one of the software code statements fails.

a. Do the system software code statements operate in series or in parallel? Explain.
b. Give an expression for the reliability of the system software.

17.28 *Detecting intruders to a computer system.* The Center for High Assurance Computer Systems at the Naval Research Laboratory in Washington, D.C. has developed several theoretical models for detecting intruder attacks on high-consequence computer systems (*International Information Assurance Workshop*, March 2005). Consider a naive attacker against a computer system with two servers. Server A will detect the intruder with probability .95. Server B detects the intruder with probability .99.

a. If the two servers operate in series, find the probability that the intruder is detected.
b. If the two servers operate in parallel, find the probability that the intruder is detected.
c. Suppose the system is designed so that only one of the two servers is operating at a time. Server A is in operation 1/3 of the time, and Server B is in operation 2/3 of the time. Find the probability that the system will detect the naive intruder.

17.29 *Reliability of an eight-component system.* A system consists of eight components, as shown in the accompanying diagram. Find the reliability of the system, given that the individual probabilities of functioning are $p_A = .90$, $p_B = .95$, $p_C = .85$, $p_D = .85$, $p_E = .98$, $p_F = .80$, $p_G = .95$, and $p_H = .95$.




17.30 *Reliability of an electrical circuit.* Consider an electrical circuit consisting of two subcircuits, the first of which involves components A, B, and C in parallel and the second of which involves components D and E in parallel. Suppose that the individual reliabilities of the components are given by $p_A = .95$, $p_B = .95$, $p_C = .90$, $p_D = .90$, and $p_E = .98$.

a. Find the reliability of the system if the two subcircuits are connected in series.
b. Find the reliability of the system if the two subcircuits are connected in parallel.

17.31 *Reliability of a three-component system.* The reliability of a system consisting of three identical components is .95. What must be the probability of functioning for each component if:

a. The components are connected in series?
b. The components are connected in parallel?

## *STATISTICS IN ACTION REVISITED*

### Modeling the Hazard Rate of Reinforced Concrete Bridge Deck Deterioration

We now return to the *Journal of Infrastructure Systems* (June 2001) study of the distribution of deterioration times for reinforced concrete bridge decks in Indiana. Recall that the goal was to predict the probability that a bridge deck will undergo a significant change in condition-state (deterioration) at a given time. Using data on the characteristics of concrete bridges obtained from the Indiana Bridge Inventory (IBI) database, the researchers fit a model that was a cross between a *parametric model* (in which the underlying hazard rate is restricted to follow a specific probability density function) and a *nonparametric model* (which, although flexible, does not directly relate deterioration time to relevant explanatory factors). The technique, called **semiparametric regression modeling**, allows the hazard function to be determined solely from the available data.

The most common semiparametric regression model is the **Cox proportional hazards model**. The hazard function, $h(t)$, used in the model can be expressed as follows:

$$h(t) = \lambda_0(t) \cdot \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\}$$

where $\lambda_0(t)$ is the baseline hazard function that is empirically obtained from the data and $x_1, x_2, \ldots, x_k$ are the explanatory independent variables. According to the researchers, this "hazard function quantifies the instantaneous risk that the bridge deck will experience a [deterioration] at time $t$." The Cox regression approach maximizes the natural logarithm of the likelihood function for $h(t)$ to obtain estimates of the $\beta$ parameters.

The explanatory variables used in the analysis, obtained from the IBI database, are listed in Table SIA17.1. One model was developed for bridge decks in each of three states: 6 (fair), 7 (generally fair), and 8 (good). The dependent variable in all models is the hazard associated with dropping to a lower deck deterioration condition from the current condition. The statistically significant (at $\alpha = .01$) parameter estimates for the Cox proportional hazards regressions are shown in Table SIA 17.2.

To interpret a parameter estimate for a dummy variable, we first compute the antilogarithm of the estimate, $e^{\beta}$. For example, for bridge condition 6, the antilog of the estimate for the dummy variable REGION is $e^{.85} = 2.34$. This value represents the ratio of the estimated hazard for bridge decks in the north (REGION = 1) to the estimated hazard for those in the south (REGION = 0). Thus, for condition 6 the model estimates that bridge decks in the northern Indiana region have a hazard rate that is over twice as high as bridge decks

**TABLE SIA17.1 IBI Bridge Deck Variables Used in Cox Hazards Model**

| Variable Name | Description |
|---|---|
| TYPECONT | Deck structural type (1 = continuous, 0 = otherwise) |
| PRESTRES | Deck prestress concrete (1 = true, 0 = false) |
| REGION | Climatic region (1 = north, 0 = south) |
| HWCLASS | Highway system classification (1 = interstate/rural, 2 = interstate/urban, 3 = other/rural, 4 = other/urban, 5 = secondary/rural, 6 = secondary/urban, 7 = secondary/local) |
| NUMSPANS | Number of spans in main unit |
| LANEDIR | Number of traffic lanes per direction |
| ADT | Average daily traffic (number of vehicles) |
| ADTYR | Year of ADT count |
| AVGADT | Mean ADT for all annual ADT counts |
| PROTSYS | Wearing surface code-protective system (1 = true, 0 = false) |
| DECKWID | Deck width (tenths of a foot) |
| STATE | Deck deterioration condition rating (9 = new, 8 = good, 7 = generally good, 6 = fair, 5 = generally fair, 4 = marginal, 3 = poor, 2 = critical/needs repair, 1 = critical/closed, 0 = critical/beyond repair) |
| DROP | Number of deterioration condition ratings dropped since last inspection |
| AGE | Deck age (years) |
| TIS | Time (years) deck has been in state condition |
| RCENSOR | Right censored data (1 = true, 0 = false) |
| SECONDARY | Secondary roadway bridge (1 = true, 0 = false) |

**TABLE SIA17.2 Hazard Rate Model Parameter Estimates for Three Bridge Conditions**

| | State = 6 (Fair) | State = 7 (Generally Fair) | State = 8 (Good) |
|---|---|---|---|
| REGION | .85 | .80 | .81 |
| DECKWID | — | — | −.0036 |
| NUMSPANS | — | .11 | — |
| PROTSYS | — | −.14 | — |
| PRESTRES | — | — | 1.19 |
| TYPECONT | — | −.46 | — |
| AGE | — | — | −.52 |
| AVGADT | .00004 | — | — |
| AGE × AVGADT | — | — | .00002 |
| SECONDARY | .84 | −.55 | −.82 |

in the southern region. The researchers explain that this difference is due to the heavy use of salts with sand for deicing roadways in the north, whereas in the south deicing salts are rarely used. Similarly, the antilog of the estimate for the dummy variable SECONDARY is $e^{.84} = 2.32$. Thus, the hazard rate for secondary roadway bridges (SECONDARY = 1) is more than twice as high as the hazard rate for bridges on primary highways. (According to the engineers, this result is due to secondary roadway bridges having lower design or maintenance standards.)

For parameter estimates for quantitative explanatory variables, compute the antilog of the estimate and subtract 1, then multiply by 100. The result represents the percentage change in the hazard rate for every 1-unit increase in the quantitative variable. For example, for bridge condition 6, the $\beta$ estimate for mean average daily traffic count (AVGADT) is .00004. Therefore, we compute $(e^{.00004} - 1) \times 100 = .004$. Thus, for every 1 vehicle increase in daily traffic count, the hazard rate for the bridge deck increases by .004%. Similarly, for bridge condition 7, a 1-span increase in NUMSPANS yields an estimated increase in the hazard of $(e^{.11} - 1) \times 100 = 11.6\%$. Inferences like these on the parameters of the Cox regression models gave the engineers insight into the factors that impact the hazard rate of reinforced concrete bridges.

Graphs of the predicted hazard functions over time were presented by the researchers. They discovered an approximately linearly increasing trend for bridges with condition 6, but almost flat (constant) trends for bridges with conditions 7 and 8. This suggests that for conditions 7 and 8, the deterioration time may be "memoryless" and represented by an exponential distribution.

# Quick Review

## Key Terms

Accelerated life tests, 958
Censored data, 958
Censored sampling, 958
Cox proportional hazards model, 972
Failure time, 952
Failure time distribution, 953
Hazard rate, 954
Life test, 958
Life testing with replacement, 958
Memoryless distribution 955
Parallel system, 967
Parallel system reliability, 968
Reliability, 952
Semiparametric modeling, 972
Series system, 967
Series system reliability, 968
Survival function, 953

## Key Formulas

| | |
|---|---|
| Reliability (Survival function): $R(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function | 953 |
| Hazard rate: $z(t) = f(t)/R(t)$ | 954 |
| Estimated Mean Life $\hat{\beta}$ for an Exponential Distribution: | |
| Uncensored life testing: $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} t_i$ | 959 |
| Censored sampling with $r$ fixed: $\hat{\beta} = \frac{1}{r}\left[\sum_{i=1}^{r} t_i + (n - r)t_r\right]$ | 960 |
| Censored sampling with test time $t_c$ fixed: $\hat{\beta} = \frac{1}{r}\left[\sum_{i=1}^{r}[t_i + (n - r)t_c\right]$ | 960 |
| Confidence Interval: $2(\text{Total life})/\chi^2_{\alpha/2} \le \beta \le 2(\text{Total life}))/\chi^2_{(1-\alpha/2)}$ | 960 |
| Estimated Parameters of a Weibull Life Distribution: | |
| $\hat{\alpha} = \hat{\beta}_1$ and $\hat{\beta} = e^{-\hat{\beta}_0}$, where $E(y) = \beta_0 + \beta_1 x$, $y_i = \ln[-\ln(\hat{R}(i))]$, $x_i = \ln(i)$ and $\hat{R}(i) = n_i/n$ | 964 |
| Reliability of a series system: $p_A p_B \cdots p_K$, where $p_i = P(i\text{th component functions})$ | 968 |
| Reliability of a parallel system: $1 - (1 - p_A)(1 - p_B)\cdots(1 - p_K)$, where $p_i = P(i\text{th component functions})$ | 968 |

## LANGUAGE LAB

| Symbol | Pronounciation | Description |
|---|---|---|
| $F(t)$ | Cap-F-of-t | Cumulative distribution function |
| $R(t)$ | R-of-t | Reliability (survival) function |
| $z(t)$ | z-of-t | Hazard rate |

## Chapter Summary

- The **reliability** of a product is the probability that the product will function a specified length of time.
- The **failure time** of a product is the time at which the product fails.
- Commonly applied **failure time distributions** are the **exponential, Weibull**, and **normal** distributions.
- The **hazard rate** for a product is proportional to the probability that the product will fail in a small fixed interval of time, given that the product has survived to time $t$.
- A **life test** involves placing a number of product items on test and recording the observed time to failure of each.
- The length of time for a life test is sometimes shortened by **censoring** the sample—that is, stopping the life test either after a specified number of failures have been observed or after a specified amount of time has elapsed.
- A **series system** is one that fails if any one of its components fails.
- A **parallel system** is one that fails only if all of its components fail.

## Applied Supplementary Exercises

17.32 *Exponential failure time.* A certain component has an exponential failure time distribution with mean $\beta = 3$ hours.

a. Find the probability that the component will fail before time $t = 2$ hours.

b. What is the reliability of the component at time $t = 5$ hours? Interpret this value.

c. Find and graph the hazard rate for the component. Interpret the results.

17.33 *Weibull failure time.* Suppose the lifelength (in hours) of a fluorescent light has a Weibull failure time distribution with parameters $\alpha = .05$ and $\beta = .70$.

a. Find the probability that the fluorescent light will fail before time $t = 1{,}000$ hours.

b. Find the reliability of the fluorescent light at time $t = 500$ hours and interpret its value.
c. Find and graph the hazard rate for the fluorescent light at time $t = 500$ hours. Interpret the results.

17.34 *Gamma failure time.* Consider the gamma failure time distribution with $\alpha = 2$ and $\beta = 1$ given by the density function

$$f(t) = \begin{cases} te^{-t} & 0 \leq t < \infty \\ 0 & \text{elsewhere} \end{cases}$$

a. Find $F(t)$. (*Hint*: $\int te^{-t}\,dt = -te^{-t} + \int e^{-t}\,dt$)
b. Find expressions for the reliability $R(t)$ and the hazard rate $z(t)$.
c. Use the results of part **b** to find $R(3)$ and $z(3)$. Interpret these values.

17.35 *Uniform failure time.* Consider the uniform failure time distribution given by the density function

$$f(t) = \begin{cases} \dfrac{1}{\beta} & 0 \leq t \leq \beta \\ 0 & \text{elsewhere} \end{cases}$$

a. Find $F(t)$, $R(t)$, and $z(t)$.
b. Graph the hazard rate $z(t)$ for $t = 0, 1, 2, \ldots, 5$ when $\beta = 10$.
c. Compute the reliability of the system at $t = 4$ when $\beta = 10$.

17.36 *CPU failures.* To investigate the performance of the central processing unit (CPU) of a certain type of microcomputer, 20 CPUs were placed on test for a period of 5,000 hours. When the test was terminated, four CPUs had failed with failure times of 1,850, 2,090, 3,440, and 3,970 hours. Assume a negative exponential failure time distribution.

a. Find a 90% confidence interval for the mean time $\beta$ until failure of the microcomputer's CPU.
b. Find a 90% confidence interval for the reliability of the CPU at time $t = 2{,}000$ hours.
c. Find a 90% confidence interval for the hazard rate of the CPU.

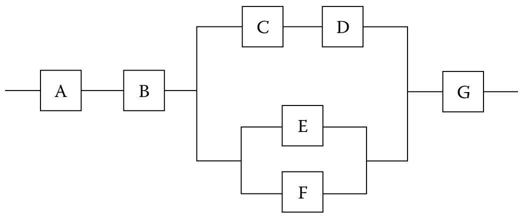
17.37 *Corrosion resistance of pipes.* A certain type of coating for pipes is designed to resist corrosion. Five hundred pieces of coated pipe were placed on test and subjected to a 90% solution of hydrochloric acid. At the end of each hour, for a period of 5 hours, the number of pipe specimens that had resisted corrosion was recorded, as shown in the accompanying table.

**PIPESPEC**

| *Hour* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| *Number Resisting Corrosion* | 438 | 280 | 146 | 51 | 15 |

a. Use the data in the table to estimate the parameters $\alpha$ and $\beta$ of a Weibull failure time distribution.
b. Use the estimates obtained in part **a** to find expression for the hazard rate $z(t)$ and the reliability $R(t)$ of the coated pipes.
c. Find the probability that a piece of coated pipe will resist corrosion under similar experimental conditions for at least 1 hour.

17.38 *System with seven tubes.* A piece of equipment consists of seven tubes connected as shown in the diagram below. Find the reliability of the system if the tubes have probabilities of functioning given by $p_A = .80$, $p_B = .90$, $p_C = .85$, $p_D = .85$, $p_E = .75$, $p_F = .75$, and $p_G = .95$.




17.39 *Resistors in parallel.* Two resistors connected in series have exponential failure time distributions with mean $\beta = 1{,}000$ hours. At time $t = 1{,}400$ hours, what is the reliability of the system?

17.40 *Components in parallel.* Four components, A, B, C, and D, are connected in parallel. Suppose that components A and B have normal failure time distributions with parameters $\mu = 500$ hours and $\sigma = 100$ hours, whereas components C and D have Weibull failure time distributions with parameters $\alpha = .5$ and $\beta = 100$. Find the reliability of the system at time $t = 300$ hours.

17.41 *Life tests of semiconductors.* The service life (in hours) of a semiconductor has an approximate exponential failure time distribution. Ten semiconductors are placed on life test until four fail. The failure times for these four semiconductors are 585, 972, 1,460, and 2,266 hours.

a. Construct a 95% confidence interval for the mean time $\beta$ until failure for the semiconductors.
b. What is the probability that a semiconductor will still be in operation after 4,000 hours? Find a 95% confidence interval for this probability.
c. Compute and interpret the hazard rate for the semiconductors. Construct a 95% confidence interval for this hazard rate.

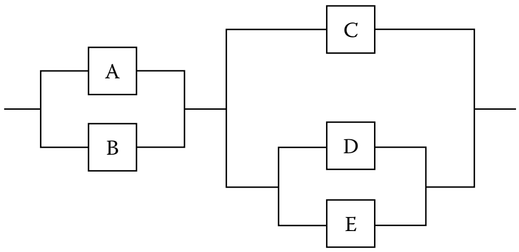
17.42 *Missile guidance failure.* The failure times (in hours) of electronic components in a guidance system for a missile have a Weibull distribution with unknown parameters $\alpha$ and $\beta$. To derive estimates of these parameters, 1,000 components were placed on life test and every 50 hours the number of components still in operation was recorded. The data are provided in the table.

**MISSILE**

| *Hours* | 50 | 100 | 150 | 200 | 250 | 300 | 350 |
|---|---|---|---|---|---|---|---|
| *Number in Operation* | 611 | 362 | 231 | 136 | 84 | 53 | 17 |

a. Find estimates of $\alpha$ and $\beta$. If you have access to a regression computer package, find 99% confidence intervals for both $\alpha$ and $\beta$.
b. Calculate the reliability of the electronic components at $t = 200$ hours.
c. Find and graph the hazard rate for $t = 50, 100, 150, \ldots$.

17.43 *Reliability of a system.* Consider the product system shown in the diagram. Given the individual component reliabilities $p_A = .85$, $p_B = .75$, $p_C = .75$, $p_D = .90$, and $p_E = .95$, find the overall reliability of the system.




## Theoretical Supplementary Exercises

17.44 Show that the hazard rate $z(t)$ can be expressed as

$$z(t) = \frac{-d[\ln R(t)]}{dt}$$

[*Hint*: Make use of the fact that $R(t) = 1 - F(t)$ and, hence, that

$$f(t) = \frac{dF(t)}{dt} = \frac{-dR(t)}{dt}$$

Then substitute these expressions into the formula given in Definition 17.5.]

17.45 Use the result of Exercise 17.44 and the relation $f(t) = z(t)R(t)$ to show that the failure time density can be expressed as

$$f(t) = z(t)e^{-\int_0^t z(y)\,dy}$$

[*Hint*: The differential equation

$$z(t) = \frac{-d[\ln R(t)]}{dt}$$

has

$$R(t) = e^{-\int_0^t z(y)\,dy}$$

as its solution.]

17.46 Suppose we are concerned only with the initial failure of a component. That is, once the component has survived past a certain time $t = \alpha$, we treat the component (for all practical purposes) as if it never failed. In this situation it is reasonable to use the hazard rate

$$z(t) = \begin{cases} \beta(1 - t/\alpha) & 0 < t < \alpha \\ 0 & \text{elsewhere} \end{cases}$$

a. Use the result of Exercise 17.45 to find expression for $f(t)$, $F(t)$, and $R(t)$.
b. Show that the probability of initial failure, i.e., the probability that the component fails before time $t = \alpha$, is $1 - e^{-\alpha\beta/2}$.

# APPENDIX A Matrix Algebra

**CONTENTS**



## A.1 Matrices and Matrix Multiplication

For some statistical procedures (e.g., multiple regression), the formulas for conducting the analysis are more easily given using **matrix algebra** instead of ordinary algebra. By arranging the data in particular rectangular patterns called **matrices** and performing various operations with them, we can obtain the results of the analyses much more quickly. In this appendix, we will define a matrix and explain various operations that can be performed with matrices. (We explained how to use this information to conduct a regression analysis in Section 11.4.)

Three matrices, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, are shown here. Note that each matrix is a rectangular arrangement of numbers with one number in every row–column position.

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 \\ 0 & 1 \\ -1 & 6 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 3 & 0 & 1 \\ -1 & 0 & 1 \\ 4 & 2 & 0 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

**Definition A.1**

A **matrix** is a rectangular array of numbers.*

The numbers that appear in a matrix are called **elements** of the matrix. If a matrix contains $r$ rows and $c$ columns, there will be an element in each of the row–column positions of the matrix, and the matrix will have $r \times c$ elements. For example, the matrix $\boldsymbol{A}$ shown above contains $r = 3$ rows, $c = 2$ columns, and $rc = (3)(2) = 6$ elements, one in each of the 6 row–column positions.

**Definition A.2**

A number in a particular row–column position is called an **element** of the matrix.

Notice that the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ contain different numbers of rows and columns. The numbers of rows and columns give the **dimensions** of a matrix.

**Definition A.3**

A matrix containing $r$ rows and $c$ columns is said to be an $\boldsymbol{r \times c}$ **matrix**, where $r$ and $c$ are the **dimensions** of the matrix.

**Definition A.4**

If $r = c$, a matrix is said to be a **square matrix**.

When we give a formula in matrix notation, the elements of a matrix will be represented by symbols. For example, if we have a matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

*For our purpose, we assume that the numbers are real.

the symbol $a_{ij}$ will denote the element in the $i$th row and $j$th column of the matrix. The first subscript always identifies the row and the second identifies the column in which the element is located. For example, the element $a_{12}$ is in the first row and second column of matrix A. The rows are always numbered from top to bottom, and the columns are always numbered from left to right.

Matrices are usually identified by capital letters, such as $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, corresponding to the letters of the alphabet employed in ordinary algebra. The difference is that in ordinary algebra, a letter is used to denote a single real number, whereas in matrix algebra, a letter denotes a rectangular array of numbers. The operations of matrix algebra are very similar to those of ordinary algebra—you can add matrices, subtract them, multiply them, and so on. However, there are a few operations that are unique to matrices, such as the **transpose of a matrix**. For example, if

$$\boldsymbol{A} = \begin{bmatrix} 5 \\ 1 \\ 0 \\ 4 \\ 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 4 \\ 1 & 2 \\ 1 & 6 \end{bmatrix}$$

then the transpose matrices of the $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices, denoted as $\boldsymbol{A}'$ and $\boldsymbol{B}'$, respectively, are

$$\boldsymbol{A}' = \begin{bmatrix} 5 & 1 & 0 & 4 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 4 & 2 & 6 \end{bmatrix}$$

**Definition A.5**

The **transpose of a matrix A**, denoted as $\boldsymbol{A}'$, is obtained by interchanging corresponding rows and columns of the $\boldsymbol{A}$ matrix. That is, the $i$th row of the $\boldsymbol{A}$ matrix becomes the $i$th column of the $\boldsymbol{A}'$ matrix.

Since we are concerned mainly with the applications of matrix algebra to the solution of the least-squares equations in multiple regression (see Chapter 11), we will define only the operations and types of matrices that are pertinent to that subject.

The most important operation for us is matrix multiplication, which requires **row–column multiplication.** To illustrate this process, suppose we wish to find the product $\boldsymbol{AB}$, where

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 4 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 2 & 0 & 3 \\ -1 & 4 & 0 \end{bmatrix}$$

We will always multiply the rows of $\boldsymbol{A}$ (the matrix on the left) by the columns of $\boldsymbol{B}$ (the matrix on the right). The product formed by the first row of $\boldsymbol{A}$ times the first column of $\boldsymbol{B}$ is obtained by multiplying the elements in corresponding positions and summing these products. Thus, the first row, first column product, shown diagrammatically below, is

$$(2)(2) + (1)(-1) = 4 - 1 = 3$$

$$\boldsymbol{AB} = \begin{bmatrix} 2 & 1 \\ 4 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 3 \\ -1 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 3 & & \\ & & \end{bmatrix}$$

Similarly, the first row, second column product is

$$(2)(0) + (1)(4) = 0 + 4 = 4$$

So far we have

$$\boldsymbol{AB} = \begin{bmatrix} 3 & 4 & \\ & & \end{bmatrix}$$

To find the complete matrix product $\boldsymbol{AB}$, all we need to do is find each element in the $\boldsymbol{AB}$ matrix. Thus, we will define an element in the $i$th row, $j$th column of $\boldsymbol{AB}$ as the product of the $i$th row of $\boldsymbol{A}$ and the $j$th column of $\boldsymbol{B}$. We complete the process in Example A.1.

**Example A.1**

Find the product $\boldsymbol{AB}$, where

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 4 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 2 & 0 & 3 \\ -1 & 4 & 0 \end{bmatrix}$$

**Solution**

If we represent the product $\boldsymbol{AB}$ as

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

we have already found $c_{11} = 3$ and $c_{12} = 4$. Similarly, the element $c_{21}$, the element in the second row, first column of $\boldsymbol{AB}$, is the product of the second row of $\boldsymbol{A}$ and the first column of $\boldsymbol{B}$:

$$(4)(2) + (-1)(-1) = 8 + 1 = 9$$

Proceeding in a similar manner to find the remaining elements of $\boldsymbol{AB}$, we have

$$\boldsymbol{AB} = \begin{bmatrix} 2 & 1 \\ 4 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 3 \\ -1 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 6 \\ 9 & -4 & 12 \end{bmatrix}$$

Now, try to find the product $\boldsymbol{BA}$, using matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ from Example A.1. You will observe two very important differences between multiplication in matrix algebra and multiplication in ordinary algebra:

1. You cannot find the product $\boldsymbol{BA}$, because you cannot perform row–column multiplication. You can see that the dimensions do not match by placing the matrices side-by-side.

$$\underset{2\times 3 \quad 2\times 2}{\boldsymbol{BA}} \qquad \text{does not exist}$$

The number of elements (3) in a row of $\boldsymbol{B}$ (the matrix on the left) does not match the number of elements (2) in a column of $\boldsymbol{A}$ (the matrix on the right). Therefore, you cannot perform row–column multiplication, and the matrix product $\boldsymbol{BA}$ does not exist. The point is, not all matrices can be multiplied. You can find products for matrices $\boldsymbol{AB}$, only where $\boldsymbol{A}$ is $r \times d$ and $\boldsymbol{B}$ is $d \times c$. That is:

### Requirement for Multiplication

$$\underset{r\times d \quad d\times c}{\boldsymbol{AB}}$$

The two inner-dimension numbers must be equal. The dimensions of the product will always be given by the outer-dimension numbers. (See the following box.)

Dimensions of ***AB*** are $r \times c$

$$\boldsymbol{AB}$$

$$r \times d \quad d \times c$$

2. The second difference between ordinary and matrix multiplication is that in ordinary algebra, $ab = ba$. In matrix algebra, $\boldsymbol{AB}$ usually does not equal $\boldsymbol{BA}$. In fact, as noted in item 1 above, it may not even exist.

**Definition A.6**

The product ***AB*** of an $r \times d$ matrix ***A*** and a $d \times c$ matrix ***B*** is an $r \times c$ matrix ***C***, where the element $c_{ij}$ $(i = 1, 2, \ldots, r; j = 1, 2, \ldots, c)$ of ***C*** is the product of the $i$th row of ***A*** and the $j$th column of ***B***.

**Example A.2**

Given the matrices shown, find ***IA*** and ***IB***.

$$\boldsymbol{A} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 3 & 0 \\ 1 & 2 \\ 4 & -1 \end{bmatrix} \qquad \boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Solution Notice that the product

$$\boldsymbol{IA}$$

$$3 \times 3 \quad 3 \times 1$$

exists and that it will be of dimensions $3 \times 1$:

$$\boldsymbol{IA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Similarly,

$$\boldsymbol{IB}$$

$$3 \times 3 \quad 3 \times 2$$

exists and is of dimensions $3 \times 2$:

$$\boldsymbol{IB} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 1 & 2 \\ 4 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & 2 \\ 4 & -1 \end{bmatrix}$$

Notice that the $\boldsymbol{I}$ matrix possesses a special property. We have $\boldsymbol{IA} = \boldsymbol{A}$ and $\boldsymbol{IB} = \boldsymbol{B}$. We will comment further on this property in Section A.2.

## Exercises

A.1 Consider the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 \\ -1 & 4 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 3 \\ -2 & 1 & 2 \end{bmatrix}$$

a. Find $\boldsymbol{AB}$. b. Find $\boldsymbol{AC}$. c. Find $\boldsymbol{BA}$.

A.2 Consider the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 1 & 3 \\ 2 & 0 & 4 \\ -4 & 1 & 2 \end{bmatrix} \qquad \boldsymbol{B} = [1 \quad 0 \quad 2] \qquad \boldsymbol{C} = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$$

a. Find $\boldsymbol{AC}$.
b. Find $\boldsymbol{BC}$.
c. Is it possible to find $\boldsymbol{AB}$? Explain.

A.3 Assume that $\boldsymbol{A}$ is a $3 \times 2$ matrix and $\boldsymbol{B}$ is a $2 \times 4$ matrix.
a. What are the dimensions of $\boldsymbol{AB}$?
b. Is it possible to find the product $\boldsymbol{BA}$? Explain.

A.4 Assume that matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ are of dimensions $1 \times 3$ and $3 \times 1$, respectively.
a. What are the dimensions of the product $\boldsymbol{BC}$?
b. What are the dimensions of $\boldsymbol{CB}$?
c. If $\boldsymbol{B}$ and $\boldsymbol{C}$ are the matrices shown in Exercise A.2, find $\boldsymbol{CB}$.

A.5 Consider the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 2 & 3 \\ -3 & 0 \\ 4 & -1 \end{bmatrix} \qquad \boldsymbol{C} = [3 \quad 0 \quad 2]$$

a. Find $\boldsymbol{AB}$. b. Find $\boldsymbol{CA}$. c. Find $\boldsymbol{CB}$.

A.6 Consider the matrices:

$$\boldsymbol{A} = [3 \quad 0 \quad -1 \quad 2] \qquad \boldsymbol{B} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 3 \end{bmatrix}$$

a. Find $\boldsymbol{AB}$. b. Find $\boldsymbol{BA}$.

## A.2 Identity Matrices and Matrix Inversion

In ordinary algebra, the number 1 is the identity element for the multiplication operation. That is, 1 is the element such that any other number, say, $c$, multiplied by the identity element is equal to $c$. Thus, $4(1) = 4$, $(-5)(1) = -5$, etc.

The corresponding identity element for multiplication in matrix algebra, identified by the symbol $\boldsymbol{I}$, is a matrix such that

$$\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A} \quad \text{for any matrix } \boldsymbol{A}$$

The difference between identity elements in ordinary algebra and matrix algebra is that in ordinary algebra, there is only one identity element, the number 1. In matrix algebra, the identity matrix must possess the correct dimensions for the product $\boldsymbol{IA}$ to exist. Thus, there is an infinitely large number of identity matrices—all square and all possessing the same pattern. The $1 \times 1$, $2 \times 2$, and $3 \times 3$ identity matrices are

$$\underset{1 \times 1}{\boldsymbol{I}} = [1] \qquad \underset{2 \times 2}{\boldsymbol{I}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \underset{3 \times 3}{\boldsymbol{I}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In Example A.2, we demonstrated the fact that this matrix satisfies the property

$$\boldsymbol{IA} = \boldsymbol{A}$$

**Example A.3**

If $\boldsymbol{A}$ is the matrix shown, find $\boldsymbol{IA}$ and $\boldsymbol{AI}$.

$$\boldsymbol{A} = \begin{bmatrix} 3 & 4 & -1 \\ 1 & 0 & 2 \end{bmatrix}$$

Solution

$$\underset{2 \times 2 \;\; 2 \times 3}{\boldsymbol{IA}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 4 & -1 \\ 1 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 4 & -1 \\ 1 & 0 & 2 \end{bmatrix} = \boldsymbol{A}$$

$$\underset{2 \times 3 \ \ 3 \times 3}{\boldsymbol{AI}} = \begin{bmatrix} 3 & 4 & -1 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 4 & -1 \\ 1 & 0 & 2 \end{bmatrix} = \boldsymbol{A}$$

Notice that the identity matrices used to find the products $\boldsymbol{IA}$ and $\boldsymbol{AI}$ were of different dimensions. This was necessary for the products to exist.

**Definition A.7**

If $\boldsymbol{A}$ is any matrix, then a matrix $\boldsymbol{I}$ is defined to be an **identity matrix** if $\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A}$. The matrices that satisfy this definition possess the pattern

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

The identity element assumes importance when we consider the process of division and its role in the solution of equations. In ordinary algebra, division is essentially multiplication using the reciprocals of elements. For example, the equation

$$2X = 6$$

can be solved by dividing both sides of the equation by 2, *or* it can be solved by *multiplying* both sides of the equation by $\frac{1}{2}$, which is the reciprocal of 2. Thus,

$$\left(\frac{1}{2}\right) 2X = \frac{1}{2}(6)$$
$$X = 3$$

What is the reciprocal of an element? It is the element such that the reciprocal times the element is equal to the identity element. Thus, the reciprocal of 3 is $\frac{1}{3}$ because

$$3\left(\frac{1}{3}\right) = 1$$

The identity matrix plays the same role in matrix algebra. Thus, the reciprocal of a matrix $\boldsymbol{A}$, called $\boldsymbol{A}$**-inverse** and denoted by the symbol $\boldsymbol{A}^{-1}$, is a matrix such that $\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$.

Inverses are defined only for square matrices, but not all square matrices possess inverses. (Those that do play an important role in solving the least-squares equations and in other aspects of a regression analysis.) We will show you one important application of the inverse matrix in Section A.3. The procedure for finding the inverse of a matrix is demonstrated in Section A.4.

**Definition A.8**

The square matrix $\boldsymbol{A}^{-1}$ is said to be the **inverse** of the square matrix $\boldsymbol{A}$ if

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{AA}^{-1} = \boldsymbol{I}$$

The procedure for finding an inverse matrix is computationally quite tedious and is performed most often using a computer. There is one exception. Finding the inverse of one type of matrix, called a **diagonal matrix**, is easy. A diagonal matrix is one that has nonzero elements down the **main diagonal** (running top left of the matrix to bottom right) and 0 elements elsewhere. For example, the identity matrix is a diagonal matrix (with 1's along the main diagonal), as are the following matrices:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

**Definition A.9**

A **diagonal matrix** is one that contains nonzero elements on the main diagonal and 0 elements elsewhere.

You can verify that the inverse of

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{is} \quad \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$$

i.e., $\boldsymbol{AA}^{-1} = \boldsymbol{I}$. The inverse of a diagonal matrix is given by Theorem A.1.

**THEOREM A.1**

The inverse of a diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ 0 & 0 & d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \cdots & d_{nn} \end{bmatrix} \quad \text{is} \quad \boldsymbol{D}^{-1} = \begin{bmatrix} 1/d_{11} & 0 & 0 & \cdots & 0 \\ 0 & 1/d_{22} & 0 & \cdots & 0 \\ 0 & 0 & 1/d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1/d_{nn} \end{bmatrix}$$

## Exercises

A.7 Consider the following matrix:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 2 \\ -1 & 1 & 4 \end{bmatrix}$$

a. Give the identity matrix that will be used to obtain the product $\boldsymbol{IA}$.
b. Show that $\boldsymbol{IA} = \boldsymbol{A}$.
c. Give the identity matrix that will be used to find the product $\boldsymbol{AI}$.
d. Show that $\boldsymbol{AI} = \boldsymbol{A}$.

A.8 For the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ given here, show that $\boldsymbol{AB} = \boldsymbol{I}$ and that $\boldsymbol{BA} = \boldsymbol{I}$, and, consequently, verify that $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

A.9 If

$$\boldsymbol{A} = \begin{bmatrix} 12 & 0 & 0 & 8 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 8 & 0 & 0 & 8 \end{bmatrix} \text{ verify that } \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 & -\frac{1}{4} \\ 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \\ -\frac{1}{4} & 0 & 0 & \frac{3}{8} \end{bmatrix}$$

A.10 If

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix} \text{ show that } \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix}$$

A.11 Verify Theorem A.1.

## A.3 Solving Systems of Simultaneous Linear Equations

Consider the following set of simultaneous linear equations in two unknowns:

$$2v_1 + v_2 = 7$$
$$v_1 - v_2 = 2$$

Note that the solution for these equations is $v_1 = 3$, $v_2 = 1$.

Now define the matrices

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \boldsymbol{G} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$$

Thus, $\boldsymbol{A}$ is the matrix of coefficients of $v_1$ and $v_2$, $\boldsymbol{V}$ is a column matrix containing the unknowns (written in order, top to bottom), and $\boldsymbol{G}$ is a column matrix containing the numbers on the right-hand side of the equal signs.

Now, the system of simultaneous equations shown above can be rewritten as a matrix equation:

$$\boldsymbol{AV} = \boldsymbol{G}$$

By a matrix equation, we mean that the product matrix, $\boldsymbol{AV}$, is equal to the matrix $\boldsymbol{G}$. *Equality of matrices means that corresponding elements are equal*. You can see that this is true for the expression $\boldsymbol{AV} = \boldsymbol{G}$, since

$$\underset{\substack{\nearrow\ \nwarrow \\ 2 \times 2\ \ 2 \times 1}}{\boldsymbol{AV}} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} (2v_1 + v_2) \\ (v_1 - v_2) \end{bmatrix} = \underset{2 \times 1}{\boldsymbol{G}}$$

The matrix procedure for expressing a system of two simultaneous linear equations in two unknowns can be extended to express a set of $k$ simultaneous equations in $k$ unknowns. If the equations are written in the orderly pattern

$$\begin{matrix} a_{11}v_1 & + & a_{12}v_2 & + & \cdots & + & a_{1k}v_k & = & g_1 \\ a_{21}v_1 & + & a_{22}v_2 & + & \cdots & + & a_{2k}v_k & = & g_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{k1}v_1 & + & a_{k2}v_2 & + & \cdots & + & a_{kk}v_k & = & g_k \end{matrix}$$

then the set of simultaneous linear equations can be expressed as the matrix equation $\boldsymbol{AV} = \boldsymbol{G}$, where

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & & \cdots & a_{2k} \\ \vdots & & & \vdots \\ a_{k1} & & \cdots & a_{kk} \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} \qquad \boldsymbol{G} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{bmatrix}$$

Now let us solve this system of simultaneous equations. (If they are uniquely solvable, it can be shown that $\boldsymbol{A}^{-1}$ exists.) Multiplying both sides of the matrix equation by $\boldsymbol{A}^{-1}$, we have

$$(\boldsymbol{A}^{-1})\boldsymbol{AV} = (\boldsymbol{A}^{-1})\boldsymbol{G}$$

But since $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, we have

$$(\boldsymbol{I})\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{G}$$
$$\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{G}$$

In other words, if we know $\boldsymbol{A}^{-1}$, we can find the solution to the set of simultaneous linear equations by obtaining the product $\boldsymbol{A}^{-1}\boldsymbol{G}$.

Matrix Solution to a Set of Simultaneous Linear Equations, $AV = G$

*Solution:* $\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{G}$

**Example A.4**

Apply the boxed result to find the solution to the set of simultaneous linear equations

$$\begin{aligned} 2v_1 + v_2 &= 7 \\ v_1 - v_2 &= 2 \end{aligned}$$

**Solution** The first step is to obtain the inverse of the coefficient matrix,

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}$$

namely,

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} \end{bmatrix}$$

(This matrix can be found using a packaged computer program for matrix inversion or, for this simple case, you could use the procedure explained in Section A.4.) As a check, note that

$$\begin{aligned} \boldsymbol{A}^{-1}\boldsymbol{A} &= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{I} \end{aligned}$$

The second step is to obtain the product $\boldsymbol{A}^{-1}\boldsymbol{G}$. Thus,

$$\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{G} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} \end{bmatrix} \begin{bmatrix} 7 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Since

$$\boldsymbol{V} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

it follows that $v_1 = 3$ and $v_2 = 1$. You can see that these values of $v_1$ and $v_2$ satisfy the simultaneous linear equations and are the values that we specified as a solution at the beginning of this section.

## Exercises

A.12 Suppose the simultaneous linear equations

$$\begin{aligned} 3v_1 + v_2 &= 5 \\ v_1 - v_2 &= 3 \end{aligned}$$

are expressed as a matrix equation,

$\boldsymbol{AV} = \boldsymbol{G}$

a. Find the matrices $\boldsymbol{A}$, $\boldsymbol{V}$, and $\boldsymbol{G}$.

b. Verify that

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{3}{4} \end{bmatrix}$$

(*Note*: A procedure for finding $A^{-1}$ is given in Section A.4.)

c. Solve the equations by finding $V = A^{-1}G$ .

A.13 For the simultaneous linear equations

$$\begin{aligned} 10v_1 + 20v_3 - 60 &= 0 \\ 20v_2 - 60 &= 0 \\ 20v_1 + 68v_3 - 176 &= 0 \end{aligned}$$

a. Find the matrices $\boldsymbol{A}$, $\boldsymbol{V}$, and $\boldsymbol{G}$.

b. Verify that

$$A^{-1} = \begin{bmatrix} \frac{17}{70} & 0 & -\frac{1}{14} \\ 0 & \frac{1}{20} & 0 \\ -\frac{1}{14} & 0 & \frac{1}{28} \end{bmatrix}$$

c. Solve the equations by finding $V = A^{-1}G$.

## A.4 A Procedure for Inverting a Matrix

There are several different methods for inverting matrices. All are tedious and time-consuming. Consequently, in practice, you will invert almost all matrices using computer software. The purpose of this section is to present one method so that you will be able to invert small ($2 \times 2$ or $3 \times 3$) matrices manually and so that you will appreciate the enormous computing problem involved in inverting large matrices (and, consequently, in fitting linear models containing many terms to a set of data). Particularly, you will be able to understand why rounding errors creep into the inversion process and, consequently, why two different computer programs might invert the same matrix and produce inverse matrices with slightly different corresponding elements.

The procedure we will demonstrate to invert a matrix $\boldsymbol{A}$ requires that we perform a series of operations on the rows of the $\boldsymbol{A}$ matrix. For example, suppose

$$A = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}$$

We will identify two different ways to operate on a row of a matrix:*

1. We can multiply every element in one particular row by a constant, $c$. For example, we could operate on the first row of the $\boldsymbol{A}$ matrix by multiplying every element in the row by a constant, say, 2. Then the resulting row would be $[2 \quad -4]$.
2. We can operate on a row by multiplying another row of the matrix by a constant and then adding (or subtracting) the elements of that row to elements in corresponding positions in the row operated upon. For example, we could operate on the first row of the $\boldsymbol{A}$ matrix by multiplying the second row by a constant, say, 2:

   $$2[-2 \quad 6] = [-4 \quad 12]$$

   Then we add this row to row 1:

   $$[(1-4) \quad (-2+12)] = [-3 \quad 10]$$

   Note one important point. We operated on the *first* row of the $\boldsymbol{A}$ matrix. Although we used the second row of the matrix to perform the operation, *the second row would remain unchanged*. Therefore, the row operation on the $\boldsymbol{A}$ matrix that we have just described would produce the new matrix,

   $$\begin{bmatrix} -3 & 10 \\ -2 & 6 \end{bmatrix}$$

Matrix inversion using row operations is based on an elementary result from matrix algebra. It can be shown (proof omitted) that performing a series of row operations on

*We omit a third row operation, because it would add little and could be confusing.

a matrix $\boldsymbol{A}$ is equivalent to multiplying $\boldsymbol{A}$ by a matrix $\boldsymbol{B}$, i.e., row operations produce a new matrix, $\boldsymbol{BA}$. This result is used as follows: Place the $\boldsymbol{A}$ matrix and an identity matrix $\boldsymbol{I}$ of the same dimensions side by side. Then perform the same series of row operations on both $\boldsymbol{A}$ and $\boldsymbol{I}$ until the $\boldsymbol{A}$ matrix has been changed into the identity matrix $\boldsymbol{I}$. This means that you have multiplied both $\boldsymbol{A}$ and $\boldsymbol{I}$ by some matrix $\boldsymbol{B}$ such that

$$\boldsymbol{A} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \qquad \boldsymbol{I} = \begin{bmatrix} 1 & 0 & & \cdots & 0 \\ 0 & 1 & & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdot & \ddots & \vdots \\ 0 & & & \cdots & 1 \end{bmatrix}$$

$$\downarrow \longleftarrow \text{Row operations change } \boldsymbol{A} \text{ to } \boldsymbol{I} \longrightarrow \downarrow$$

$$\boldsymbol{I} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

$$\boldsymbol{BA} = \boldsymbol{I} \quad \text{and} \quad \boldsymbol{BI} = \boldsymbol{B}$$

Since $\boldsymbol{BA} = \boldsymbol{I}$, it follows that $\boldsymbol{B} = \boldsymbol{A}^{-1}$. Therefore, as the $\boldsymbol{A}$ matrix is transformed by row operations into the identity matrix $\boldsymbol{I}$, the identity matrix $\boldsymbol{I}$ is transformed into $\boldsymbol{A}^{-1}$, i.e.,

$$\boldsymbol{BI} = \boldsymbol{B} = \boldsymbol{A}^{-1}$$

We will show you how this procedure works with two examples.

**Example A.5**

Find the inverse of the matrix,

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}$$

Solution

Place the $\boldsymbol{A}$ matrix and a $2 \times 2$ identity matrix side by side and then perform the following series of row operations (we will indicate by arrow the row operated upon in each operation):

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} \qquad \boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

**OPERATION 1** Multiply the first row by 2 and add to the second row:

$$\begin{matrix} \\ \rightarrow \end{matrix} \begin{bmatrix} 1 & -2 \\ 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

**OPERATION 2** Multiply the second row by $\frac{1}{2}$:

$$\begin{matrix} \\ \rightarrow \end{matrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 1 & \frac{1}{2} \end{bmatrix}$$

**OPERATION 3** Multiply the second row by 2 and add it to the first row:

$$\begin{matrix} \rightarrow \\ \end{matrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 3 & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$$

Thus,

$$A^{-1} = \begin{bmatrix} 3 & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$$

The final step in finding an inverse is to check your solution by finding the product $A^{-1}A$ to see if it equals the identity matrix $I$. To check:

$$A^{-1}A = \begin{bmatrix} 3 & 1 \\ 1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Since this product is equal to the identity matrix, it follows that our solution for $A^{-1}$ is correct.

## Example A.6

Find the inverse of the matrix,

$$A = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 4 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

Solution

Place an identity matrix alongside the $A$ matrix and perform the row operations:

**OPERATION 1** Multiply row 1 by $\frac{1}{2}$:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 4 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

**OPERATION 2** Multiply row 1 by 3 and subtract from row 3:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 4 & 1 \\ 0 & 1 & -\frac{5}{2} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{3}{2} & 0 & 1 \end{bmatrix}$$

**OPERATION 3** Multiply row 2 by $\frac{1}{4}$:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 1 & \frac{1}{4} \\ 0 & 1 & -\frac{5}{2} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ -\frac{3}{2} & 0 & 1 \end{bmatrix}$$

**OPERATION 4** Subtract row 2 from row 3:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 1 & \frac{1}{4} \\ 0 & 0 & -\frac{11}{4} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ -\frac{3}{2} & -\frac{1}{4} & 1 \end{bmatrix}$$

**OPERATION 5** Multiply row 3 by $-\frac{4}{11}$:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 1 & \frac{1}{4} \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ \frac{12}{22} & \frac{1}{11} & -\frac{4}{11} \end{bmatrix}$$

**OPERATION 6** Operate on row 2 by subtracting $\frac{1}{4}$ of row 3:

$$\rightarrow \begin{bmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{3}{22} & \frac{5}{22} & \frac{1}{11} \\ \frac{12}{22} & \frac{1}{11} & -\frac{4}{11} \end{bmatrix}$$

**OPERATION 7** Operate on row 1 by subtracting $\frac{3}{2}$ of row 3:

$$\rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -\frac{7}{22} & -\frac{3}{22} & \frac{6}{11} \\ -\frac{3}{22} & \frac{5}{22} & \frac{1}{11} \\ \frac{6}{11} & \frac{1}{11} & -\frac{4}{11} \end{bmatrix} = \boldsymbol{A}^{-1}$$

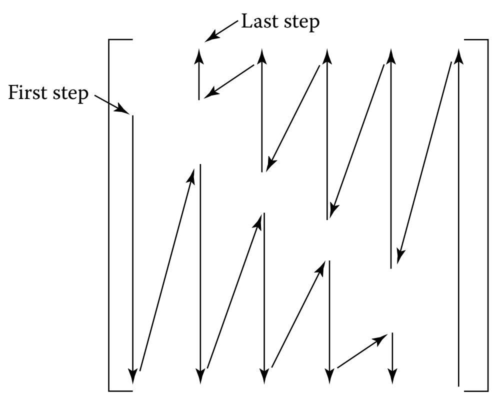
To check the solution, we find the product,

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \begin{bmatrix} -\frac{7}{22} & -\frac{3}{22} & \frac{6}{11} \\ -\frac{3}{22} & \frac{5}{22} & \frac{1}{11} \\ \frac{6}{11} & \frac{1}{11} & -\frac{4}{11} \end{bmatrix} \begin{bmatrix} 2 & 0 & 3 \\ 0 & 4 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since the product $\boldsymbol{A}^{-1}\boldsymbol{A}$ is equal to the identity matrix, it follows that our solution for $\boldsymbol{A}^{-1}$ is correct.

Examples A.5 and A.6 indicate the strategy employed when performing row operations on the $\boldsymbol{A}$ matrix to change it into an identity matrix. Multiply the first row by a constant to change the element in the top left row into a 1. Then perform operations to change all elements in the first column into 0's. Then operate on the second row and change the second diagonal element into a 1. Then operate to change all elements in the second column beneath row 2 into 0's. Then operate on the diagonal element in row 3, etc. When all elements on the main diagonal are 1's and all below the main diagonal are 0's, perform row operations to change the last column to 0; then the next-to-last column, etc., until you get back to the first row. The procedure for changing the off-diagonal elements to 0's is indicated diagrammatically as shown:




The preceding instructions on how to invert a matrix using row operations suggest that the inversion of a large matrix would involve many multiplications, subtractions, and additions and, consequently, could produce large rounding errors in the calculations unless you carry a large number of significant figures in the calculations. This explains why two different multiple regression analysis computer programs may produce different estimates of the same $\beta$ parameters, and it emphasizes the importance of carrying a large number of significant figures in all computations when inverting a matrix.

## Exercise

A.14 Invert the following matrices and check your answers to make certain that $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$:

a. $\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 4 & 5 \end{bmatrix}$

b. $\boldsymbol{A} = \begin{bmatrix} 3 & 0 & -2 \\ 1 & 4 & 2 \\ 5 & 1 & 1 \end{bmatrix}$

c. $\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}$

d. $\boldsymbol{A} = \begin{bmatrix} 4 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 5 \end{bmatrix}$

(*Note*: No answers are given to these exercises. You will know your answers are correct if $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$.)

# APPENDIX B Useful Statistical Tables

CONTENTS

**TABLE 1 Random Numbers**

| Row \ Column | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10480 | 15011 | 01536 | 02011 | 81647 | 91646 | 69179 | 14194 | 62590 | 36207 | 20969 | 99570 | 91291 | 90700 |
| 2 | 22368 | 46573 | 25595 | 85393 | 30995 | 89198 | 27982 | 53402 | 93965 | 34095 | 52666 | 19174 | 39615 | 99505 |
| 3 | 24130 | 48360 | 22527 | 97265 | 76393 | 64809 | 15179 | 24830 | 49340 | 32081 | 30680 | 19655 | 63348 | 58629 |
| 4 | 42167 | 93093 | 06243 | 61680 | 07856 | 16376 | 39440 | 53537 | 71341 | 57004 | 00849 | 74917 | 97758 | 16379 |
| 5 | 37570 | 39975 | 81837 | 16656 | 06121 | 91782 | 60468 | 81305 | 49684 | 60672 | 14110 | 06927 | 01263 | 54613 |
| 6 | 77921 | 06907 | 11008 | 42751 | 27756 | 53498 | 18602 | 70659 | 90655 | 15053 | 21916 | 81825 | 44394 | 42880 |
| 7 | 99562 | 72905 | 56420 | 69994 | 98872 | 31016 | 71194 | 18738 | 44013 | 48840 | 63213 | 21069 | 10634 | 12952 |
| 8 | 96301 | 91977 | 05463 | 07972 | 18876 | 20922 | 94595 | 56869 | 69014 | 60045 | 18425 | 84903 | 42508 | 32307 |
| 9 | 89579 | 14342 | 63661 | 10281 | 17453 | 18103 | 57740 | 84378 | 25331 | 12566 | 58678 | 44947 | 05585 | 56941 |
| 10 | 85475 | 36857 | 53342 | 53988 | 53060 | 59533 | 38867 | 62300 | 08158 | 17983 | 16439 | 11458 | 18593 | 64952 |
| 11 | 28918 | 69578 | 88231 | 33276 | 70997 | 79936 | 56865 | 05859 | 90106 | 31595 | 01547 | 85590 | 91610 | 78188 |
| 12 | 63553 | 40961 | 48235 | 03427 | 49626 | 69445 | 18663 | 72695 | 52180 | 20847 | 12234 | 90511 | 33703 | 90322 |
| 13 | 09429 | 93969 | 52636 | 92737 | 88974 | 33488 | 36320 | 17617 | 30015 | 08272 | 84115 | 27156 | 30613 | 74952 |
| 14 | 10365 | 61129 | 87529 | 85689 | 48237 | 52267 | 67689 | 93394 | 01511 | 26358 | 85104 | 20285 | 29975 | 89868 |
| 15 | 07119 | 97336 | 71048 | 08178 | 77233 | 13916 | 47564 | 81056 | 97735 | 85977 | 29372 | 74461 | 28551 | 90707 |
| 16 | 51085 | 12765 | 51821 | 51259 | 77452 | 16308 | 60756 | 92144 | 49442 | 53900 | 70960 | 63990 | 75601 | 40719 |
| 17 | 02368 | 21382 | 52404 | 60268 | 89368 | 19885 | 55322 | 44819 | 01188 | 65255 | 64835 | 44919 | 05944 | 55157 |
| 18 | 01011 | 54092 | 33362 | 94904 | 31273 | 04146 | 18594 | 29852 | 71585 | 85030 | 51132 | 01915 | 92747 | 64951 |
| 19 | 52162 | 53916 | 46369 | 58586 | 23216 | 14513 | 83149 | 98736 | 23495 | 64350 | 94738 | 17752 | 35156 | 35749 |
| 20 | 07056 | 97628 | 33787 | 09998 | 42698 | 06691 | 76988 | 13602 | 51851 | 46104 | 88916 | 19509 | 25625 | 58104 |
| 21 | 48663 | 91245 | 85828 | 14346 | 09172 | 30168 | 90229 | 04734 | 59193 | 22178 | 30421 | 61666 | 99904 | 32812 |
| 22 | 54164 | 58492 | 22421 | 74103 | 47070 | 25306 | 76468 | 26384 | 58151 | 06646 | 21524 | 15227 | 96909 | 44592 |
| 23 | 32639 | 32363 | 05597 | 24200 | 13363 | 38005 | 94342 | 28728 | 35806 | 06912 | 17012 | 64161 | 18296 | 22851 |
| 24 | 29334 | 27001 | 87637 | 87308 | 58731 | 00256 | 45834 | 15398 | 46557 | 41135 | 10367 | 07684 | 36188 | 18510 |
| 25 | 02488 | 33062 | 28834 | 07351 | 19731 | 92420 | 60952 | 61280 | 50001 | 67658 | 32586 | 86679 | 50720 | 94953 |
| 26 | 81525 | 72295 | 04839 | 96423 | 24878 | 82651 | 66566 | 14778 | 76797 | 14780 | 13300 | 87074 | 79666 | 95725 |
| 27 | 29676 | 20591 | 68086 | 26432 | 46901 | 20849 | 89768 | 81536 | 86645 | 12659 | 92259 | 57102 | 80428 | 25280 |
| 28 | 00742 | 57392 | 39064 | 66432 | 84673 | 40027 | 32832 | 61362 | 98947 | 96067 | 64760 | 64584 | 96096 | 98253 |

**TABLE 1 Random Numbers (continued)**

| Row \ Column | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | 05366 | 04213 | 25669 | 26422 | 44407 | 44048 | 37937 | 63904 | 45766 | 66134 | 75470 | 66520 | 34693 | 90449 |
| 30 | 91921 | 26418 | 64117 | 94305 | 26766 | 25940 | 39972 | 22209 | 71500 | 64568 | 91402 | 42416 | 07844 | 69618 |
| 31 | 00582 | 04711 | 87917 | 77341 | 42206 | 35126 | 74087 | 99547 | 81817 | 42607 | 43808 | 76655 | 62028 | 76630 |
| 32 | 00725 | 69884 | 62797 | 56170 | 86324 | 88072 | 76222 | 36086 | 84637 | 93161 | 76038 | 65855 | 77919 | 88006 |
| 33 | 69011 | 65795 | 95876 | 55293 | 18988 | 27354 | 26575 | 08625 | 40801 | 59920 | 29841 | 80150 | 12777 | 48501 |
| 34 | 25976 | 57948 | 29888 | 88604 | 67917 | 48708 | 18912 | 82271 | 65424 | 69774 | 33611 | 54262 | 85963 | 03547 |
| 35 | 09763 | 83473 | 73577 | 12908 | 30883 | 18317 | 28290 | 35797 | 05998 | 41688 | 34952 | 37888 | 38917 | 88050 |
| 36 | 91576 | 42595 | 27958 | 30134 | 04024 | 86385 | 29880 | 99730 | 55536 | 84855 | 29080 | 09250 | 79656 | 73211 |
| 37 | 17955 | 56349 | 90999 | 49127 | 20044 | 59931 | 06115 | 20542 | 18059 | 02008 | 73708 | 83517 | 36103 | 42791 |
| 38 | 46503 | 18584 | 18845 | 49618 | 02304 | 51038 | 20655 | 58727 | 28168 | 15475 | 56942 | 53389 | 20562 | 87338 |
| 39 | 92157 | 89634 | 94824 | 78171 | 84610 | 82834 | 09922 | 25417 | 44137 | 48413 | 25555 | 21246 | 35509 | 20468 |
| 40 | 14577 | 62765 | 35605 | 81263 | 39667 | 47358 | 56873 | 56307 | 61607 | 49518 | 89656 | 20103 | 77490 | 18062 |
| 41 | 98427 | 07523 | 33362 | 64270 | 01638 | 92477 | 66969 | 98420 | 04880 | 45585 | 46565 | 04102 | 46880 | 45709 |
| 42 | 34914 | 63976 | 88720 | 82765 | 34476 | 17032 | 87589 | 40836 | 32427 | 70002 | 70663 | 88863 | 77775 | 69348 |
| 43 | 70060 | 28277 | 39475 | 46473 | 23219 | 53416 | 94970 | 25832 | 69975 | 94884 | 19661 | 72828 | 00102 | 66794 |
| 44 | 53976 | 54914 | 06990 | 67245 | 68350 | 82948 | 11398 | 42878 | 80287 | 88267 | 47363 | 46634 | 06541 | 97809 |
| 45 | 76072 | 29515 | 40980 | 07391 | 58745 | 25774 | 22987 | 80059 | 39911 | 96189 | 41151 | 14222 | 60697 | 59583 |
| 46 | 90725 | 52210 | 83974 | 29992 | 65831 | 38857 | 50490 | 83765 | 55657 | 14361 | 31720 | 57375 | 56228 | 41546 |
| 47 | 64364 | 67412 | 33339 | 31926 | 14883 | 24413 | 59744 | 92351 | 97473 | 89286 | 35931 | 04110 | 23726 | 51900 |
| 48 | 08962 | 00358 | 31662 | 25388 | 61642 | 34072 | 81249 | 35648 | 56891 | 69352 | 48373 | 45578 | 78547 | 81788 |
| 49 | 95012 | 68379 | 93526 | 70765 | 10592 | 04542 | 76463 | 54328 | 02349 | 17247 | 28865 | 14777 | 62730 | 92277 |
| 50 | 15664 | 10493 | 20492 | 38391 | 91132 | 21999 | 59516 | 81652 | 27195 | 48223 | 46751 | 22923 | 32261 | 85653 |
| 51 | 16408 | 81899 | 04153 | 53381 | 79401 | 21438 | 83035 | 92350 | 36693 | 31238 | 59649 | 91754 | 72772 | 02338 |
| 52 | 18629 | 81953 | 05520 | 91962 | 04739 | 13092 | 97662 | 24822 | 94730 | 06496 | 35090 | 04822 | 86774 | 98289 |
| 53 | 73115 | 35101 | 47498 | 87637 | 99016 | 71060 | 88824 | 71013 | 18735 | 20286 | 23153 | 72924 | 35165 | 43040 |
| 54 | 57491 | 16703 | 23167 | 49323 | 45021 | 33132 | 12544 | 41035 | 80780 | 45393 | 44812 | 12515 | 98931 | 91202 |
| 55 | 30405 | 83946 | 23792 | 14422 | 15059 | 45799 | 22716 | 19792 | 09983 | 74353 | 68668 | 30429 | 70735 | 25499 |
| 56 | 16631 | 35006 | 85900 | 98275 | 32388 | 52390 | 16815 | 69298 | 82732 | 38480 | 73817 | 32523 | 41961 | 44437 |

**TABLE 1 Random Numbers (continued)**

| Row \ Column | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 96773 | 20206 | 42559 | 78985 | 05300 | 22164 | 24369 | 54224 | 35083 | 19687 | 11052 | 91491 | 60383 | 19746 |
| 58 | 38935 | 64202 | 14349 | 82674 | 66523 | 44133 | 00697 | 35552 | 35970 | 19124 | 63318 | 29686 | 03387 | 59846 |
| 59 | 31624 | 76384 | 17403 | 53363 | 44167 | 64486 | 64758 | 75366 | 76554 | 31601 | 12614 | 33072 | 60332 | 92325 |
| 60 | 78919 | 19474 | 23632 | 27889 | 47914 | 02584 | 37680 | 20801 | 72152 | 39339 | 34806 | 08930 | 85001 | 87820 |
| 61 | 03931 | 33309 | 57047 | 74211 | 63445 | 17361 | 62825 | 39908 | 05607 | 91284 | 68833 | 25570 | 38818 | 46920 |
| 62 | 74426 | 33278 | 43972 | 10119 | 89917 | 15665 | 52872 | 73823 | 73144 | 88662 | 88970 | 74492 | 51805 | 99378 |
| 63 | 09066 | 00903 | 20795 | 95452 | 92648 | 45454 | 09552 | 88815 | 16553 | 51125 | 79375 | 97596 | 16296 | 66092 |
| 64 | 42238 | 12426 | 87025 | 14267 | 20979 | 04508 | 64535 | 31355 | 86064 | 29472 | 47689 | 05974 | 52468 | 16834 |
| 65 | 16153 | 08002 | 26504 | 41744 | 81959 | 65642 | 74240 | 56302 | 00033 | 67107 | 77510 | 70625 | 28725 | 34191 |
| 66 | 21457 | 40742 | 29820 | 96783 | 29400 | 21840 | 15035 | 34537 | 33310 | 06116 | 95240 | 15957 | 16572 | 06004 |
| 67 | 21581 | 57802 | 02050 | 89728 | 17937 | 37621 | 47075 | 42080 | 97403 | 48626 | 68995 | 43805 | 33386 | 21597 |
| 68 | 55612 | 78095 | 83197 | 33732 | 05810 | 24813 | 86902 | 60397 | 16489 | 03264 | 88525 | 42786 | 05269 | 92532 |
| 69 | 44657 | 66999 | 99324 | 51281 | 84463 | 60563 | 79312 | 93454 | 68876 | 25471 | 93911 | 25650 | 12682 | 73572 |
| 70 | 91340 | 84979 | 46949 | 81973 | 37949 | 61023 | 43997 | 15263 | 80644 | 43942 | 89203 | 71795 | 99533 | 50501 |
| 71 | 91227 | 21199 | 31935 | 27022 | 84067 | 05462 | 35216 | 14486 | 29891 | 68607 | 41867 | 14951 | 91696 | 85065 |
| 72 | 50001 | 38140 | 66321 | 19924 | 72163 | 09538 | 12151 | 06878 | 91903 | 18749 | 34405 | 56087 | 82790 | 70925 |
| 73 | 65390 | 05224 | 72958 | 28609 | 81406 | 39147 | 25549 | 48542 | 42627 | 45233 | 57202 | 94617 | 23772 | 07896 |
| 74 | 27504 | 96131 | 83944 | 41575 | 10573 | 08619 | 64482 | 73923 | 36152 | 05184 | 94142 | 25299 | 84387 | 34925 |
| 75 | 37169 | 94851 | 39117 | 89632 | 00959 | 16487 | 65536 | 49071 | 39782 | 17095 | 02330 | 74301 | 00275 | 48280 |
| 76 | 11508 | 70225 | 51111 | 38351 | 19444 | 66499 | 71945 | 05422 | 13442 | 78675 | 84081 | 66938 | 93654 | 59894 |
| 77 | 37449 | 30362 | 06694 | 54690 | 04052 | 53115 | 62757 | 95348 | 78662 | 11163 | 81651 | 50245 | 34971 | 52924 |
| 78 | 46515 | 70331 | 85922 | 38329 | 57015 | 15765 | 97161 | 17869 | 45349 | 61796 | 66345 | 81073 | 49106 | 79860 |
| 79 | 30986 | 81223 | 42416 | 58353 | 21532 | 30502 | 32305 | 86482 | 05174 | 07901 | 54339 | 58861 | 74818 | 46942 |
| 80 | 63798 | 64995 | 46583 | 09785 | 44160 | 78128 | 83991 | 42865 | 92520 | 83531 | 80377 | 35909 | 81250 | 54238 |
| 81 | 82486 | 84846 | 99254 | 67632 | 43218 | 50076 | 21361 | 64816 | 51202 | 88124 | 41870 | 52689 | 51275 | 83556 |
| 82 | 21885 | 32906 | 92431 | 09060 | 64297 | 51674 | 64126 | 62570 | 26123 | 05155 | 59194 | 52799 | 28225 | 85762 |
| 83 | 60336 | 98782 | 07408 | 53458 | 13564 | 59089 | 26445 | 29789 | 85205 | 41001 | 12535 | 12133 | 14645 | 23541 |
| 84 | 43937 | 46891 | 24010 | 25560 | 86355 | 33941 | 25786 | 54990 | 71899 | 15475 | 95434 | 98227 | 21824 | 19585 |

**TABLE 1 Random Numbers (continued)**

| Row \ Column | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 85 | 97656 | 63175 | 89303 | 16275 | 07100 | 92063 | 21942 | 18611 | 47348 | 20203 | 18534 | 03862 | 78095 | 50136 |
| 86 | 03299 | 01221 | 05418 | 38982 | 55758 | 92237 | 26759 | 86367 | 21216 | 98442 | 08303 | 56613 | 91511 | 75928 |
| 87 | 79626 | 06486 | 03574 | 17668 | 07785 | 76020 | 79924 | 25651 | 83325 | 88428 | 85076 | 72811 | 22717 | 50585 |
| 88 | 85636 | 68335 | 47539 | 03129 | 65651 | 11977 | 02510 | 26113 | 99447 | 68645 | 34327 | 15152 | 55230 | 93448 |
| 89 | 18039 | 14367 | 61337 | 06177 | 12143 | 46609 | 32989 | 74014 | 64708 | 00533 | 35398 | 58408 | 13261 | 47908 |
| 90 | 08362 | 15656 | 60627 | 36478 | 65648 | 16764 | 53412 | 09013 | 07832 | 41574 | 17639 | 82163 | 60859 | 75567 |
| 91 | 79556 | 29068 | 04142 | 16268 | 15387 | 12856 | 66227 | 38358 | 22478 | 73373 | 88732 | 09443 | 82558 | 05250 |
| 92 | 92608 | 82674 | 27072 | 32534 | 17075 | 27698 | 98204 | 63863 | 11951 | 34648 | 88022 | 56148 | 34925 | 57031 |
| 93 | 23982 | 25835 | 40055 | 67006 | 12293 | 02753 | 14827 | 23235 | 35071 | 99704 | 37543 | 11601 | 35503 | 85171 |
| 94 | 09915 | 96306 | 05908 | 97901 | 28395 | 14186 | 00821 | 80703 | 70426 | 75647 | 76310 | 88717 | 37890 | 40129 |
| 95 | 59037 | 33300 | 26695 | 62247 | 69927 | 76123 | 50842 | 43834 | 86654 | 70959 | 79725 | 93872 | 28117 | 19233 |
| 96 | 42488 | 78077 | 69882 | 61657 | 34136 | 79180 | 97526 | 43092 | 04098 | 73571 | 80799 | 76536 | 71255 | 64239 |
| 97 | 46764 | 86273 | 63003 | 93017 | 31204 | 36692 | 40202 | 35275 | 57306 | 55543 | 53203 | 18098 | 47625 | 88684 |
| 98 | 03237 | 45430 | 55417 | 63282 | 90816 | 17349 | 88298 | 90183 | 36600 | 78406 | 06216 | 95787 | 42579 | 90730 |
| 99 | 86591 | 81482 | 52667 | 61582 | 14972 | 90053 | 89534 | 76036 | 49199 | 43716 | 97548 | 04379 | 46370 | 28672 |
| 100 | 38534 | 01715 | 94964 | 87288 | 65680 | 43772 | 39560 | 12918 | 86537 | 62738 | 19636 | 51132 | 25739 | 56947 |

*Source:* Abridged from Beyer W. H. (ed.), *CRC Standard Mathematical Tables*, 24th ed. (Cleveland: The Chemical Rubber Company), 1976. Reproduced by permission of the publisher.

**TABLE 2 Cumulative Binomial Probabilities**

*Tabulated values are* $\sum_{y=o}^{k} p(y)$.

a. $n = 5$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .9510 | .7738 | .5905 | .3277 | .1681 | .0778 | .0313 | .0102 | .0024 | .0003 | .0000 | .0000 | .0000 |
| 1 | .9990 | .9774 | .9185 | .7373 | .5282 | .3370 | .1875 | .0870 | .0308 | .0067 | .0005 | .0000 | .0000 |
| 2 | 1.0000 | .9988 | .9914 | .9421 | .8369 | .6826 | .5000 | .3174 | .1631 | .0579 | .0086 | .0012 | .0000 |
| 3 | 1.0000 | 1.0000 | .9995 | .9933 | .9692 | .9130 | .8125 | .6630 | .4718 | .2627 | .0815 | .0226 | .0010 |
| 4 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9976 | .9898 | .9687 | .9222 | .8319 | .6723 | .4095 | .2262 | .0490 |

b. $n = 6$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .9415 | .7351 | .5314 | .2621 | .1176 | .0467 | .0156 | .0041 | .0007 | .0001 | .0000 | .0000 | .0000 |
| 1 | .9985 | .9672 | .8857 | .6554 | .4202 | .2333 | .1094 | .0410 | .0109 | .0016 | .0001 | .0000 | .0000 |
| 2 | 1.0000 | .9978 | .9841 | .9011 | .7443 | .5443 | .3437 | .1792 | .0705 | .0170 | .0013 | .0001 | .0000 |
| 3 | 1.0000 | .9999 | .9987 | .9830 | .9295 | .8208 | .6562 | .4557 | .2557 | .0989 | .0158 | .0022 | .0000 |
| 4 | 1.0000 | 1.0000 | .9999 | .9984 | .9891 | .9590 | .8906 | .7667 | .5798 | .3446 | .1143 | .0328 | .0015 |
| 5 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9993 | .9959 | .9844 | .9533 | .8824 | .7379 | .4686 | .2649 | .0585 |

c. $n = 7$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .9321 | .6983 | .4783 | .2097 | .0824 | .0280 | .0078 | .0016 | .0002 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9980 | .9556 | .8503 | .5767 | .3294 | .1586 | .0625 | .0188 | .0038 | .0004 | .0000 | .0000 | .0000 |
| 2 | 1.0000 | .9962 | .9743 | .8520 | .6471 | .4199 | .2266 | .0963 | .0288 | .0047 | .0002 | .0000 | .0000 |
| 3 | 1.0000 | .9998 | .9973 | .9667 | .8740 | .7102 | .5000 | .2898 | .1260 | .0333 | .0027 | .0002 | .0000 |
| 4 | 1.0000 | 1.0000 | .9998 | .9953 | .9712 | .9037 | .7734 | .5801 | .3529 | .1480 | .0257 | .0038 | .0000 |
| 5 | 1.0000 | 1.0000 | 1.0000 | .9996 | .9962 | .9812 | .9375 | .8414 | .6706 | .4233 | .1497 | .0444 | .0020 |
| 6 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9998 | .9984 | .9922 | .9720 | .9176 | .7903 | .5217 | .3017 | .0679 |

d. $n = 8$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .9227 | .6634 | .4305 | .1678 | .0576 | .0168 | .0039 | .0007 | .0001 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9973 | .9423 | .8131 | .5033 | .2553 | .1064 | .0352 | .0085 | .0013 | .0001 | .0000 | .0000 | .0000 |
| 2 | .9999 | .9942 | .9619 | .7969 | .5518 | .3154 | .1445 | .0498 | .0113 | .0012 | .0000 | .0000 | .0000 |
| 3 | 1.0000 | .9996 | .9950 | .9437 | .8059 | .5941 | .3633 | .1737 | .0580 | .0104 | .0004 | .0000 | .0000 |
| 4 | 1.0000 | 1.0000 | .9996 | .9896 | .9420 | .8263 | .6367 | .4059 | .1941 | .0563 | .0050 | .0004 | .0000 |
| 5 | 1.0000 | 1.0000 | 1.0000 | .9988 | .9887 | .9502 | .8555 | .6346 | .4482 | .2031 | .0381 | .0058 | .0001 |
| 6 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9987 | .9915 | .9648 | .8936 | .7447 | .4967 | .1869 | .0572 | .0027 |
| 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9993 | .9961 | .9832 | .9424 | .8322 | .5695 | .3366 | .0773 |

**TABLE 2 Cumulative Binomial Probabilities (continued)**

e. $n = 9$

| | | | | | | | $p$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | .01 | .05 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .95 | .99 |
| 0 | .9135 | .6302 | .3874 | .1342 | .0404 | .0101 | .0020 | .0003 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9966 | .9288 | .7748 | .4362 | .1960 | .0705 | .0195 | .0038 | .0004 | .0000 | .0000 | .0000 | .0000 |
| 2 | .9999 | .9916 | .9470 | .7382 | .4623 | .2318 | .0898 | .0250 | .0043 | .0003 | .0000 | .0000 | .0000 |
| 3 | 1.0000 | .9994 | .9917 | .9144 | .7297 | .4826 | .2539 | .0994 | .0253 | .0031 | .0001 | .0000 | .0000 |
| 4 | 1.0000 | 1.0000 | .9991 | .9804 | .9012 | .7334 | .5000 | .2666 | .0988 | .0196 | .0009 | .0000 | .0000 |
| 5 | 1.0000 | 1.0000 | .9999 | .9969 | .9747 | .9006 | .7461 | .5174 | .2703 | .0856 | .0083 | .0006 | .0000 |
| 6 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9957 | .9750 | .9102 | .7682 | .5372 | .2618 | .0530 | .0084 | .0001 |
| 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9996 | .9962 | .9805 | .9295 | .8040 | .5638 | .2252 | .0712 | .0034 |
| 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9980 | .9899 | .9596 | .8658 | .6126 | .3698 | .0865 |

f. $n = 10$

| | | | | | | | $p$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | .01 | .05 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .95 | .99 |
| 0 | .9044 | .5987 | .3487 | .1074 | .0282 | .0060 | .0010 | .0001 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9957 | .9139 | .7361 | .3758 | .1493 | .0464 | .0107 | .0017 | .0001 | .0000 | .0000 | .0000 | .0000 |
| 2 | .9999 | .9885 | .9298 | .6778 | .3828 | .1673 | .0547 | .0123 | .0016 | .0001 | .0000 | .0000 | .0000 |
| 3 | 1.0000 | .9990 | .9872 | .8791 | .6496 | .3823 | .1719 | .0548 | .0106 | .0009 | .0000 | .0000 | .0000 |
| 4 | 1.0000 | .9999 | .9984 | .9672 | .8497 | .6331 | .3770 | .1662 | .0473 | .0064 | .0001 | .0000 | .0000 |
| 5 | 1.0000 | 1.0000 | .9999 | .9936 | .9527 | .8338 | .6230 | .3669 | .1503 | .0328 | .0016 | .0001 | .0000 |
| 6 | 1.0000 | 1.0000 | 1.0000 | .9991 | .9894 | .9452 | .8281 | .6177 | .3504 | .1209 | .0128 | .0010 | .0000 |
| 7 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9984 | .9877 | .9453 | .8327 | .6172 | .3222 | .0702 | .0115 | .0001 |
| 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9983 | .9893 | .9536 | .8507 | .6242 | .2639 | .0861 | .0043 |
| 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9990 | .9940 | .9718 | .8926 | .6513 | .4013 | .0956 |

**TABLE 2 Cumulative Binomial Probabilities (continued)**

g. $n = 15$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .8601 | .4633 | .2059 | .0352 | .0047 | .0005 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9904 | .8290 | .5490 | .1671 | .0353 | .0052 | .0005 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 2 | .9996 | .9638 | .8159 | .3980 | .1268 | .0271 | .0037 | .0003 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 3 | 1.0000 | .9945 | .9444 | .6482 | .2969 | .0905 | .0176 | .0019 | .0001 | .0000 | .0000 | .0000 | .0000 |
| 4 | 1.0000 | .9994 | .9873 | .8358 | .5155 | .2173 | .0592 | .0093 | .0007 | .0000 | .0000 | .0000 | .0000 |
| 5 | 1.0000 | .9999 | .9978 | .9389 | .7216 | .4032 | .1509 | .0338 | .0037 | .0001 | .0000 | .0000 | .0000 |
| 6 | 1.0000 | 1.0000 | .9997 | .9819 | .8689 | .6098 | .3036 | .0950 | .0152 | .0008 | .0000 | .0000 | .0000 |
| 7 | 1.0000 | 1.0000 | 1.0000 | .9958 | .9500 | .7869 | .5000 | .2131 | .0500 | .0042 | .0000 | .0000 | .0000 |
| 8 | 1.0000 | 1.0000 | 1.0000 | .9992 | .9848 | .9050 | .6964 | .3902 | .1311 | .0181 | .0003 | .0000 | .0000 |
| 9 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9963 | .9662 | .8491 | .5968 | .2784 | .0611 | .0022 | .0001 | .0000 |
| 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9993 | .9907 | .9408 | .7827 | .4845 | .1642 | .0127 | .0006 | .0000 |
| 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9981 | .9824 | .9095 | .7031 | .3518 | .0556 | .0055 | .0000 |
| 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9963 | .9729 | .8732 | .6020 | .1841 | .0362 | .0004 |
| 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9995 | .9948 | .9647 | .8329 | .4510 | .1710 | .0096 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9995 | .9953 | .9648 | .7941 | .5367 | .1399 |

h. $n = 20$

| k | .01 | .05 | .1 | .2 | .3 | .4 | p .5 | .6 | .7 | .8 | .9 | .95 | .99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .8179 | .3585 | .1216 | .0115 | .0008 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9831 | .7358 | .3917 | .0692 | .0076 | .0005 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 2 | .9990 | .9245 | .6769 | .2061 | .0355 | .0036 | .0002 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 3 | 1.0000 | .9841 | .8670 | .4114 | .1071 | .0160 | .0013 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 4 | 1.0000 | .9974 | .9568 | .6296 | .2375 | .0510 | .0059 | .0003 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 5 | 1.0000 | .9997 | .9887 | .8042 | .4164 | .1256 | .0207 | .0016 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 6 | 1.0000 | 1.0000 | .9976 | .9133 | .6080 | .2500 | .0577 | .0065 | .0003 | .0000 | .0000 | .0000 | .0000 |
| 7 | 1.0000 | 1.0000 | .9996 | .9679 | .7723 | .4159 | .1316 | .0210 | .0013 | .0000 | .0000 | .0000 | .0000 |
| 8 | 1.0000 | 1.0000 | .9999 | .9900 | .8867 | .5956 | .2517 | .0565 | .0051 | .0001 | .0000 | .0000 | .0000 |
| 9 | 1.0000 | 1.0000 | 1.0000 | .9974 | .9520 | .7553 | .4119 | .1275 | .0171 | .0006 | .0000 | .0000 | .0000 |
| 10 | 1.0000 | 1.0000 | 1.0000 | .9994 | .9829 | .8725 | .5881 | .2447 | .0480 | .0026 | .0000 | .0000 | .0000 |
| 11 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9949 | .9435 | .7483 | .4044 | .1133 | .0100 | .0001 | .0000 | .0000 |
| 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9987 | .9790 | .8684 | .5841 | .2277 | .0321 | .0004 | .0000 | .0000 |
| 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9935 | .9423 | .7500 | .3920 | .0867 | .0024 | .0000 | .0000 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9984 | .9793 | .8744 | .5836 | .1958 | .0113 | .0003 | .0000 |
| 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9941 | .9490 | .7625 | .3704 | .0432 | .0026 | .0000 |
| 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9987 | .9840 | .8929 | .5886 | .1330 | .0159 | .0000 |
| 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9998 | .9964 | .9645 | .7939 | .3231 | .0755 | .0010 |
| 18 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9995 | .9924 | .9308 | .6083 | .2642 | .0169 |
| 19 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9992 | .9885 | .8784 | .6415 | .1821 |

**TABLE 2 Cumulative Binomial Probabilities (continued)**

i. $n = 25$

| | | | | | | | *p* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *k* | .01 | .05 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .95 | .99 |
| 0 | .7778 | .2774 | .0718 | .0038 | .0001 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 1 | .9742 | .6424 | .2712 | .0274 | .0016 | .0001 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 2 | .9980 | .8729 | .5371 | .0982 | .0090 | .0004 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 3 | .9999 | .9659 | .7636 | .2340 | .0332 | .0024 | .0001 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 4 | 1.0000 | .9928 | .9020 | .4207 | .0905 | .0095 | .0005 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 5 | 1.0000 | .9988 | .9666 | .6167 | .1935 | .0294 | .0020 | .0001 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 6 | 1.0000 | .9998 | .9905 | .7800 | .3407 | .0736 | .0073 | .0003 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 7 | 1.0000 | 1.0000 | .9977 | .8909 | .5118 | .1536 | .0216 | .0012 | .0000 | .0000 | .0000 | .0000 | .0000 |
| 8 | 1.0000 | 1.0000 | .9995 | .9532 | .6769 | .2735 | .0539 | .0043 | .0001 | .0000 | .0000 | .0000 | .0000 |
| 9 | 1.0000 | 1.0000 | .9999 | .9827 | .8106 | .4246 | .1148 | .0132 | .0005 | .0000 | .0000 | .0000 | .0000 |
| 10 | 1.0000 | 1.0000 | 1.0000 | .9944 | .9022 | .5858 | .2122 | .0344 | .0018 | .0000 | .0000 | .0000 | .0000 |
| 11 | 1.0000 | 1.0000 | 1.0000 | .9985 | .9558 | .7323 | .3450 | .0778 | .0060 | .0001 | .0000 | .0000 | .0000 |
| 12 | 1.0000 | 1.0000 | 1.0000 | .9996 | .9825 | .8462 | .5000 | .1538 | .0175 | .0004 | .0000 | .0000 | .0000 |
| 13 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9940 | .9222 | .6550 | .2677 | .0442 | .0015 | .0000 | .0000 | .0000 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9982 | .9656 | .7878 | .4142 | .0978 | .0056 | .0000 | .0000 | .0000 |
| 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9995 | .9868 | .8852 | .5754 | .1894 | .0173 | .0001 | .0000 | .0000 |
| 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9957 | .9461 | .7265 | .3231 | .0468 | .0005 | .0000 | .0000 |
| 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9988 | .9784 | .8464 | .4882 | .1091 | .0023 | .0000 | .0000 |
| 18 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9927 | .9264 | .6593 | .2200 | .0095 | .0002 | .0000 |
| 19 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9980 | .9706 | .8065 | .3833 | .0334 | .0012 | .0000 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9995 | .9905 | .9095 | .5793 | .0980 | .0072 | .0000 |
| 21 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9976 | .9668 | .7660 | .2364 | .0341 | .0001 |
| 22 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9996 | .9910 | .9018 | .4629 | .1271 | .0020 |
| 23 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9984 | .9726 | .7288 | .3576 | .0258 |
| 24 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9962 | .9282 | .7226 | .2222 |

**TABLE 3 Exponentials**

| λ | $e^{-\lambda}$ | λ | $e^{-\lambda}$ | λ | $e^{-\lambda}$ | λ | $e^{-\lambda}$ | λ | $e^{-\lambda}$ |
|---|---|---|---|---|---|---|---|---|---|
| .00 | 1.000000 | 2.05 | .128735 | 4.05 | .017422 | 6.05 | .002358 | 8.05 | .000319 |
| .05 | .951229 | 2.10 | .122456 | 4.10 | .016573 | 6.10 | .002243 | 8.10 | .000304 |
| .10 | .904837 | 2.15 | .116484 | 4.15 | .015764 | 6.15 | .002133 | 8.15 | .000289 |
| .15 | .860708 | 2.20 | .110803 | 4.20 | .014996 | 6.20 | .002029 | 8.20 | .000275 |
| .20 | .818731 | 2.25 | .105399 | 4.25 | .014264 | 6.25 | .001930 | 8.25 | .000261 |
| .25 | .778801 | 2.30 | .100259 | 4.30 | .013569 | 6.30 | .001836 | 8.30 | .000249 |
| .30 | .740818 | 2.35 | .095369 | 4.35 | .012907 | 6.35 | .001747 | 8.35 | .000236 |
| .35 | .704688 | 2.40 | .090718 | 4.40 | .012277 | 6.40 | .001661 | 8.40 | .000225 |
| .40 | .670320 | 2.45 | .086294 | 4.45 | .011679 | 6.45 | .001581 | 8.45 | .000214 |
| .45 | .637628 | 2.50 | .082085 | 4.50 | .011109 | 6.50 | .001503 | 8.50 | .000204 |
| .50 | .606531 | 2.55 | .078082 | 4.55 | .010567 | 6.55 | .001430 | 8.55 | .000194 |
| .55 | .576950 | 2.60 | .074274 | 4.60 | .010052 | 6.60 | .001360 | 8.60 | .000184 |
| .60 | .548812 | 2.65 | .070651 | 4.65 | .009562 | 6.65 | .001294 | 8.65 | .000175 |
| .65 | .522046 | 2.70 | .067206 | 4.70 | .009095 | 6.70 | .001231 | 8.70 | .000167 |
| .70 | .496585 | 2.75 | .063928 | 4.75 | .008652 | 6.75 | .001171 | 8.75 | .000158 |
| .75 | .472367 | 2.80 | .060810 | 4.80 | .008230 | 6.80 | .001114 | 8.80 | .000151 |
| .80 | .449329 | 2.85 | .057844 | 4.85 | .007828 | 6.85 | .001059 | 8.85 | .000143 |
| .85 | .427415 | 2.90 | .055023 | 4.90 | .007447 | 6.90 | .001008 | 8.90 | .000136 |
| .90 | .406570 | 2.95 | .052340 | 4.95 | .007083 | 6.95 | .000959 | 8.95 | .000130 |
| .95 | .386741 | 3.00 | .049787 | 5.00 | .006738 | 7.00 | .000912 | 9.00 | .000123 |
| 1.00 | .367879 | 3.05 | .047359 | 5.05 | .006409 | 7.05 | .000867 | 9.05 | .000117 |
| 1.05 | .349938 | 3.10 | .045049 | 5.10 | .006097 | 7.10 | .000825 | 9.10 | .000112 |
| 1.10 | .332871 | 3.15 | .042852 | 5.15 | .005799 | 7.15 | .000785 | 9.15 | .000106 |
| 1.15 | .316637 | 3.20 | .040762 | 5.20 | .005517 | 7.20 | .000747 | 9.20 | .000101 |
| 1.20 | .301194 | 3.25 | .038774 | 5.25 | .005248 | 7.25 | .000710 | 9.25 | .000096 |
| 1.25 | .286505 | 3.30 | .036883 | 5.30 | .004992 | 7.30 | .000676 | 9.30 | .000091 |
| 1.30 | .272532 | 3.35 | .035084 | 5.35 | .004748 | 7.35 | .000643 | 9.35 | .000087 |
| 1.35 | .259240 | 3.40 | .033373 | 5.40 | .004517 | 7.40 | .000611 | 9.40 | .000083 |
| 1.40 | .246597 | 3.45 | .031746 | 5.45 | .004296 | 7.45 | .000581 | 9.45 | .000079 |
| 1.45 | .234570 | 3.50 | .030197 | 5.50 | .004087 | 7.50 | .000553 | 9.50 | .000075 |
| 1.50 | .223130 | 3.55 | .028725 | 5.55 | .003887 | 7.55 | .000526 | 9.55 | .000071 |
| 1.55 | .212248 | 3.60 | .027324 | 5.60 | .003698 | 7.60 | .000501 | 9.60 | .000068 |
| 1.60 | .201897 | 3.65 | .025991 | 5.65 | .003518 | 7.65 | .000476 | 9.65 | .000064 |
| 1.65 | .192050 | 3.70 | .024724 | 5.70 | .003346 | 7.70 | .000453 | 9.70 | .000061 |
| 1.70 | .182684 | 3.75 | .023518 | 5.75 | .003183 | 7.75 | .000431 | 9.75 | .000058 |
| 1.75 | .173774 | 3.80 | .022371 | 5.80 | .003028 | 7.80 | .000410 | 9.80 | .000056 |
| 1.80 | .165299 | 3.85 | .021280 | 5.85 | .002880 | 7.85 | .000390 | 9.85 | .000053 |
| 1.85 | .157237 | 3.90 | .020242 | 5.90 | .002739 | 7.90 | .000371 | 9.90 | .000050 |
| 1.90 | .149569 | 3.95 | .019255 | 5.95 | .002606 | 7.95 | .000353 | 9.95 | .000048 |
| 1.95 | .142274 | 4.00 | .018316 | 6.00 | .002479 | 8.00 | .000336 | 10.00 | .000045 |
| 2.00 | .135335 | | | | | | | | |

## TABLE 4 Cumulative Poisson Probabilities

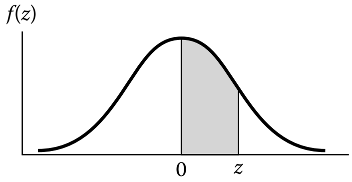
*Tabulated values are* $\sum_{y=0}^{k} p(y)$

| | Poisson Mean $\mu$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | .5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 |
| 0 | .6065 | .3679 | .2231 | .1353 | .0821 | .0498 | .0302 | .0183 | .0111 | .0067 |
| 1 | .9098 | .7358 | .5578 | .4060 | .2873 | .1991 | .1359 | .0916 | .0611 | .0404 |
| 2 | .9856 | .9197 | .8088 | .6767 | .5438 | .4232 | .3208 | .2381 | .1736 | .1247 |
| 3 | .9982 | .9810 | .9344 | .8571 | .7576 | .6472 | .5366 | .4335 | .3423 | .2650 |
| 4 | .9998 | .9963 | .9814 | .9473 | .8912 | .8153 | .7254 | .6288 | .5321 | .4405 |
| 5 | 1.0000 | .9994 | .9955 | .9834 | .9580 | .9161 | .8576 | .7851 | .7029 | .6160 |
| 6 | 1.0000 | .9999 | .9991 | .9955 | .9858 | .9665 | .9347 | .8893 | .8311 | .7622 |
| 7 | 1.0000 | 1.0000 | .9998 | .9989 | .9958 | .9881 | .9733 | .9489 | .9134 | .8666 |
| 8 | 1.0000 | 1.0000 | 1.0000 | .9998 | .9989 | .9962 | .9901 | .9786 | .9597 | .9319 |
| 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9997 | .9989 | .9967 | .9919 | .9829 | .9682 |
| 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9997 | .9990 | .9972 | .9933 | .9863 |
| 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9997 | .9991 | .9976 | .9945 |
| 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9997 | .9992 | .9980 |
| 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9997 | .9993 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9998 |
| 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 |
| 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 18 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 19 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**TABLE 4 Cumulative Poisson Probabilities (continued)**

| | Poisson Mean $\mu$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 |
| 0 | .0041 | .0025 | .0015 | .0009 | .0006 | .0003 | .0002 | .0001 | .0001 | .0000 |
| 1 | .0266 | .0174 | .0113 | .0073 | .0047 | .0030 | .0019 | .0012 | .0008 | .0005 |
| 2 | .0884 | .0620 | .0430 | .0296 | .0203 | .0138 | .0093 | .0062 | .0042 | .0028 |
| 3 | .2017 | .1512 | .1118 | .0818 | .0591 | .0424 | .0301 | .0212 | .0149 | .0103 |
| 4 | .3575 | .2851 | .2237 | .1730 | .1321 | .0996 | .0744 | .0550 | .0403 | .0293 |
| 5 | .5289 | .4457 | .3690 | .3007 | .2414 | .1912 | .1496 | .1157 | .0885 | .0671 |
| 6 | .6860 | .6063 | .5265 | .4497 | .3782 | .3134 | .2562 | .2068 | .1649 | .1301 |
| 7 | .8095 | .7440 | .6728 | .5987 | .5246 | .4530 | .3856 | .3239 | .2687 | .2202 |
| 8 | .8944 | .8472 | .7916 | .7291 | .6620 | .5925 | .5231 | .4557 | .3918 | .3328 |
| 9 | .9462 | .9161 | .8774 | .8305 | .7764 | .7166 | .6530 | .5874 | .5218 | .4579 |
| 10 | .9747 | .9574 | .9332 | .9015 | .8622 | .8159 | .7634 | .7060 | .6453 | .5830 |
| 11 | .9890 | .9799 | .9661 | .9467 | .9208 | .8881 | .8487 | .8030 | .7520 | .6968 |
| 12 | .9955 | .9912 | .9840 | .9730 | .9573 | .9362 | .9091 | .8758 | .8364 | .7916 |
| 13 | .9983 | .9964 | .9929 | .9872 | .9784 | .9658 | .9486 | .9261 | .8981 | .8645 |
| 14 | .9994 | .9986 | .9970 | .9943 | .9897 | .9827 | .9726 | .9585 | .9400 | .9165 |
| 15 | .9998 | .9995 | .9988 | .9976 | .9954 | .9918 | .9862 | .9780 | .9665 | .9513 |
| 16 | .9999 | .9998 | .9996 | .9990 | .9980 | .9963 | .9934 | .9889 | .9823 | .9730 |
| 17 | 1.0000 | .9999 | .9998 | .9996 | .9992 | .9984 | .9970 | .9947 | .9911 | .9857 |
| 18 | 1.0000 | 1.0000 | .9999 | .9999 | .9997 | .9993 | .9987 | .9976 | .9957 | .9928 |
| 19 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9997 | .9995 | .9989 | .9980 | .9965 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | .9999 | .9998 | .9996 | .9991 | .9984 |

## TABLE 5 Normal Curve Areas




| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .0 | .0000 | .0040 | .0080 | .0120 | .0160 | .0199 | .0239 | .0279 | .0319 | .0359 |
| .1 | .0398 | .0438 | .0478 | .0517 | .0557 | .0596 | .0636 | .0675 | .0714 | .0753 |
| .2 | .0793 | .0832 | .0871 | .0910 | .0948 | .0987 | .1026 | .1064 | .1103 | .1141 |
| .3 | .1179 | .1217 | .1255 | .1293 | .1331 | .1368 | .1406 | .1443 | .1480 | .1517 |
| .4 | .1554 | .1591 | .1628 | .1664 | .1700 | .1736 | .1772 | .1808 | .1844 | .1879 |
| .5 | .1915 | .1950 | .1985 | .2019 | .2054 | .2088 | .2123 | .2157 | .2190 | .2224 |
| .6 | .2257 | .2291 | .2324 | .2357 | .2389 | .2422 | .2454 | .2486 | .2517 | .2549 |
| .7 | .2580 | .2611 | .2642 | .2673 | .2704 | .2734 | .2764 | .2794 | .2823 | .2852 |
| .8 | .2881 | .2910 | .2939 | .2967 | .2995 | .3023 | .3051 | .3078 | .3106 | .3133 |
| .9 | .3159 | .3186 | .3212 | .3238 | .3264 | .3289 | .3315 | .3340 | .3365 | .3389 |
| 1.0 | .3413 | .3438 | .3461 | .3485 | .3508 | .3531 | .3554 | .3577 | .3599 | .3621 |
| 1.1 | .3643 | .3665 | .3686 | .3708 | .3729 | .3749 | .3770 | .3790 | .3810 | .3830 |
| 1.2 | .3849 | .3869 | .3888 | .3907 | .3925 | .3944 | .3962 | .3980 | .3997 | .4015 |
| 1.3 | .4032 | .4049 | .4066 | .4082 | .4099 | .4115 | .4131 | .4147 | .4162 | .4177 |
| 1.4 | .4192 | .4207 | .4222 | .4236 | .4251 | .4265 | .4279 | .4292 | .4306 | .4319 |
| 1.5 | .4332 | .4345 | .4357 | .4370 | .4382 | .4394 | .4406 | .4418 | .4429 | .4441 |
| 1.6 | .4452 | .4463 | .4474 | .4484 | .4495 | .4505 | .4515 | .4525 | .4535 | .4545 |
| 1.7 | .4554 | .4564 | .4573 | .4582 | .4591 | .4599 | .4608 | .4616 | .4625 | .4633 |
| 1.8 | .4641 | .4649 | .4656 | .4664 | .4671 | .4678 | .4686 | .4693 | .4699 | .4706 |
| 1.9 | .4713 | .4719 | .4726 | .4732 | .4738 | .4744 | .4750 | .4756 | .4761 | .4767 |
| 2.0 | .4772 | .4778 | .4783 | .4788 | .4793 | .4798 | .4803 | .4808 | .4812 | .4817 |
| 2.1 | .4821 | .4826 | .4830 | .4834 | .4838 | .4842 | .4846 | .4850 | .4854 | .4857 |
| 2.2 | .4861 | .4864 | .4868 | .4871 | .4875 | .4878 | .4881 | .4884 | .4887 | .4890 |
| 2.3 | .4893 | .4896 | .4898 | .4901 | .4904 | .4906 | .4909 | .4911 | .4913 | .4916 |
| 2.4 | .4918 | .4920 | .4922 | .4925 | .4927 | .4929 | .4931 | .4932 | .4934 | .4936 |
| 2.5 | .4938 | .4940 | .4941 | .4943 | .4945 | .4946 | .4948 | .4949 | .4951 | .4952 |
| 2.6 | .4953 | .4955 | .4956 | .4957 | .4959 | .4960 | .4961 | .4962 | .4963 | .4964 |
| 2.7 | .4965 | .4966 | .4967 | .4968 | .4969 | .4970 | .4971 | .4972 | .4973 | .4974 |
| 2.8 | .4974 | .4975 | .4976 | .4977 | .4977 | .4978 | .4979 | .4979 | .4980 | .4981 |
| 2.9 | .4981 | .4982 | .4982 | .4983 | .4984 | .4984 | .4985 | .4985 | .4986 | .4986 |
| 3.0 | .4987 | .4987 | .4987 | .4988 | .4988 | .4989 | .4989 | .4989 | .4990 | .4990 |

*Source:* Abridged from Table 1 of Hald, A. *Statistical Tables and Formulas* (New York: Wiley), 1952, Reproduced by permission of A. Hald and the publisher. John Wiley & Sons, Inc.

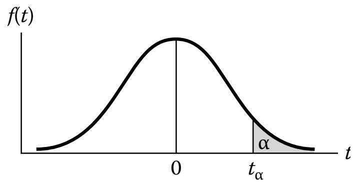
**TABLE 6 Gamma Function**

*Value of* $\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx;\ \Gamma(n+1) = n\Gamma(n)$

| $n$ | $\Gamma(n)$ | $n$ | $\Gamma(n)$ | $n$ | $\Gamma(n)$ | $n$ | $\Gamma(n)$ |
|---|---|---|---|---|---|---|---|
| 1.00 | 1.00000 | 1.25 | .90640 | 1.50 | .88623 | 1.75 | .91906 |
| 1.01 | .99433 | 1.26 | .90440 | 1.51 | .88659 | 1.76 | .92137 |
| 1.02 | .98884 | 1.27 | .90250 | 1.52 | .88704 | 1.77 | .92376 |
| 1.03 | .98355 | 1.28 | .90072 | 1.53 | .88757 | 1.78 | .92623 |
| 1.04 | .97844 | 1.29 | .89904 | 1.54 | .88818 | 1.79 | .92877 |
| 1.05 | .97350 | 1.30 | .89747 | 1.55 | .88887 | 1.80 | .93138 |
| 1.06 | .96874 | 1.31 | .89600 | 1.56 | .88964 | 1.81 | .93408 |
| 1.07 | .96415 | 1.32 | .89464 | 1.57 | .89049 | 1.82 | .93685 |
| 1.08 | .95973 | 1.33 | .89338 | 1.58 | .89142 | 1.83 | .93969 |
| 1.09 | .95546 | 1.34 | .89222 | 1.59 | .89243 | 1.84 | .94261 |
| 1.10 | .95135 | 1.35 | .89115 | 1.60 | .89352 | 1.85 | .94561 |
| 1.11 | .94739 | 1.36 | .89018 | 1.61 | .89468 | 1.86 | .94869 |
| 1.12 | .94359 | 1.37 | .88931 | 1.62 | .89592 | 1.87 | .95184 |
| 1.13 | .93993 | 1.38 | .88854 | 1.63 | .89724 | 1.88 | .95507 |
| 1.14 | .93642 | 1.39 | .88785 | 1.64 | .89864 | 1.89 | .95838 |
| 1.15 | .93304 | 1.40 | .88726 | 1.65 | .90012 | 1.90 | .96177 |
| 1.16 | .92980 | 1.41 | .88676 | 1.66 | .90167 | 1.91 | .96523 |
| 1.17 | .92670 | 1.42 | .88636 | 1.67 | .90330 | 1.92 | .96878 |
| 1.18 | .92373 | 1.43 | .88604 | 1.68 | .90500 | 1.93 | .97240 |
| 1.19 | .92088 | 1.44 | .88580 | 1.69 | .90678 | 1.94 | .97610 |
| 1.20 | .91817 | 1.45 | .88565 | 1.70 | .90864 | 1.95 | .97988 |
| 1.21 | .91558 | 1.46 | .88560 | 1.71 | .91057 | 1.96 | .98374 |
| 1.22 | .91311 | 1.47 | .88563 | 1.72 | .91258 | 1.97 | .98768 |
| 1.23 | .91075 | 1.48 | .88575 | 1.73 | .91466 | 1.98 | .99171 |
| 1.24 | .90852 | 1.49 | .88595 | 1.74 | .91683 | 1.99 | .99581 |
| | | | | | | 2.00 | 1.00000 |

*Source:* Abridged from Beyer, W. H. (ed.) *Handbook of Tables for Probability and Statistics*, 1966. Reproduced by permission of the publisher, The Chemical Rubber Company.

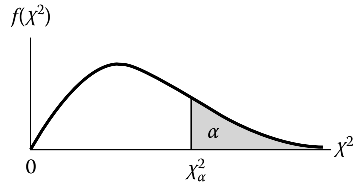
**TABLE 7 Critical Values for Student's *T***




| $\nu$ | $t_{.100}$ | $t_{.050}$ | $t_{.025}$ | $t_{.010}$ | $t_{.005}$ | $t_{.001}$ | $t_{.0005}$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.31 | 636.62 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.326 | 31.598 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.213 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.767 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.160 | 3.373 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |

*Source:* This table is reproduced with the kind permission of the Trustees of Biometrika from Pearson, E. S., and Hartley, H. O. (eds.) *The Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., *Biometrika*, 1966.

**TABLE 8 Critical Values of $\chi^2$**



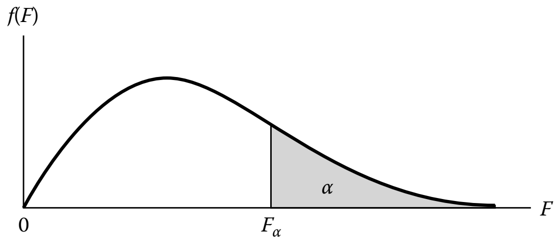

| Degrees of Freedom | $\chi^2_{.995}$ | $\chi^2_{.990}$ | $\chi^2_{.975}$ | $\chi^2_{.950}$ | $\chi^2_{.900}$ |
|---|---|---|---|---|---|
| 1 | .0000393 | .0001571 | .0009821 | .0039321 | .0157908 |
| 2 | .0100251 | .0201007 | .0506356 | .102587 | .210720 |
| 3 | .0717212 | .114832 | .215795 | .351846 | .584375 |
| 4 | .206990 | .297110 | .484419 | .710721 | 1.063623 |
| 5 | .411740 | .554300 | .831211 | 1.145476 | 1.61031 |
| 6 | .675727 | 0.872085 | 1.237347 | 1.63539 | 2.20413 |
| 7 | .989265 | 1.239043 | 1.68987 | 2.16735 | 2.83311 |
| 8 | 1.344419 | 1.646482 | 2.17973 | 2.73264 | 3.48954 |
| 9 | 1.734926 | 2.087912 | 2.70039 | 3.32511 | 4.16816 |
| 10 | 2.15585 | 2.55821 | 3.24697 | 3.94030 | 4.86518 |
| 11 | 2.60321 | 3.05347 | 3.81575 | 4.57481 | 5.57779 |
| 12 | 3.07382 | 3.57056 | 4.40379 | 5.22603 | 6.30380 |
| 13 | 3.56503 | 4.10691 | 5.00874 | 5.89186 | 7.04150 |
| 14 | 4.07468 | 4.66043 | 5.62872 | 6.57063 | 7.78953 |
| 15 | 4.60094 | 5.22935 | 6.26214 | 7.26094 | 8.54675 |
| 16 | 5.14224 | 5.81221 | 6.90766 | 7.96164 | 9.31223 |
| 17 | 5.69724 | 6.40776 | 7.56418 | 8.67176 | 10.0852 |
| 18 | 6.26481 | 7.01491 | 8.23075 | 9.39046 | 10.8649 |
| 19 | 6.84398 | 7.63273 | 8.90655 | 10.1170 | 11.6509 |
| 20 | 7.43386 | 8.26040 | 9.59083 | 10.8508 | 12.4426 |
| 21 | 8.03366 | 8.89720 | 10.28293 | 11.5913 | 13.2396 |
| 22 | 8.64272 | 9.54249 | 10.9823 | 12.3380 | 14.0415 |
| 23 | 9.26042 | 10.19567 | 11.6885 | 13.0905 | 14.8479 |
| 24 | 9.88623 | 10.8564 | 12.4011 | 13.8484 | 15.6587 |
| 25 | 10.5197 | 11.5240 | 13.1197 | 14.6114 | 16.4734 |
| 26 | 11.1603 | 12.1981 | 13.8439 | 15.3791 | 17.2919 |
| 27 | 11.8076 | 12.8786 | 14.5733 | 16.1513 | 18.1138 |
| 28 | 12.4613 | 13.5648 | 15.3079 | 16.9279 | 18.9392 |
| 29 | 13.1211 | 14.2565 | 16.0471 | 17.7083 | 19.7677 |
| 30 | 13.7867 | 14.9535 | 16.7908 | 18.4926 | 20.5992 |
| 40 | 20.7065 | 22.1643 | 24.4331 | 26.5093 | 29.0505 |
| 50 | 27.9907 | 29.7067 | 32.3574 | 34.7642 | 37.6886 |
| 60 | 35.5346 | 37.4848 | 40.4817 | 43.1879 | 46.4589 |
| 70 | 43.2752 | 45.4418 | 48.7576 | 51.7393 | 55.3290 |
| 80 | 51.1720 | 53.5400 | 57.1532 | 60.3915 | 64.2778 |
| 90 | 59.1963 | 61.7541 | 65.6466 | 69.1260 | 73.2912 |
| 100 | 67.3276 | 70.0648 | 74.2219 | 77.9295 | 82.3581 |

**TABLE 8 Critical Values of $\chi^2$ (continued)**

| Degrees of Freedom | $\chi^2_{.100}$ | $\chi^2_{.050}$ | $\chi^2_{.025}$ | $\chi^2_{.010}$ | $\chi^2_{.005}$ |
|---|---|---|---|---|---|
| 1 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| 2 | 4.60517 | 5.99147 | 7.37776 | 9.21034 | 10.5966 |
| 3 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8381 |
| 4 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8602 |
| 5 | 9.23635 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 7 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 13.3616 | 15.5073 | 17.5346 | 20.0902 | 21.9550 |
| 9 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 10 | 15.9871 | 18.3070 | 20.4831 | 23.2093 | 25.1882 |
| 11 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7569 |
| 12 | 18.5494 | 21.0261 | 23.3367 | 26.2170 | 28.2995 |
| 13 | 19.8119 | 22.3621 | 24.7356 | 27.6883 | 29.8194 |
| 14 | 21.0642 | 23.6848 | 26.1190 | 29.1413 | 31.3193 |
| 15 | 22.3072 | 24.9958 | 27.4884 | 30.5779 | 32.8013 |
| 16 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 |
| 17 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 |
| 18 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1564 |
| 19 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5822 |
| 20 | 28.4120 | 31.4104 | 34.1696 | 37.5662 | 39.9968 |
| 21 | 29.6151 | 32.6705 | 35.4789 | 38.9321 | 41.4010 |
| 22 | 30.8133 | 33.9244 | 36.7807 | 40.2894 | 42.7956 |
| 23 | 32.0069 | 35.1725 | 38.0757 | 41.6384 | 44.1813 |
| 24 | 33.1963 | 36.4151 | 39.3641 | 42.9798 | 45.5585 |
| 25 | 34.3816 | 37.6525 | 40.6465 | 44.3141 | 46.9278 |
| 26 | 35.5631 | 38.8852 | 41.9232 | 45.6417 | 48.2899 |
| 27 | 36.7412 | 40.1133 | 43.1944 | 46.9630 | 49.6449 |
| 28 | 37.9159 | 41.3372 | 44.4607 | 48.2782 | 50.9933 |
| 29 | 39.0875 | 42.5569 | 45.7222 | 49.5879 | 52.3356 |
| 30 | 40.2560 | 43.7729 | 46.9792 | 50.8922 | 53.6720 |
| 40 | 51.8050 | 55.7585 | 59.3417 | 63.6907 | 66.7659 |
| 50 | 63.1671 | 67.5048 | 71.4202 | 76.1539 | 79.4900 |
| 60 | 74.3970 | 79.0819 | 83.2976 | 88.3794 | 91.9517 |
| 70 | 85.5271 | 90.5312 | 95.0231 | 100.425 | 104.215 |
| 80 | 96.5782 | 101.879 | 106.629 | 112.329 | 116.321 |
| 90 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 100 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

*Source:* From Thompson, C. M. "Tables of the percentage points of the $\chi^2$-distribution." *Biometrika*, 1941, Vol. 32, pp. 188–189. Reproduced by permisson of the *Biometrika* Trustees.

**TABLE 9 Percentage Points of the *F* Distribution, $\alpha = .10$**




| $\nu_2$ \ $\nu_1$ | Numerator Degrees of Freedom 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 | 58.20 | 58.91 | 59.44 | 59.86 |
| 2 | 8.53 | 9.00 | 9.16 | 9.24 | 9.29 | 9.33 | 9.35 | 9.37 | 9.38 |
| 3 | 5.54 | 5.46 | 5.39 | 5.34 | 5.31 | 5.28 | 5.27 | 5.25 | 5.24 |
| 4 | 4.54 | 4.32 | 4.19 | 4.11 | 4.05 | 4.01 | 3.98 | 3.95 | 3.94 |
| 5 | 4.06 | 3.78 | 3.62 | 3.52 | 3.45 | 3.40 | 3.37 | 3.34 | 3.32 |
| 6 | 3.78 | 3.46 | 3.29 | 3.18 | 3.11 | 3.05 | 3.01 | 2.98 | 2.96 |
| 7 | 3.59 | 3.26 | 3.07 | 2.96 | 2.88 | 2.83 | 2.78 | 2.75 | 2.72 |
| 8 | 3.46 | 3.11 | 2.92 | 2.81 | 2.73 | 2.67 | 2.62 | 2.59 | 2.56 |
| 9 | 3.36 | 3.01 | 2.81 | 2.69 | 2.61 | 2.55 | 2.51 | 2.47 | 2.44 |
| 10 | 3.29 | 2.92 | 2.73 | 2.61 | 2.52 | 2.46 | 2.41 | 2.38 | 2.35 |
| 11 | 3.23 | 2.86 | 2.66 | 2.54 | 2.45 | 2.39 | 2.34 | 2.30 | 2.27 |
| 12 | 3.18 | 2.81 | 2.61 | 2.48 | 2.39 | 2.33 | 2.28 | 2.24 | 2.21 |
| 13 | 3.14 | 2.76 | 2.56 | 2.43 | 2.35 | 2.28 | 2.23 | 2.20 | 2.16 |
| 14 | 3.10 | 2.73 | 2.52 | 2.39 | 2.31 | 2.24 | 2.19 | 2.15 | 2.12 |
| 15 | 3.07 | 2.70 | 2.49 | 2.36 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 |
| 16 | 3.05 | 2.67 | 2.46 | 2.33 | 2.24 | 2.18 | 2.13 | 2.09 | 2.06 |
| 17 | 3.03 | 2.64 | 2.44 | 2.31 | 2.22 | 2.15 | 2.10 | 2.06 | 2.03 |
| 18 | 3.01 | 2.62 | 2.42 | 2.29 | 2.20 | 2.13 | 2.08 | 2.04 | 2.00 |
| 19 | 2.99 | 2.61 | 2.40 | 2.27 | 2.18 | 2.11 | 2.06 | 2.02 | 1.98 |
| 20 | 2.97 | 2.59 | 2.38 | 2.25 | 2.16 | 2.09 | 2.04 | 2.00 | 1.96 |
| 21 | 2.96 | 2.57 | 2.36 | 2.23 | 2.14 | 2.08 | 2.02 | 1.98 | 1.95 |
| 22 | 2.95 | 2.56 | 2.35 | 2.22 | 2.13 | 2.06 | 2.01 | 1.97 | 1.93 |
| 23 | 2.94 | 2.55 | 2.34 | 2.21 | 2.11 | 2.05 | 1.99 | 1.95 | 1.92 |
| 24 | 2.93 | 2.54 | 2.33 | 2.19 | 2.10 | 2.04 | 1.98 | 1.94 | 1.91 |
| 25 | 2.92 | 2.53 | 2.32 | 2.18 | 2.09 | 2.02 | 1.97 | 1.93 | 1.89 |
| 26 | 2.91 | 2.52 | 2.31 | 2.17 | 2.08 | 2.01 | 1.96 | 1.92 | 1.88 |
| 27 | 2.90 | 2.51 | 2.30 | 2.17 | 2.07 | 2.00 | 1.95 | 1.91 | 1.87 |
| 28 | 2.89 | 2.50 | 2.29 | 2.16 | 2.06 | 2.00 | 1.94 | 1.90 | 1.87 |
| 29 | 2.89 | 2.50 | 2.28 | 2.15 | 2.06 | 1.99 | 1.93 | 1.89 | 1.86 |
| 30 | 2.88 | 2.49 | 2.28 | 2.14 | 2.05 | 1.98 | 1.93 | 1.88 | 1.85 |
| 40 | 2.84 | 2.44 | 2.23 | 2.09 | 2.00 | 1.93 | 1.87 | 1.83 | 1.79 |
| 60 | 2.79 | 2.39 | 2.18 | 2.04 | 1.95 | 1.87 | 1.82 | 1.77 | 1.74 |
| 120 | 2.75 | 2.35 | 2.13 | 1.99 | 1.90 | 1.82 | 1.77 | 1.72 | 1.68 |
| ∞ | 2.71 | 2.30 | 2.08 | 1.94 | 1.85 | 1.77 | 1.72 | 1.67 | 1.63 |

*Denominator Degrees of Freedom*

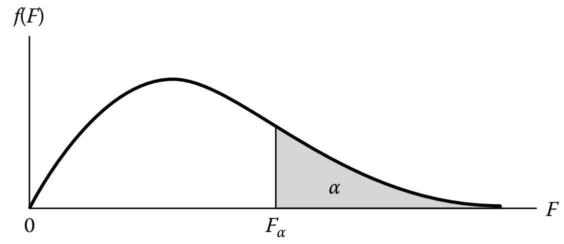
**TABLE 9 Percentage Points of the *F* Distribution, $\alpha$ = .10 (continued)**

| $\nu_2$ \ $\nu_1$ | Numerator Degrees of Freedom 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 60.19 | 60.71 | 61.22 | 61.74 | 62.00 | 62.26 | 62.53 | 62.79 | 63.06 | 63.33 |
| 2 | 9.39 | 9.41 | 9.42 | 9.44 | 9.45 | 9.46 | 9.47 | 9.47 | 9.48 | 9.49 |
| 3 | 5.23 | 5.22 | 5.20 | 5.18 | 5.18 | 5.17 | 5.16 | 5.15 | 5.14 | 5.13 |
| 4 | 3.92 | 3.90 | 3.87 | 3.84 | 3.83 | 3.82 | 3.80 | 3.79 | 3.78 | 3.76 |
| 5 | 3.30 | 3.27 | 3.24 | 3.21 | 3.19 | 3.17 | 3.16 | 3.14 | 3.12 | 3.10 |
| 6 | 2.94 | 2.90 | 2.87 | 2.84 | 2.82 | 2.80 | 2.78 | 2.76 | 2.74 | 2.72 |
| 7 | 2.70 | 2.67 | 2.63 | 2.59 | 2.58 | 2.56 | 2.54 | 2.51 | 2.49 | 2.47 |
| 8 | 2.54 | 2.50 | 2.46 | 2.42 | 2.40 | 2.38 | 2.36 | 2.34 | 2.32 | 2.29 |
| 9 | 2.42 | 2.38 | 2.34 | 2.30 | 2.28 | 2.25 | 2.23 | 2.21 | 2.18 | 2.16 |
| 10 | 2.32 | 2.28 | 2.24 | 2.20 | 2.18 | 2.16 | 2.13 | 2.11 | 2.08 | 2.06 |
| 11 | 2.25 | 2.21 | 2.17 | 2.12 | 2.10 | 2.08 | 2.05 | 2.03 | 2.00 | 1.97 |
| 12 | 2.19 | 2.15 | 2.10 | 2.06 | 2.04 | 2.01 | 1.99 | 1.96 | 1.93 | 1.90 |
| 13 | 2.14 | 2.10 | 2.05 | 2.01 | 1.98 | 1.96 | 1.93 | 1.90 | 1.88 | 1.85 |
| 14 | 2.10 | 2.05 | 2.01 | 1.96 | 1.94 | 1.91 | 1.89 | 1.86 | 1.83 | 1.80 |
| 15 | 2.06 | 2.02 | 1.97 | 1.92 | 1.90 | 1.87 | 1.85 | 1.82 | 1.79 | 1.76 |
| 16 | 2.03 | 1.99 | 1.94 | 1.89 | 1.87 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 |
| 17 | 2.00 | 1.96 | 1.91 | 1.86 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 |
| 18 | 1.98 | 1.93 | 1.89 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 |
| 19 | 1.96 | 1.91 | 1.86 | 1.81 | 1.79 | 1.76 | 1.73 | 1.70 | 1.67 | 1.63 |
| 20 | 1.94 | 1.89 | 1.84 | 1.79 | 1.77 | 1.74 | 1.71 | 1.68 | 1.64 | 1.61 |
| 21 | 1.92 | 1.87 | 1.83 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 |
| 22 | 1.90 | 1.86 | 1.81 | 1.76 | 1.73 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 |
| 23 | 1.89 | 1.84 | 1.80 | 1.74 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 | 1.55 |
| 24 | 1.88 | 1.83 | 1.78 | 1.73 | 1.70 | 1.67 | 1.64 | 1.61 | 1.57 | 1.53 |
| 25 | 1.87 | 1.82 | 1.77 | 1.72 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 |
| 26 | 1.86 | 1.81 | 1.76 | 1.71 | 1.68 | 1.65 | 1.61 | 1.58 | 1.54 | 1.50 |
| 27 | 1.85 | 1.80 | 1.75 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 | 1.53 | 1.49 |
| 28 | 1.84 | 1.79 | 1.74 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 | 1.48 |
| 29 | 1.83 | 1.78 | 1.73 | 1.68 | 1.65 | 1.62 | 1.58 | 1.55 | 1.51 | 1.47 |
| 30 | 1.82 | 1.77 | 1.72 | 1.67 | 1.64 | 1.61 | 1.57 | 1.54 | 1.50 | 1.46 |
| 40 | 1.76 | 1.71 | 1.66 | 1.61 | 1.57 | 1.54 | 1.51 | 1.47 | 1.42 | 1.38 |
| 60 | 1.71 | 1.66 | 1.60 | 1.54 | 1.51 | 1.48 | 1.44 | 1.40 | 1.35 | 1.29 |
| 120 | 1.65 | 1.60 | 1.55 | 1.48 | 1.45 | 1.41 | 1.37 | 1.32 | 1.26 | 1.19 |
| $\infty$ | 1.60 | 1.55 | 1.49 | 1.42 | 1.38 | 1.34 | 1.30 | 1.24 | 1.17 | 1.00 |

Denominator Degrees of Freedom

*Source:* From Merrington, M., and Thompson, C. M. "Tables of percentage points of the inverted beta (*F*)-distribution." *Biometrika*, 1943, Vol. 33, pp. 73–88. Reproduced by permission of the *Biometrika* Trustees.

**TABLE 10 Percentage Points of the $F$ Distribution, $\alpha = .05$**




| $\nu_2$ \ $\nu_1$ | Numerator Degrees of Freedom | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Denominator Degrees of Freedom | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.17 | 2.09 | 2.02 | 1.96 |
| ∞ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 |

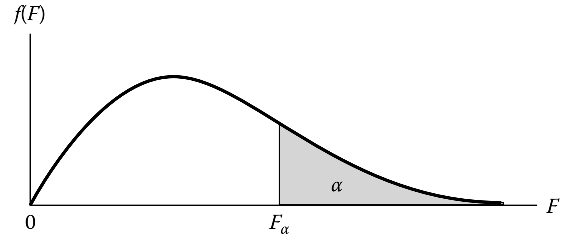
**TABLE 10 Percentage Points of the *F* Distribution, $\alpha$ = .05 (continued)**

| $v_2$ \ $v_1$ (Numerator Degrees of Freedom) | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 241.9 | 243.9 | 245.9 | 248.0 | 249.1 | 250.1 | 251.1 | 252.2 | 253.3 | 254.3 |
| 2 | 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.50 |
| 3 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 |
| 4 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| 5 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.36 |
| 6 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 |
| 7 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 |
| 8 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 |
| 9 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 |
| 10 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 |
| 11 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 |
| 12 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 |
| 13 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 |
| 14 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 |
| 15 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 |
| 16 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 |
| 17 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 |
| 18 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 |
| 19 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 |
| 20 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 |
| 21 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 |
| 22 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 |
| 23 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 |
| 24 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 |
| 25 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 |
| 26 | 2.22 | 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 1.69 |
| 27 | 2.20 | 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 1.67 |
| 28 | 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 1.65 |
| 29 | 2.18 | 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 1.64 |
| 30 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 |
| 40 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 |
| 60 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.39 |
| 120 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 |
| ∞ | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | 1.00 |

*Denominator Degrees of Freedom* ($v_2$)

*Source:* From Merrington, M., and Thompson, C. M. "Tables of percentage points of the inverted beta (*F*)-distribution". *Biometrika*, 1943, Vol. 33, pp. 73–88. Reproduced by permission of the *Biometrika* Trustees.

**TABLE 11 Percentage Points of the *F* Distribution, $\alpha = .025$**




| $\nu_2$ \ $\nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| | Numerator Degrees of Freedom | | | | | | | | |
| 1 | 647.8 | 799.5 | 864.2 | 899.6 | 921.8 | 937.1 | 948.2 | 956.7 | 963.3 |
| 2 | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 |
| 3 | 17.44 | 16.04 | 15.44 | 15.10 | 14.88 | 14.73 | 14.62 | 14.54 | 14.47 |
| 4 | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 |
| 5 | 10.01 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 |
| 6 | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 |
| 7 | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 |
| 8 | 7.57 | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 |
| 9 | 7.21 | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 |
| 10 | 6.94 | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 |
| 11 | 6.72 | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 |
| 12 | 6.55 | 5.10 | 4.47 | 4.12 | 3.89 | 3.73 | 3.61 | 3.51 | 3.44 |
| 13 | 6.41 | 4.97 | 4.35 | 4.00 | 3.77 | 3.60 | 3.48 | 3.39 | 3.31 |
| 14 | 6.30 | 4.86 | 4.24 | 3.89 | 3.66 | 3.50 | 3.38 | 3.29 | 3.21 |
| 15 | 6.20 | 4.77 | 4.15 | 3.80 | 3.58 | 3.41 | 3.29 | 3.20 | 3.12 |
| 16 | 6.12 | 4.69 | 4.08 | 3.73 | 3.50 | 3.34 | 3.22 | 3.12 | 3.05 |
| 17 | 6.04 | 4.62 | 4.01 | 3.66 | 3.44 | 3.28 | 3.16 | 3.06 | 2.98 |
| 18 | 5.98 | 4.56 | 3.95 | 3.61 | 3.38 | 3.22 | 3.10 | 3.01 | 2.93 |
| 19 | 5.92 | 4.51 | 3.90 | 3.56 | 3.33 | 3.17 | 3.05 | 2.96 | 2.88 |
| 20 | 5.87 | 4.46 | 3.86 | 3.51 | 3.29 | 3.13 | 3.01 | 2.91 | 2.84 |
| 21 | 5.83 | 4.42 | 3.82 | 3.48 | 3.25 | 3.09 | 2.97 | 2.87 | 2.80 |
| 22 | 5.79 | 4.38 | 3.78 | 3.44 | 3.22 | 3.05 | 2.93 | 2.84 | 2.76 |
| 23 | 5.75 | 4.35 | 3.75 | 3.41 | 3.18 | 3.02 | 2.90 | 2.81 | 2.73 |
| 24 | 5.72 | 4.32 | 3.72 | 3.38 | 3.15 | 2.99 | 2.87 | 2.78 | 2.70 |
| 25 | 5.69 | 4.29 | 3.69 | 3.35 | 3.13 | 2.97 | 2.85 | 2.75 | 2.68 |
| 26 | 5.66 | 4.27 | 3.67 | 3.33 | 3.10 | 2.94 | 2.82 | 2.73 | 2.65 |
| 27 | 5.63 | 4.24 | 3.65 | 3.31 | 3.08 | 2.92 | 2.80 | 2.71 | 2.63 |
| 28 | 5.61 | 4.22 | 3.63 | 3.29 | 3.06 | 2.90 | 2.78 | 2.69 | 2.61 |
| 29 | 5.59 | 4.20 | 3.61 | 3.27 | 3.04 | 2.88 | 2.76 | 2.67 | 2.59 |
| 30 | 5.57 | 4.18 | 3.59 | 3.25 | 3.03 | 2.87 | 2.75 | 2.65 | 2.57 |
| 40 | 5.42 | 4.05 | 3.46 | 3.13 | 2.90 | 2.74 | 2.62 | 2.53 | 2.45 |
| 60 | 5.29 | 3.93 | 3.34 | 3.01 | 2.79 | 2.63 | 2.51 | 2.41 | 2.33 |
| 120 | 5.15 | 3.80 | 3.23 | 2.89 | 2.67 | 2.52 | 2.39 | 2.30 | 2.22 |
| ∞ | 5.02 | 3.69 | 3.12 | 2.79 | 2.57 | 2.41 | 2.29 | 2.19 | 2.11 |

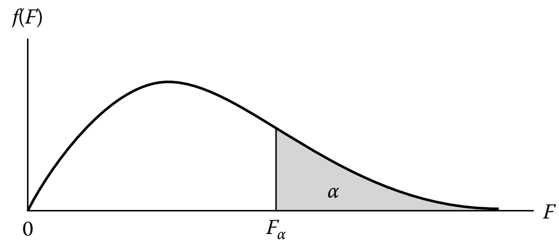
*Denominator Degrees of Freedom*

**TABLE 11 Percentage Points of the $F$ Distribution, $\alpha = .025$ (continued)**

| $\nu_2$ \ $\nu_1$ (Denominator Degrees of Freedom \ Numerator Degrees of Freedom) | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 968.6 | 976.7 | 984.9 | 993.1 | 997.2 | 1,001 | 1,006 | 1,010 | 1,014 | 1,108 |
| 2 | 39.40 | 39.41 | 39.43 | 39.45 | 39.46 | 39.46 | 39.47 | 39.48 | 39.49 | 39.50 |
| 3 | 14.42 | 14.34 | 14.25 | 14.17 | 14.12 | 14.08 | 14.04 | 13.99 | 13.95 | 13.90 |
| 4 | 8.84 | 8.75 | 8.66 | 8.56 | 8.51 | 8.46 | 8.41 | 8.36 | 8.31 | 8.26 |
| 5 | 6.62 | 6.52 | 6.43 | 6.33 | 6.28 | 6.23 | 6.18 | 6.12 | 6.07 | 6.02 |
| 6 | 5.46 | 5.37 | 5.27 | 5.17 | 5.12 | 5.07 | 5.01 | 4.96 | 4.90 | 4.85 |
| 7 | 4.76 | 4.67 | 4.57 | 4.47 | 4.42 | 4.36 | 4.31 | 4.25 | 4.20 | 4.14 |
| 8 | 4.30 | 4.20 | 4.10 | 4.00 | 3.95 | 3.89 | 3.84 | 3.78 | 3.73 | 3.67 |
| 9 | 3.96 | 3.87 | 3.77 | 3.67 | 3.61 | 3.56 | 3.51 | 3.45 | 3.39 | 3.33 |
| 10 | 3.72 | 3.62 | 3.52 | 3.42 | 3.37 | 3.31 | 3.26 | 3.20 | 3.14 | 3.08 |
| 11 | 3.53 | 3.43 | 3.33 | 3.23 | 3.17 | 3.12 | 3.06 | 3.00 | 2.94 | 2.88 |
| 12 | 3.37 | 3.28 | 3.18 | 3.07 | 3.02 | 2.96 | 2.91 | 2.85 | 2.79 | 2.72 |
| 13 | 3.25 | 3.15 | 3.05 | 2.95 | 2.89 | 2.84 | 2.78 | 2.72 | 2.66 | 2.60 |
| 14 | 3.15 | 3.05 | 2.95 | 2.84 | 2.79 | 2.73 | 2.67 | 2.61 | 2.55 | 2.49 |
| 15 | 3.06 | 2.96 | 2.86 | 2.76 | 2.70 | 2.64 | 2.59 | 2.52 | 2.46 | 2.40 |
| 16 | 2.99 | 2.89 | 2.79 | 2.68 | 2.63 | 2.57 | 2.51 | 2.45 | 2.38 | 2.32 |
| 17 | 2.92 | 2.82 | 2.72 | 2.62 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.25 |
| 18 | 2.87 | 2.77 | 2.67 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.26 | 2.19 |
| 19 | 2.82 | 2.72 | 2.62 | 2.51 | 2.45 | 2.39 | 2.33 | 2.27 | 2.20 | 2.13 |
| 20 | 2.77 | 2.68 | 2.57 | 2.46 | 2.41 | 2.35 | 2.29 | 2.22 | 2.16 | 2.09 |
| 21 | 2.73 | 2.64 | 2.53 | 2.42 | 2.37 | 2.31 | 2.25 | 2.18 | 2.11 | 2.04 |
| 22 | 2.70 | 2.60 | 2.50 | 2.39 | 2.33 | 2.27 | 2.21 | 2.14 | 2.08 | 2.00 |
| 23 | 2.67 | 2.57 | 2.47 | 2.36 | 2.30 | 2.24 | 2.18 | 2.11 | 2.04 | 1.97 |
| 24 | 2.64 | 2.54 | 2.44 | 2.33 | 2.27 | 2.21 | 2.15 | 2.08 | 2.01 | 1.94 |
| 25 | 2.61 | 2.51 | 2.41 | 2.30 | 2.24 | 2.18 | 2.12 | 2.05 | 1.98 | 1.91 |
| 26 | 2.59 | 2.49 | 2.39 | 2.28 | 2.22 | 2.16 | 2.09 | 2.03 | 1.95 | 1.88 |
| 27 | 2.57 | 2.47 | 2.36 | 2.25 | 2.19 | 2.13 | 2.07 | 2.00 | 1.93 | 1.85 |
| 28 | 2.55 | 2.45 | 2.34 | 2.23 | 2.17 | 2.11 | 2.05 | 1.98 | 1.91 | 1.83 |
| 29 | 2.53 | 2.43 | 2.32 | 2.21 | 2.15 | 2.09 | 2.03 | 1.96 | 1.89 | 1.81 |
| 30 | 2.51 | 2.41 | 2.31 | 2.20 | 2.14 | 2.07 | 2.01 | 1.94 | 1.87 | 1.79 |
| 40 | 2.39 | 2.29 | 2.18 | 2.07 | 2.01 | 1.94 | 1.88 | 1.80 | 1.72 | 1.64 |
| 60 | 2.27 | 2.17 | 2.06 | 1.94 | 1.88 | 1.82 | 1.74 | 1.67 | 1.58 | 1.48 |
| 120 | 2.16 | 2.05 | 1.94 | 1.82 | 1.76 | 1.69 | 1.61 | 1.53 | 1.43 | 1.31 |
| ∞ | 2.05 | 1.94 | 1.83 | 1.71 | 1.64 | 1.57 | 1.48 | 1.39 | 1.27 | 1.00 |

*Source:* From Merrington, M., and Thompson, C. M. "Tables of percentage points of the inverted beta ($F$)-distribution". *Biometrika*, 1943, Vol. 33, pp. 73–88. Reproduced by permission of the *Biometrika* Trustees.

**TABLE 12 Percentage Points of the *F* Distribution, $\alpha = .01$**




| $\nu_2$ \ $\nu_1$ | Numerator Degrees of Freedom | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Denominator Degrees of Freedom | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 4,052 | 4,999.5 | 5,403 | 5,625 | 5,764 | 5,859 | 5,928 | 5,982 | 6,022 |
| 2 | 98.50 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.36 | 99.37 | 99.39 |
| 3 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.35 |
| 4 | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 |
| 5 | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 |
| 6 | 13.75 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 |
| 7 | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 |
| 8 | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 |
| 9 | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 |
| 10 | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 |
| 17 | 8.40 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 |
| 19 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 |
| $\infty$ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 |

**TABLE 12 Percentage Points of the *F* Distribution, $\alpha = .01$ (continued)**

| $\nu_2$ \ $\nu_1$ (Numerator Degrees of Freedom) | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6,056 | 6,106 | 6,157 | 6,209 | 6,235 | 6,261 | 6,287 | 6,313 | 6,339 | 6,366 |
| 2 | 99.40 | 99.42 | 99.43 | 99.45 | 99.46 | 99.47 | 99.47 | 99.48 | 99.49 | 99.50 |
| 3 | 27.23 | 27.05 | 26.87 | 26.69 | 26.60 | 26.50 | 26.41 | 26.32 | 26.22 | 26.13 |
| 4 | 14.55 | 14.37 | 14.20 | 14.02 | 13.93 | 13.84 | 13.75 | 13.65 | 13.56 | 13.46 |
| 5 | 10.05 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.02 |
| 6 | 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.88 |
| 7 | 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.65 |
| 8 | 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.86 |
| 9 | 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.31 |
| 10 | 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.91 |
| 11 | 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.60 |
| 12 | 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.36 |
| 13 | 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 3.17 |
| 14 | 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.00 |
| 15 | 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.87 |
| 16 | 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.75 |
| 17 | 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.65 |
| 18 | 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.57 |
| 19 | 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.49 |
| 20 | 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.42 |
| 21 | 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 2.36 |
| 22 | 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.31 |
| 23 | 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.26 |
| 24 | 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 2.21 |
| 25 | 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 | 2.36 | 2.27 | 2.17 |
| 26 | 3.09 | 2.96 | 2.81 | 2.66 | 2.58 | 2.50 | 2.42 | 2.33 | 2.23 | 2.13 |
| 27 | 3.06 | 2.93 | 2.78 | 2.63 | 2.55 | 2.47 | 2.38 | 2.29 | 2.20 | 2.10 |
| 28 | 3.03 | 2.90 | 2.75 | 2.60 | 2.52 | 2.44 | 2.35 | 2.26 | 2.17 | 2.06 |
| 29 | 3.00 | 2.87 | 2.73 | 2.57 | 2.49 | 2.41 | 2.33 | 2.23 | 2.14 | 2.03 |
| 30 | 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.01 |
| 40 | 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.80 |
| 60 | 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.60 |
| 120 | 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.38 |
| ∞ | 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | 1.00 |

*Denominator Degrees of Freedom* ($\nu_2$)

*Source:* From Merrington, M., and Thompson, C. M. "Tables of percentage points of the inverted beta (*F*)-distribution". *Biometrika*, 1943, Vol. 33, pp. 73–88. Reproduced by permission of the *Biometrika* Trustees.

**TABLE 13 Percentage Points of the Studentized Range $q(p, \nu)$, $\alpha = .05$**

| $\nu$ \ $p$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 17.97 | 26.98 | 32.82 | 37.08 | 40.41 | 43.12 | 45.40 | 47.36 | 49.07 | 50.59 |
| 2 | 6.08 | 8.33 | 9.80 | 10.88 | 11.74 | 12.44 | 13.03 | 13.54 | 13.99 | 14.39 |
| 3 | 4.50 | 5.91 | 6.82 | 7.50 | 8.04 | 8.48 | 8.85 | 9.18 | 9.46 | 9.72 |
| 4 | 3.93 | 5.04 | 5.76 | 6.29 | 6.71 | 7.05 | 7.35 | 7.60 | 7.83 | 8.03 |
| 5 | 3.64 | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 | 6.80 | 6.99 | 7.17 |
| 6 | 3.46 | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 | 6.32 | 6.49 | 6.65 |
| 7 | 3.34 | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 | 6.00 | 6.16 | 6.30 |
| 8 | 3.26 | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92 | 6.05 |
| 9 | 3.20 | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 | 5.59 | 5.74 | 5.87 |
| 10 | 3.15 | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60 | 5.72 |
| 11 | 3.11 | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49 | 5.61 |
| 12 | 3.08 | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.39 | 5.51 |
| 13 | 3.06 | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32 | 5.43 |
| 14 | 3.03 | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25 | 5.36 |
| 15 | 3.01 | 3.67 | 4.08 | 4.37 | 4.60 | 4.78 | 4.94 | 5.08 | 5.20 | 5.31 |
| 16 | 3.00 | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15 | 5.26 |
| 17 | 2.98 | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11 | 5.21 |
| 18 | 2.97 | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 | 4.96 | 5.07 | 5.17 |
| 19 | 2.96 | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 | 4.92 | 5.04 | 5.14 |
| 20 | 2.95 | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01 | 5.11 |
| 24 | 2.92 | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.01 |
| 30 | 2.89 | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 | 4.72 | 4.82 | 4.92 |
| 40 | 2.86 | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.73 | 4.82 |
| 60 | 2.83 | 3.40 | 3.74 | 3.98 | 4.16 | 4.31 | 4.44 | 4.55 | 4.65 | 4.73 |
| 120 | 2.80 | 3.36 | 3.68 | 3.92 | 4.10 | 4.24 | 4.36 | 4.47 | 4.56 | 4.64 |
| $\infty$ | 2.77 | 3.31 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.55 |

**TABLE 13 Percentage Points of the Studentized Range $q(p, \nu)$, $\alpha = .05$ (continued)**

| $\nu$ \ $p$ | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 51.96 | 53.20 | 54.33 | 55.36 | 56.32 | 57.22 | 58.04 | 58.83 | 59.56 |
| 2 | 14.75 | 15.08 | 15.38 | 15.65 | 15.91 | 16.14 | 16.37 | 16.57 | 16.77 |
| 3 | 9.95 | 10.15 | 10.35 | 10.52 | 10.69 | 10.84 | 10.98 | 11.11 | 11.24 |
| 4 | 8.21 | 8.37 | 8.52 | 8.66 | 8.79 | 8.91 | 9.03 | 9.13 | 9.23 |
| 5 | 7.32 | 7.47 | 7.60 | 7.72 | 7.83 | 7.93 | 8.03 | 8.12 | 8.21 |
| 6 | 6.79 | 6.92 | 7.03 | 7.14 | 7.24 | 7.34 | 7.43 | 7.51 | 7.59 |
| 7 | 6.43 | 6.55 | 6.66 | 6.76 | 6.85 | 6.94 | 7.02 | 7.10 | 7.17 |
| 8 | 6.18 | 6.29 | 6.39 | 6.48 | 6.57 | 6.65 | 6.73 | 6.80 | 6.87 |
| 9 | 5.98 | 6.09 | 6.19 | 6.28 | 6.36 | 6.44 | 6.51 | 6.58 | 6.64 |
| 10 | 5.83 | 5.93 | 6.03 | 6.11 | 6.19 | 6.27 | 6.34 | 6.40 | 6.47 |
| 11 | 5.71 | 5.81 | 5.90 | 5.98 | 6.06 | 6.13 | 6.20 | 6.27 | 6.33 |
| 12 | 5.61 | 5.71 | 5.80 | 5.88 | 5.95 | 6.02 | 6.09 | 6.15 | 6.21 |
| 13 | 5.53 | 5.63 | 5.71 | 5.79 | 5.86 | 5.93 | 5.99 | 6.05 | 6.11 |
| 14 | 5.46 | 5.55 | 5.64 | 5.71 | 5.79 | 5.85 | 5.91 | 5.97 | 6.03 |
| 15 | 5.40 | 5.49 | 5.57 | 5.65 | 5.72 | 5.78 | 5.85 | 5.90 | 5.96 |
| 16 | 5.35 | 5.44 | 5.52 | 5.59 | 5.66 | 5.73 | 5.79 | 5.84 | 5.90 |
| 17 | 5.31 | 5.39 | 5.47 | 5.54 | 5.61 | 5.67 | 5.73 | 5.79 | 5.84 |
| 18 | 5.27 | 5.35 | 5.43 | 5.50 | 5.57 | 5.63 | 5.69 | 5.74 | 5.79 |
| 19 | 5.23 | 5.31 | 5.39 | 5.46 | 5.53 | 5.59 | 5.65 | 5.70 | 5.75 |
| 20 | 5.20 | 5.28 | 5.36 | 5.43 | 5.49 | 5.55 | 5.61 | 5.66 | 5.71 |
| 24 | 5.10 | 5.18 | 5.25 | 5.32 | 5.38 | 5.44 | 5.49 | 5.55 | 5.59 |
| 30 | 5.00 | 5.08 | 5.15 | 5.21 | 5.27 | 5.33 | 5.38 | 5.43 | 5.47 |
| 40 | 4.90 | 4.98 | 5.04 | 5.11 | 5.16 | 5.22 | 5.27 | 5.31 | 5.36 |
| 60 | 4.81 | 4.88 | 4.94 | 5.00 | 5.06 | 5.11 | 5.15 | 5.20 | 5.24 |
| 120 | 4.71 | 4.78 | 4.84 | 4.90 | 4.95 | 5.00 | 5.04 | 5.09 | 5.13 |
| $\infty$ | 4.62 | 4.68 | 4.74 | 4.80 | 4.85 | 4.89 | 4.93 | 4.97 | 5.01 |

*Source: Biometrika Tables for Statisticians*, Vol. I, 3rd ed., edited by E. S. Pearson and H. O. Hartley. Cambridge: Cambridge University Press, 1966. Reproduced by permission of Professor E. S. Pearson and the *Biometrika* Trustees.

**TABLE 14 Percentage Points of the Studentized Range $q(p, \nu)$, $\alpha = .01$**

| $\nu$ \ p | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 90.03 | 135.0 | 164.3 | 185.6 | 202.2 | 215.8 | 227.2 | 237.0 | 245.6 | 253.2 |
| 2 | 14.04 | 19.02 | 22.29 | 24.72 | 26.63 | 28.20 | 29.53 | 30.68 | 31.69 | 32.59 |
| 3 | 8.26 | 10.62 | 12.17 | 13.33 | 14.24 | 15.00 | 15.64 | 16.20 | 16.69 | 17.13 |
| 4 | 6.51 | 8.12 | 9.17 | 9.96 | 10.58 | 11.10 | 11.55 | 11.93 | 12.27 | 12.57 |
| 5 | 5.70 | 6.98 | 7.80 | 8.42 | 8.91 | 9.32 | 9.67 | 9.97 | 10.24 | 10.48 |
| 6 | 5.24 | 6.33 | 7.03 | 7.56 | 7.97 | 8.32 | 8.61 | 8.87 | 9.10 | 9.30 |
| 7 | 4.95 | 5.92 | 6.54 | 7.01 | 7.37 | 7.68 | 7.94 | 8.17 | 8.37 | 8.55 |
| 8 | 4.75 | 5.64 | 6.20 | 6.62 | 6.96 | 7.24 | 7.47 | 7.68 | 7.86 | 8.03 |
| 9 | 4.60 | 5.43 | 5.96 | 6.35 | 6.66 | 6.91 | 7.13 | 7.33 | 7.49 | 7.65 |
| 10 | 4.48 | 5.27 | 5.77 | 6.14 | 6.43 | 6.67 | 6.87 | 7.05 | 7.21 | 7.36 |
| 11 | 4.39 | 5.15 | 5.62 | 5.97 | 6.25 | 6.48 | 6.67 | 6.84 | 6.99 | 7.13 |
| 12 | 4.32 | 5.05 | 5.50 | 5.84 | 6.10 | 6.32 | 6.51 | 6.67 | 6.81 | 6.94 |
| 13 | 4.26 | 4.96 | 5.40 | 5.73 | 5.98 | 6.19 | 6.37 | 6.53 | 6.67 | 6.79 |
| 14 | 4.21 | 4.89 | 5.32 | 5.63 | 5.88 | 6.08 | 6.26 | 6.41 | 6.54 | 6.66 |
| 15 | 4.17 | 4.84 | 5.25 | 5.56 | 5.80 | 5.99 | 6.16 | 6.31 | 6.44 | 6.55 |
| 16 | 4.13 | 4.79 | 5.19 | 5.49 | 5.72 | 5.92 | 6.08 | 6.22 | 6.35 | 6.46 |
| 17 | 4.10 | 4.74 | 5.14 | 5.43 | 5.66 | 5.85 | 6.01 | 6.15 | 6.27 | 6.38 |
| 18 | 4.07 | 4.70 | 5.09 | 5.38 | 5.60 | 5.79 | 5.94 | 6.08 | 6.20 | 6.31 |
| 19 | 4.05 | 4.67 | 5.05 | 5.33 | 5.55 | 5.73 | 5.89 | 6.02 | 6.14 | 6.25 |
| 20 | 4.02 | 4.64 | 5.02 | 5.29 | 5.51 | 5.69 | 5.84 | 5.97 | 6.09 | 6.19 |
| 24 | 3.96 | 4.55 | 4.91 | 5.17 | 5.37 | 5.54 | 5.69 | 5.81 | 5.92 | 6.02 |
| 30 | 3.89 | 4.45 | 4.80 | 5.05 | 5.24 | 5.40 | 5.54 | 5.65 | 5.76 | 5.85 |
| 40 | 3.82 | 4.37 | 4.70 | 4.93 | 5.11 | 5.26 | 5.39 | 5.50 | 5.60 | 5.69 |
| 60 | 3.76 | 4.28 | 4.59 | 4.82 | 4.99 | 5.13 | 5.25 | 5.36 | 5.45 | 5.53 |
| 120 | 3.70 | 4.20 | 4.50 | 4.71 | 4.87 | 5.01 | 5.12 | 5.21 | 5.30 | 5.37 |
| $\infty$ | 3.64 | 4.12 | 4.40 | 4.60 | 4.76 | 4.88 | 4.99 | 5.08 | 5.16 | 5.23 |

**TABLE 14 Percentage Points of the Studentized Range $q(p, \nu)$, $\alpha = .01$ (continued)**

| $\nu$ \ $p$ | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 260.0 | 266.2 | 271.8 | 277.0 | 281.8 | 286.3 | 290.0 | 294.3 | 298.0 |
| 2 | 33.40 | 34.13 | 34.81 | 35.43 | 36.00 | 36.53 | 37.03 | 37.50 | 37.95 |
| 3 | 17.53 | 17.89 | 18.22 | 18.52 | 18.81 | 19.07 | 19.32 | 19.55 | 19.77 |
| 4 | 12.84 | 13.09 | 13.32 | 13.53 | 13.73 | 13.91 | 14.08 | 14.24 | 14.40 |
| 5 | 10.70 | 10.89 | 11.08 | 11.24 | 11.40 | 11.55 | 11.68 | 11.81 | 11.93 |
| 6 | 9.48 | 9.65 | 9.81 | 9.95 | 10.08 | 10.21 | 10.32 | 10.43 | 10.54 |
| 7 | 8.71 | 8.86 | 9.00 | 9.12 | 9.24 | 9.35 | 9.46 | 9.55 | 9.65 |
| 8 | 8.18 | 8.31 | 8.44 | 8.55 | 8.66 | 8.76 | 8.85 | 8.94 | 9.03 |
| 9 | 7.78 | 7.91 | 8.03 | 8.13 | 8.23 | 8.33 | 8.41 | 8.49 | 8.57 |
| 10 | 7.49 | 7.60 | 7.71 | 7.81 | 7.91 | 7.99 | 8.08 | 8.15 | 8.23 |
| 11 | 7.25 | 7.36 | 7.46 | 7.56 | 7.65 | 7.73 | 7.81 | 7.88 | 7.95 |
| 12 | 7.06 | 7.17 | 7.26 | 7.36 | 7.44 | 7.52 | 7.59 | 7.66 | 7.73 |
| 13 | 6.90 | 7.01 | 7.10 | 7.19 | 7.27 | 7.35 | 7.42 | 7.48 | 7.55 |
| 14 | 6.77 | 6.87 | 6.96 | 7.05 | 7.13 | 7.20 | 7.27 | 7.33 | 7.39 |
| 15 | 6.66 | 6.76 | 6.84 | 6.93 | 7.00 | 7.07 | 7.14 | 7.20 | 7.26 |
| 16 | 6.56 | 6.66 | 6.74 | 6.82 | 6.90 | 6.97 | 7.03 | 7.09 | 7.15 |
| 17 | 6.48 | 6.57 | 6.66 | 6.73 | 6.81 | 6.87 | 6.94 | 7.00 | 7.05 |
| 18 | 6.41 | 6.50 | 6.58 | 6.65 | 6.72 | 6.79 | 6.85 | 6.91 | 6.97 |
| 19 | 6.34 | 6.43 | 6.51 | 6.58 | 6.65 | 6.72 | 6.78 | 6.84 | 6.89 |
| 20 | 6.28 | 6.37 | 6.45 | 6.52 | 6.59 | 6.65 | 6.71 | 6.77 | 6.82 |
| 24 | 6.11 | 6.19 | 6.26 | 6.33 | 6.39 | 6.45 | 6.51 | 6.56 | 6.61 |
| 30 | 5.93 | 6.01 | 6.08 | 6.14 | 6.20 | 6.26 | 6.31 | 6.36 | 6.41 |
| 40 | 5.76 | 5.83 | 5.90 | 5.96 | 6.02 | 6.07 | 6.12 | 6.16 | 6.21 |
| 60 | 5.60 | 5.67 | 5.73 | 5.78 | 5.84 | 5.89 | 5.93 | 5.97 | 6.01 |
| 120 | 5.44 | 5.50 | 5.56 | 5.61 | 5.66 | 5.71 | 5.75 | 5.79 | 5.83 |
| $\infty$ | 5.29 | 5.35 | 5.40 | 5.45 | 5.49 | 5.54 | 5.57 | 5.61 | 5.65 |

*Source: Biometrika Tables for Statisticians*, Vol. I, 3d ed., edited by E. S. Pearson and H. O. Hartley. Cambridge: Cambridge University Press, 1966. Reproduced by permission of Professor E. S. Pearson and the *Biometrika* Trustees.

## TABLE 15 Critical Values of $T_L$ and $T_U$ for the Wilcoxon Rank Sum Test: Independent Samples

*Test statistic is the rank sum associated with the smaller sample (if equal sample sizes, either rank sum can be used).*

a. $\alpha = .025$ one-tailed; $\alpha = .05$ two-tailed

| $n_2$ \ $n_1$ | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ |
| 3 | 5 | 16 | 6 | 18 | 6 | 21 | 7 | 23 | 7 | 26 | 8 | 28 | 8 | 31 | 9 | 33 |
| 4 | 6 | 18 | 11 | 25 | 12 | 28 | 12 | 32 | 13 | 35 | 14 | 38 | 15 | 41 | 16 | 44 |
| 5 | 6 | 21 | 12 | 28 | 18 | 37 | 19 | 41 | 20 | 45 | 21 | 49 | 22 | 53 | 24 | 56 |
| 6 | 7 | 23 | 12 | 32 | 19 | 41 | 26 | 52 | 28 | 56 | 29 | 61 | 31 | 65 | 32 | 70 |
| 7 | 7 | 26 | 13 | 35 | 20 | 45 | 28 | 56 | 37 | 68 | 39 | 73 | 41 | 78 | 43 | 83 |
| 8 | 8 | 28 | 14 | 38 | 21 | 49 | 29 | 61 | 39 | 73 | 49 | 87 | 51 | 93 | 54 | 98 |
| 9 | 8 | 31 | 15 | 41 | 22 | 53 | 31 | 65 | 41 | 78 | 51 | 93 | 63 | 108 | 66 | 114 |
| 10 | 9 | 33 | 16 | 44 | 24 | 56 | 32 | 70 | 43 | 83 | 54 | 98 | 66 | 114 | 79 | 131 |

$\alpha = .05$ one-tailed; $\alpha = .10$ two-tailed

| $n_2$ \ $n_1$ | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ | $T_L$ | $T_U$ |
| 3 | 6 | 15 | 7 | 17 | 7 | 20 | 8 | 22 | 9 | 24 | 9 | 27 | 10 | 29 | 11 | 31 |
| 4 | 7 | 17 | 12 | 24 | 13 | 27 | 14 | 30 | 15 | 33 | 16 | 36 | 17 | 39 | 18 | 42 |
| 5 | 7 | 20 | 13 | 27 | 19 | 36 | 20 | 40 | 22 | 43 | 24 | 46 | 25 | 50 | 26 | 54 |
| 6 | 8 | 22 | 14 | 30 | 20 | 40 | 28 | 50 | 30 | 54 | 32 | 58 | 33 | 63 | 35 | 67 |
| 7 | 9 | 24 | 15 | 33 | 22 | 43 | 30 | 54 | 39 | 66 | 41 | 71 | 43 | 76 | 46 | 80 |
| 8 | 9 | 27 | 16 | 36 | 24 | 46 | 32 | 58 | 41 | 71 | 52 | 84 | 53 | 90 | 57 | 95 |
| 9 | 10 | 29 | 17 | 39 | 25 | 50 | 33 | 63 | 43 | 76 | 54 | 90 | 66 | 105 | 69 | 111 |
| 10 | 11 | 31 | 18 | 42 | 26 | 54 | 35 | 67 | 46 | 80 | 57 | 95 | 69 | 111 | 83 | 127 |

*Source:* From Wilcoxon, F., and Wilcox, R. A. "Some rapid approximate statistical procedures". 1964, pp. 20–23. Reproduced with the permission of American Cyanamid Company.

**TABLE 16 Critical Values of $T_0$ in the Wilcoxon Matched-Pairs Signed Rank Test**

| One-Tailed | Two-Tailed | $n = 5$ | $n = 6$ | $n = 7$ | $n = 8$ | $n = 9$ | $n = 10$ |
|---|---|---|---|---|---|---|---|
| $\alpha = .05$ | $\alpha = .10$ | 1 | 2 | 4 | 6 | 8 | 11 |
| $\alpha = .025$ | $\alpha = .05$ | | 1 | 2 | 4 | 6 | 8 |
| $\alpha = .01$ | $\alpha = .02$ | | | 0 | 2 | 3 | 5 |
| $\alpha = .005$ | $\alpha = .01$ | | | | 0 | 2 | 3 |
| | | $n = 11$ | $n = 12$ | $n = 13$ | $n = 14$ | $n = 15$ | $n = 16$ |
| $\alpha = .05$ | $\alpha = .10$ | 14 | 17 | 21 | 26 | 30 | 36 |
| $\alpha = .025$ | $\alpha = .05$ | 11 | 14 | 17 | 21 | 25 | 30 |
| $\alpha = .01$ | $\alpha = .02$ | 7 | 10 | 13 | 16 | 20 | 24 |
| $\alpha = .005$ | $\alpha = .01$ | 5 | 7 | 10 | 13 | 16 | 19 |
| | | $n = 17$ | $n = 18$ | $n = 19$ | $n = 20$ | $n = 21$ | $n = 22$ |
| $\alpha = .05$ | $\alpha = .10$ | 41 | 47 | 54 | 60 | 68 | 75 |
| $\alpha = .025$ | $\alpha = .05$ | 35 | 40 | 46 | 52 | 59 | 66 |
| $\alpha = .01$ | $\alpha = .02$ | 28 | 33 | 38 | 43 | 49 | 56 |
| $\alpha = .005$ | $\alpha = .01$ | 23 | 28 | 32 | 37 | 43 | 49 |
| | | $n = 23$ | $n = 24$ | $n = 25$ | $n = 26$ | $n = 27$ | $n = 28$ |
| $\alpha = .05$ | $\alpha = .10$ | 83 | 92 | 101 | 110 | 120 | 130 |
| $\alpha = .025$ | $\alpha = .05$ | 73 | 81 | 90 | 98 | 107 | 117 |
| $\alpha = .01$ | $\alpha = .02$ | 62 | 69 | 77 | 85 | 93 | 102 |
| $\alpha = .005$ | $\alpha = .01$ | 55 | 61 | 68 | 76 | 84 | 92 |
| | | $n = 29$ | $n = 30$ | $n = 31$ | $n = 32$ | $n = 33$ | $n = 34$ |
| $\alpha = .05$ | $\alpha = .10$ | 141 | 152 | 163 | 175 | 188 | 201 |
| $\alpha = .025$ | $\alpha = .05$ | 127 | 137 | 148 | 159 | 171 | 183 |
| $\alpha = .01$ | $\alpha = .02$ | 111 | 120 | 130 | 141 | 151 | 162 |
| $\alpha = .005$ | $\alpha = .01$ | 100 | 109 | 118 | 128 | 138 | 149 |
| | | $n = 35$ | $n = 36$ | $n = 37$ | $n = 38$ | $n = 39$ | |
| $\alpha = .05$ | $\alpha = .10$ | 214 | 228 | 242 | 256 | 271 | |
| $\alpha = .025$ | $\alpha = .05$ | 195 | 208 | 222 | 235 | 250 | |
| $\alpha = .01$ | $\alpha = .02$ | 174 | 186 | 198 | 211 | 224 | |
| $\alpha = .005$ | $\alpha = .01$ | 160 | 171 | 183 | 195 | 208 | |
| | | $n = 40$ | $n = 41$ | $n = 42$ | $n = 43$ | $n = 44$ | $n = 45$ |
| $\alpha = .05$ | $\alpha = .10$ | 287 | 303 | 319 | 336 | 353 | 371 |
| $\alpha = .025$ | $\alpha = .05$ | 264 | 279 | 295 | 311 | 327 | 344 |
| $\alpha = .01$ | $\alpha = .02$ | 238 | 252 | 267 | 281 | 297 | 313 |
| $\alpha = .005$ | $\alpha = .01$ | 221 | 234 | 248 | 262 | 277 | 292 |
| | | $n = 46$ | $n = 47$ | $n = 48$ | $n = 49$ | $n = 50$ | |
| $\alpha = .05$ | $\alpha = .10$ | 389 | 408 | 427 | 446 | 466 | |
| $\alpha = .025$ | $\alpha = .05$ | 361 | 379 | 397 | 415 | 434 | |
| $\alpha = .01$ | $\alpha = .02$ | 329 | 345 | 362 | 380 | 398 | |
| $\alpha = .005$ | $\alpha = .01$ | 307 | 323 | 339 | 356 | 373 | |

*Source:* From Wilcoxon, F., and Wilcox, R. A. "Some rapid approximate statistical procedures." 1964, p. 28. Reproduced with the permission of American Cyanamid Company.

**TABLE 17 Critical Values of Spearman's Rank Correlation Coefficient**

*The $\alpha$ values correspond to a one-tailed test of $H_0$: $\rho_s = 0$. The value should be doubled for two-tailed tests.*

| $n$ | $\alpha = .05$ | $\alpha = .025$ | $\alpha = .01$ | $\alpha = .005$ | $n$ | $\alpha = .05$ | $\alpha = .025$ | $\alpha = .01$ | $\alpha = .005$ |
|---|---|---|---|---|---|---|---|---|---|
| 5 | .900 | — | — | — | 18 | .399 | .476 | .564 | .625 |
| 6 | .829 | .886 | .943 | — | 19 | .388 | .462 | .549 | .608 |
| 7 | .714 | .786 | .893 | — | 20 | .377 | .450 | .534 | .591 |
| 8 | .643 | .738 | .833 | .881 | 21 | .368 | .438 | .521 | .576 |
| 9 | .600 | .683 | .783 | .833 | 22 | .359 | .428 | .508 | .562 |
| 10 | .564 | .648 | .745 | .794 | 23 | .351 | .418 | .496 | .549 |
| 11 | .523 | .623 | .736 | .818 | 24 | .343 | .409 | .485 | .537 |
| 12 | .497 | .591 | .703 | .780 | 25 | .336 | .400 | .475 | .526 |
| 13 | .475 | .566 | .673 | .745 | 26 | .329 | .392 | .465 | .515 |
| 14 | .457 | .545 | .646 | .716 | 27 | .323 | .385 | .456 | .505 |
| 15 | .441 | .525 | .623 | .689 | 28 | .317 | .377 | .448 | .496 |
| 16 | .425 | .507 | .601 | .666 | 29 | .311 | .370 | .440 | .487 |
| 17 | .412 | .490 | .582 | .645 | 30 | .305 | .364 | .432 | .478 |

*Source:* From Olds, E. G. "Distribution of sums of squares of rank differences for small samples". *Annals of Mathematical Statistics*, 1938, p. 9. Reproduced with the permission of the Editor, *Annals of Mathematical Statistics*.

**TABLE 18 Critical Values of $C$ for the Theil Zero-Slope Test**

| | $n$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| x | 4 | 5 | 8 | 9 | 12 | 13 | 16 | 17 | 20 |
| 0 | .625 | .592 | .548 | .540 | .527 | .524 | .518 | .516 | .513 |
| 2 | .375 | .408 | .452 | .460 | .473 | .476 | .482 | .484 | .487 |
| 4 | .167 | .242 | .360 | .381 | .420 | .429 | .447 | .452 | .462 |
| 6 | .042 | .117 | .274 | .306 | .369 | .383 | .412 | .420 | .436 |
| 8 | | .042 | .199 | .238 | .319 | .338 | .378 | .388 | .411 |
| 10 | | .008 | .138 | .179 | .273 | .295 | .345 | .358 | .387 |
| 12 | | | .089 | .130 | .230 | .255 | .313 | .328 | .362 |
| 14 | | | .054 | .090 | .190 | .218 | .282 | .299 | .339 |
| 16 | | | .031 | .060 | .155 | .184 | .253 | .271 | .315 |
| 18 | | | .016 | .038 | .125 | .153 | .225 | .245 | .293 |
| 20 | | | .007 | .022 | .098 | .126 | .199 | .220 | .271 |
| 22 | | | .002 | .012 | .076 | .102 | .175 | .196 | .250 |
| 24 | | | .001 | .006 | .058 | .082 | .153 | .174 | .230 |
| 26 | | | .000 | .003 | .043 | .064 | .133 | .154 | .211 |
| 28 | | | | .001 | .031 | .050 | .114 | .135 | .193 |
| 30 | | | | .000 | .022 | .038 | .097 | .118 | .176 |
| 32 | | | | | .016 | .029 | .083 | .102 | .159 |
| 34 | | | | | .010 | .021 | .070 | .088 | .144 |
| 36 | | | | | .007 | .015 | .058 | .076 | .130 |
| 38 | | | | | .004 | .011 | .048 | .064 | .117 |
| 40 | | | | | .003 | .007 | .039 | .054 | .104 |
| 42 | | | | | .002 | .005 | .032 | .046 | .093 |
| 44 | | | | | .001 | .003 | .026 | .038 | .082 |
| 46 | | | | | .000 | .002 | .021 | .032 | .073 |
| 48 | | | | | | .001 | .016 | .026 | .064 |
| 50 | | | | | | .001 | .013 | .021 | .056 |
| 52 | | | | | | .000 | .010 | .017 | .049 |
| 54 | | | | | | | .008 | .014 | .043 |
| 56 | | | | | | | .006 | .011 | .037 |
| 58 | | | | | | | .004 | .009 | .032 |
| 60 | | | | | | | .003 | .007 | .027 |
| 62 | | | | | | | .002 | .005 | .023 |
| 64 | | | | | | | .002 | .004 | .020 |
| 66 | | | | | | | .001 | .003 | .017 |
| 68 | | | | | | | .001 | .002 | .014 |
| 70 | | | | | | | .001 | .002 | .012 |
| 72 | | | | | | | .000 | .001 | .010 |
| 74 | | | | | | | | .001 | .008 |
| 76 | | | | | | | | .001 | .007 |
| 78 | | | | | | | | .000 | .006 |
| 80 | | | | | | | | | .005 |
| 82 | | | | | | | | | .004 |
| 84 | | | | | | | | | .003 |
| 86 | | | | | | | | | .002 |
| 88 | | | | | | | | | .002 |
| 90 | | | | | | | | | .002 |
| 92 | | | | | | | | | .001 |
| 94 | | | | | | | | | .001 |
| 96 | | | | | | | | | .001 |
| 98 | | | | | | | | | .001 |
| 100 | | | | | | | | | .000 |

**TABLE 18 Critical Values of *C* for the Theil Zero-Slope Test (continued)**

| | *n* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *x* | 21 | 24 | 25 | 28 | 29 | 32 | 33 | 36 | 37 | 40 |
| 0 | .512 | .510 | .509 | .508 | .507 | .506 | .506 | .505 | .505 | .505 |
| 2 | .488 | .490 | .491 | .492 | .493 | .494 | .494 | .495 | .495 | .495 |
| 4 | .464 | .471 | .472 | .477 | .478 | .481 | .482 | .484 | .484 | .486 |
| 6 | .441 | .451 | .454 | .461 | .463 | .468 | .469 | .473 | .474 | .477 |
| 8 | .417 | .432 | .436 | .446 | .448 | .455 | .457 | .462 | .464 | .468 |
| 10 | .394 | .413 | .418 | .430 | .434 | .442 | .445 | .452 | .453 | .459 |
| 12 | .371 | .394 | .400 | .415 | .419 | .430 | .433 | .441 | .443 | .449 |
| 14 | .349 | .375 | .382 | .400 | .405 | .417 | .421 | .430 | .433 | .440 |
| 16 | .327 | .356 | .364 | .385 | .390 | .405 | .409 | .420 | .423 | .431 |
| 18 | .306 | .338 | .347 | .370 | .376 | .392 | .397 | .409 | .413 | .422 |
| 20 | .285 | .320 | .330 | .355 | .362 | .380 | .385 | .399 | .403 | .413 |
| 22 | .265 | .303 | .314 | .341 | .348 | .368 | .373 | .388 | .393 | .404 |
| 24 | .246 | .286 | .297 | .326 | .334 | .356 | .362 | .378 | .383 | .395 |
| 26 | .228 | .270 | .282 | .312 | .321 | .344 | .350 | .368 | .373 | .386 |
| 28 | .210 | .254 | .266 | .298 | .308 | .332 | .339 | .358 | .363 | .377 |
| 30 | .193 | .238 | .251 | .285 | .295 | .320 | .328 | .347 | .353 | .369 |
| 32 | .177 | .223 | .237 | .272 | .282 | .309 | .317 | .338 | .344 | .360 |
| 34 | .162 | .209 | .222 | .259 | .270 | .298 | .306 | .328 | .334 | .351 |
| 36 | .147 | .195 | .209 | .246 | .257 | .287 | .295 | .318 | .325 | .343 |
| 38 | .134 | .181 | .196 | .234 | .246 | .276 | .285 | .308 | .315 | .334 |
| 40 | .121 | .169 | .183 | .222 | .234 | .265 | .274 | .299 | .306 | .326 |
| 42 | .109 | .156 | .171 | .211 | .223 | .255 | .264 | .290 | .297 | .318 |
| 44 | .098 | .145 | .159 | .200 | .212 | .244 | .254 | .280 | .288 | .309 |
| 46 | .088 | .134 | .148 | .189 | .201 | .234 | .244 | .271 | .279 | .301 |
| 48 | .079 | .123 | .138 | .178 | .191 | .224 | .235 | .262 | .271 | .293 |
| 50 | .070 | .113 | .128 | .168 | .181 | .215 | .225 | .254 | .262 | .285 |
| 52 | .062 | .104 | .118 | .158 | .171 | .206 | .216 | .245 | .254 | .277 |
| 54 | .055 | .095 | .109 | .149 | .162 | .197 | .207 | .237 | .245 | .270 |
| 56 | .049 | .087 | .101 | .140 | .153 | .188 | .199 | .228 | .237 | .262 |
| 58 | .043 | .079 | .093 | .131 | .144 | .179 | .190 | .220 | .229 | .255 |
| 60 | .037 | .072 | .085 | .123 | .136 | .171 | .182 | .212 | .222 | .247 |
| 62 | .032 | .066 | .078 | .115 | .128 | .163 | .174 | .204 | .214 | .240 |
| 64 | .028 | .059 | .071 | .108 | .120 | .155 | .166 | .197 | .206 | .233 |
| 66 | .024 | .054 | .065 | .101 | .112 | .147 | .158 | .189 | .199 | .226 |
| 68 | .021 | .048 | .059 | .094 | .105 | .140 | .151 | .182 | .192 | .219 |
| 70 | .018 | .044 | .054 | .087 | .099 | .133 | .144 | .175 | .185 | .212 |
| 72 | .015 | .039 | .049 | .081 | .092 | .126 | .137 | .168 | .178 | .205 |
| 74 | .013 | .035 | .044 | .075 | .086 | .119 | .130 | .161 | .171 | .199 |
| 76 | .011 | .031 | .040 | .070 | .080 | .113 | .124 | .155 | .165 | .192 |
| 78 | .009 | .028 | .036 | .065 | .075 | .107 | .117 | .148 | .158 | .186 |
| 80 | .008 | .025 | .032 | .060 | .070 | .101 | .111 | .142 | .152 | .180 |
| 82 | .007 | .022 | .029 | .055 | .065 | .095 | .106 | .136 | .146 | .174 |
| 84 | .005 | .019 | .026 | .051 | .060 | .090 | .100 | .130 | .140 | .168 |
| 86 | .005 | .017 | .023 | .047 | .056 | .085 | .095 | .124 | .134 | .162 |
| 88 | .004 | .015 | .021 | .043 | .052 | .080 | .090 | .119 | .129 | .156 |
| 90 | .003 | .013 | .018 | .039 | .048 | .075 | .085 | .114 | .123 | .151 |
| 92 | .002 | .011 | .016 | .036 | .044 | .070 | .080 | .108 | .118 | .146 |
| 94 | .002 | .010 | .014 | .033 | .041 | .066 | .075 | .103 | .113 | .140 |
| 96 | .002 | .009 | .013 | .030 | .037 | .062 | .071 | .099 | .108 | .135 |
| 98 | .001 | .007 | .011 | .027 | .034 | .058 | .067 | .094 | .103 | .130 |
| 100 | .001 | .006 | .010 | .025 | .031 | .054 | .063 | .089 | .098 | .125 |

**TABLE 18 Critical Values of *C* for the Theil Zero-Slope Test (continued)**

| | | | | | *n* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | 6 | 7 | 10 | 11 | 14 | 15 | 18 | 19 | 22 |
| 1 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 |
| 3 | .360 | .386 | .431 | .440 | .457 | .461 | .470 | .473 | .478 |
| 5 | .235 | .281 | .364 | .381 | .415 | .423 | .441 | .445 | .456 |
| 7 | .136 | .191 | .300 | .324 | .374 | .385 | .411 | .418 | .434 |
| 9 | .068 | .119 | .242 | .271 | .334 | .349 | .383 | .391 | .412 |
| 11 | .028 | .068 | .190 | .223 | .295 | .313 | .354 | .365 | .390 |
| 13 | .008 | .035 | .146 | .179 | .259 | .279 | .327 | .339 | .369 |
| 15 | .001 | .015 | .108 | .141 | .225 | .248 | .300 | .314 | .348 |
| 17 | | .005 | .078 | .109 | .194 | .218 | .275 | .290 | .328 |
| 19 | | .001 | .054 | .082 | .165 | .190 | .250 | .267 | .308 |
| 21 | | .000 | .036 | .060 | .140 | .164 | .227 | .245 | .289 |
| 23 | | | .023 | .043 | .117 | .141 | .205 | .223 | .270 |
| 25 | | | .014 | .030 | .096 | .120 | .184 | .203 | .252 |
| 27 | | | .008 | .020 | .079 | .101 | .165 | .184 | .234 |
| 29 | | | .005 | .013 | .063 | .084 | .147 | .166 | .217 |
| 31 | | | .002 | .008 | .050 | .070 | .130 | .149 | .201 |
| 33 | | | .001 | .005 | .040 | .057 | .115 | .133 | .186 |
| 35 | | | .000 | .003 | .031 | .046 | .100 | .119 | .171 |
| 37 | | | | .002 | .024 | .037 | .088 | .105 | .157 |
| 39 | | | | .001 | .018 | .029 | .076 | .093 | .144 |
| 41 | | | | .000 | .013 | .023 | .066 | .082 | .131 |
| 43 | | | | | .010 | .018 | .056 | .072 | .120 |
| 45 | | | | | .007 | .014 | .048 | .062 | .109 |
| 47 | | | | | .005 | .010 | .041 | .054 | .099 |
| 49 | | | | | .003 | .008 | .034 | .047 | .089 |
| 51 | | | | | .002 | .006 | .029 | .040 | .080 |
| 53 | | | | | .002 | .004 | .024 | .034 | .072 |
| 55 | | | | | .001 | .003 | .020 | .029 | .064 |
| 57 | | | | | .001 | .002 | .016 | .025 | .058 |
| 59 | | | | | .000 | .001 | .013 | .021 | .051 |
| 61 | | | | | | .001 | .011 | .017 | .045 |
| 63 | | | | | | .001 | .009 | .014 | .040 |
| 65 | | | | | | .000 | .007 | .012 | .035 |
| 67 | | | | | | | .005 | .010 | .031 |
| 69 | | | | | | | .004 | .008 | .027 |
| 71 | | | | | | | .003 | .006 | .024 |
| 73 | | | | | | | .003 | .005 | .021 |
| 75 | | | | | | | .002 | .004 | .018 |
| 77 | | | | | | | .001 | .003 | .015 |
| 79 | | | | | | | .001 | .003 | .013 |
| 81 | | | | | | | .001 | .002 | .011 |
| 83 | | | | | | | .001 | .002 | .010 |
| 85 | | | | | | | .000 | .001 | .008 |
| 87 | | | | | | | | .001 | .007 |
| 89 | | | | | | | | .001 | .006 |
| 91 | | | | | | | | .001 | .005 |
| 93 | | | | | | | | .000 | .004 |
| 95 | | | | | | | | | .003 |
| 97 | | | | | | | | | .003 |
| 99 | | | | | | | | | .002 |
| 101 | | | | | | | | | .002 |

**TABLE 18 Critical Values of $C$ for the Theil Zero-Slope Test (continued)**

| | $n$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ | 23 | 26 | 27 | 30 | 31 | 34 | 35 | 38 | 39 |
| 1 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 |
| 3 | .479 | .483 | .484 | .486 | .487 | .488 | .489 | .490 | .490 |
| 5 | .458 | .465 | .467 | .472 | .473 | .477 | .478 | .480 | .481 |
| 7 | .438 | .448 | .451 | .458 | .460 | .465 | .466 | .470 | .472 |
| 9 | .417 | .431 | .434 | .444 | .446 | .453 | .455 | .460 | .462 |
| 11 | .397 | .414 | .418 | .430 | .433 | .442 | .444 | .450 | .452 |
| 13 | .377 | .397 | .402 | .416 | .420 | .430 | .433 | .440 | .443 |
| 15 | .357 | .380 | .386 | .402 | .407 | .418 | .422 | .431 | .433 |
| 17 | .338 | .363 | .371 | .389 | .394 | .407 | .411 | .421 | .424 |
| 19 | .319 | .347 | .355 | .375 | .381 | .396 | .400 | .411 | .414 |
| 21 | .301 | .331 | .340 | .362 | .368 | .384 | .389 | .401 | .405 |
| 23 | .283 | .316 | .325 | .349 | .355 | .373 | .378 | .392 | .396 |
| 25 | .265 | .300 | .310 | .336 | .343 | .362 | .368 | .382 | .387 |
| 27 | .248 | .285 | .296 | .323 | .331 | .351 | .357 | .373 | .377 |
| 29 | .232 | .270 | .281 | .310 | .318 | .340 | .347 | .363 | .368 |
| 31 | .216 | .256 | .268 | .298 | .306 | .329 | .336 | .354 | .359 |
| 33 | .201 | .242 | .254 | .286 | .295 | .319 | .326 | .345 | .350 |
| 35 | .187 | .229 | .241 | .274 | .283 | .308 | .316 | .336 | .341 |
| 37 | .173 | .216 | .228 | .262 | .272 | .298 | .306 | .327 | .333 |
| 39 | .160 | .203 | .216 | .251 | .261 | .288 | .296 | .318 | .324 |
| 41 | .147 | .191 | .204 | .239 | .250 | .278 | .286 | .309 | .315 |
| 43 | .135 | .179 | .192 | .228 | .239 | .268 | .277 | .300 | .307 |
| 45 | .124 | .168 | .181 | .218 | .229 | .259 | .267 | .291 | .298 |
| 47 | .114 | .157 | .170 | .208 | .219 | .249 | .258 | .283 | .290 |
| 49 | .104 | .147 | .160 | .198 | .209 | .240 | .249 | .274 | .282 |
| 51 | .094 | .137 | .150 | .188 | .199 | .231 | .240 | .266 | .274 |
| 53 | .086 | .127 | .141 | .178 | .190 | .222 | .232 | .258 | .266 |
| 55 | .078 | .118 | .132 | .169 | .181 | .213 | .223 | .250 | .258 |
| 57 | .070 | .110 | .123 | .160 | .172 | .205 | .215 | .242 | .250 |
| 59 | .063 | .102 | .115 | .152 | .164 | .196 | .206 | .234 | .243 |
| 61 | .057 | .094 | .107 | .144 | .155 | .188 | .198 | .227 | .235 |
| 63 | .051 | .087 | .099 | .136 | .147 | .180 | .191 | .219 | .228 |
| 65 | .046 | .080 | .092 | .128 | .140 | .173 | .183 | .212 | .221 |
| 67 | .041 | .073 | .085 | .121 | .132 | .165 | .176 | .205 | .214 |
| 69 | .036 | .067 | .079 | .114 | .125 | .158 | .168 | .198 | .207 |
| 71 | .032 | .062 | .073 | .107 | .118 | .151 | .161 | .191 | .200 |
| 73 | .028 | .057 | .067 | .100 | .112 | .144 | .154 | .184 | .193 |
| 75 | .025 | .052 | .062 | .094 | .105 | .137 | .148 | .177 | .187 |
| 77 | .022 | .047 | .057 | .088 | .099 | .131 | .141 | .171 | .180 |
| 79 | .019 | .043 | .052 | .083 | .093 | .125 | .135 | .165 | .174 |
| 81 | .017 | .039 | .048 | .077 | .088 | .119 | .129 | .158 | .168 |
| 83 | .015 | .035 | .044 | .072 | .082 | .113 | .123 | .152 | .162 |
| 85 | .013 | .032 | .040 | .067 | .077 | .107 | .117 | .147 | .156 |
| 87 | .011 | .029 | .036 | .063 | .072 | .102 | .112 | .141 | .150 |
| 89 | .009 | .026 | .033 | .059 | .068 | .097 | .107 | .135 | .145 |
| 91 | .008 | .023 | .030 | .054 | .063 | .092 | .101 | .130 | .139 |
| 93 | .007 | .021 | .027 | .051 | .059 | .087 | .096 | .125 | .134 |
| 95 | .006 | .019 | .025 | .047 | .055 | .082 | .092 | .120 | .129 |
| 97 | .005 | .017 | .022 | .043 | .052 | .078 | .087 | .115 | .124 |
| 99 | .004 | .015 | .020 | .040 | .048 | .074 | .083 | .110 | .119 |
| 101 | .004 | .013 | .018 | .037 | .045 | .070 | .078 | .105 | .114 |

**TABLE 19 Factors Used When Constructing Control Charts**

| *Number of Observations in Sample* *n* | *Chart for Averages* $A_2$ | $d_2$ | *Chart for Ranges* $d_3$ | $D_3$ | $D_4$ |
|---|---|---|---|---|---|
| 2 | 1.880 | 1.128 | .853 | 0 | 3.276 |
| 3 | 1.023 | 1.693 | .888 | 0 | 2.575 |
| 4 | .729 | 2.059 | .880 | 0 | 2.282 |
| 5 | .577 | 2.326 | .864 | 0 | 2.115 |
| 6 | .483 | 2.534 | .848 | 0 | 2.004 |
| 7 | .419 | 2.704 | .833 | .076 | 1.924 |
| 8 | .373 | 2.847 | .820 | .136 | 1.864 |
| 9 | .337 | 2.970 | .808 | .184 | 1.816 |
| 10 | .308 | 3.078 | .797 | .223 | 1.777 |
| 11 | .285 | 3.173 | .787 | .256 | 1.744 |
| 12 | .266 | 3.258 | .778 | .284 | 1.719 |
| 13 | .249 | 3.336 | .770 | .308 | 1.692 |
| 14 | .235 | 3.407 | .762 | .329 | 1.671 |
| 15 | .223 | 3.472 | .755 | .348 | 1.652 |
| 16 | .212 | 3.532 | .749 | .364 | 1.636 |
| 17 | .203 | 3.588 | .743 | .379 | 1.621 |
| 18 | .194 | 3.640 | .738 | .392 | 1.608 |
| 19 | .187 | 3.689 | .733 | .404 | 1.596 |
| 20 | .180 | 3.735 | .729 | .414 | 1.586 |
| 21 | .173 | 3.778 | .724 | .425 | 1.575 |
| 22 | .167 | 3.819 | .720 | .434 | 1.566 |
| 23 | .162 | 3.858 | .716 | .443 | 1.557 |
| 24 | .157 | 3.895 | .712 | .452 | 1.548 |
| 25 | .153 | 3.931 | .709 | .459 | 1.541 |

*Source: ASTM Manual on Quality Control of Materials*, American Society for Testing Materials, Philadelphia, PA, 1951. Copyright ASTM. Reprinted with permission.

**TABLE 20 Values of $K$ for Tolerance Limits for Normal Distributions**

| $\gamma$ / $n$ | $1-\alpha = .95$ | | | $1-\alpha = .99$ | | |
|---|---|---|---|---|---|---|
| | .90 | .95 | .99 | .90 | .95 | .99 |
| 2 | 32.019 | 37.674 | 48.430 | 160.193 | 188.491 | 242.300 |
| 3 | 8.380 | 9.916 | 12.861 | 18.930 | 22.401 | 29.055 |
| 4 | 5.369 | 6.370 | 8.299 | 9.398 | 11.150 | 14.527 |
| 5 | 4.275 | 5.079 | 6.634 | 6.612 | 7.855 | 10.260 |
| 6 | 3.712 | 4.414 | 5.775 | 5.337 | 6.345 | 8.301 |
| 7 | 3.369 | 4.007 | 5.248 | 4.613 | 5.488 | 7.187 |
| 8 | 3.136 | 3.732 | 4.891 | 4.147 | 4.936 | 6.468 |
| 9 | 2.967 | 3.532 | 4.631 | 3.822 | 4.550 | 5.966 |
| 10 | 2.839 | 3.379 | 4.433 | 3.582 | 4.265 | 5.594 |
| 11 | 2.737 | 3.259 | 4.277 | 3.397 | 4.045 | 5.308 |
| 12 | 2.655 | 3.162 | 4.150 | 3.250 | 3.870 | 5.079 |
| 13 | 2.587 | 3.081 | 4.044 | 3.130 | 3.727 | 4.893 |
| 14 | 2.529 | 3.012 | 3.955 | 3.029 | 3.608 | 4.737 |
| 15 | 2.480 | 2.954 | 3.878 | 2.945 | 3.507 | 4.605 |
| 16 | 2.437 | 2.903 | 3.812 | 2.872 | 3.421 | 4.492 |
| 17 | 2.400 | 2.858 | 3.754 | 2.808 | 3.345 | 4.393 |
| 18 | 2.366 | 2.819 | 3.702 | 2.753 | 3.279 | 4.307 |
| 19 | 2.337 | 2.784 | 3.656 | 2.703 | 3.221 | 4.230 |
| 20 | 2.310 | 2.752 | 3.615 | 2.659 | 3.168 | 4.161 |
| 25 | 2.208 | 2.631 | 3.457 | 2.494 | 2.972 | 3.904 |
| 30 | 2.140 | 2.549 | 3.350 | 2.385 | 2.841 | 3.733 |
| 35 | 2.090 | 2.490 | 3.272 | 2.306 | 2.748 | 3.611 |
| 40 | 2.052 | 2.445 | 3.213 | 2.247 | 2.677 | 3.518 |
| 45 | 2.021 | 2.408 | 3.165 | 2.200 | 2.621 | 3.444 |
| 50 | 1.996 | 2.379 | 3.126 | 2.162 | 2.576 | 3.385 |
| 55 | 1.976 | 2.354 | 3.094 | 2.130 | 2.538 | 3.335 |
| 60 | 1.958 | 2.333 | 3.066 | 2.103 | 2.506 | 3.293 |
| 65 | 1.943 | 2.315 | 3.042 | 2.080 | 2.478 | 3.257 |
| 70 | 1.929 | 2.299 | 3.021 | 2.060 | 2.454 | 3.225 |
| 75 | 1.917 | 2.285 | 3.002 | 2.042 | 2.433 | 3.197 |
| 80 | 1.907 | 2.272 | 2.986 | 2.026 | 2.414 | 3.173 |
| 85 | 1.897 | 2.261 | 2.971 | 2.012 | 2.397 | 3.150 |
| 90 | 1.889 | 2.251 | 2.958 | 1.999 | 2.382 | 3.130 |
| 95 | 1.881 | 2.241 | 2.945 | 1.987 | 2.368 | 3.112 |
| 100 | 1.874 | 2.233 | 2.934 | 1.977 | 2.355 | 3.096 |
| 150 | 1.825 | 2.175 | 2.859 | 1.905 | 2.270 | 2.983 |
| 200 | 1.798 | 2.143 | 2.816 | 1.865 | 2.222 | 2.921 |
| 250 | 1.780 | 2.121 | 2.788 | 1.839 | 2.191 | 2.880 |
| 300 | 1.767 | 2.106 | 2.767 | 1.820 | 2.169 | 2.850 |
| 400 | 1.749 | 2.084 | 2.739 | 1.794 | 2.138 | 2.809 |
| 500 | 1.737 | 2.070 | 2.721 | 1.777 | 2.117 | 2.783 |
| 600 | 1.729 | 2.060 | 2.707 | 1.764 | 2.102 | 2.763 |
| 700 | 1.722 | 2.052 | 2.697 | 1.755 | 2.091 | 2.748 |
| 800 | 1.717 | 2.046 | 2.688 | 1.747 | 2.082 | 2.736 |
| 900 | 1.712 | 2.040 | 2.682 | 1.741 | 2.075 | 2.726 |
| 1000 | 1.709 | 2.036 | 2.676 | 1.736 | 2.068 | 2.718 |
| $\infty$ | 1.645 | 1.960 | 2.576 | 1.645 | 1.960 | 2.576 |

*Source:* From *Techniques of Statistical Analysis* by C. Eisenhart, M. W. Hastay, and W. A. Wallis. 

**TABLE 21 Sample Size $n$ for Nonparametric Tolerance Limits**

| | $1-\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| $\gamma$ | .50 | .70 | .90 | .95 | .99 | .995 |
| .995 | 336 | 488 | 777 | 947 | 1,325 | 1,483 |
| .99 | 168 | 244 | 388 | 473 | 662 | 740 |
| .95 | 34 | 49 | 77 | 93 | 130 | 146 |
| .90 | 17 | 24 | 38 | 46 | 64 | 72 |
| .85 | 11 | 16 | 25 | 30 | 42 | 47 |
| .80 | 9 | 12 | 18 | 22 | 31 | 34 |
| .75 | 7 | 10 | 15 | 18 | 24 | 27 |
| .70 | 6 | 8 | 12 | 14 | 20 | 22 |
| .60 | 4 | 6 | 9 | 10 | 14 | 16 |
| .50 | 3 | 5 | 7 | 8 | 11 | 12 |

*Source:* Tables A-25d of Wilfrid J. Dixon and Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3rd ed., McGraw-Hill Book Company, New York, 1969. Used with permission of McGraw-Hill Book Company.

**TABLE 22 Sample Size Code Letters: MIL-STD-105D**

| | Special Inspection Levels | | | | General Inspection Levels | | |
|---|---|---|---|---|---|---|---|
| Lot of Batch Size | S-1 | S-2 | S-3 | S-4 | I | II | III |
| 2–8 | A | A | A | A | A | A | B |
| 9–15 | A | A | A | A | A | B | C |
| 16–25 | A | A | B | B | B | C | D |
| 26–50 | A | B | B | C | C | D | E |
| 51–90 | B | B | C | C | C | E | F |
| 91–150 | B | B | C | D | D | F | G |
| 151–280 | B | C | D | E | E | G | H |
| 281–500 | B | C | D | E | F | H | J |
| 501–1,200 | C | C | E | F | G | J | K |
| 1,201–3,200 | C | D | E | G | H | K | L |
| 3,201–10,000 | C | D | F | G | J | L | M |
| 10,001–35,000 | C | D | F | H | K | M | N |
| 35,001–150,000 | D | E | G | J | L | N | P |
| 150,001–500,000 | D | E | G | J | M | P | Q |
| 500,001 and over | D | E | H | K | N | Q | R |

**TABLE 23 A Portion of the Master Table for Normal Inspection (Single Sampling): MIL-STD-105D**

| Sample Size Code Letter | Sample Size | Acceptable Quality Level (Normal Inspection) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | .010 | .015 | .025 | .040 | .065 | .10 | .15 | .25 | .40 | .65 | 1.0 | 1.5 | 2.5 | 4.0 | 6.5 | 10 | 15 | 25 | 40 | 65 |
| | | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re | Ac Re |
| A | 2 | | | | | | | | | | | | | | ↓ | 0 1 | | ↓ | 1 2 | 2 3 | 3 4 |
| B | 3 | | | | | | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 |
| C | 5 | | | | | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 |
| D | 8 | | | | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 |
| E | 13 | | | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 |
| F | 20 | | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 |
| G | 32 | | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ |
| H | 50 | | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | |
| J | 80 | | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | |
| K | 125 | | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | |
| L | 200 | | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | |
| M | 315 | | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | | |
| N | 500 | | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | | | |
| P | 800 | ↓ | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | | | | |
| Q | 1,250 | 0 1 | ↑ | ↓ | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | | | | | |
| R | 2,000 | ↑ | | 1 2 | 2 3 | 3 4 | 5 6 | 7 8 | 10 11 | 14 15 | 21 22 | ↑ | | | | | | | | | |

↓ = Use first sampling plan below arrow. If sample size equals or exceeds lot or batch size, do 100% inspection.

↑ = Use first sampling plan above arrow.

Ac = Acceptance number.

Re = Rejection number.

APPENDIX C

# SAS FOR WINDOWS TUTORIAL

CONTENTS



## C.1 SAS Windows Environment

Upon entering into a SAS session, you will see a screen similar to Figure C.1. The window at the bottom of the screen is the SAS Editor window. The SAS program commands for creating and analyzing data are specified in this window. The window at the top of the screen is the SAS Log window, which logs whether or not each command line has been successfully executed. Once a program is run, a third window appears – the SAS Output window. This window will show the results of the analysis. The SAS printouts shown throughout this text appear in the SAS Output window.

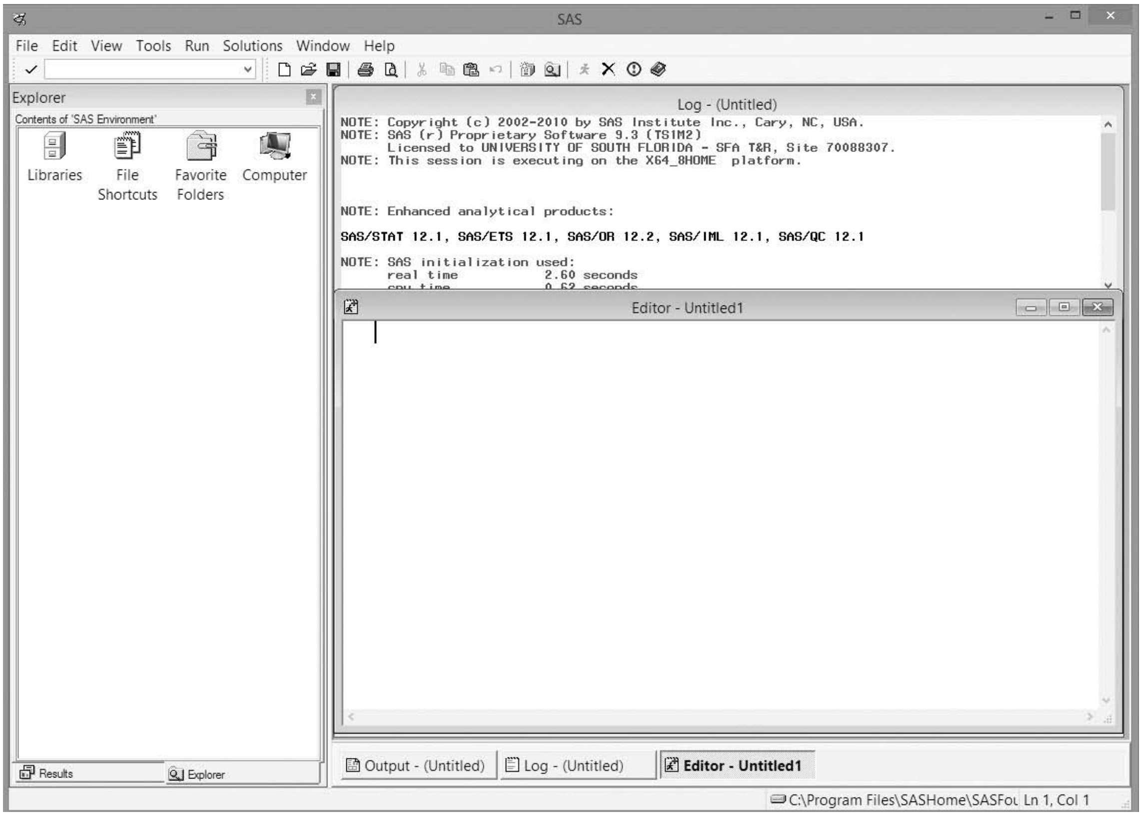



**FIGURE C.1**
Initial Screen Viewed by SAS 9.3 Windows User

## C.2 Creating a SAS Data Set Ready for Analysis

In the SAS Editor window, three basic types of instructions (commands) are utilized. (Note: All commands, except for input data values, end with a semi-colon in SAS.)

1. *DATA commands:* instructions on how the data will be accessed or entered
2. *Input data values:* the values of the variables in the data set
3. *Statistical procedural (PROC) commands:* instruction on what type of analysis is to be conducted on the data

Data sets to be analyzed are referenced with DATA commands in one of three different ways:

1. Data values entered directly into the window using an INPUT statement
2. External data sets accessed using the INFILE statement
3. Previously created SAS data files accessed using the LIBNAME and SET statement

The name of the SAS data set is specified by the user in the DATA statement.

The commands shown in Figure C.2 create a SAS data set named FUEL with direct data entry. The names of the variables (e.g., MFG, SIZE) are listed with the INPUT command. (Note: Qualitative variable names are followed by a dollar sign.) The input data values must be typed (or copied) directly into the Editor window following the DATALINES command.

```
DATA FUEL;
INPUT MODEL $ MFG $ TYPE $ SIZE CYL CITYMPG HWYMPG;
DATALINES;
TSX      Acura    Automatic   2.4   4  21  30
Jetta    VW       Automatic   2.0   4  29  40
528i     BMW      Manual      3.0   6  18  28
Fusion   Ford     Automatic   3.0   6  17  25
Camry    Toyota   Manual      2.4   4  21  31
Escalade Cadillac Automatic   6.2   8  12  19
;;;;

PROC PRINT;

RUN;
```

**FIGURE C.2**
SAS Commands for Entering Data Directly into the Editor Window

The commands shown in Figure C.3 create a SAS data set named FISH from data stored in an external data file. The INFILE command gives the folder location of the external file (called FISHDDT.DAT) and the INPUT command lists the variables (e.g., LOCATION, WEIGHT) on the data set. (Note: The program in Figure C.3 also shows how to create interaction and squared terms using the standard symbol, * , for multiplication.)

The commands shown in Figure C.4 access a SAS data (named FISHDDT) that has been previously created and saved. The LIBNAME statement gives the folder location of the SAS data file (in parentheses), identified by the user-chosen nickname DK. The SET statement gives the actual name of the SAS data file (FISHDDT) using the convention 'nickname.filename' (e.g., DK.FISHDDT).

The PRINT procedure (PROC PRINT;) is used to command SAS to create a listing of the data. To actually submit the SAS program (and obtain results in the Output window), you will need to click on the *Run* button shown on the menu bar at the top of the SAS screen. (See Figure C.1.)

```
DATA FISH;
INFILE 'C:\USERS\TERRY\DESKTOP\FISHDDT.DAT';
INPUT LOCATION $ MILE SPECIES $CHAR14. LENGTH WEIGHT DDT;
LENGTH_WT=LENGTH*WEIGHT;
LENGTHSQ=LENGTH*LENGTH;
WEIGHTSQ=WEIGHT*WEIGHT;

PROC PRINT;
RUN;
```

**FIGURE C.3**
SAS Commands for Accessing an External Data File

Program Editor - FISHPGM

```
LIBNAME DK 'C:\USERS\TERRY\DESKTOP';

DATA FISH;
SET DK.FISHDDT;

PROC PRINT;
RUN;
```

**FIGURE C.4**
SAS Commands for Accessing a SAS Data File

## C.3 Using SAS Enterprise Guide

For SAS users who are not familiar with SAS syntax, SAS has available a companion "user-friendly" menu-driven tool called SAS Enterprise Guide (SAS EG). In SAS EG, you do not need to know any SAS commands. You obtain results by simply clicking on the appropriate menu options.

Once you have entered into an EG session, you will see the screen shown in Figure C.5.

To access a SAS data set for analysis, click on the "*File*" button on the menu bar, then click on "*Open*", and "*Data*". You will see an "Open Data" screen as shown in Figure C.6. Specify the folder where the data set resides, then select the data file by double-clicking on the file name. Now the data table will appear, as shown in Figure C.7.

The variable names in the SAS data set appear at the top of each column, the actual data in the rows. Once you access the data in this fashion, you are ready to analyze it using the menu-driven features of SAS EG.

## C.4 Listing Data

To access a listing (printout) of your data using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Describe*", and "*List Data*". The resulting menu, or dialog box, appears as in Figure C.8. Move the variables you want to print into the "List variables" box on the right side of the menu. Then click "Run". The printout will show up on your screen.

## C.5 Graphing Data

To obtain graphical descriptions of your data (e.g., bar charts, histograms, scattergrams, etc.) using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Graph*", and select the type of graph (e.g., pie chart) you desire. (See Figure C.9.) One or more dialog boxes will appear requesting you make selections (e.g., variable to be graphed). Make the appropriate variable selections (see, for example, Figure C.10 for a scatterplot) and click "Run" to view the graph.

## C.6 Descriptive Statistics and Correlations

To obtain numerical descriptive measures for a quantitative variable (e.g., mean, standard deviation, etc.) using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Describe*", and "*Summary Statistics*". Move the variable you want to analyze into the "Analysis variables" box on the right side of the menu. (As an option, you can

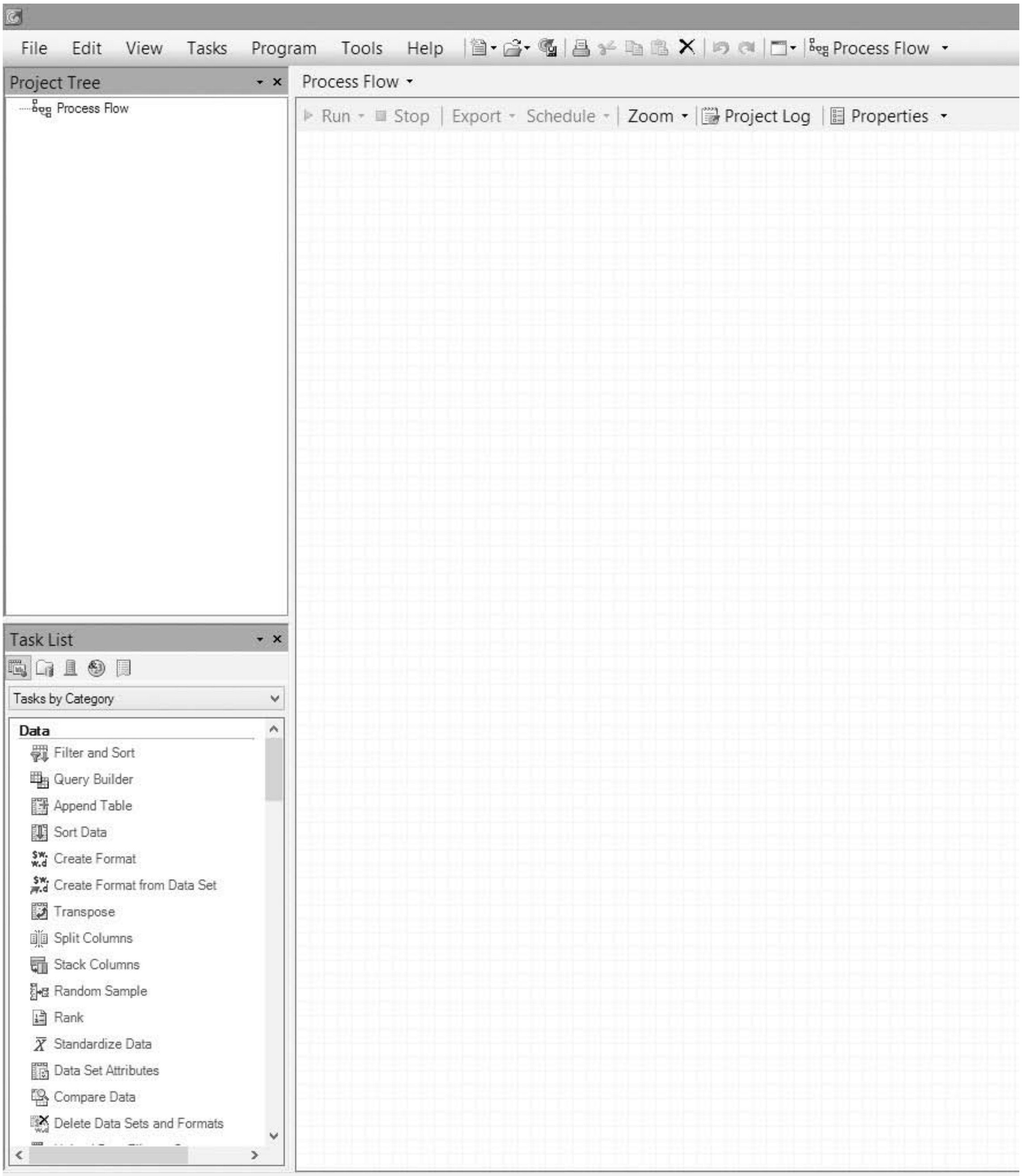
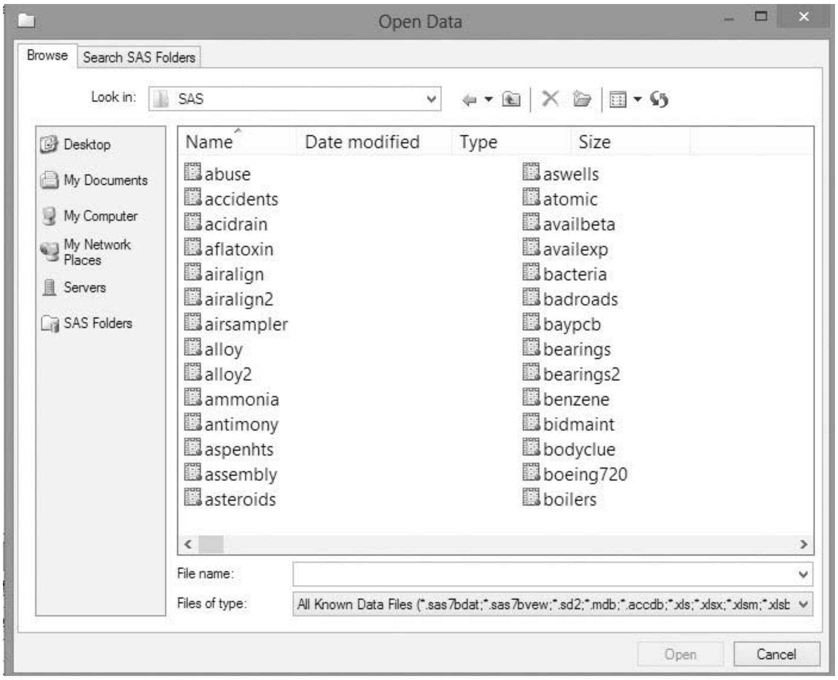



**FIGURE C.5**
Initial Screen Viewed by SAS Enterprise Guide User

obtain summary statistics on this quantitative variable for different levels of a qualitative variable by placing the qualitative variable in the "Classification variables" box on the right side — see Figure C.11). Click the "*Statistics*" button at the top left of the menu to select what descriptive statistics (e.g, mean) you want to compute. For percentiles, click the "*Percentiles*" button. After you have made your selections (see Figure C.12), click "Run". The printout will show up on your screen.

To obtain Pearson product moment correlations for pairs of quantitative variables using SAS EG, click on the "*Tasks*" button on the menu bar, then click on

**FIGURE C.6**
Selecting the SAS Data Table to Open in SAS EG



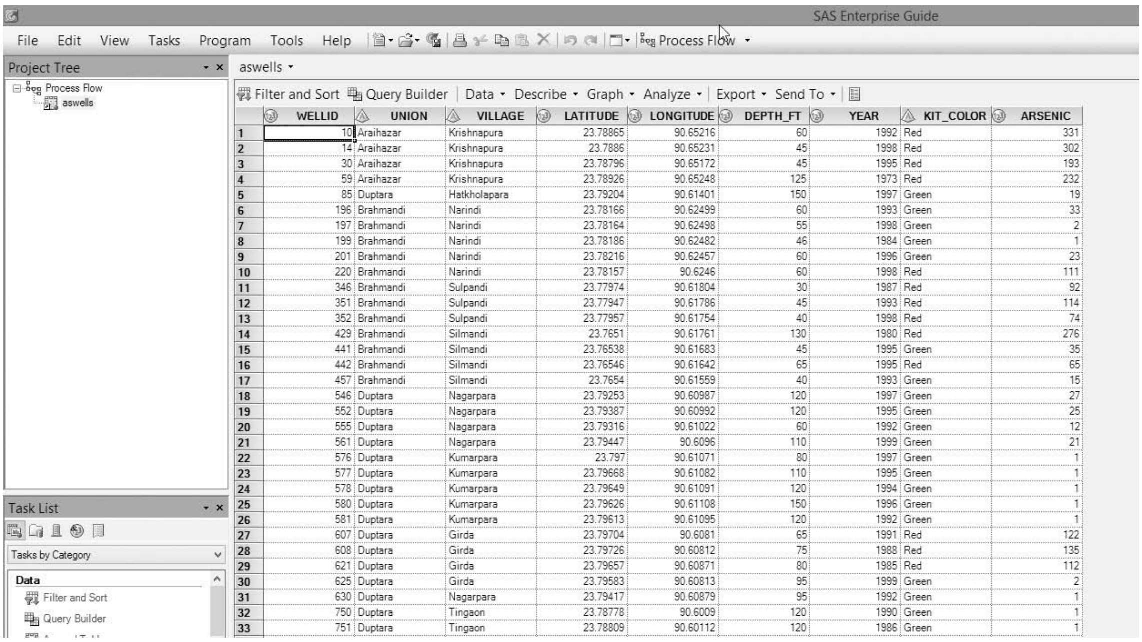




**FIGURE C.7**
SAS Data Table Opened in SAS EG

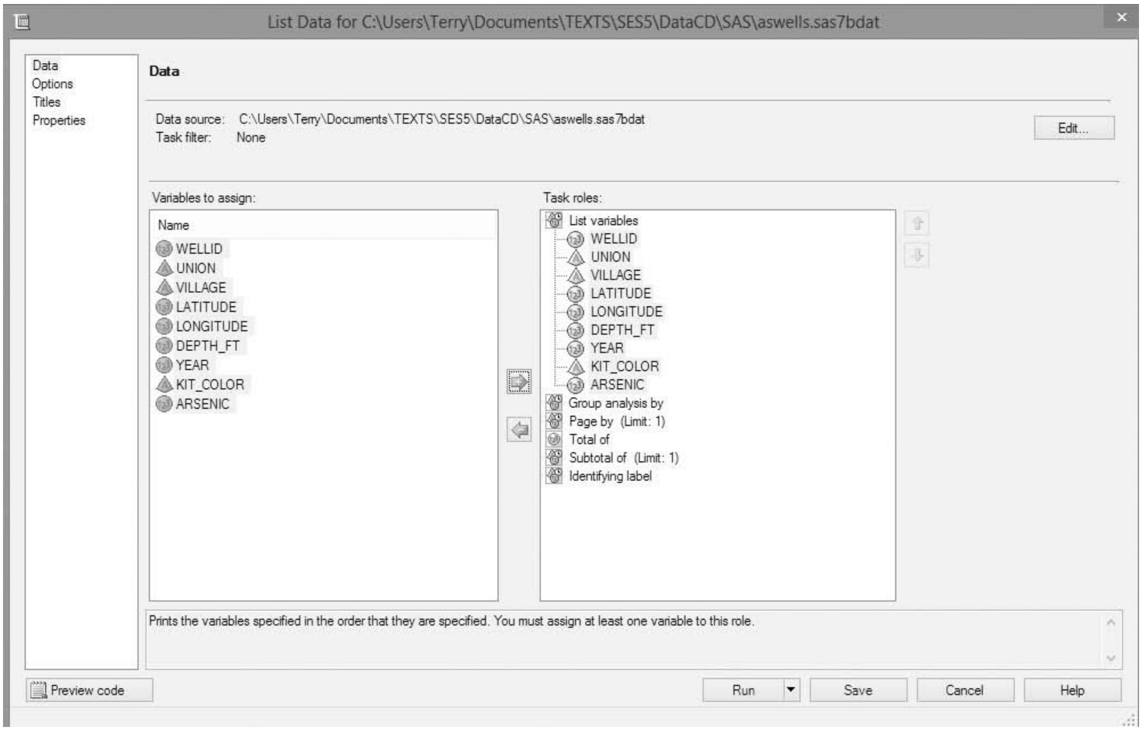



**FIGURE C.8**
SAS EG List Data Menu

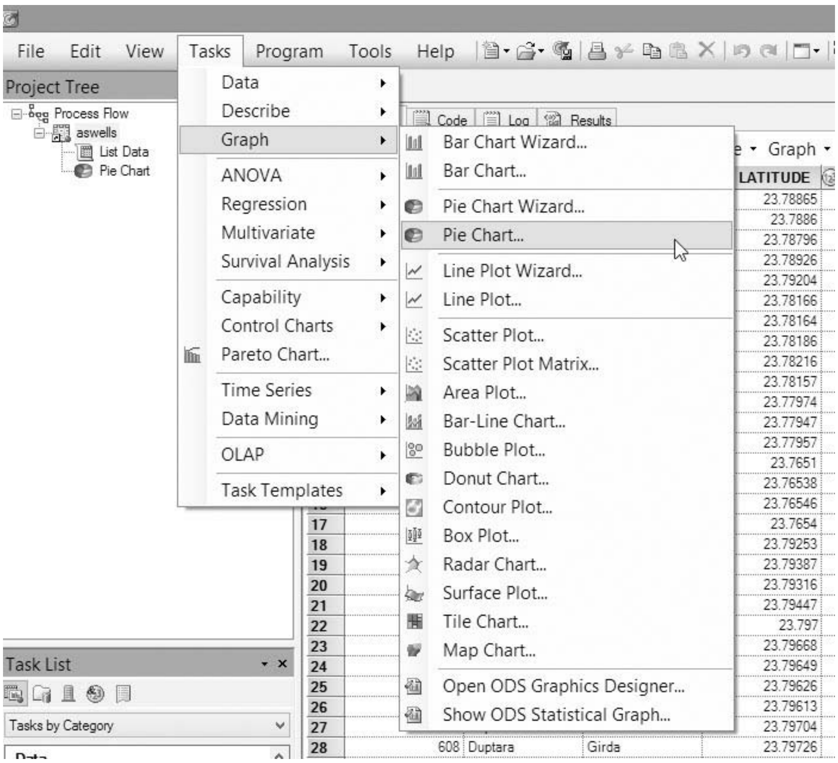



**FIGURE C.9**
SAS EG Options for Graphing Your Data

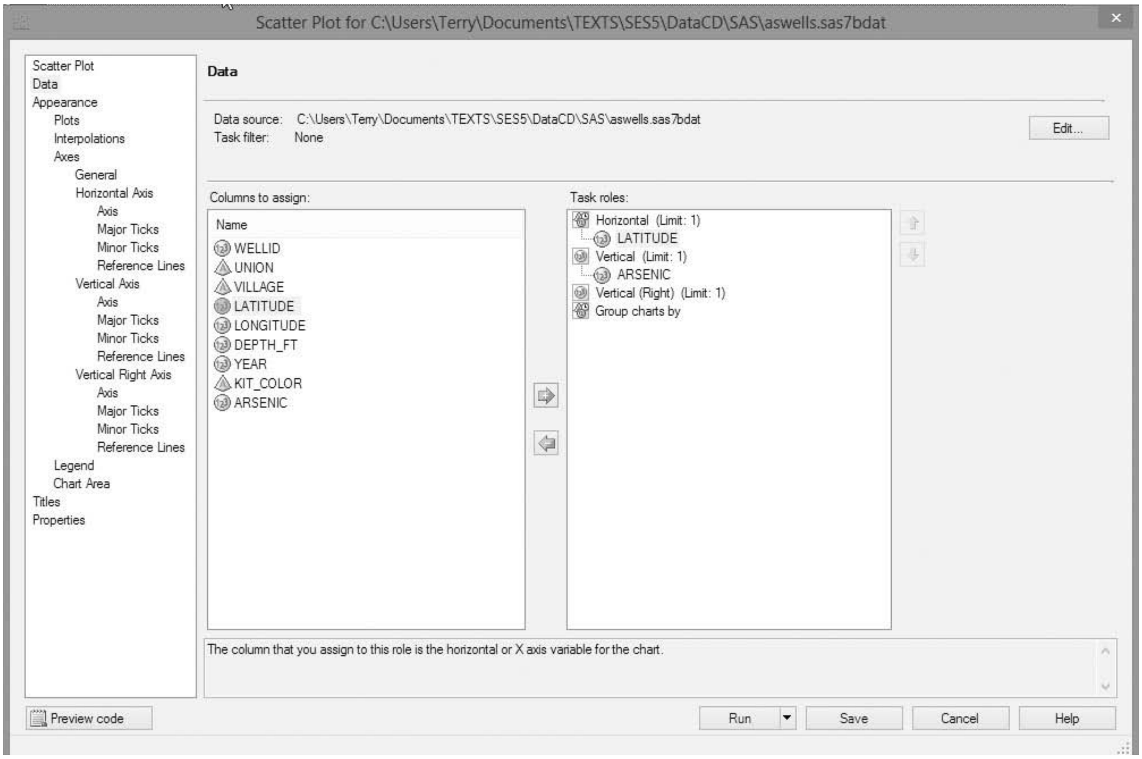



**FIGURE C.10**
SAS EG Options for Obtaining a Scatterplot

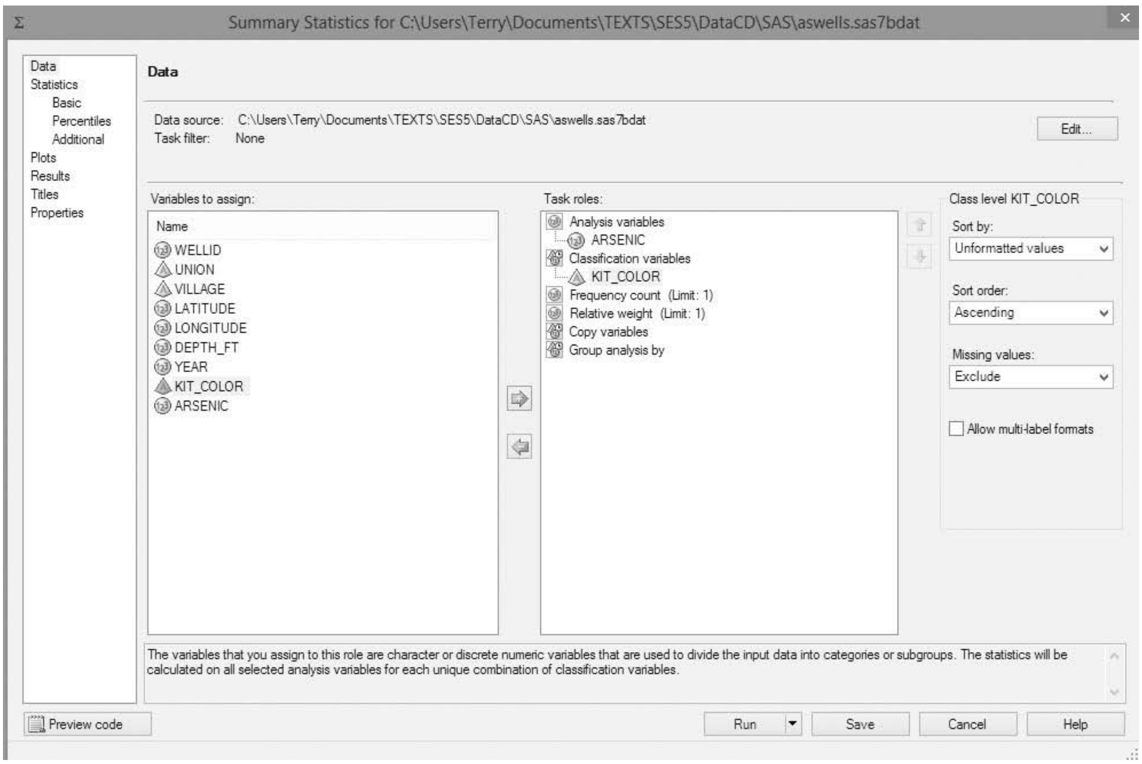



**FIGURE C.11**
SAS EG Summary Statistics Dialog Box

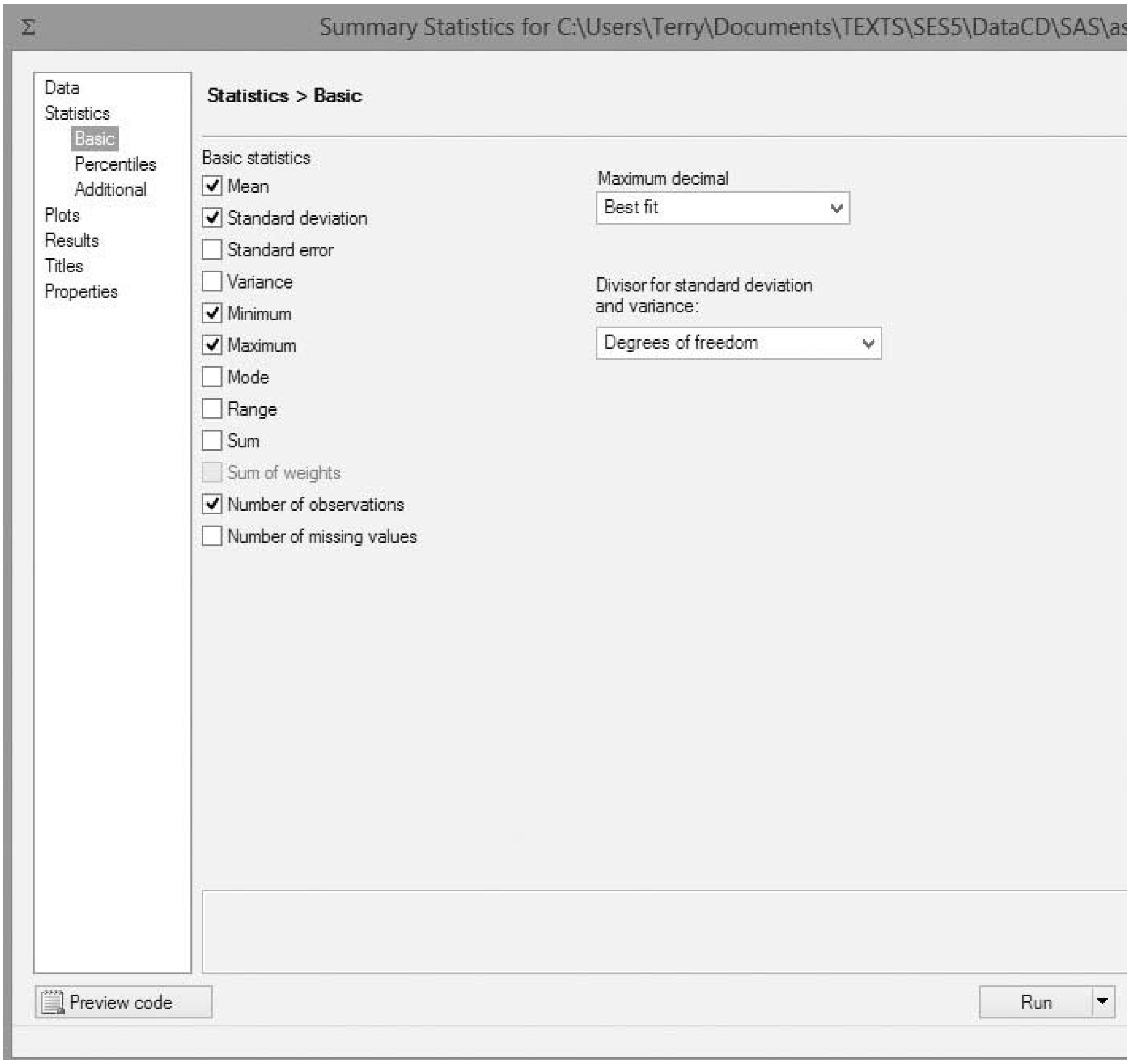



**FIGURE C.12**
SAS EG Options for Selecting Descriptive Statistics

"*Multivariate*", and "*Correlations*". (See Figure C.13.) Move the variables you want to analyze into the "Analysis variables" box on the right side of the menu. Click "Run" to obtain a printout of the correlations.

## C.7 Confidence Intervals and Hypothesis Tests for a Single Mean

To conduct a test of hypothesis and form a confidence interval for a single population mean of a quantitative variable using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*ANOVA*", and "*t Test*". (See Figure C.14.) On the resulting screen, select "*One Sample*" as the *t*-test type, then select the "*Data*" option and move the variable you want to analyze into the "Analysis variables" box on the right side of the menu. Now click the "Analysis" option. On the resulting screen (see Figure C.15), specify the null hypothesis value of the mean and the confidence level. Then click "Run" to obtain a printout of the results.

**FIGURE C.13**
SAS EG Menu Selections to Obtain Correlations

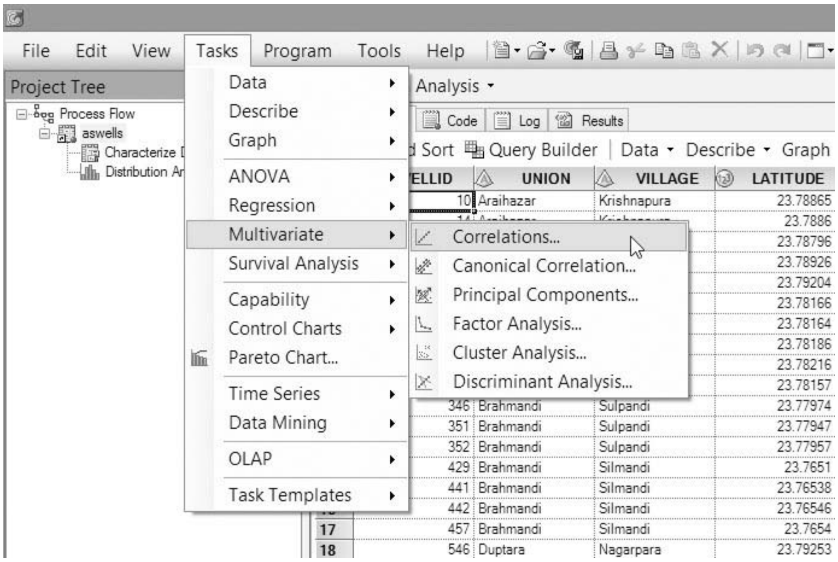
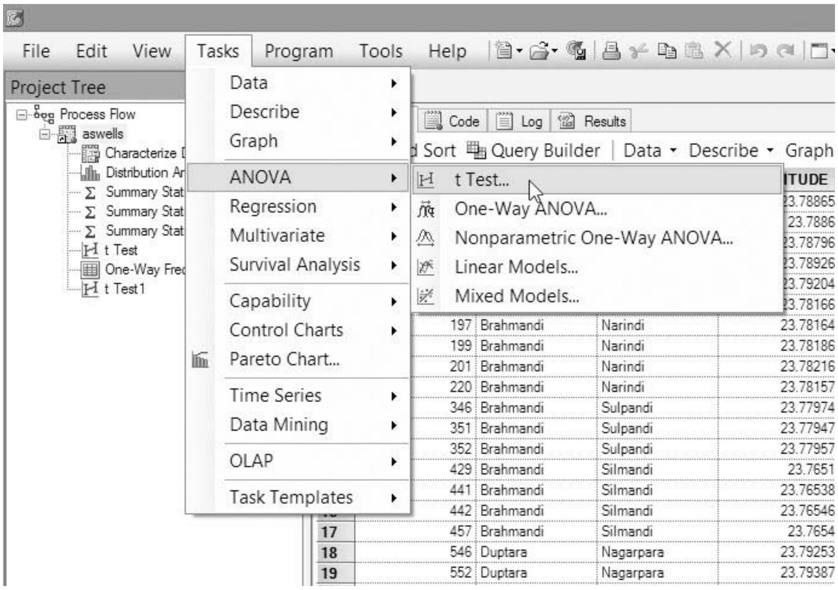



**FIGURE C.14**
SAS EG Options for Inferences on a Single Mean




## C.8 Confidence Intervals and Hypothesis Tests for the Difference Between Two Means — Independent Samples

To conduct a test of hypothesis and form a confidence interval for the difference between two population means based on independent samples, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*ANOVA*", and "*t* Test". (See, again, Figure C.14.) On the resulting screen, select "*Two Sample*" as the *t*-test type, then

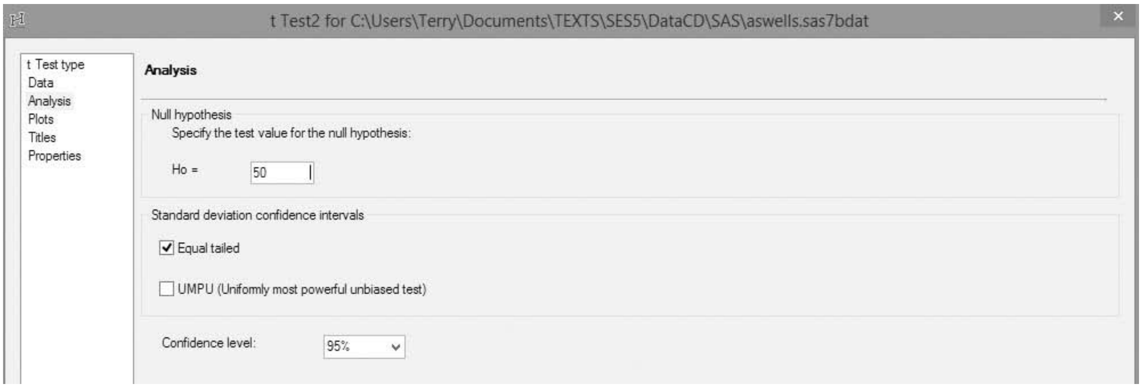



**FIGURE C.15**
SAS EG Dialog Box for Inferences on a Single Mean

select the "*Data*" option and move the quantitative variable you want to analyze into the "Analysis variables" box on the right side of the menu, and move the qualitative variable with values that represent the two populations into the "*Classification variables*" box. (See Figure C.16.) Now click the "Analysis" option. On the resulting screen, specify the null hypothesis value of the difference in means (default is 0) and the confidence level. Then click "Run" to obtain a printout of the results.

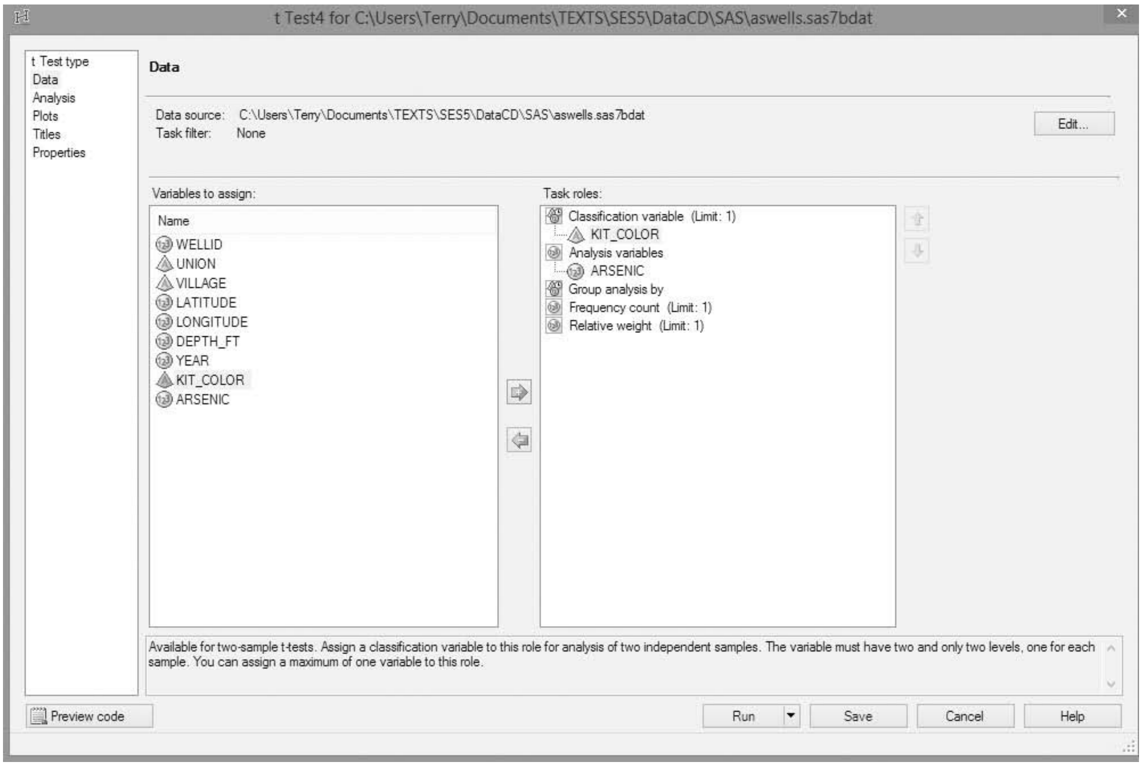



**FIGURE C.16**
SAS EG Dialog Box for Comparing Two Means

## C.9 Confidence Intervals and Hypothesis Tests for the Difference Between Two Means — Matched Pairs

To conduct a test of hypothesis and form a confidence interval for the difference between two population means based on matched pairs, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*ANOVA*", and "t Test". (See, again, Figure C.14.) On the resulting screen, select "*Paired*" as the t-test type, then select the "*Data*" option. Move the two quantitative variables that you want to compare into the "Paired variables" boxes on the right side of the menu. (See Figure C.17.) Now click the "Analysis" option. On the resulting screen, specify the null hypothesis value of the difference in means (default is 0) and the confidence level. Then click "Run" to obtain a printout of the results.

## C.10 Hypothesis Test for the Ratio of Two Variances — Independent Samples

To conduct a test of hypothesis for the ratio of two population variances based on independent samples using SAS EG, follow the instructions under Section C.8 above. The *F* test for comparing variances will appear at the bottom of the printout.

## C.11 Categorical Data Analysis

SAS EG can produce a frequency table for a single qualitative variable (i.e., a one-way table) and can conduct a chi-square test for independence of two qualitative variables in a two-way (contingency) table.

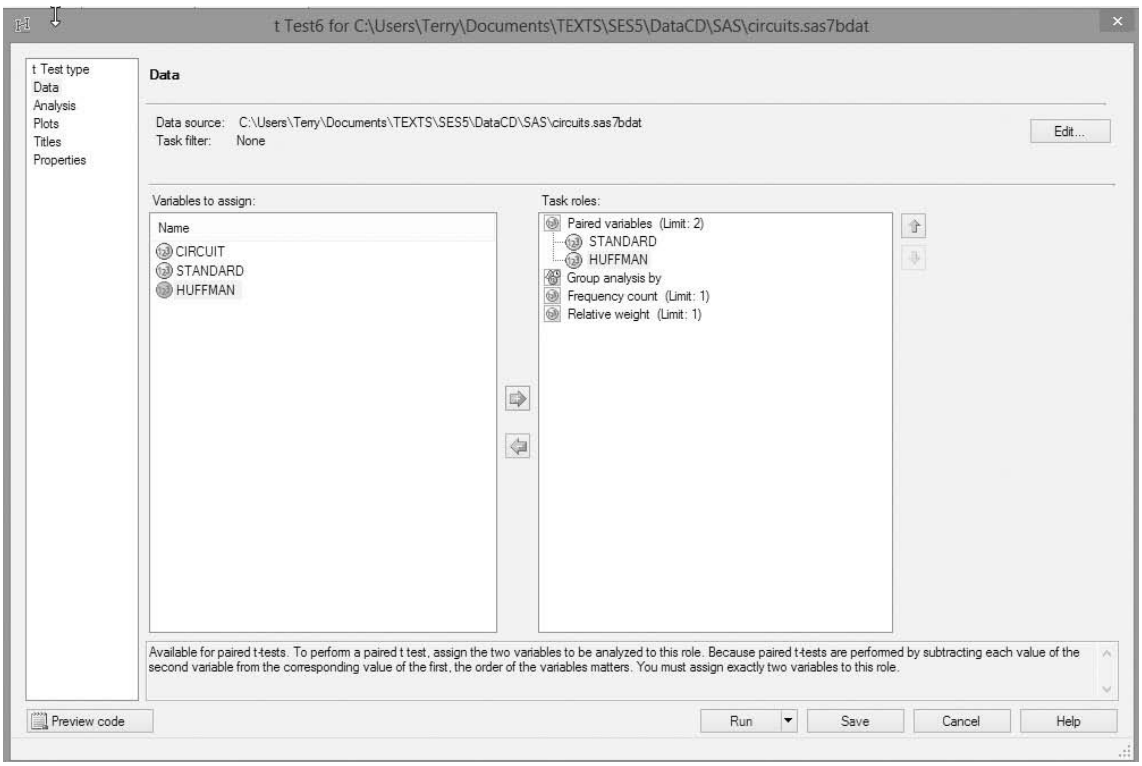



**FIGURE C.17**
SAS EG Dialog Box for Matched-Pairs Data

## One-Way Table

For a one-way table, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Describe*", and "*One-Way Frequencies*". (See Figure C.18.) On the resulting "*Data*" dialog box, move the qualitative variable you want to analyze into the "Analysis variables" box on the right side of the menu. Now click the "Statistics" option at the top left of the dialog box. On the resulting screen, check the "*Chi-square goodness of fit, Asymptotic test*" box (see Figure C.19). Then click "Run" to obtain a printout of the results.

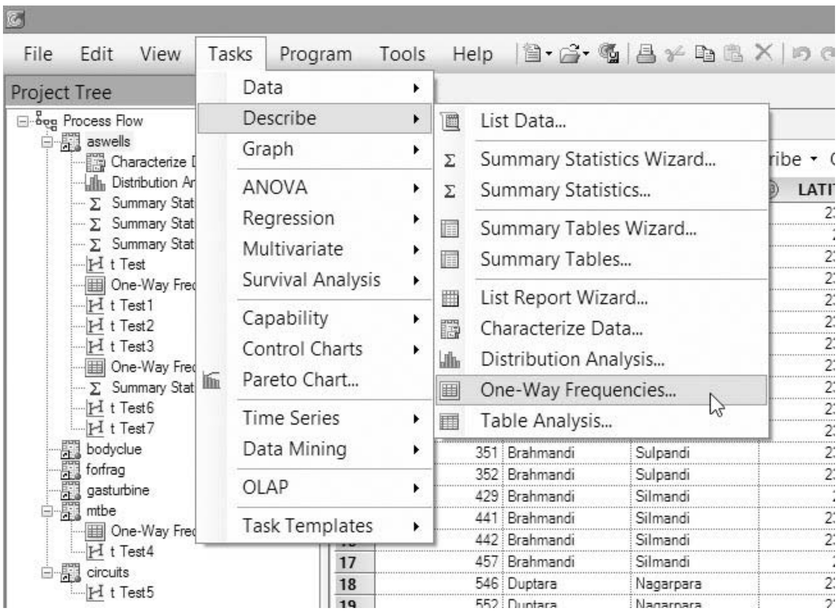
*Note 1*: If the qualitative variable of interest is a binomial (two-level) categorical variable, you can use SAS EG to generate a confidence interval and test for a proportion associated with one of the levels. On the one-way frequency dialog box (Figure C.19), check the "*Asymptotic test*" box under "*Binomial proportions*", then specify the values of the "*Test proportion*" (i.e., the hypothesized proportion) and "Confidence level" in the appropriate boxes. Click "Run" to generate the printout.

*Note 2*: The chi-square goodness of fit test produced using SAS EG tests the null hypothesis of equal proportions. If you desire a test where the hypothesized proportions are not the same (e.g., $H_0$: $p_1=.2$, $p_2=.3$, $p_3=.5$, you cannot use SAS EG menu options. Rather, you need to specify the appropriate SAS programming commands in the SAS Editor window. The commands (PROC SURVEYFREQ) shown in Figure C.20 will produce a chi-square test for a one-way table on the 3-level categorical variable called ICETYPE. The values following "TESTP=" are the null hypothesized percentages associated with the three categories.

## Two-Way Table

For a two-way table chi-square analysis, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Describe*", and "*Table Analysis*". (See Figure C.18.) On the resulting "*Data*" dialog box, move the two qualitative variables you want to analyze into the "*Tables variables*" box on the right side of the menu, as shown in Figure C.21. Now click the "*Tables*" option at the top left of the dialog box. On the resulting screen,

**FIGURE C.18**
SAS EG Menu Options for a One-Way Frequency Table Analysis



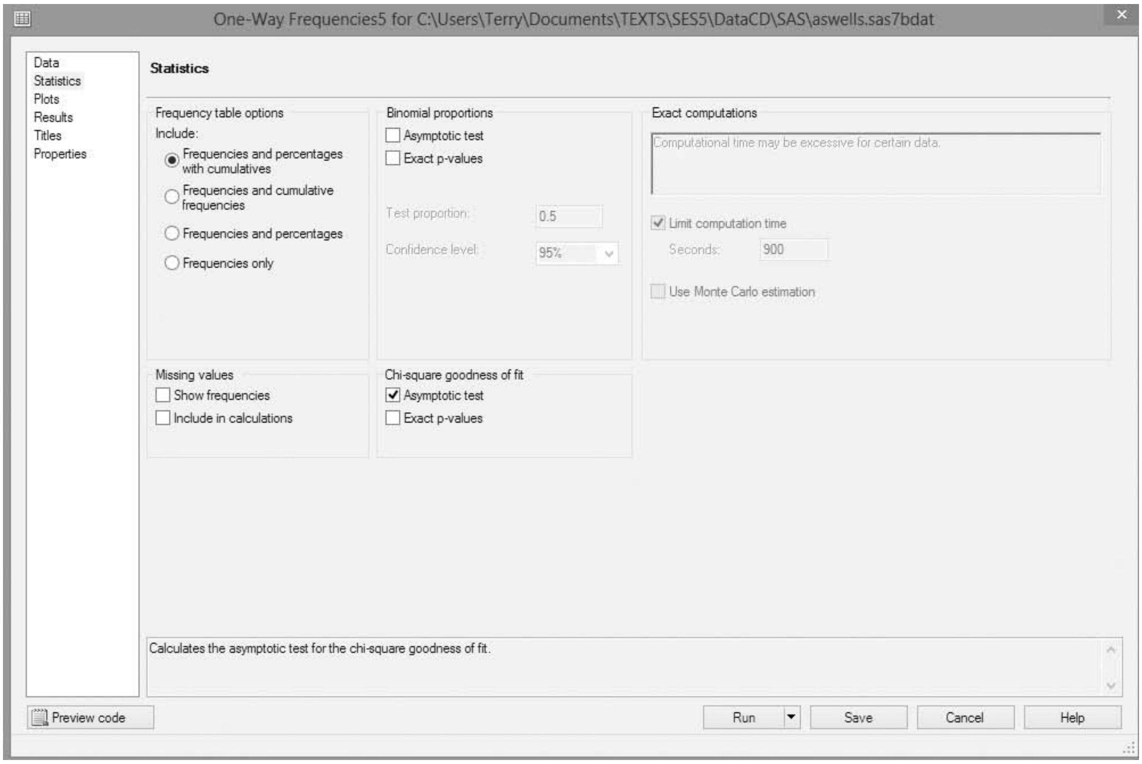
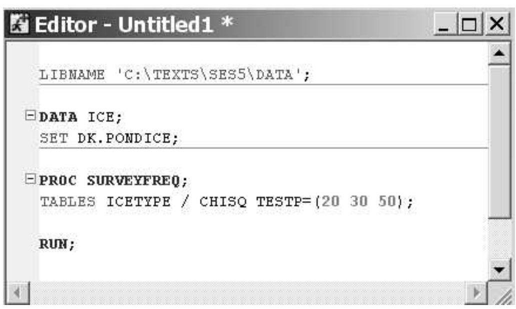



**FIGURE C.19**
SAS EG One-Way Frequency Table Dialog Box

click and drop one variable into the rows of the table and click and drop the other variable into the columns, as shown in Figure C.22. Next, click on "*Association*" under "*Table Statistics*" on the left side panel, then click the "*Chi-square tests*" box (see Figure C.23). (Note the option for selecting Fisher's exact test.) Click "Run" to obtain a printout of the results.

[*Note:* If your SAS data set contains summary information (i.e., the cell counts for the contingency table) rather than the actual categorical data values for each observation, you must specify the variable containing the cell counts on the "Two-Way Table Variable Selection" dialog box (see Figure C.21). Do this by moving the cell counts variable into the "*Frequency count*" box on the right panel.]

**FIGURE C.20**
SAS Program Commands for a One-Way Table Chi-Square




```
LIBNAME 'C:\TEXTS\SES5\DATA';

DATA ICE;
SET DK.PONDICE;

PROC SURVEYFREQ;
TABLES ICETYPE / CHISQ TESTP=(20 30 50);

RUN;
```

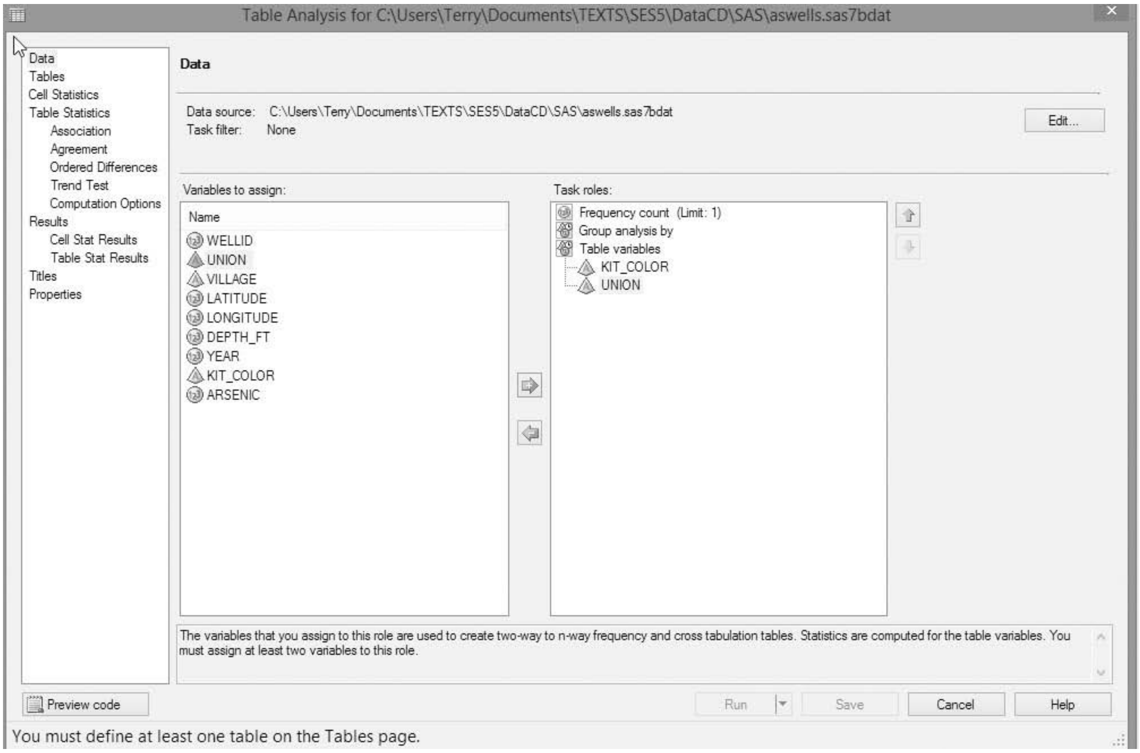



**FIGURE C.21**
SAS EG Two-Way Table Variable Selection

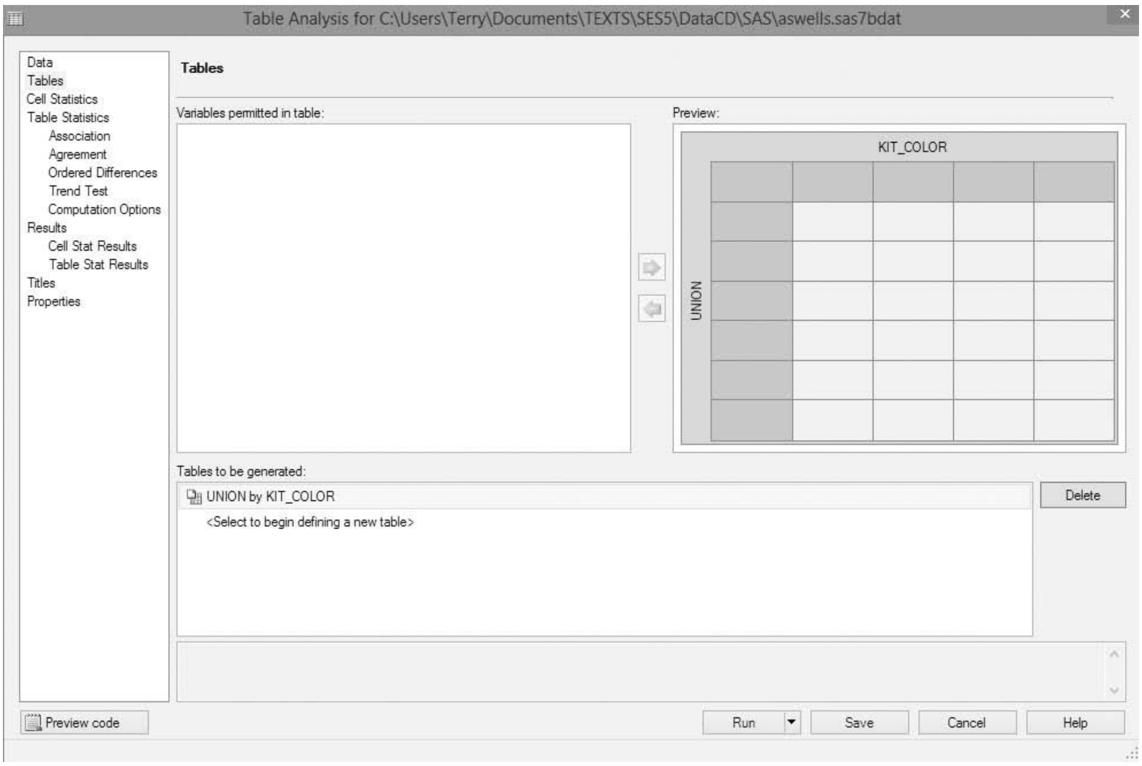



**FIGURE C.22**
SAS EG Two-Way Table Row and Column Variable Selection

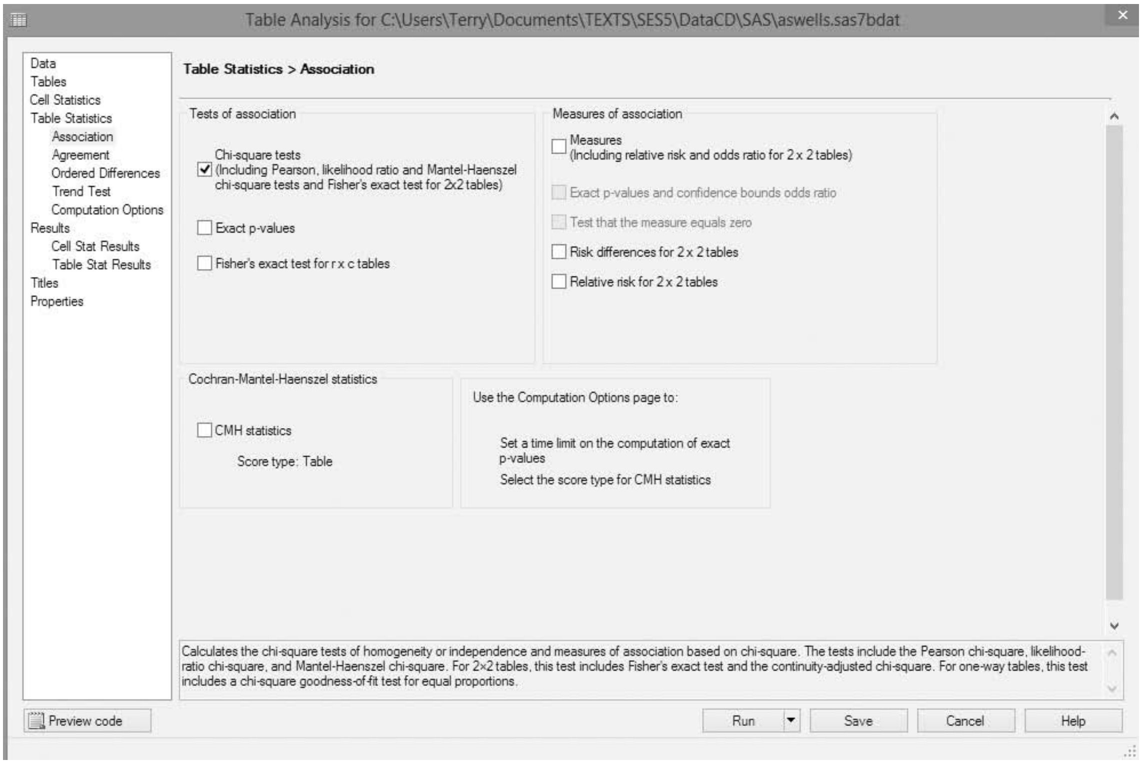



**FIGURE C.23**
SAS EG Two-Way Table Statistics Dialog Box

## C.12 Simple Linear Regression

To conduct a simple linear regression analysis, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Regression*", and "*Linear Regression*". (See Figure C.24.) On the resulting "*Data*" dialog box, move the quantitative dependent variable into the "*Dependent variable*" box and the quantitative independent variable into the "*Independent variables*" box on the right side of the menu, as shown in Figure C.25.

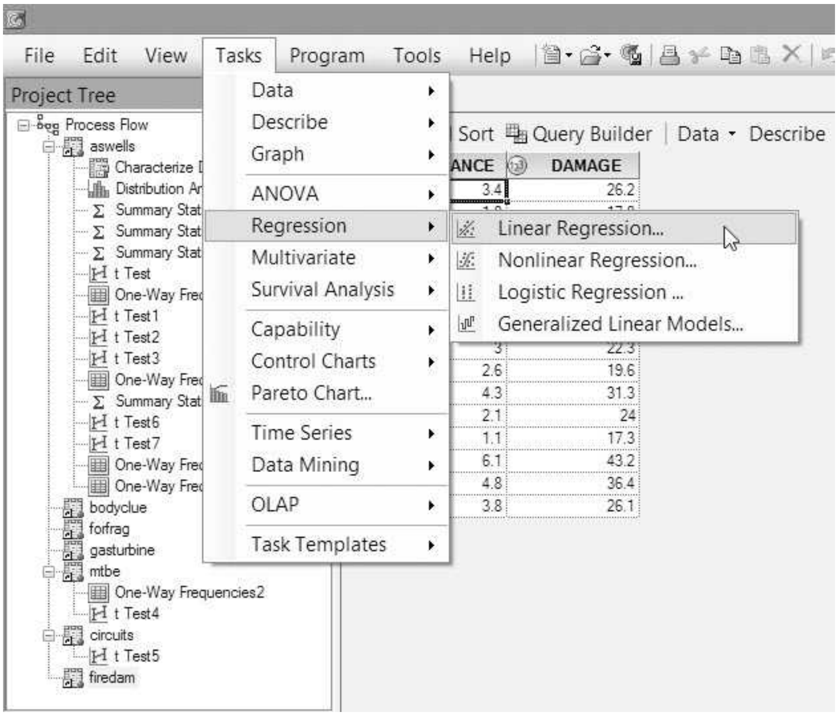
Optionally, you can get SAS EG to produce confidence intervals for the model parameters by clicking "*Statistics*" on the left panel and checking "*Confidence limits for parameter estimates*" on the resulting menu. Also, you can obtain prediction intervals and residual plots by clicking the "*Predictions*" button and "*Plots*" button, respectively, and making the appropriate selections on the resulting menus. Click "*Run*" to view the simple linear regression results.

## C.13 Multiple Regression

To conduct a multiple regression analysis, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Regression*", and "*Linear Regression*". (See Figure C.24.) On the resulting "*Data*" dialog box, move the quantitative dependent variable into the "*Dependent variable*" box and all the independent variables into the "*Independent variables*" box on the right side of the menu, as shown in Figure C.26.

Optionally, you can get SAS EG to produce confidence intervals for the model parameters by clicking "*Statistics*" on the left panel and checking "*Confidence limits*

**FIGURE C.24**
SAS EG Menu Options for Simple Linear Regression



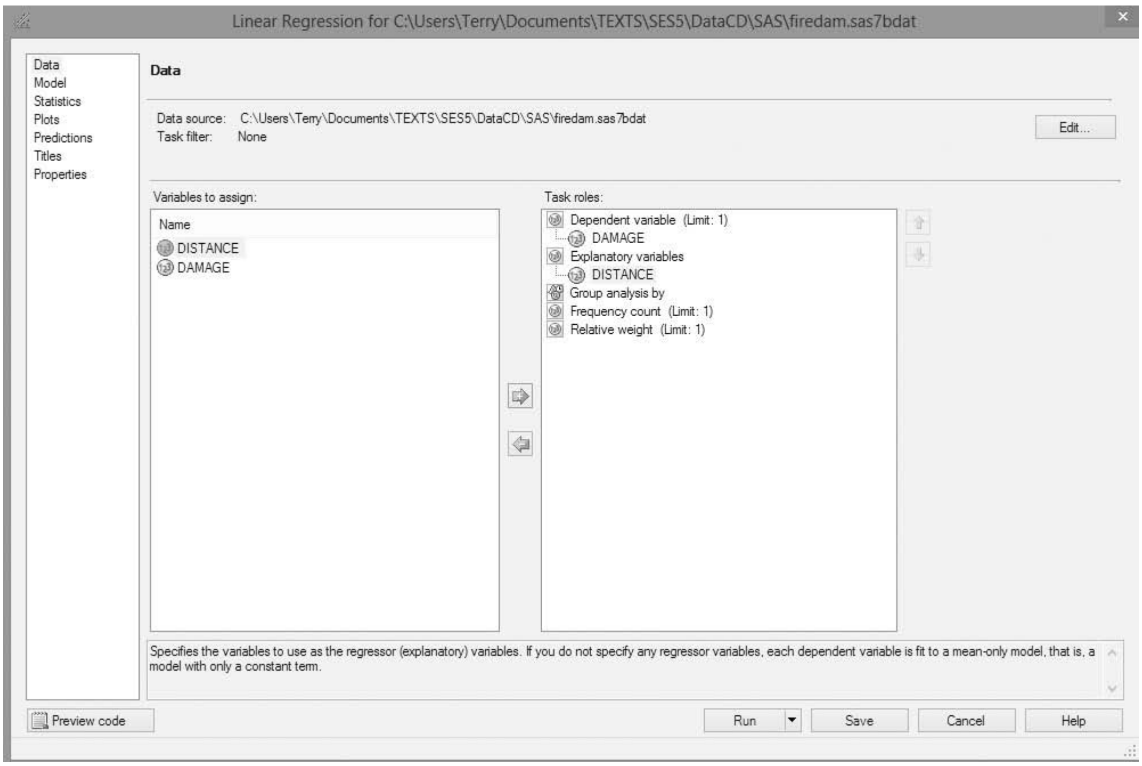




**FIGURE C.25**
SAS EG Linear Regression Data Dialog Box

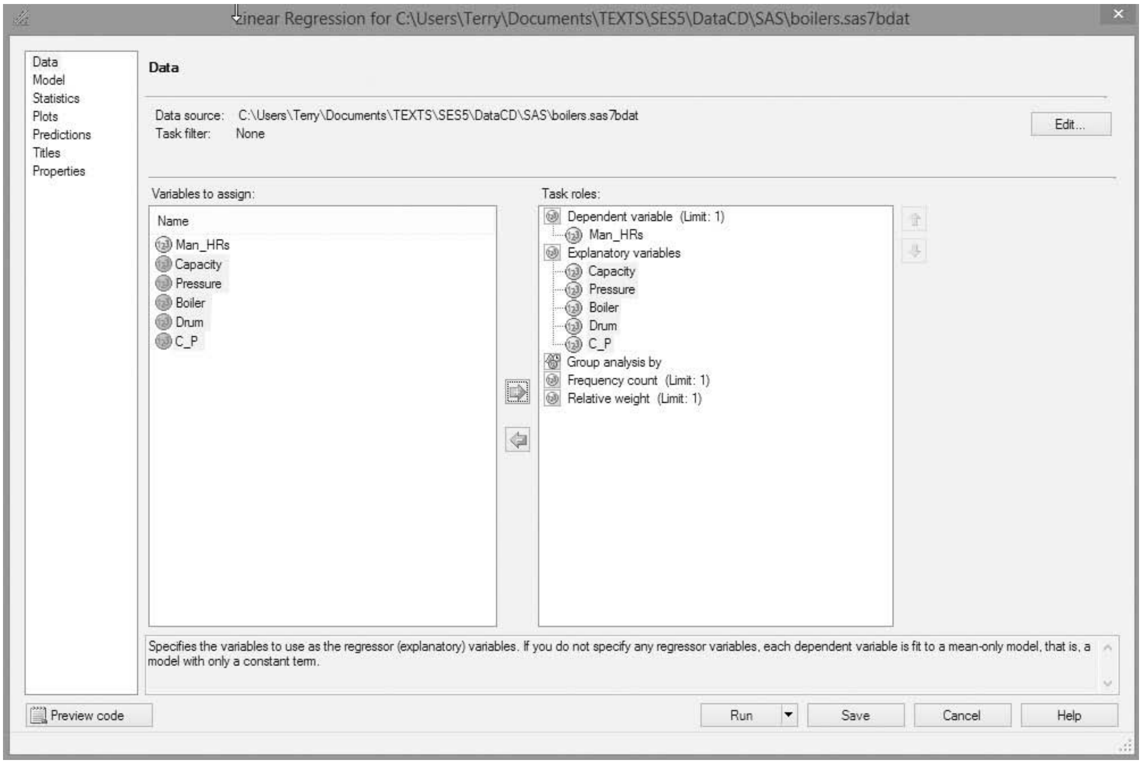



**FIGURE C.26**
SAS EG Dialog Box for Multiple Regression

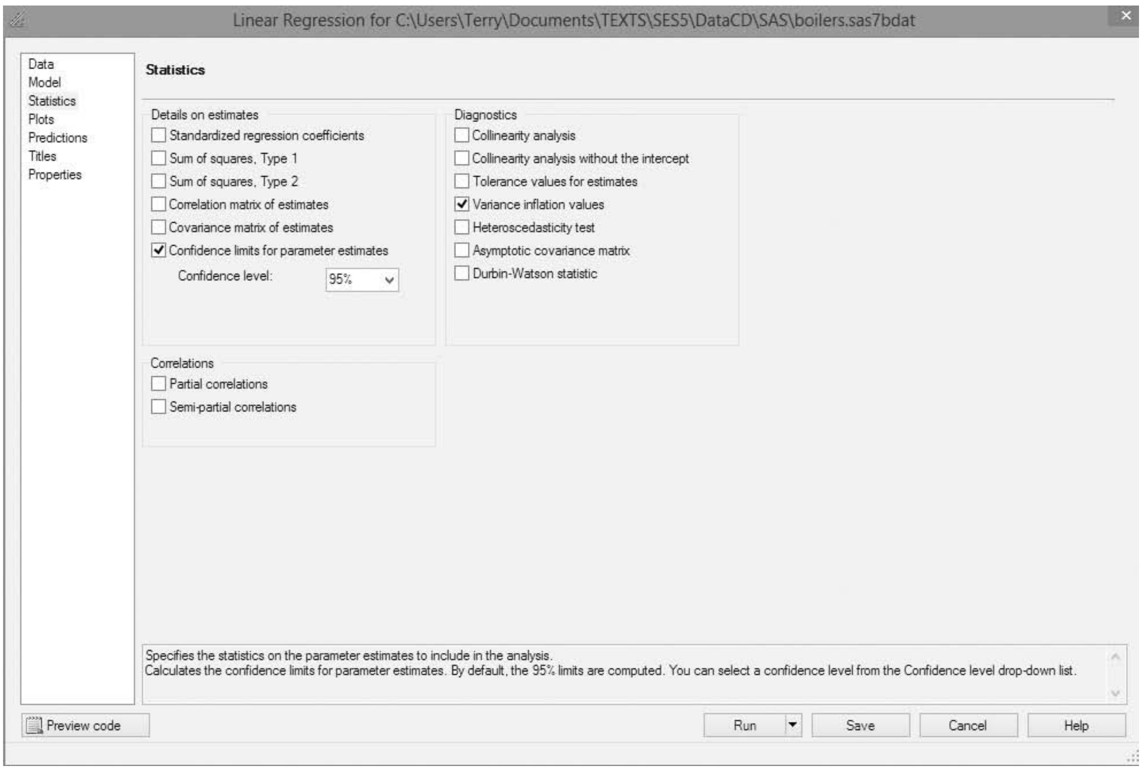
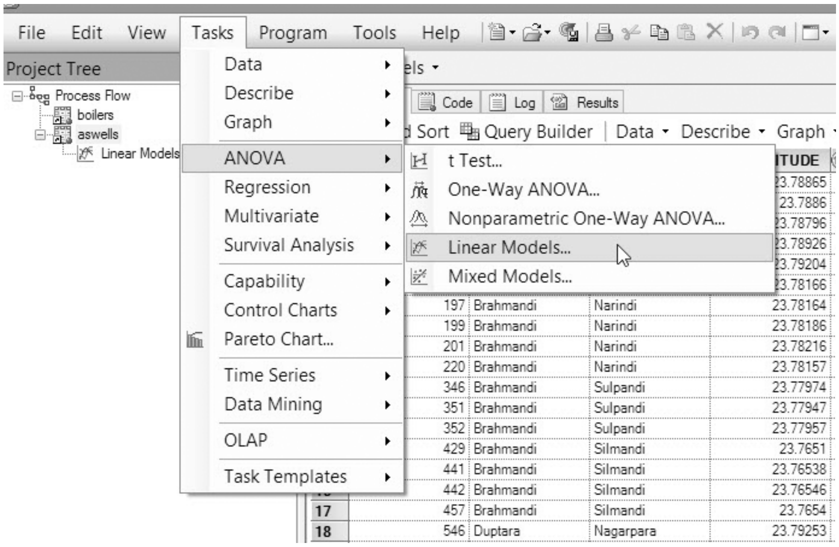



**FIGURE C.27**
SAS EG Multiple Regression Menu Options

**FIGURE C.28**
SAS EG Menu Options for General Linear Models




*for parameter estimates*" on the resulting menu (see Figure C.27). To produce variance inflation factors, check the "*Variance inflation values*" box (again, see Figure C.27). Also, you can obtain prediction intervals and residual plots by clicking the "*Predictions*" button and "*Plots*" button, respectively, and making the appropriate selections on the resulting menus. (Plots include influence diagnostics, e.g., studentized deleted residuals and Cook's D.) Click "*Run*" to view the multiple regression results.

[**Note**: If your model includes dummy variables, interactions or squared terms, you must create these variables in the DATA command lines in your SAS program *prior* to entering into a SAS EG session. See Figure C.3 for an example.]

## Fitting General Linear Models

As an alternative, you can fit general linear models using the "*ANOVA*" option available in SAS EG. To do this, click on the "*Tasks*" button on the SAS EG menu bar, then click on "ANOVA", and finally click on "Linear Models", as shown in Figure C.28. On the resulting "*Data*" dialog box, move the quantitative dependent variable into the "*Dependent variable*" box, the quantitative independent variables into the "*Quantitative variables*" box, and the qualitative independent variables into the "*Classification variables*" box on the right side of the menu, as shown in Figure C.28. (Note: SAS will automatically create the appropriate number of dummy variables for each qualitative variable specified.)

After making the variable selections, click the "Model" button on the left panel to view the dialog box shown in Figure C.29. Specify the terms in the model using the "*Main*" button (for main effects), the "*Cross*" button (for interactions) and the "*Polynomial*" button (for higher-order terms). The model terms will appear in the "*Effects*" box on the right.

Click "*Model Options*" button and check "*Show parameter estimates*" on the resulting menu to produce the estimates of the model parameters. Also, you can obtain prediction intervals and residual plots by clicking the "*Predictions*" button and "*Plots*" button, respectively, and making the appropriate selections on the resulting menus. When all the options you desire have been checked, click "*Run*" to view the multiple regression results.

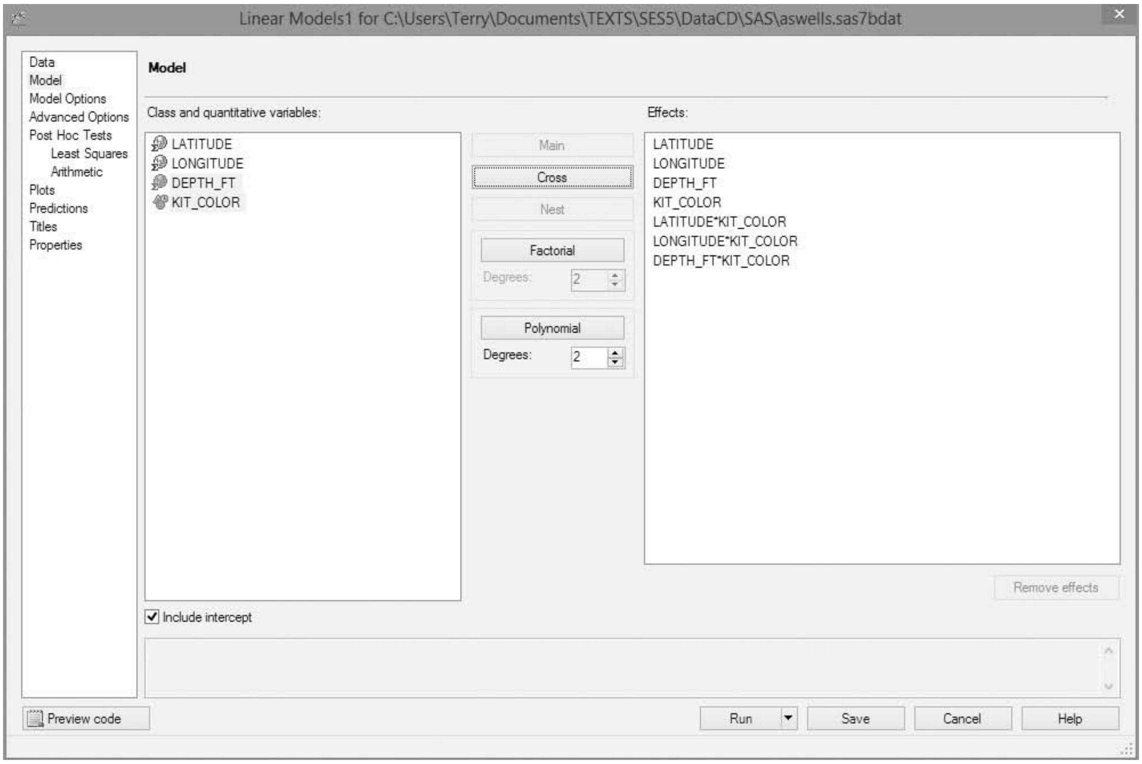



**FIGURE C.29**
SAS EG General Linear Models Dialog Box

### Stepwise Regression

To conduct a stepwise regression analysis, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Regression*", and "*Linear Regression*". (See Figure C.24.) On the resulting "*Data*" dialog box, move the quantitative dependent variable into the "*Dependent variable*" box and all the independent variables into the "*Independent variables*" box on the right side of the menu, as shown in Figure C.26. Now click on the "*Model*" button on the left panel. The resulting menu appears as shown in Figure C.30.

For the stepwise regression method, choose "*Stepwise selection*". (The default method is "*Full model fitted*".) For the all-possible-regressions-selection method, choose "*Mallows' Cp selection*", "*R-squared selection*", or "*Adjusted R-squared selection*". Once you make a selection, as an option, you may select the value of $\alpha$ to use in the analysis. (The default is $\alpha = .05$.) Click "*Run*" to view the stepwise regression results.

## C.14 One-Way Analysis of Variance

To conduct a one-way ANOVA for a completely randomized design using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*ANOVA*", and "*One-Way ANOVA*". (See Figure C.31.) On the resulting "Data" dialog box, move the quantitative dependent variable into the "*Dependent variable*" box and the qualitative variable that represents the single factor in the experiment into the "*Independent variable*" box on the right side of the menu, as shown in Figure C.32. To perform multiple comparisons of treatment means, click the "*Comparison*" button under "*Means*" on the left panel to obtain the dialog box shown in Figure C.33. On this box, select the comparison method (e.g., Bonferroni's method) and the comparison-wise error rate (e.g., confidence level).

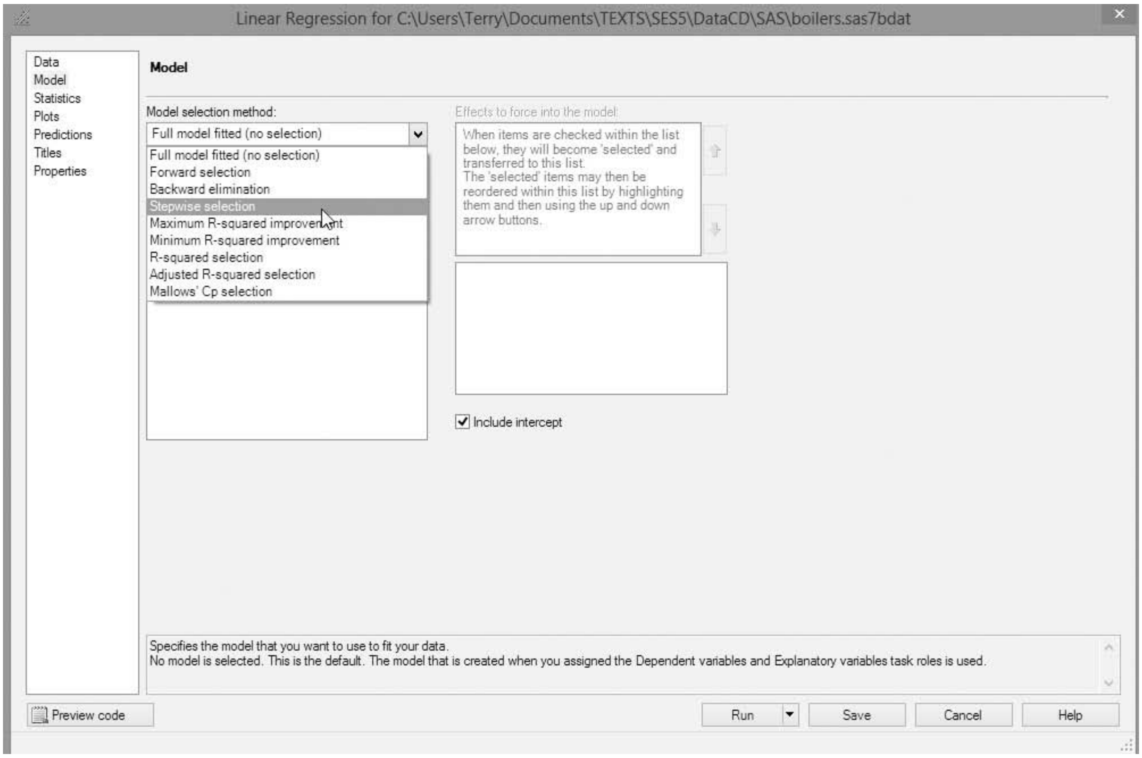



**FIGURE C.30**
SAS EG Model Menu Selection for Multiple Regression

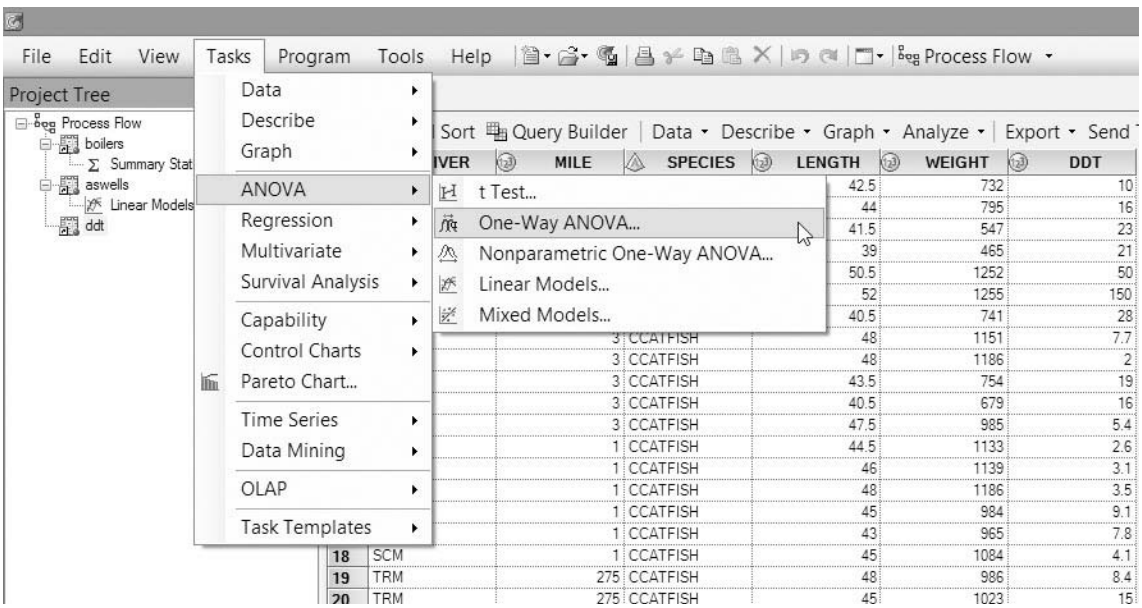



**FIGURE C.31**
SAS EG Menu Options for 1-way Analysis of Variance

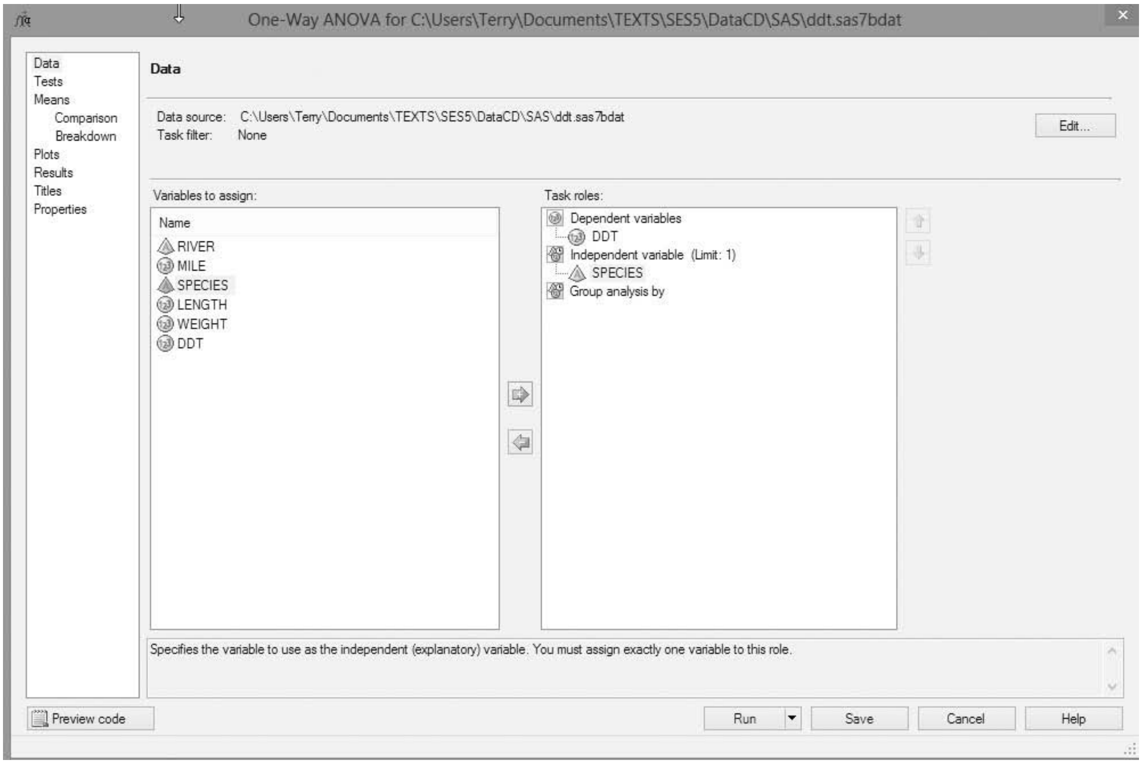



**FIGURE C.32**
SAS EG Dialog Box for One-Way ANOVA

To perform a test of equality of variances, click the "*Tests*" button on the left side panel and select the test to be performed (e.g., Levene's test) on the resulting menu. Click "*Run*" to view the ANOVA results.

## C.15 Analysis of Variance for Factorial and Other Designs

To conduct an ANOVA for designs involving two or more factors (e.g., randomized block, factorial designs) using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*ANOVA*", and finally click on "*Linear Models*". (See Figure C.28.) The resulting "Data" dialog box appears in Figure C.34.

Move the quantitative dependent variable to the "*Dependent variable*" box and the variables that represent the factors in the experiment to the "*Classification variables*" box on the right panel, as shown in Figure C.34. To specify the design model, click on the "*Model*" button on the far left panel. The dialog box shown in Figure C.35 will appear. Specify the terms in the model, using the "*Main*" button for main effects and the "*Cross*" button for interactions. The model terms will appear in the "*Effects*" box in the right side panel.

To run multiple comparisons of means for all treatment combinations, click the "*Least Squares*" button under "*Post Hoc Tests*" on the far left panel. The dialog box shown in Figure C.36 appears. Specify the interaction effect of interest by selecting "*True*" by the interaction effect on the right panel, then click "*Add*". This interaction effect should appear in the "Effects to estimate" box. Select the comparison method (e.g., Bonferroni's method) and select "All pairwise comparisons", also on the right panel. Click "*Run*" to view the ANOVA results.

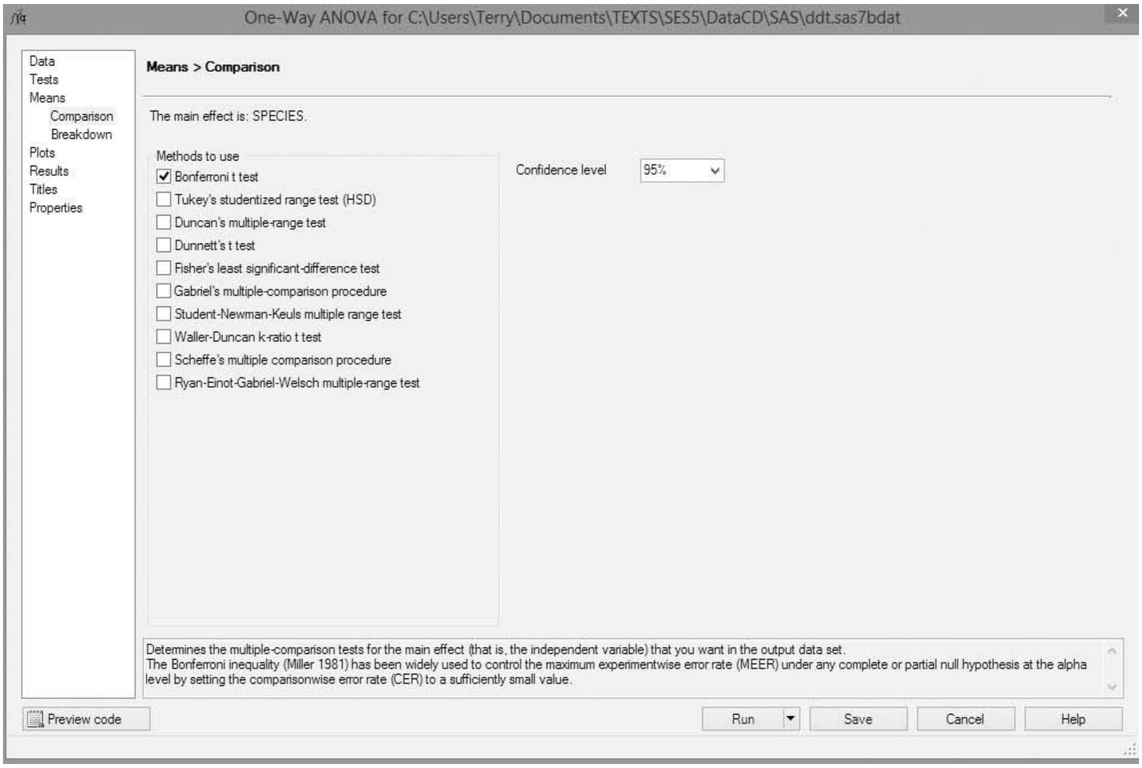



**FIGURE C.33**
SAS EG Dialog Box for Multiple Comparisons of Means

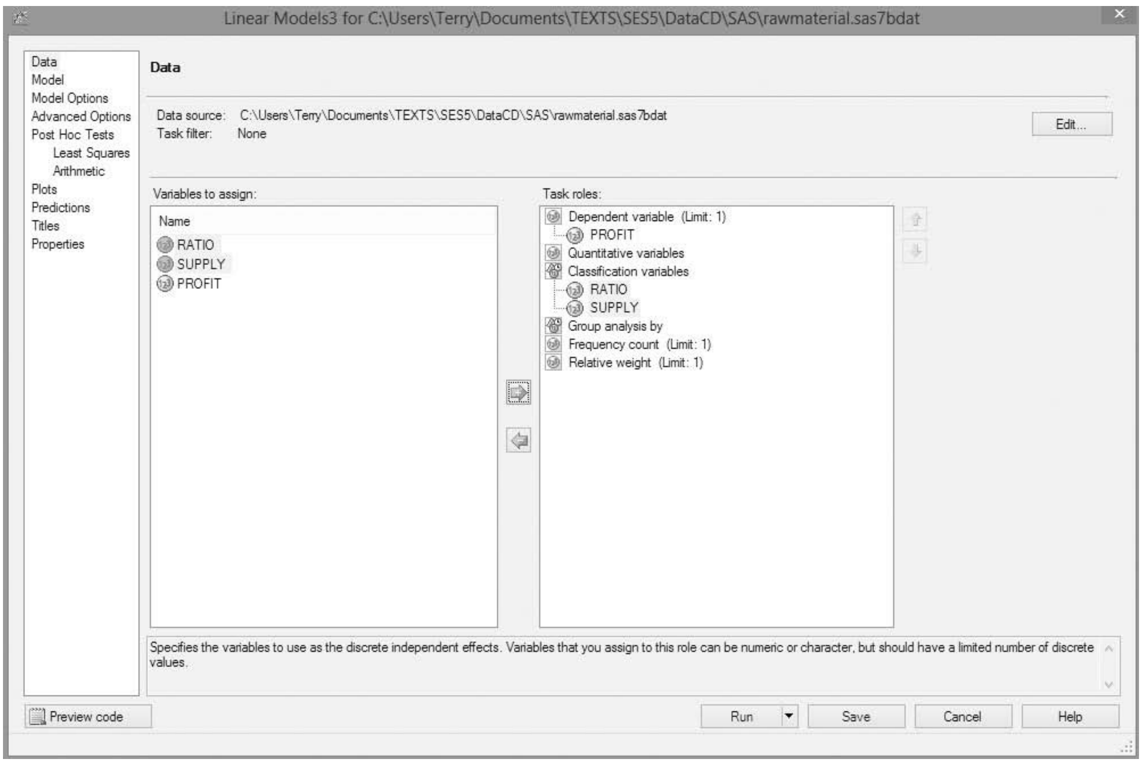



**FIGURE C.34**
SAS EG Dialog Box for Factorial ANOVAs

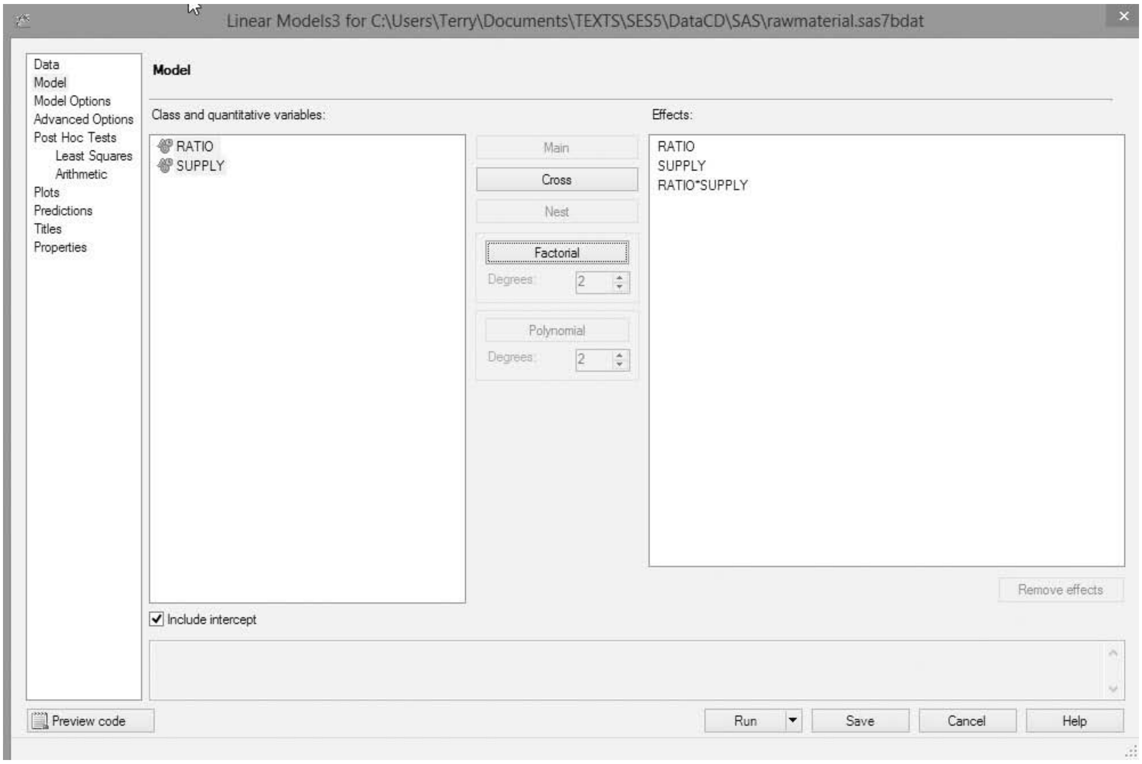



**FIGURE C.35**
SAS EG Dialog Box for Factorial Model Selection

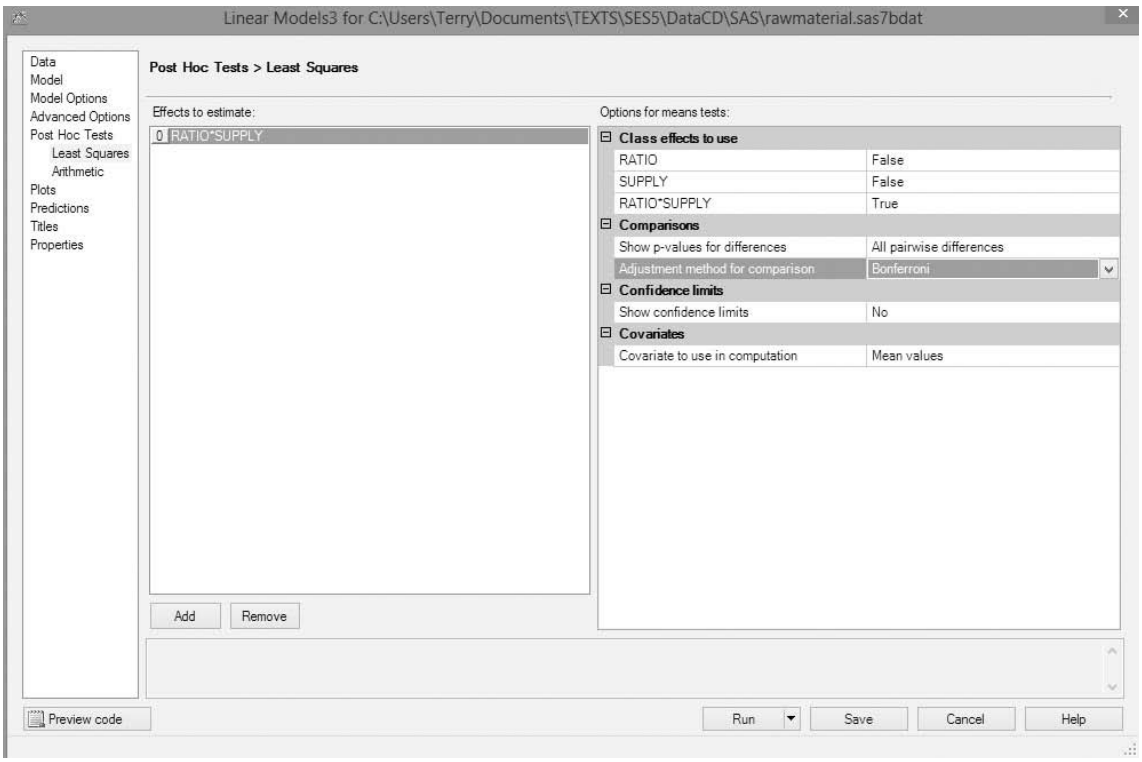



**FIGURE C.36**
SAS EG Dialog Box for Multiple Comparisons of Factorial Means

## C.16 Nonparametric Tests

Nonparametric tests in SAS are performed using either SAS Enterprise Guide or basic SAS programming commands. The choice will vary from test to test.

*Sign Test*: To run a sign test, you need to specify the appropriate SAS programming commands in the SAS Editor window. The PROC UNIVARIATE commands shown in Figure C.37 will produce a sign test (along with several other one-sample tests) for the quantitative variable, LWRATIO. The value following "MU0=" is the null hypothesized value of the population median.

*Rank Sum Test and Kruskal-Wallis Test*: You can use SAS EG to run either a Wilcoxon rank sum test to compare two populations or a Kruskal-Wallis test to compare three or more populations. Click on the "*Tasks*" button on the SAS EG menu bar, then click on "*ANOVA*", and finally click on "*Nonparametric One-Way ANOVA*". (See Figure C.38.)

The resulting "Data" dialog box appears in Figure C.39. Specify the quantitative variable to be analyzed in the "*Dependent variable*" box and the categorical variable that represents the different samples in the "*Independent variable*" box on the right panel. On the far left panel, click the "*Analysis*" button and check "*Wilcoxon*" as shown in Figure C.40. Click "*Run*" to generate the SAS printout.

*Signed Ranks Test*: To run a Wilcoxon signed ranks test, you need to specify the appropriate SAS 9.3 programming commands in the SAS Editor window. First calculate the difference between the values of the two paired quantitative variables, then run

**FIGURE C.37**
SAS Program Commands for a Sign Test

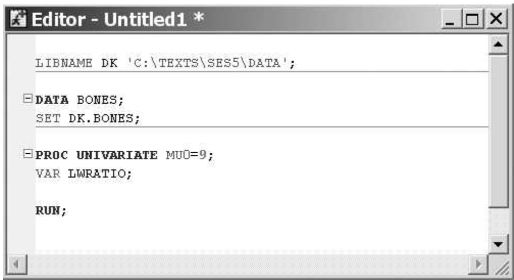
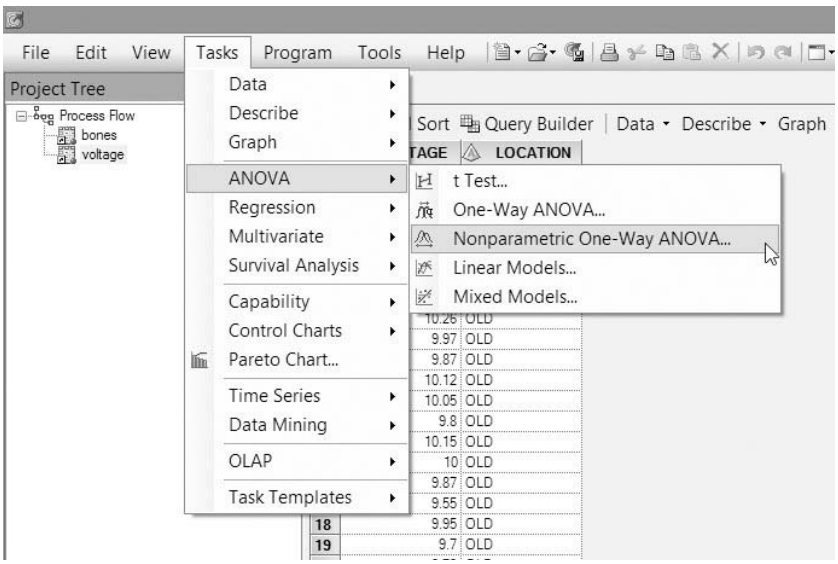



**FIGURE C.38**
SAS EG Menu Selections for Nonparametric One-Way ANOVA



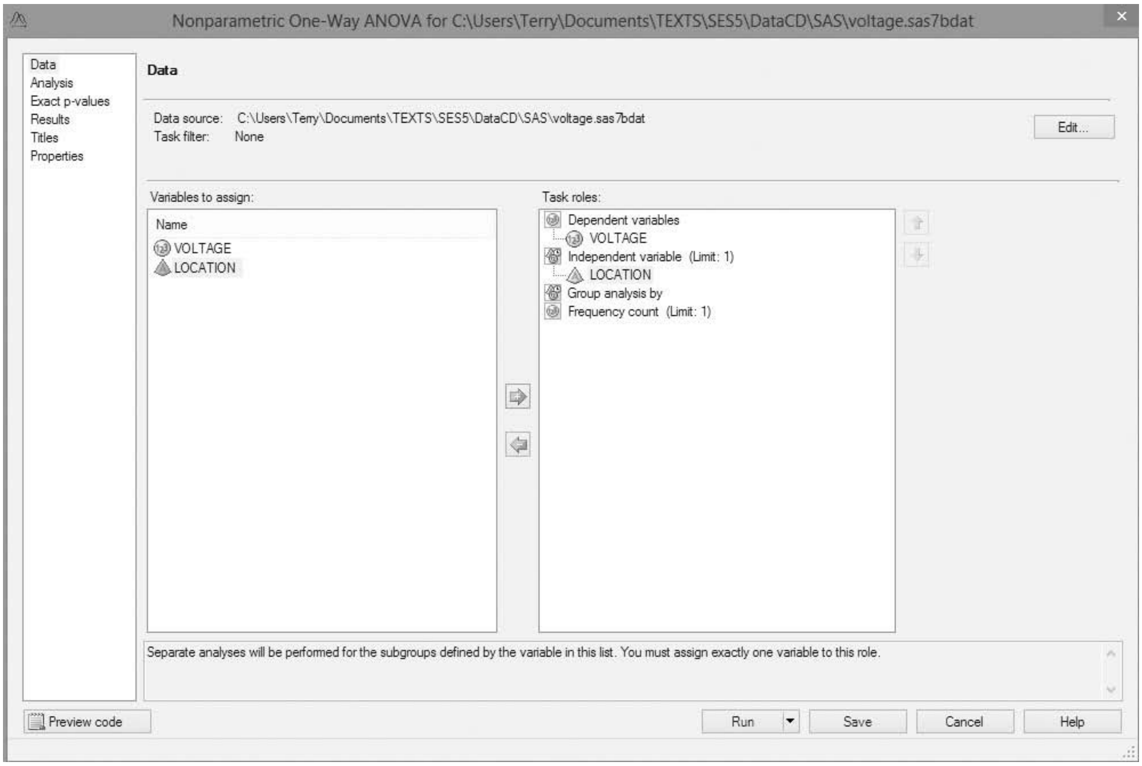



**FIGURE C.39**
SAS EG Data Dialog box for Nonparametric One-Way ANOVA

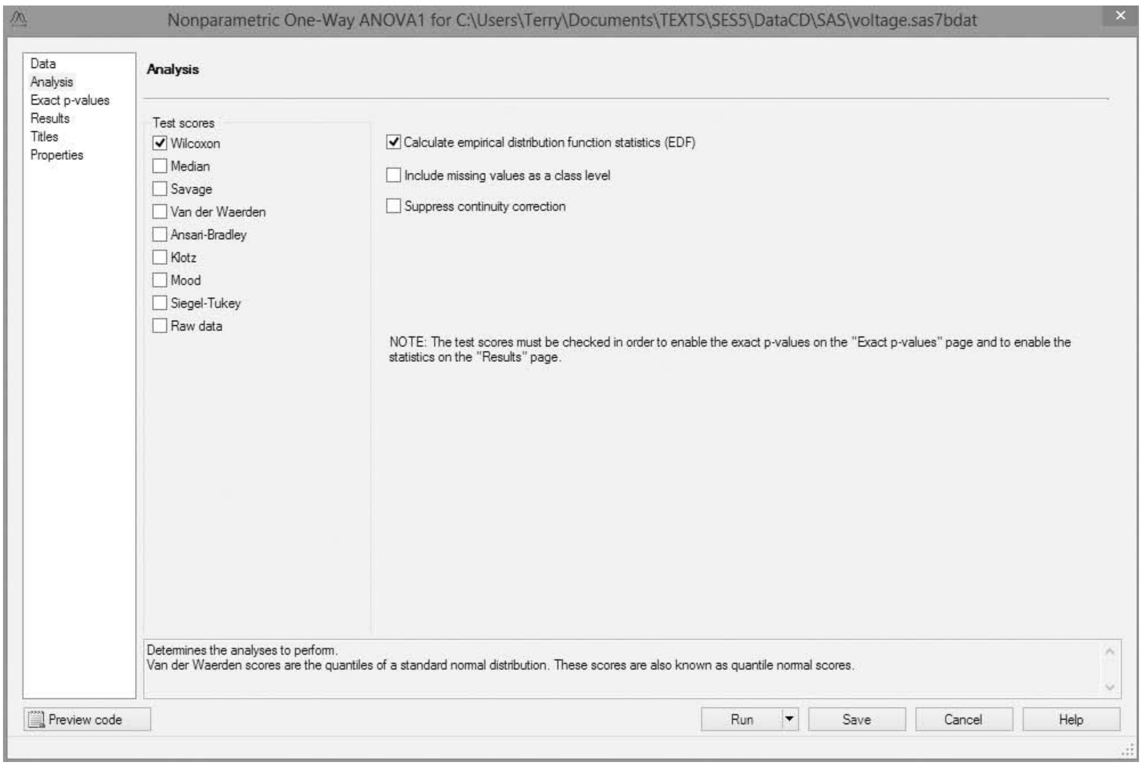
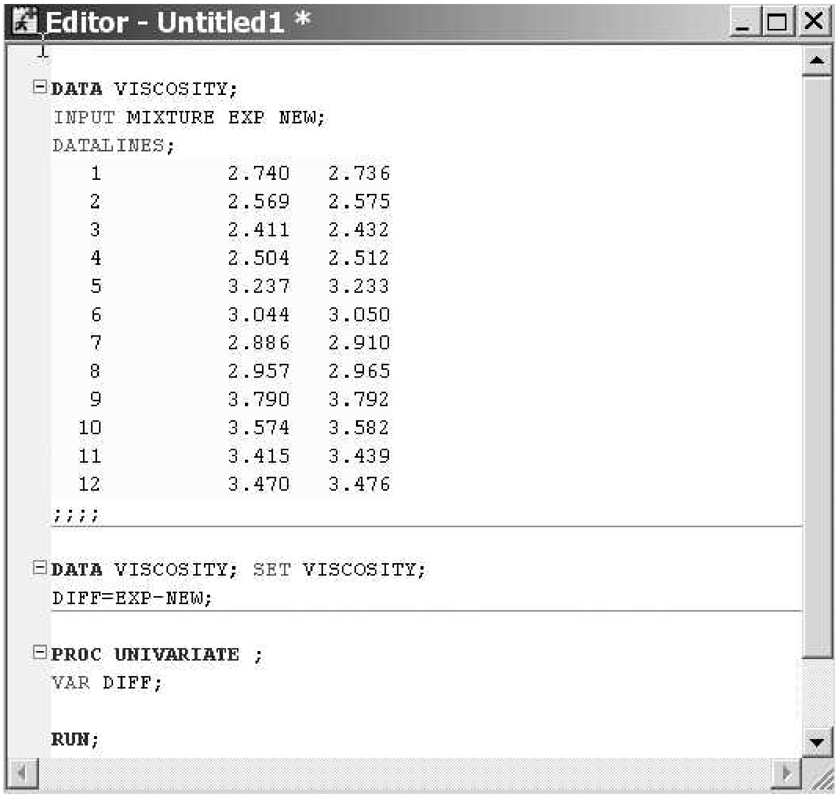



**FIGURE C.40**
Selecting the Nonparametric One-Way ANOVA Test

**FIGURE C.41**
SAS Program Commands for a Signed Rank Test

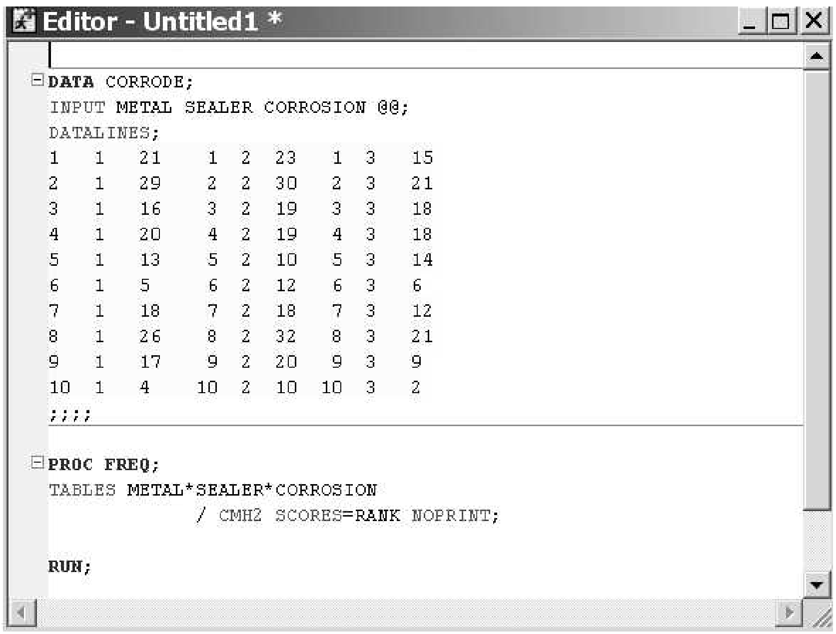

Editor - Untitled1 *

```
DATA VISCOSITY;
INPUT MIXTURE EXP NEW;
DATALINES;
   1          2.740   2.736
   2          2.569   2.575
   3          2.411   2.432
   4          2.504   2.512
   5          3.237   3.233
   6          3.044   3.050
   7          2.886   2.910
   8          2.957   2.965
   9          3.790   3.792
  10          3.574   3.582
  11          3.415   3.439
  12          3.470   3.476
;;;;

DATA VISCOSITY; SET VISCOSITY;
DIFF=EXP-NEW;

PROC UNIVARIATE ;
VAR DIFF;

RUN;
```

PROC UNIVARIATE as shown in Figure C.41. Be sure to specify the variable that represents the difference in the VAR statement.

*Friedman Test:* To run a Friedman test for a randomized block design, you need to specify the appropriate SAS 9.3 programming commands in the SAS Editor window. The test is obtained by running PROC FREQ as shown in Figure C.42. Specify the

**FIGURE C.42**
SAS Program Commands for a Friedman Test


Editor - Untitled1 *

```
DATA CORRODE;
INPUT METAL SEALER CORROSION @@;
DATALINES;
1   1   21    1  2  23    1  3   15
2   1   29    2  2  30    2  3   21
3   1   16    3  2  19    3  3   18
4   1   20    4  2  19    4  3   18
5   1   13    5  2  10    5  3   14
6   1   5     6  2  12    6  3   6
7   1   18    7  2  18    7  3   12
8   1   26    8  2  32    8  3   21
9   1   17    9  2  20    9  3   9
10  1   4    10  2  10   10  3   2
;;;;

PROC FREQ;
TABLES METAL*SEALER*CORROSION
            / CMH2 SCORES=RANK NOPRINT;

RUN;
```

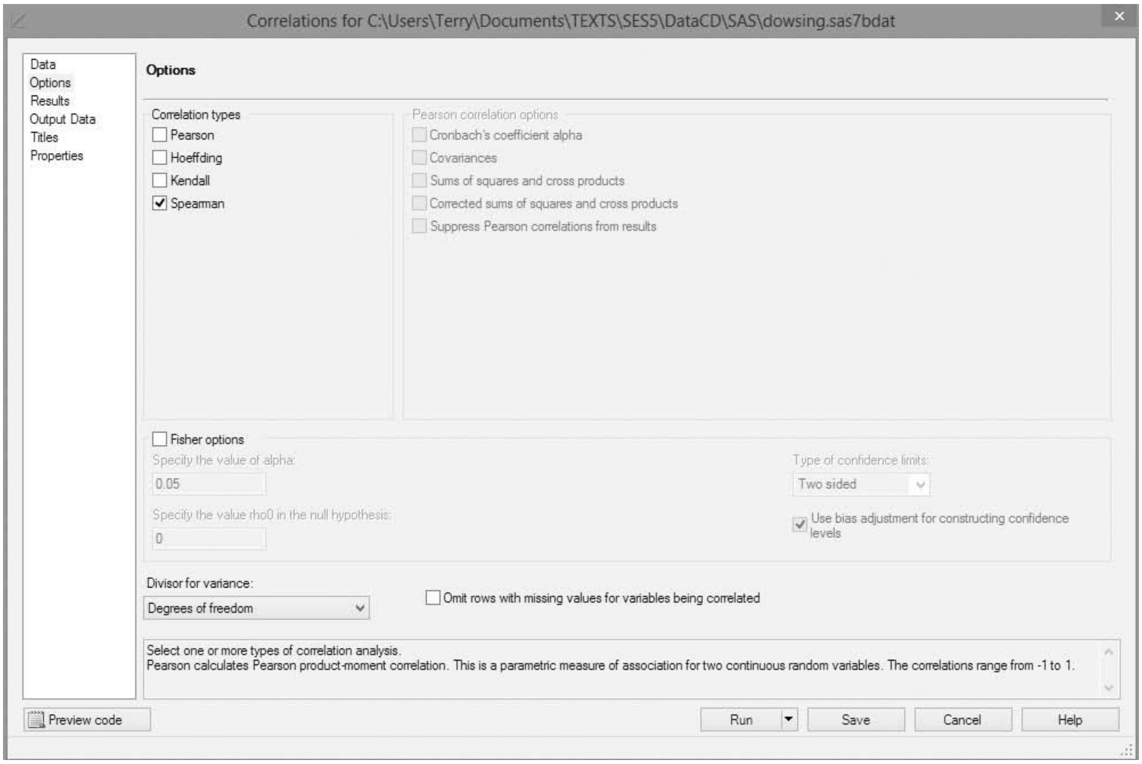



**FIGURE C.43**
Selecting the Spearman Correlation Option

block variable, treatment variable and dependent variable in the TABLES statement, placing an asterisk between the variable names. Be sure to specify the option "CMH2 SCORES=RANK NOPRINT" following the slash. The Friedman test results will appear next to "Row Mean Scores Differ" in the SAS output.

*Rank Correlation Test:* To perform Spearman's rank correlation test using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Multivariate*", and "*Correlations*". (See Figure C.13.) Move the variables you want to analyze into the "Analysis variables" box on the right side of the menu. Now click "*Options*" on the far left panel, and check "*Spearman*" under "*Correlation types*" on the resulting screen. (See Figure C.43.) Click "Run" to obtain a printout of the Spearman test.

## C.17 Control Charts and Capability Analysis

### Control Charts

To generate quality control charts using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Control Charts*", as shown in Figure C.44. The resulting menu allows you to choose an individual measurements chart, (mean) $\ddot{x}$ -chart, (range) R-chart, p-chart, or c-chart.

Once you make a selection, a control chart dialog box will appear, asking you to specify the process variable and subgroup (identifier) variable. (See Figure C.45 for the selections for an individuals chart.) Click the "*Run*" button to produce the control chart.

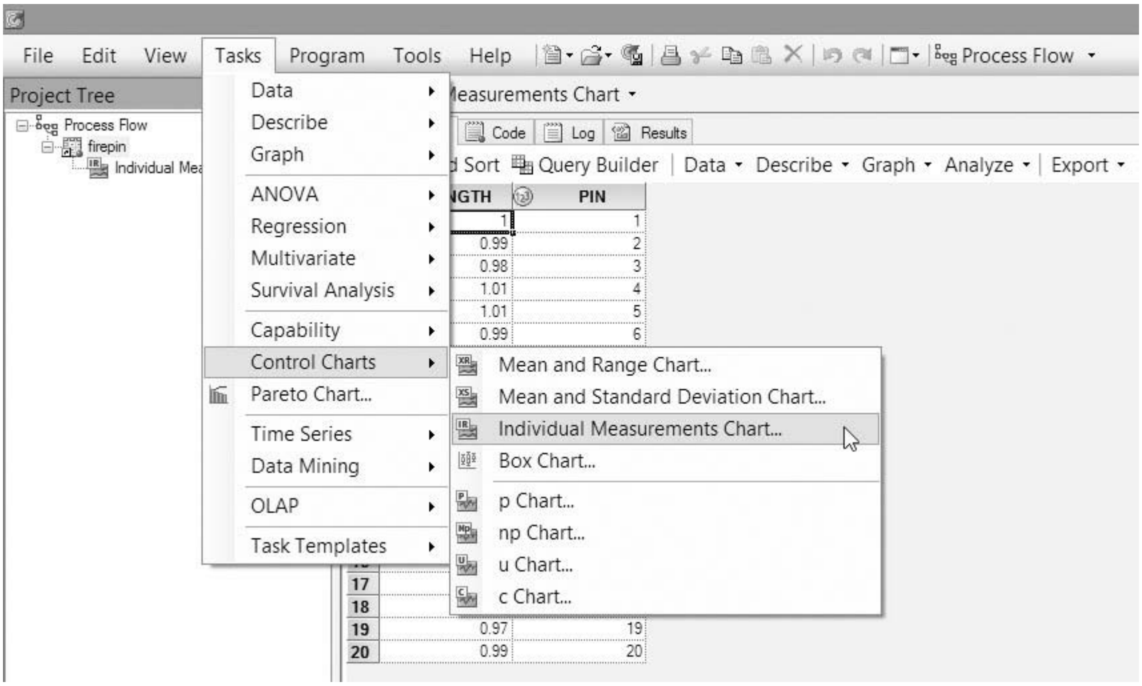



**FIGURE C.44**
SAS EG Menu Options for Control Charts

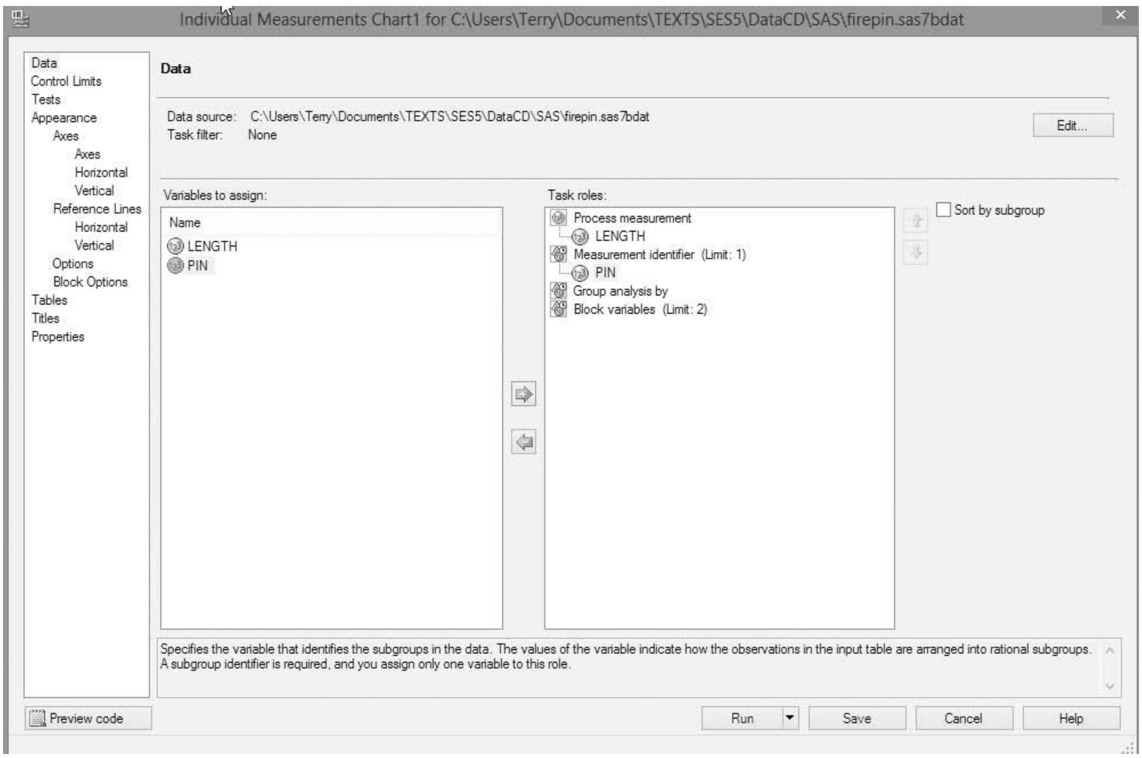



**FIGURE C.45**
SAS EG Data Dialog Box for an Individual Control Chart

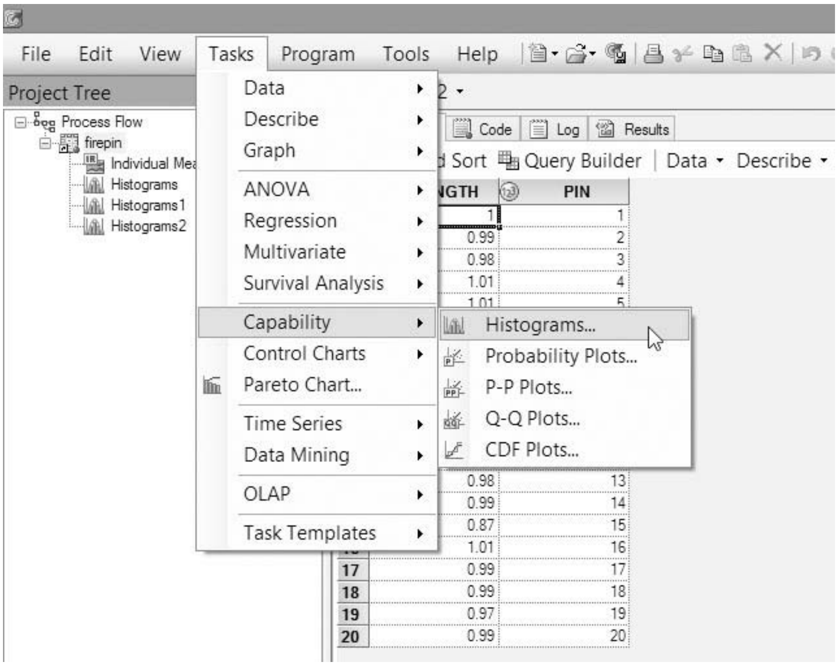



**FIGURE C.46**
SAS EG Menu Options for Process Capability Analysis

## Capability Analysis

To conduct a capability analysis using SAS EG, click on the "*Tasks*" button on the menu bar, then click on "*Capability*" and "*Histograms*", as shown in Figure C.46. The dialog box shown in Figure C.47 will be displayed. Move the process variable to the

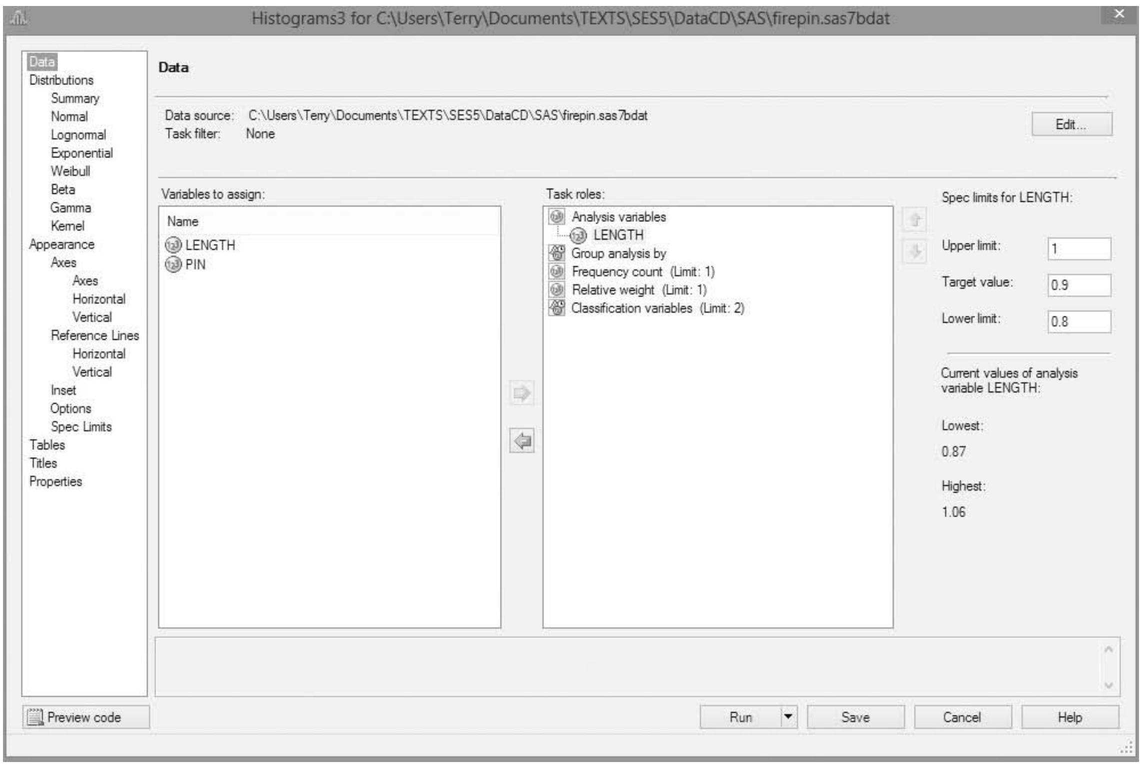



**FIGURE C.47**
SAS EG Data Dialog Box for Process Capability Analysis

"*Analysis variable*" box on the right panel. Also, on the right side of the panel, enter the target value and upper and lower specification limits. Under "Distributions" on the far left panel, click "*Normal*" (or your distribution of choice). Click the "*Run*" button to produce the capability analysis.

## C.18 Random Samples

To generate a random sample of observations from a data set using SAS EG, click on the "*Tasks*" button on the SAS EG menu bar, then click on "*Data*" and "*Random Sample*" as shown in Figure C.48. On the resulting dialog box (see Figure C.49), specify the sample size (and, optionally, a random number seed). Click "*Run*" and the values of the random sample will appear in an output data set in SAS EG.

**FIGURE C.48**
SAS EG Menu Options for Random Samples

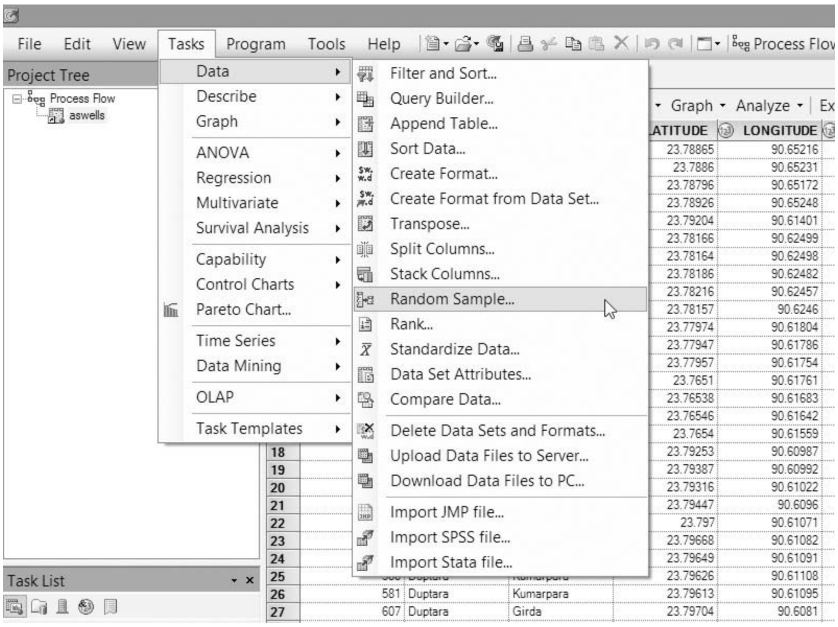
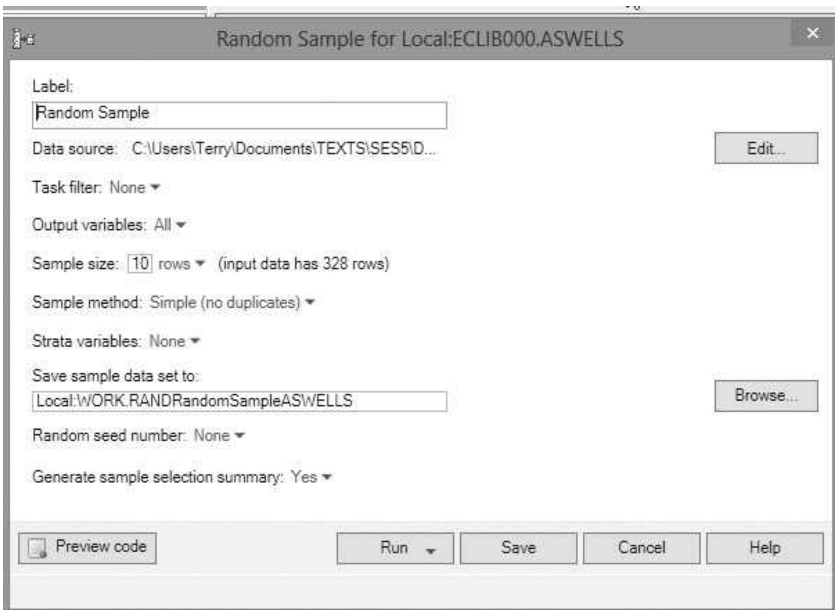



**FIGURE C.49**
SAS EG Dialog Box for Random Samples



# APPENDIX D MINITAB for Windows Tutorial

**CONTENTS**



## D.1 MINITAB Windows Environment

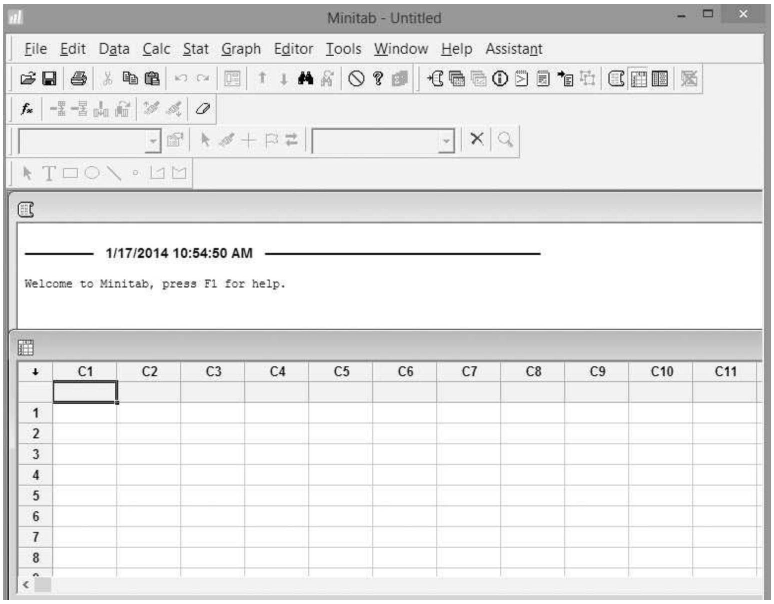
Upon entering into a MINITAB session, you will see a screen similar to Figure D.1. The bottom portion of the screen is an empty spreadsheet — called a MINITAB worksheet — with columns representing variables and rows representing observations (or cases). The very top of the screen is the MINITAB main menu bar, with buttons for the different functions and procedures available in MINITAB. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate menu buttons. The results will appear in the Session window at the top.

## D.2 Creating/Accessing a Data Set Ready for Analysis

There are three ways you can get a data set ready for analysis in MINITAB:

1. Entering data values directly into the MINITAB worksheet
2. Accessing a previously created MINITAB worksheet file
3. Accessing an external data file

**FIGURE D.1**
Initial Screen Viewed by the MINITAB User




## Direct Data Entry

Create a MINITAB data file by entering data directly into the worksheet. Figure D.2 shows data entered for a variable called "RATIO." Name the variables (columns) by typing in the name of each variable in the box below the column number.

**FIGURE D.2**
Data Entered into the MINITAB Worksheet

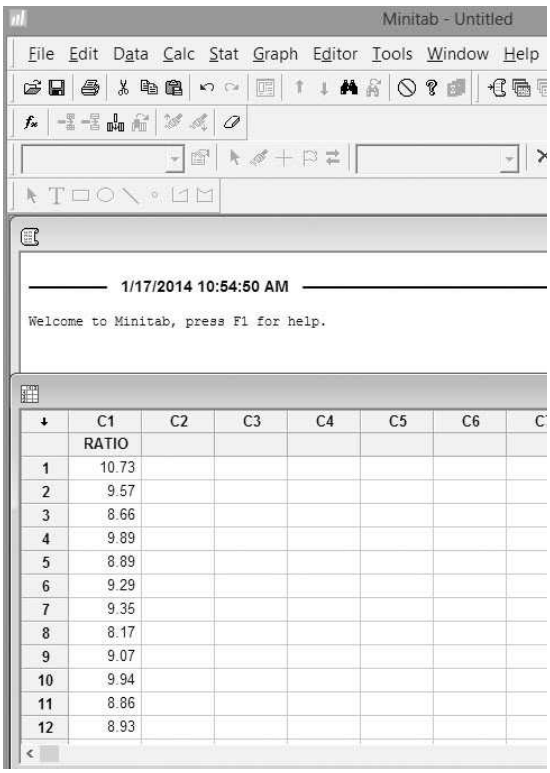
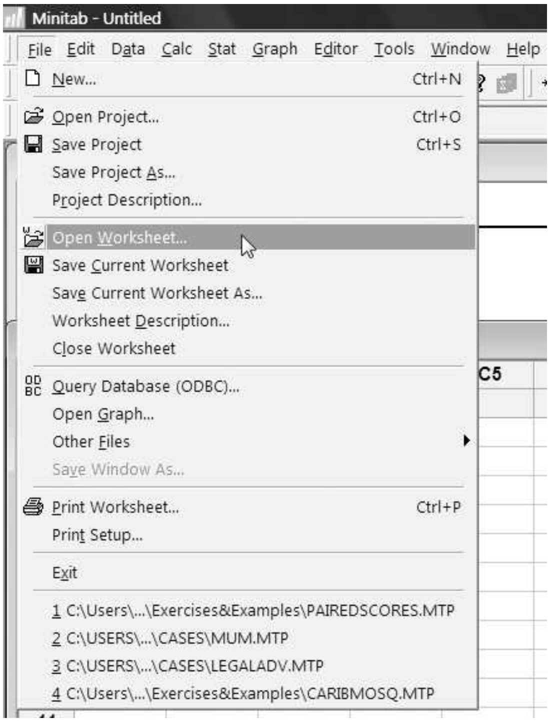


**FIGURE D.3**
Accessing a MINITAB Data File




## Getting a MINITAB File

To access data already saved as a MINITAB file, select "*File*" on the main menu bar, then "*Open Worksheet*", as shown in Figure D.3. In the resulting "Open Worksheet" dialog box (see Figure D.4), select the folder where the data file resides, then select the data set (e.g., BONES). After clicking Open, the data will appear in the spreadsheet.

**FIGURE D.4**
MINITAB Open Worksheet Dialog Box

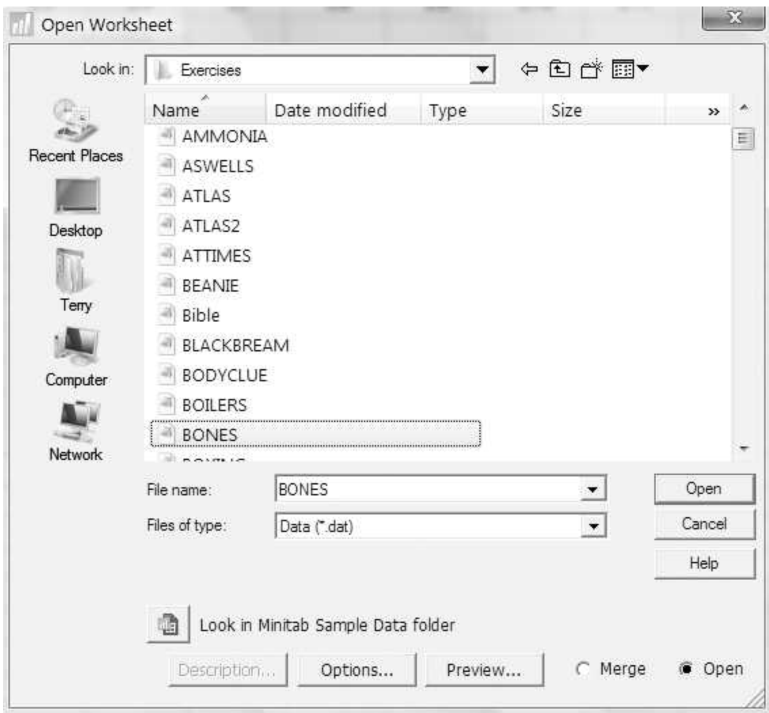
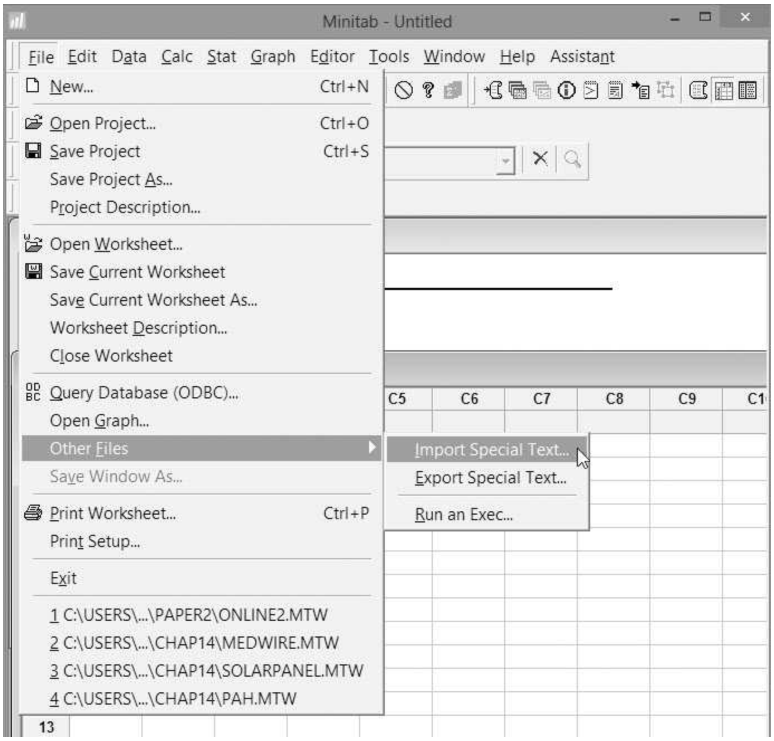


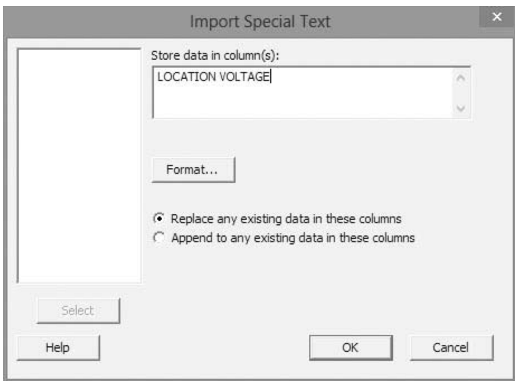
**FIGURE D.5**
MINITAB Options for Accessing an External Data File




## Getting an External File

Finally, if the data are saved in an external text file, access it by selecting "*File*" on the menu bar, click "*Other Files*", then select "*Import Special Text*" (see Figure D.5). The Import Special Text dialog box will appear, as shown in Figure D.6. Specify the variable (column) names, then click OK. On the resulting screen, specify the folder that contains the external data file, click on the file name, then select Open. The MINITAB worksheet will reappear with the data from the external text file.

**FIGURE D.6**
Import Special Text Dialog Box



## D.3 Listing Data

To access a listing (printout) of your data using MINITAB, click on the "*Data*" button on the main menu bar, and then click on "*Display Data*." The resulting menu, or dialog box, appears as in Figure D.7. Enter the names of the variables you want to print in the "*Columns, constants, and matrices to display*" box (you can do this by simply double clicking on the variables), and then click "OK." The printout will show up on your MINITAB session screen.

**FIGURE D.7**
MINITAB Menu Options for Listing Data

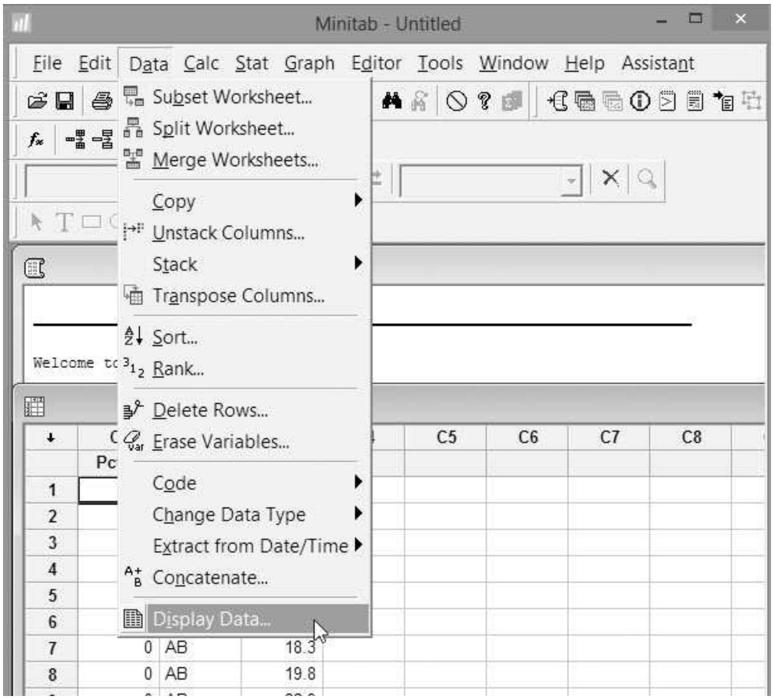
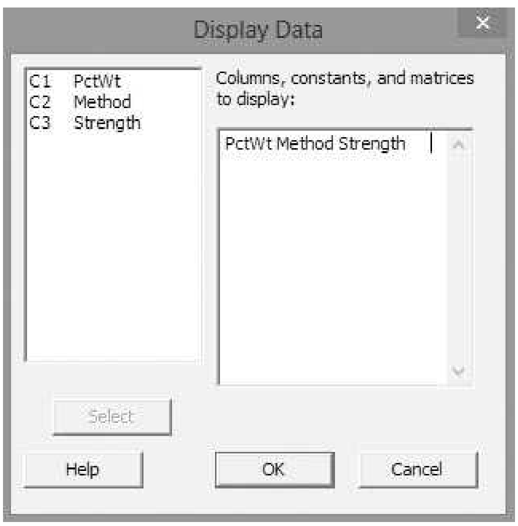


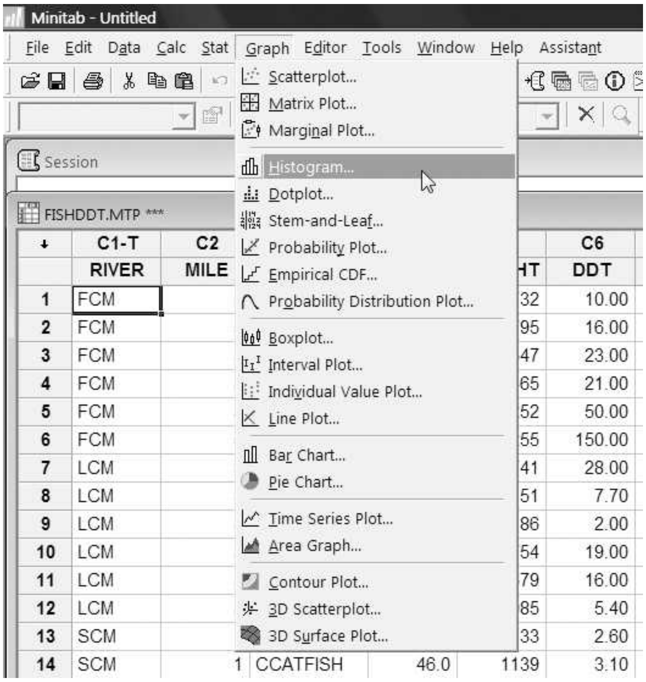

**FIGURE D.8**
Display Data Dialog Box



## D.4 Graphing Data

To obtain graphical descriptions of your data using MINITAB, click on the "Graph" button on the main menu bar, then click on the graph of your choice (Bar Chart, Pie Chart, Scatterplot, Histogram, Dot plot, or Stem-and-Leaf), as shown in Figure D.9. On the resulting dialog box(es), make the appropriate variable selections and click "OK" to view the graph. (The selections for a histogram are shown in Figure D.10.)

**FIGURE D.9**
MINITAB Menu Options for Graphing Data



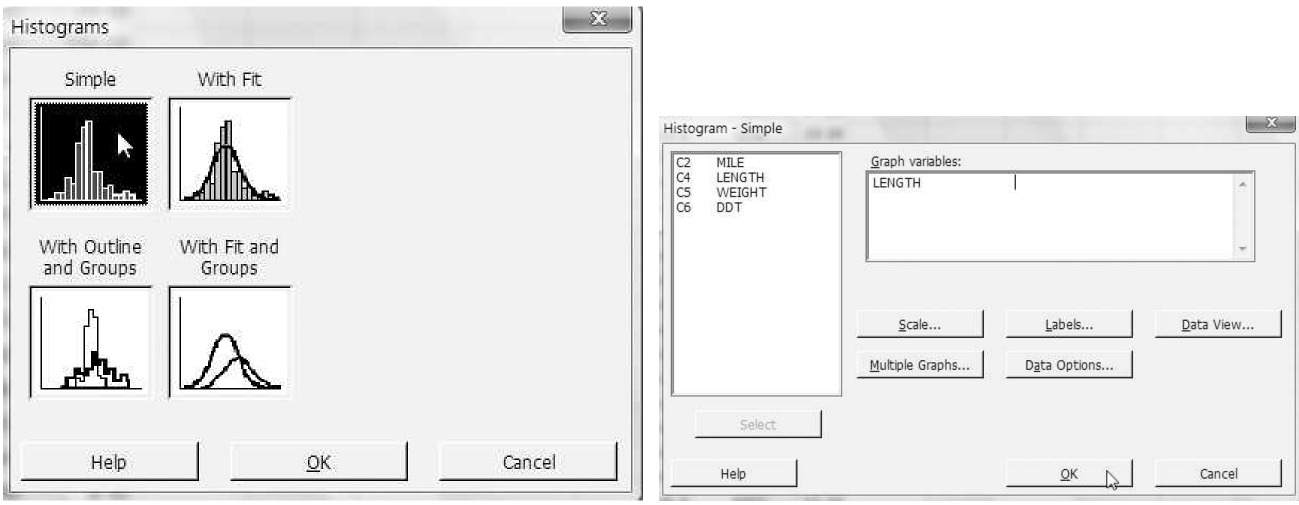




**FIGURE D.10**
Histogram Dialog Boxes

## D.5 Descriptive Statistics, Percentiles, and Correlations

### Descriptive Statistics

To obtain numerical descriptive measures for a quantitative variable (e.g., mean, median, standard deviation, etc.) using MINITAB, click on the "*Stat*" button on the main menu bar, click on "*Basic Statistics*," and then click on "*Display Descriptive Statistics*", as shown in Figure D.11. The resulting dialog box appears in Figure D.12. Select the quantitative variables you want to analyze and place them in the "*Variables*" box. You can control which descriptive statistics appear by clicking the "*Statistics*" button on the dialog box and making your selections.

**FIGURE D.11**
MINITAB Menu Options for Descriptive Statistics

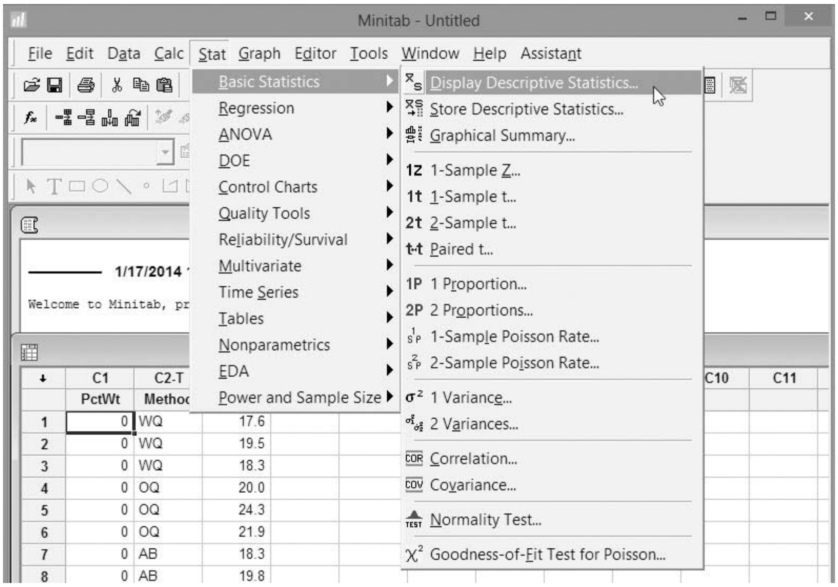
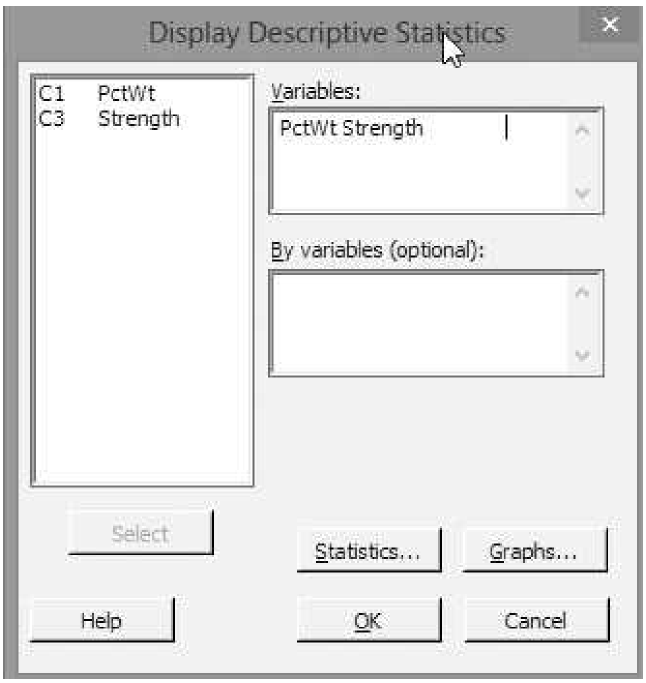



**FIGURE D.12**
Descriptive Statistics Dialog Box



## Percentiles

To obtain percentiles (e.g., 10th percentile, 95th percentile) using MINITAB, click on the "*Calc*" button on the main menu bar, then click on "*Calculator*". The resulting dialog box appears in Figure D.13. In the "*Expression*" box, specify the PERCENTILE function, where the first argument in the parentheses is the column of data you want to analyze and the second argument is the percentile value (e.g., .25, for 25th percentile). Select a column where you want to store the result on the spreadsheet, then click "OK".

## Correlations

To obtain Pearson product moment correlations for pairs of quantitative variables, click on the "*Stat*" button on the MINITAB main menu bar, click on "*Basic Statistics*," and then click on "*Correlations*" (See Figure D.11.) On the resulting dialog box, double click on the variables you want to analyze to move them into the "Variables" box on the right panel. (See Figure D.14.) Click "OK" to obtain a printout of the correlations.

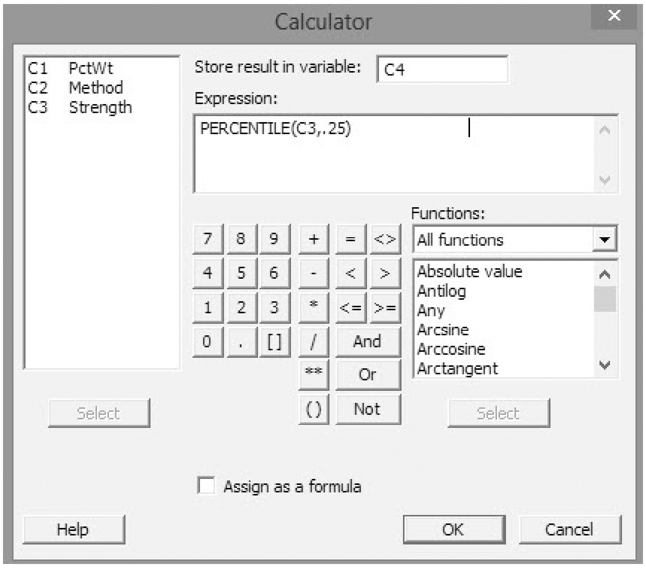



**FIGURE D.13**
Calculator Dialog Box with Percentile Function

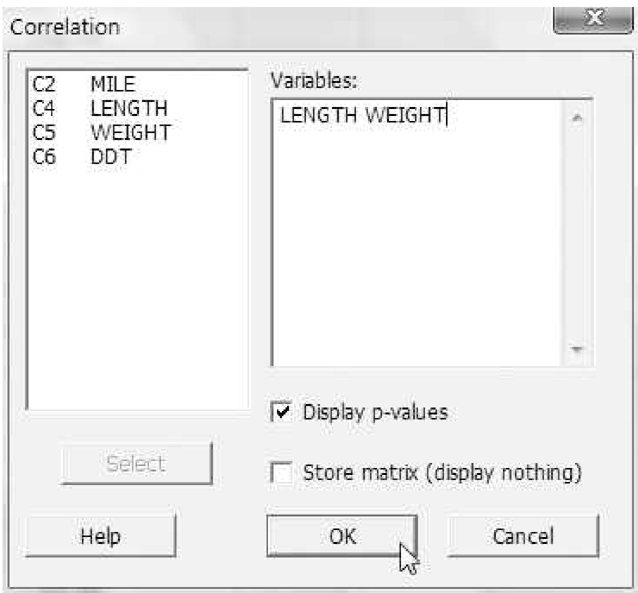



**FIGURE D.14**
Correlations Dialog Box

## D.6 Confidence Intervals and Hypothesis Tests for a Mean, Proportion, or Variance

### Population Mean

To conduct a test of hypothesis and form a confidence interval for a single population mean of a quantitative variable, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*1-Sample t*" (See Figure D.11.) On the resulting dialog box (shown in Figure D.15), click on "*Samples in Columns*," and then specify the quantitative variable of interest in the open box. Check "*Perform hypothesis test*" and specify the value of the hypothesized mean. Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

*Note:* If you want to produce a confidence interval and/or hypothesis test for the mean from summary information (e.g., the sample mean, sample standard deviation, and sample size), click on "*Summarized data*" in the "1-Sample t" dialog box (Figure D.15). Enter the values of the summary statistics and then click "OK."

### Population Proportion

To conduct a test of hypothesis and form a confidence interval for a single population proportion for a two-level (binomial) qualitative variable, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*1-Proportion*" (See Figure D.11.) On the resulting dialog box (shown in Figure D.16), click on "*Samples in Columns*," and then specify the qualitative variable of interest in the open box. Check "*Perform hypothesis test*" and specify the value of the hypothesized proportion. Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

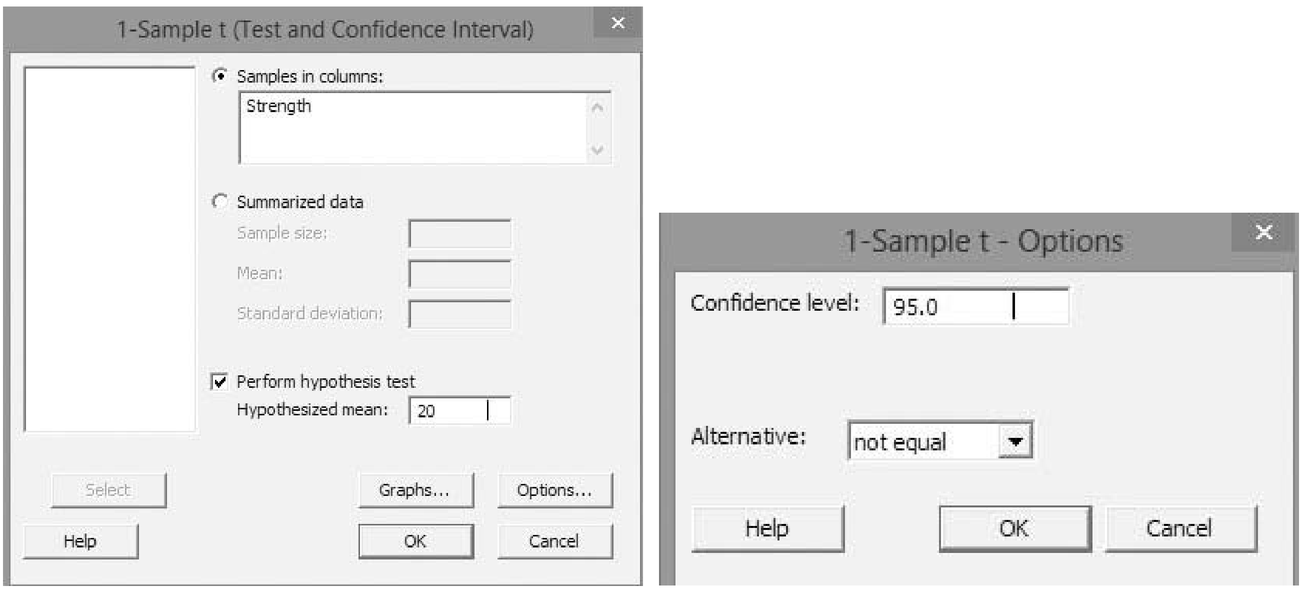



**FIGURE D.15**
One-sample t-Test for Mean Dialog Boxes

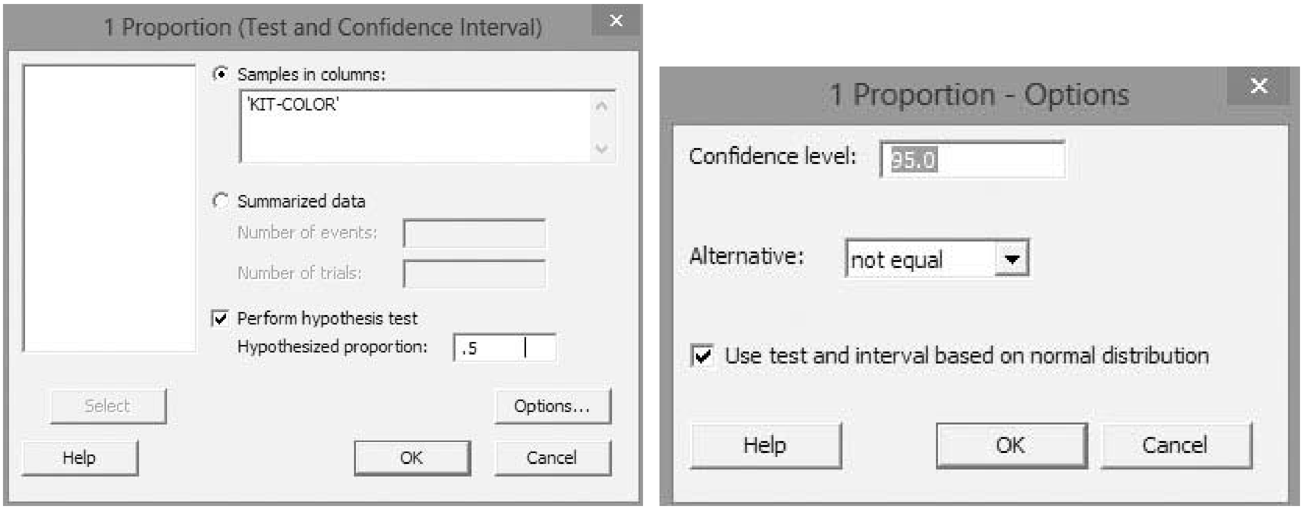



**FIGURE D.16**
One-Proportion Dialog Boxes

## Population Variance

To conduct a test of hypothesis and form a confidence interval for a single population variance of a quantitative variable, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*1 Variance*" (See Figure D.11.) On the resulting dialog box (shown in Figure D.17), click on "*Samples in Columns*," in the "*Data*" box, then specify the quantitative variable of interest in the "*Columns*" box. Check "*Perform hypothesis test*" and specify the value of the hypothesized standard deviation (or, optionally, the variance). Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

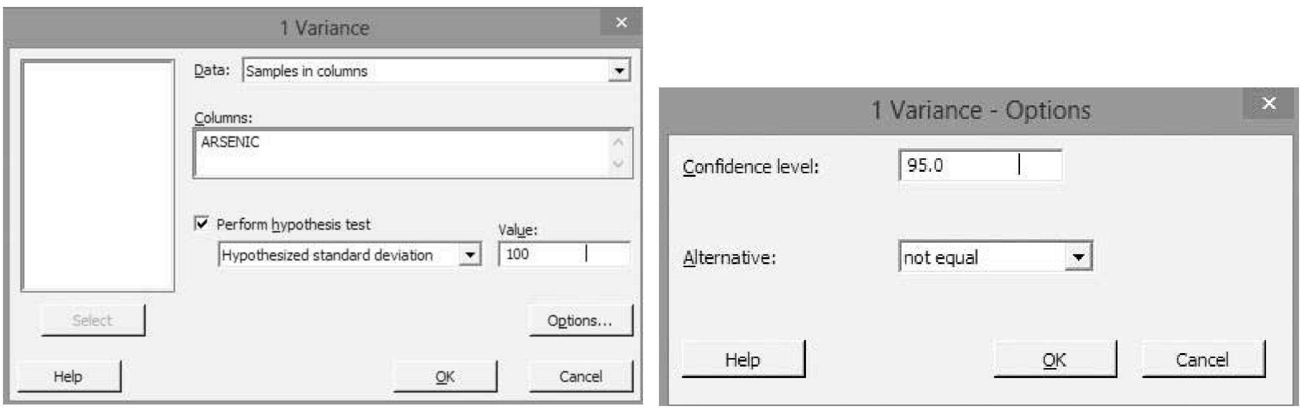



**FIGURE D.17**
One Variance Dialog Boxes

## D.7 Confidence Intervals and Hypothesis Tests for the Difference Between Means, Proportions, or Variances

### Two Means, Independent Samples

To conduct a test of hypothesis and form a confidence interval for the difference between two population means based on independent samples, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*2-Sample t*" (See Figure D.11.)

If the worksheet contains data for one quantitative variable (which the means will be computed on) and one qualitative variable (which represents the two groups or populations), select "*Samples in one column*" and then specify the quantitative variable in the "*Samples*" area and the qualitative variable in the "*Subscripts*" area. (See Figure D.18, left panel.)

If the worksheet contains the data for the first sample in one column and the data for the second sample in another column, select "*Samples in different columns*" and then specify the "*First*" and "*Second*" variables. Alternatively, if you have only summarized data (i.e., sample sizes, sample means, and sample standard deviations), select "*Summarized data*" and enter these summarized values in the appropriate boxes. (See Figure D.18, right panel.)

Note: If the sample sizes are small, be sure to check the "Assume equal variances" box. For large samples, leave this box unchecked.

Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level, the null hypothesized value of the difference, and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

### Two Means, Matched Pairs

To conduct a test of hypothesis and form a confidence interval for the difference between two population means based on matched pairs data, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*Paired t*" (See Figure D.11.)

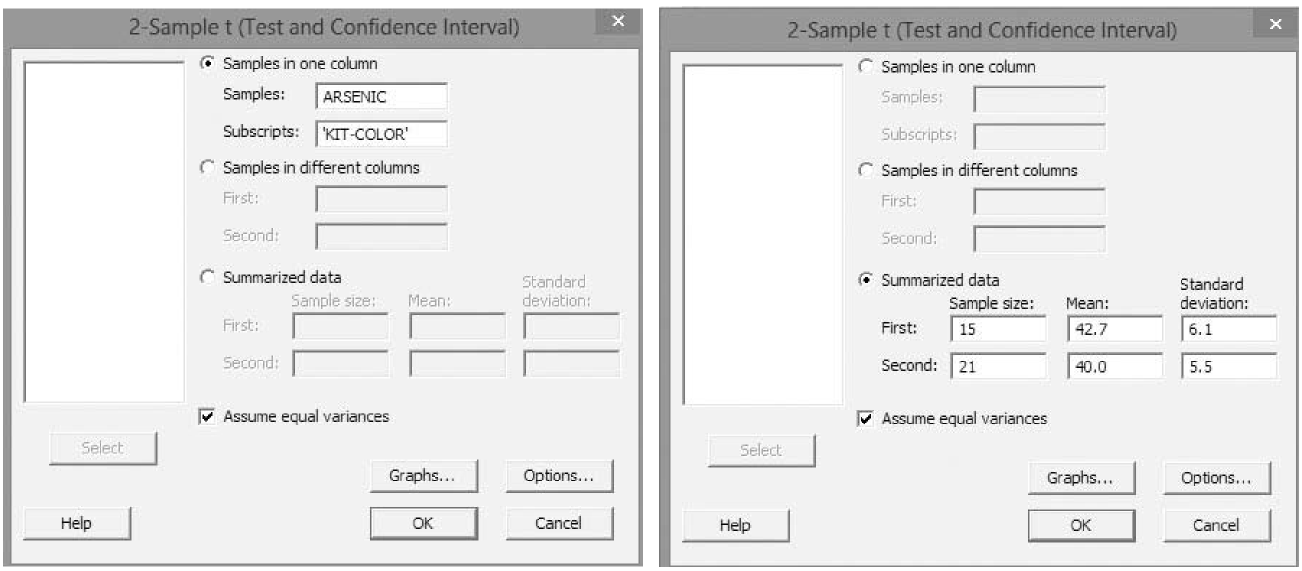



**FIGURE D.18**
Two Means Comparison Dialog Boxes

If the worksheet contains the data for the first sample in one column and the data for the second sample in another column, select "*Samples in columns*" and then specify the "*First sample*" and "*Second sample*" variables. Alternatively, if you have only summarized data (i.e., sample size, sample mean difference, and sample standard deviation of the differences), select "*Summarized data*" and enter these summarized values in the appropriate boxes.

Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level, the null hypothesized value of the difference, and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

## Two Proportions

To conduct a test of hypothesis and form a confidence interval for the difference between two population proportions based on independent samples, click on the "*Stat*" button on the MINITAB menu bar and then click on "*Basic Statistics*" and "*2 Proportions*" (See Figure D.11.) On the resulting dialog box (shown in Figure D.19, left panel), select the data option ("*Samples in different columns*" or "*Summarized data*") and make the appropriate menu choices. (Figure D.19 shows the menu options when you select "*Summarized data*.")

Click the "*Options*" button and specify the confidence level for a confidence interval, the null-hypothesized value of the difference, and the form of the alternative hypothesis (lower tailed, two tailed, or upper tailed) in the resulting dialog box, as shown in Figure D.19 (right panel). If you desire a pooled estimate of $p$ for the test, be sure to check the appropriate box. Click "*OK*" twice to produce the results.

## Two Variances

To conduct a test of hypothesis and form a confidence interval for the ratio of two population variances based on independent samples, click on the "*Stat*" button on the

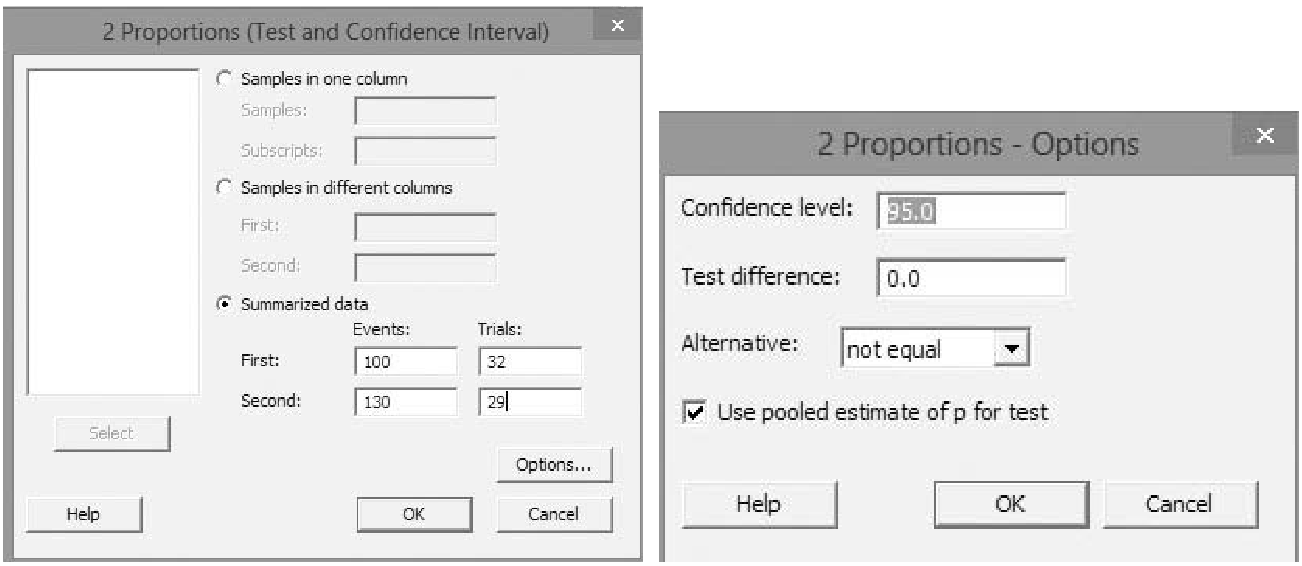



**FIGURE D.19**
Two Proportions Comparison Dialog Boxes

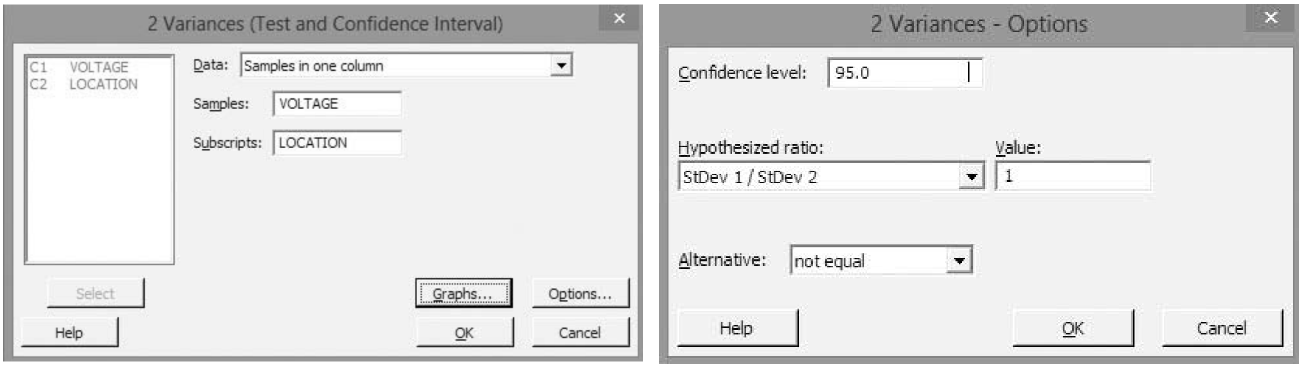



**FIGURE D.20**
Two Variances Comparison Dialog Boxes

MINITAB menu bar and then click on "*Basic Statistics*" and "*2 Variances*" (See Figure D.11.) On the resulting dialog box (shown in Figure D.20, left panel), select the data option ("*Samples in one column*", "*Samples in different columns*", "*Sample standard deviations*" or "*Sample variances*") and make the appropriate menu choices. (Figure D.20 shows the menu options when you select "*Samples in one column*.")

Click on the "*Options*" button at the bottom of the dialog box and specify the confidence level, the null hypothesized value of the ratio, and the form of the alternative hypothesis in the resulting dialog box. Click "OK" twice to obtain a printout of the results.

## D.8 Categorical Data Analysis

MINITAB can produce a frequency table for a single qualitative variable (i.e., a one-way table) and can conduct a chi-square test for independence of two qualitative variables in a two-way (contingency) table.

### One-Way Table

For a one-way table, click on the "*Stat*" button on the MINITAB menu bar, then click on "*Tables*", and "*Chi-square Goodness-of-Fit Test (One Variable)*". (See Figure D.21.) A dialog box similar to Figure D.22 will appear.

**FIGURE D.21**
MINITAB Menu Options for a One-Way Frequency Table Analysis

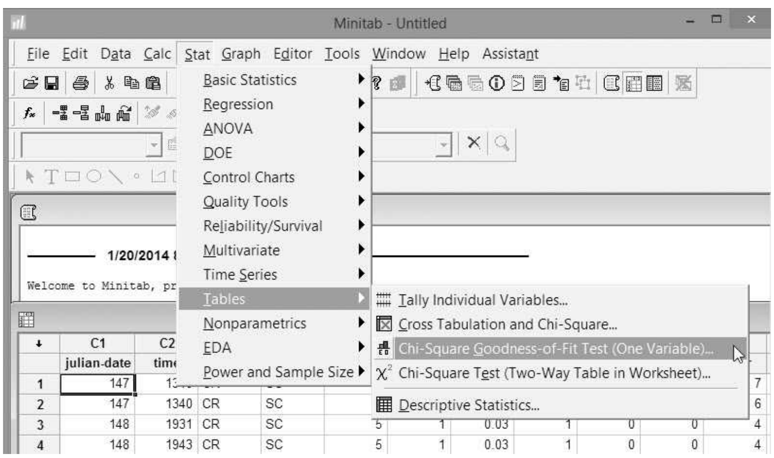
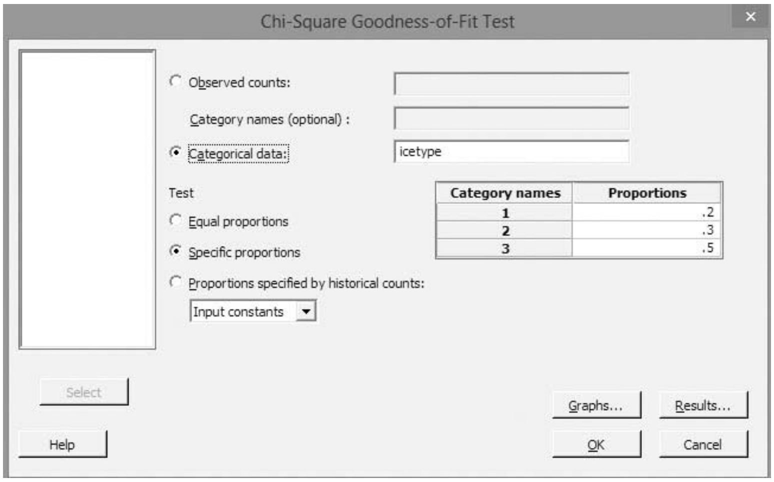


**FIGURE D.22**
One-Way Frequency Table Dialog Box



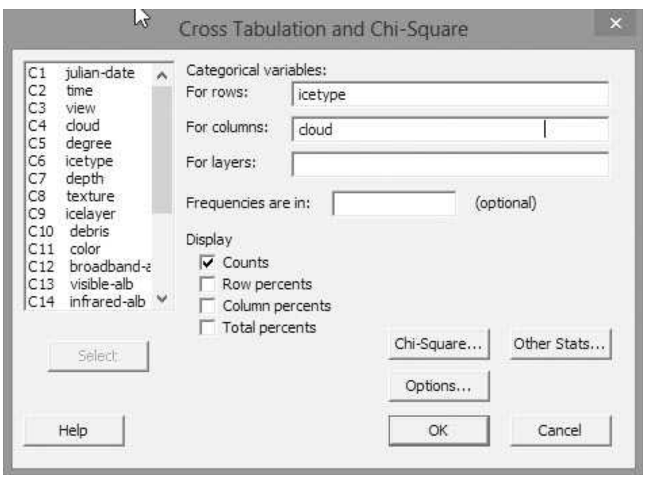

If your data have one column of values for your qualitative variable, select "Categorical data" and specify the variable name (or column) in the box. If your data have summary information in two columns — one column listing the levels of the qualitative variable and the other column with the observed counts for each level. Select "Observed counts" and specify the column with the counts and the column with the variable names in the respective boxes. Select "Equal proportions" for a test of equal proportions or select "Specific proportions" and enter the hypothesized proportion next to each level in the resulting box (as shown in Figure D.22). Click "OK" to generate the MINITAB printout.

## Two-Way Table

For a two-way table, click on the "*Stat*" button on the MINITAB menu bar, then click on "*Tables*", and "*Cross Tabulation and Chi-Square*". (See Figure D.21.) A dialog box similar to Figure D.23 will appear.

Specify one qualitative variable in the "*For rows*" box and the other qualitative variable in the "*For columns*" box, as shown in Figure D.23. [*Note:* If your worksheet contains cell counts for the categories, enter the variable with the cell counts in the "Frequencies are in" box.] Click the "*Chi-Square*" button, and then select the statistics

**FIGURE D.23**
Two-Way (Contingency) Table Dialog Box



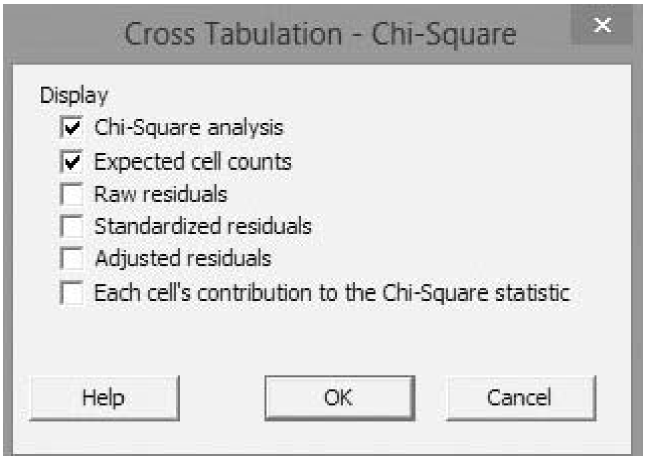



**FIGURE D.24**
Selecting Statistics for the Two-Way (Contingency) Table

you want to display in the table by making the appropriate selections in the resulting dialog box (see Figure D.24). Click "*OK*" twice to generate the MINITAB printout.

*Note:* If your MINITAB worksheet contains only the cell counts for the contingency table in columns, click the "*Chi-Square Test (Two-Way Table in Worksheet)*" menu option (see Figure D.21) and specify the columns in the "*Columns containing the table*" box. Click "*OK*" to produce the MINITAB printout.

## D.9 Simple Linear Regression

To conduct a simple linear regression analysis, click on the "*Stat*" button on the MINITAB menu bar, then click on "*Regression*", and "*Regression*" again. (See Figure D.25.) On the resulting dialog box, specify the quantitative dependent variable in the "*Response*" box and the quantitative independent variable in the "*Predictors*" box on the right side of the menu, as shown in Figure D.26.

To produce prediction intervals for $y$ and confidence intervals for $E(y)$, click the "Options" button. The resulting dialog box is shown in Figure D.27. Check "Confidence limits" and/or "Prediction limits," specify the "Confidence level," and enter the

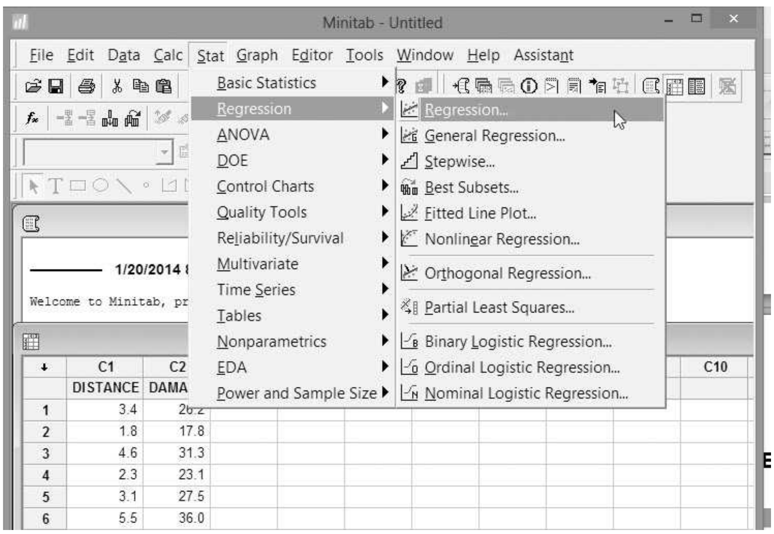



**FIGURE D.25**
MINITAB Menu Options for Simple Linear Regression

**FIGURE D.26**
Simple Linear Regression Dialog Box

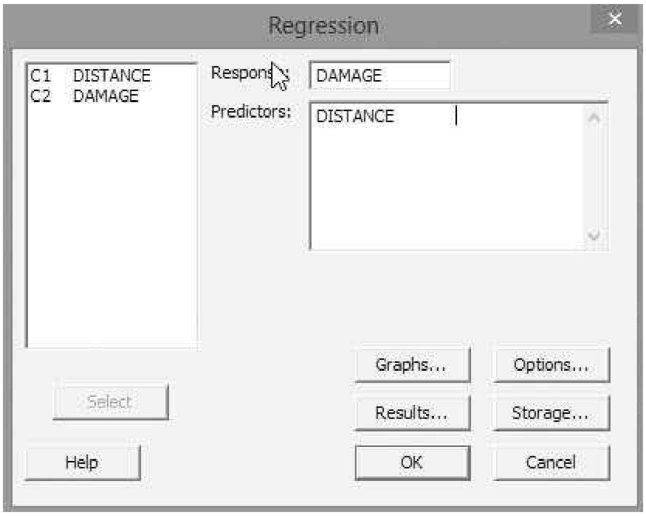
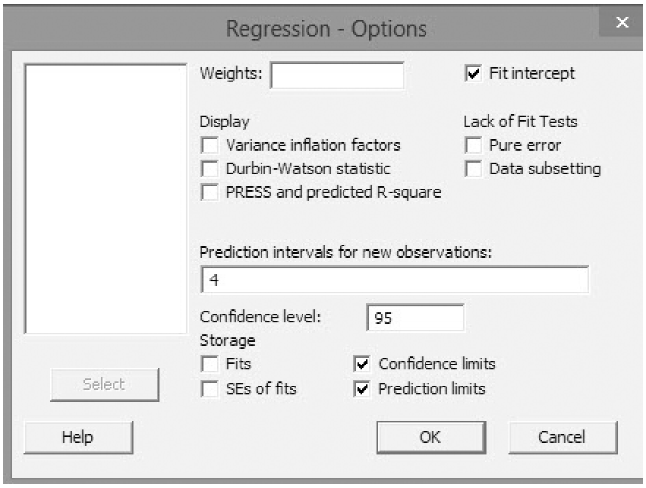



**FIGURE D.27**
MINITAB Simple Linear Regression Options




value of $x$ in the "Prediction intervals for new observations" box. Click "OK" to return to the main Regression dialog box and then click "OK" again to produce the MINITAB simple linear regression printout.

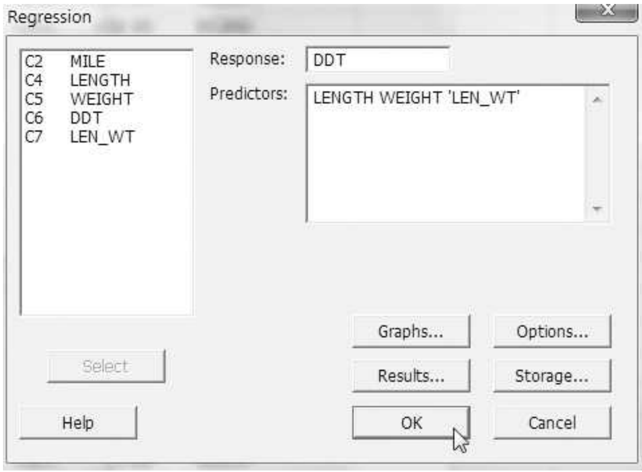
## D.10 Multiple Regression

To conduct a multiple regression analysis, click on the "*Stat*" button on the MINITAB menu bar, then click on "*Regression*", and "*Regression*" again. (See Figure D.25.) On the resulting dialog box, specify the quantitative dependent variable in the "*Response*" box and the independent variables in the "*Predictors*" box on the right side of the menu, as shown in Figure D.28.

[*Note:* If your model includes dummy variables, interactions and/or squared terms, you must create and add these variables to the MINITAB worksheet *prior* to running a regression analysis. You can do this by clicking the "Calc" button on the MINITAB main menu and selecting the "Calculator" option.]

To produce prediction intervals for $y$ and confidence intervals for $E(y)$, click the "Options" button. On the resulting dialog box (similar to Figure D.27), check "Confidence limits" and/or "Prediction limits," specify the "Confidence level," and enter the

**FIGURE D.28**
Regression Dialog Box




value of $x$ in the "Prediction intervals for new observations" box. Click "OK" to return to the main Regression dialog box.

Residual plots are obtained by clicking the "Graphs" button and making the appropriate selections on the resulting menu (see left panel, Figure D.29). Influence diagnostics (e.g., studentized deleted residuals, leverage values, Cook's distances) are

**FIGURE D.29**
Menu Selections for Residual Analysis

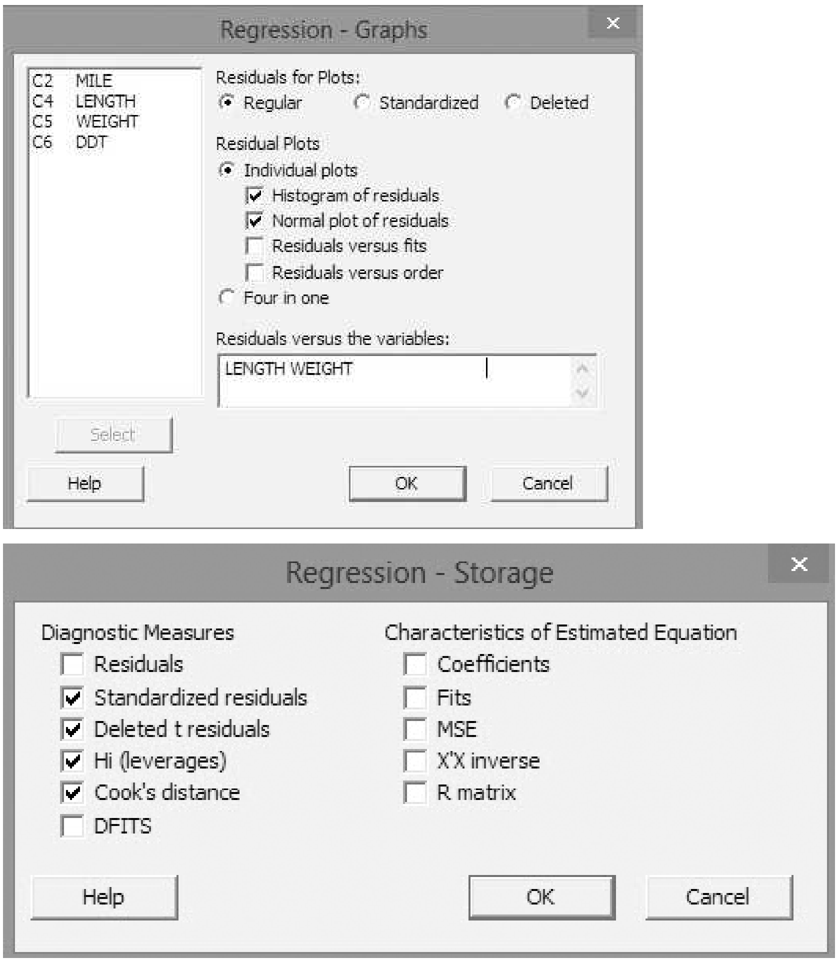
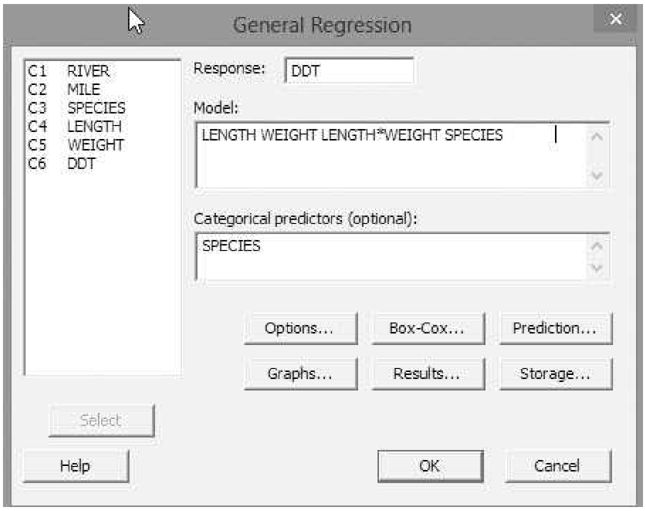


**FIGURE D.30**
General Linear Models Dialog Box




obtained by clicking the "Storage" option and checking the diagnostics on the resulting menu screen (see right panel, Figure D.29). When you have made all your selections, click "OK" on the main Regression dialog box to produce the MINITAB multiple regression printout and graphs.

## Fitting General Linear Models

As an alternative, you can fit general linear models in MINITAB without having to create dummy variables and higher-order terms in the worksheet. To do this, click on the "*Stat*" button on the main menu bar, then click on "*Regression*", and finally click on "*General Regression*" (See the menu options in Figure D.25) . On the resulting menu screen (see Figure D.30), specify the quantitative dependent variable in the "*Response*" box and any qualitative independent variables in the "*Categorical predictors*" box. (Note: MINITAB will automatically create the appropriate number of dummy variables for each qualitative variable specified.) Specify the terms in the model in the "Model" box. You specify interactions and squared terms by placing an asterisk between variable names (e.g., LENGTH*SPECIES or LENGTH*LENGTH). Click "*OK*" to produce the MINITAB printout.

## Stepwise Regression

To conduct a stepwise regression analysis, click on the "*Stat*" button on the main menu bar, then click on "*Regression*", and finally click on "*Stepwise Regression*" (see the menu options in Figure D.25). On the resulting menu screen, specify the quantitative dependent variable in the "*Response*" box and all the potential independent variables in the "*Predictors*", as shown in Figure D.31. Click on the "*Methods*" button to obtain a menu screen that will allow you to choose either stepwise selection, forward selection, or backwards elimination. (The default method is stepwise selection.) Click "OK" twice to obtain the MINITAB printout.

## All-Possible-Regressions-Selection

To run the all-possible-regressions-selection method, click on the "*Stat*" button on the main menu bar, then click on "*Regression*", and finally click on "*Best Subsets*" (see

**FIGURE D.31**
Stepwise Regression Dialog Box

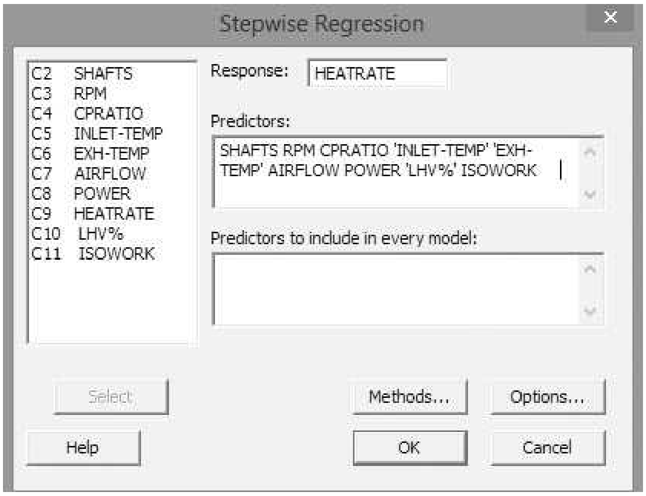
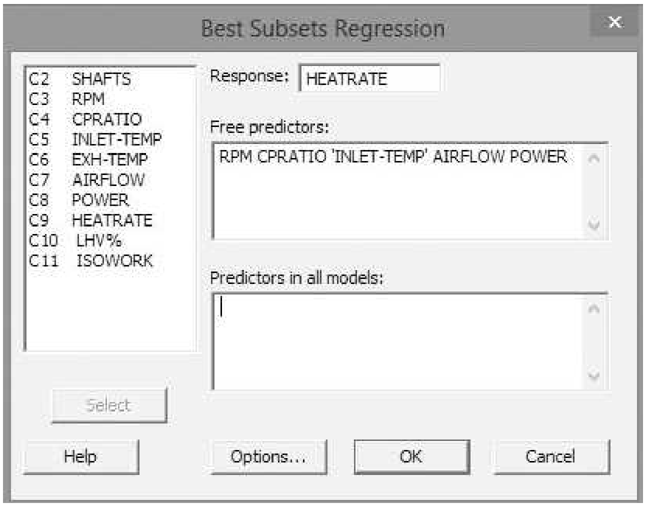



**FIGURE D.32**
All-Possible-Regressions-Selection Dialog Box




the menu options in Figure D.25). On the resulting menu screen, specify the quantitative dependent variable in the "*Response*" box and all the potential independent variables in the "*Free Predictors*" box, as shown in Figure D.32. Click "OK" to obtain the MINITAB printout.

## D.11 One-Way Analysis of Variance

To conduct a one-way ANOVA for a completely randomized design using MINITAB, click on the "*Stat*" button on the main menu bar, then click on "*ANOVA*", and "*One-Way*". (See Figure D.33.) On the resulting dialog screen (Figure D.34), specify the response variable in the "*Response*" box and the factor variable in the "Factor" box.

To perform multiple comparisons of treatment means, click the "*Comparison*" button to obtain the dialog box shown in Figure D.34. On this box, check the comparison method (e.g., "*Tukey's*" method) and specify the comparison-wise error rate (e.g., "*family error rate*"). Click "OK" twice to produce the MINITAB printout.

**FIGURE D.33**
MINITAB Menu Options for One-Way ANOVA

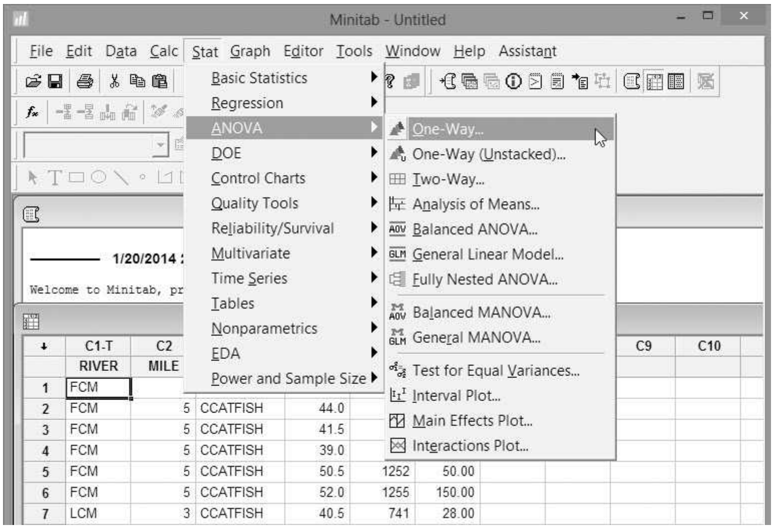
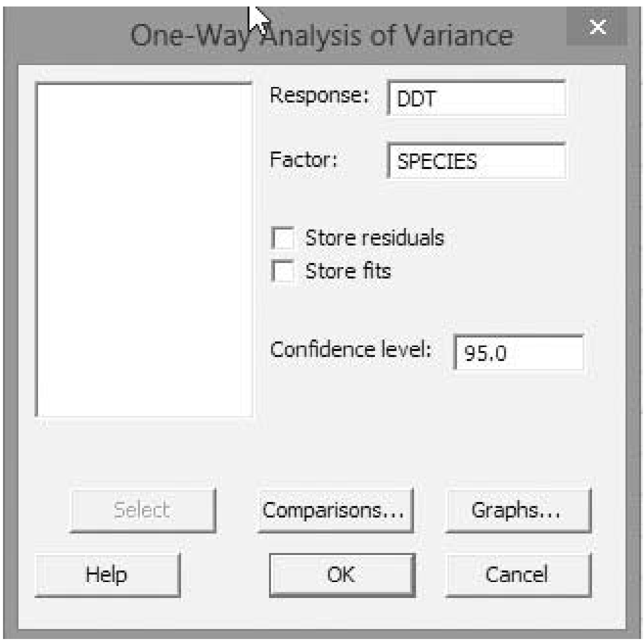



**FIGURE D.34**
One-Way ANOVA Dialog Box




To perform a test of equality of variances, click on the "*Stat*" button on the main menu bar, then click on "*ANOVA*", and "*Test for Equal Variances*". (See Figure D.33.) On the resulting dialog screen (Figure D.36), specify the response variable in the "*Response*" box, the factor variable in the "*Factor*" box, and the confidence level of the test. Click "*OK*" to view the MINITAB results (both Bartlett's and Levene's test).

**FIGURE D.35**
Multiple Comparisons Dialog Box for ANOVA

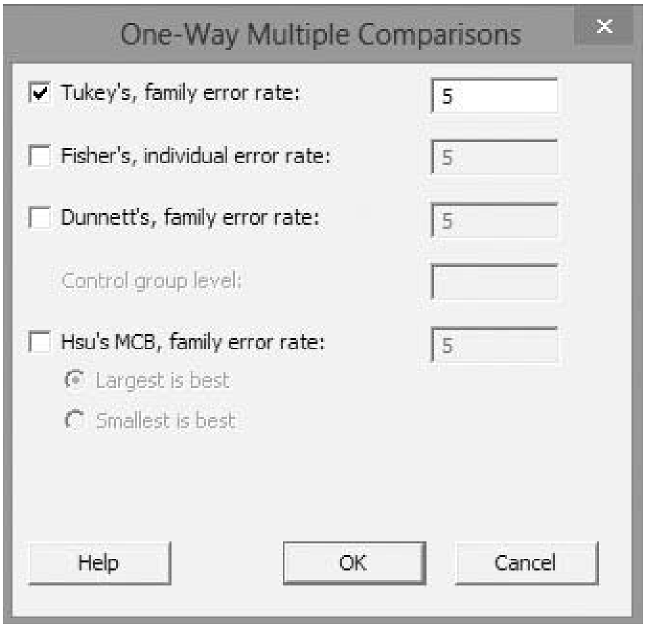
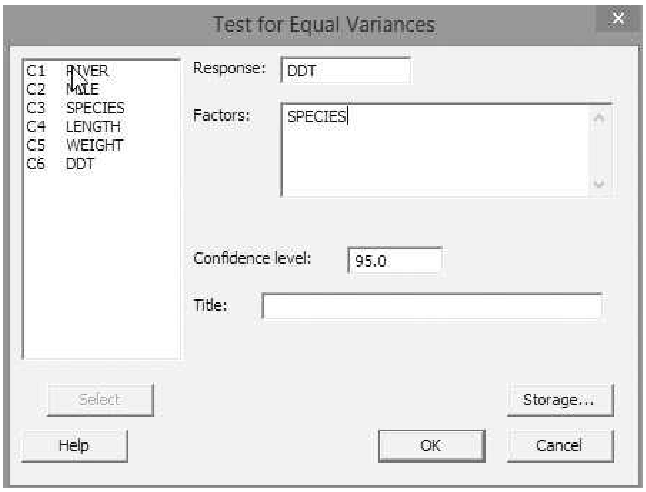



**FIGURE D.36**
Testing for Equality of Variances Dialog Box

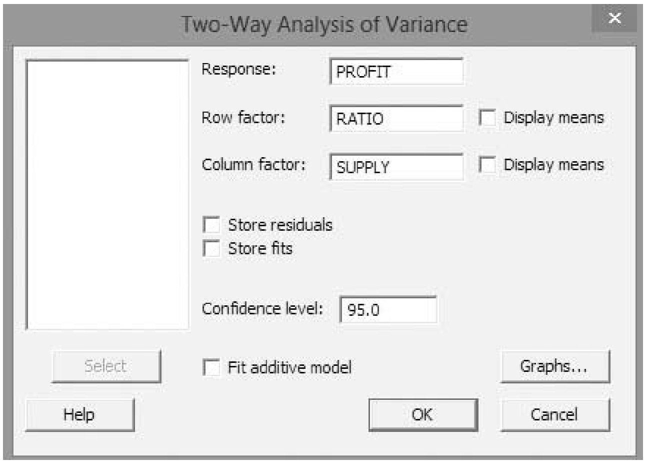



## D.12 Analysis of Variance for Factorial and Other Designs

### Two-Factor Designs

To conduct an ANOVA for designs involving two factors (e.g., randomized block, two-factor factorial designs), click on the "*Stat*" button on the main menu bar, then click on "*ANOVA*", and "*Two-Way*". (See Figure D.33.) On the resulting dialog screen (Figure D.37), specify the response variable in the "*Response*" box and the two factor variables in the "*Row factor*" and "*Column factor*" boxes. The MINITAB default is to fit a model with factor interaction. If you do not want to include interaction in the model (e.g., a model for a randomized block design, where one of the factors represents blocks), then check the "*Fit additive model*" box. Click "OK" to produce the MINITAB printout.

**FIGURE D.37**
Two-Way ANOVA Dialog Box

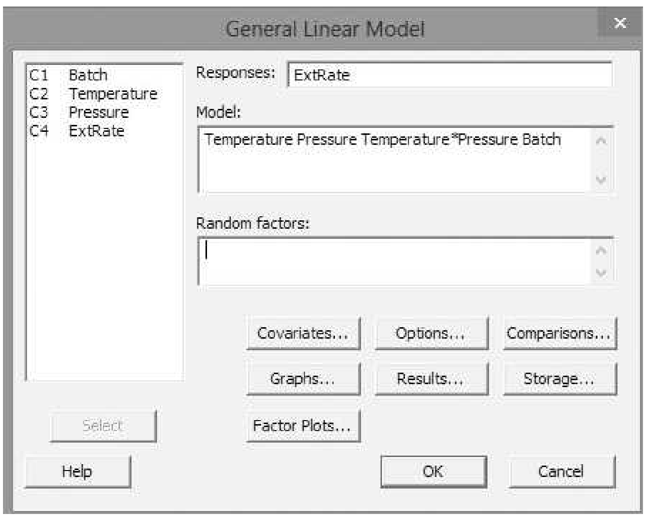



## Multi-Factor Designs

To conduct an ANOVA for designs involving more than two factors or more complex designs, click on the "*Stat*" button on the main menu bar, then click on "*ANOVA*", and "*General Linear Model*". (See Figure D.33.) On the resulting dialog screen (Figure D.38), specify the response variable in the "*Response*" box and the effects in the model in the "Model" box. You specify interactions by placing an asterisk between variable names (e.g., TEMP*PRESSURE).

To run multiple comparisons of means, click the "*Comparisons*" button. On the resulting dialog box (see Figure D.39), select "*Pairwise comparisons*", specify the effects you want to compare means on in the "*Terms*" box, check the comparison method (e.g., "*Bonferroni*") and specify the experiment-wise confidence level. Click "*OK*" twice to produce the MINITAB printout.

**FIGURE D.38**
General Linear Model Dialog Box



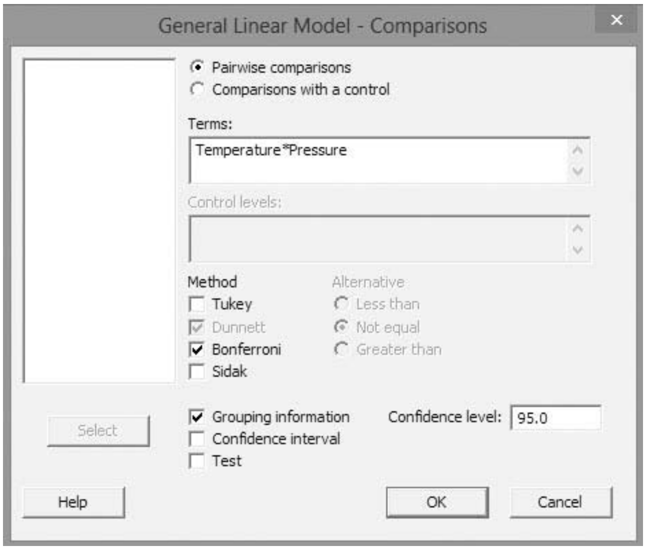



**FIGURE D.39**
Multiple Comparisons of Means Dialog Box

## D.13 Nonparametric Tests

MINITAB can perform the following nonparametric tests: sign test, Wilcoxon rank sum test, Wilcoxon signed-ranks test, Kruskal-Wallis test, Friedman test and Spearman's rank correlation test. All but Spearman's test are produced by making the following menu selections: Click on the "*Stat*" button on the MINITAB main menu bar, then click on "*Nonparametrics*", and select the test you want to run (e.g., "*1-Sample Sign*" test). These menu options are shown on Figure D.40.

### Sign Test

After selecting "*1-Sample Sign*" from the nonparametrics menu, the dialog screen shown in Figure D.41 will appear. Specify the variable to be analyzed in the

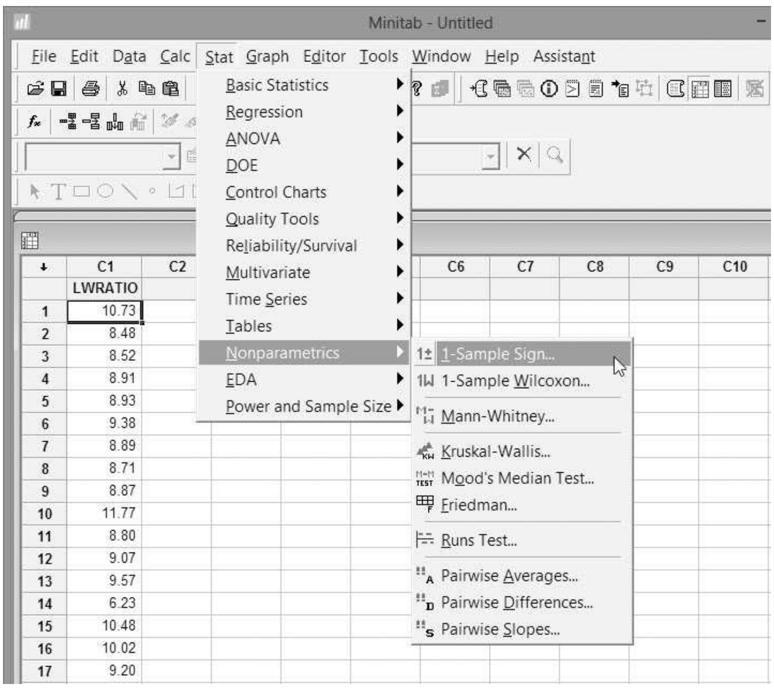
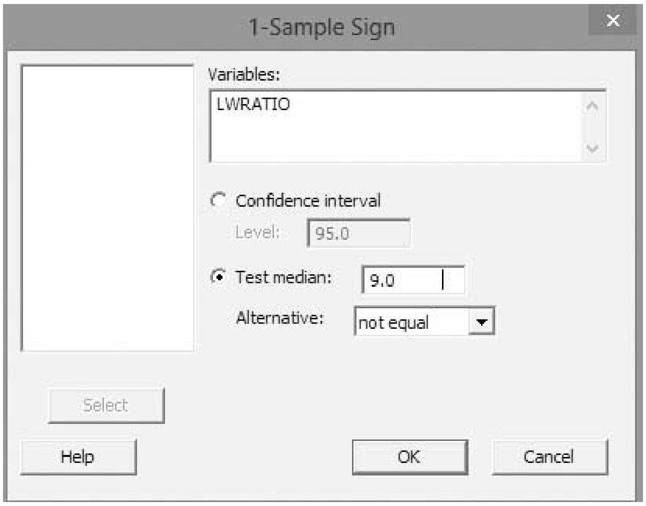



**FIGURE D.40**
MINITAB Menu Options for Nonparametric Tests

**FIGURE D.41**
Sign Test Dialog Box



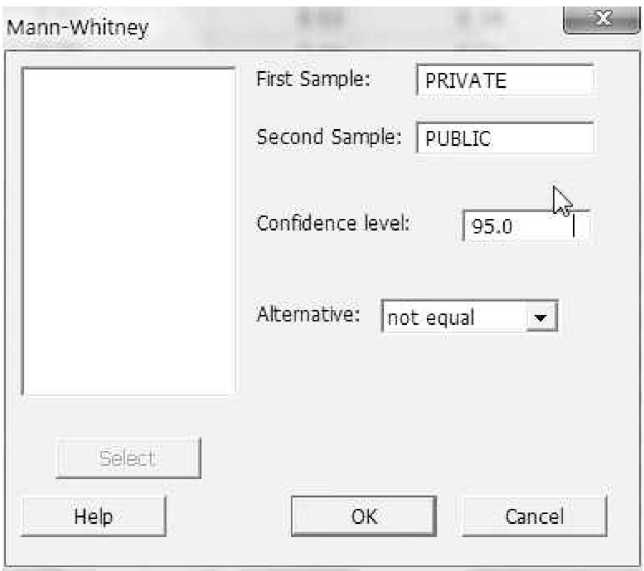

"*Variables*" box. Select "*Test median*", then enter the null hypothesized value of the median and select the form of the alternative hypothesis. Click "*OK*" to view the MINITAB printout.

## Rank Sum Test

To run a Wilcoxon rank sum test (also called the Mann-Whitney test) for independent samples, your data must be two columns on the worksheet — one column for each sample. Select "*Mann-Whitney*" from the nonparametrics menu list (see Figure D.40). The dialog screen shown in Figure D.42 will appear. Specify the variable for the first sample in the "*First Sample*" box and the variable for the second sample in the "*Second Sample*" box. Specify the confidence level and the form of the alternative hypothesis, then click "*OK*" to view the MINITAB printout.

*Signed-Ranks Test:* To run a Wilcoxon signed-ranks test for matched pairs, your paired data must be two columns on the worksheet — one column for each member of the pair. Compute the difference between these two variables and save it in a column on the worksheet. (Use the "Calc" button on the MINITAB menu bar.)

**FIGURE D.42**
Mann-Whitney Test Dialog Box



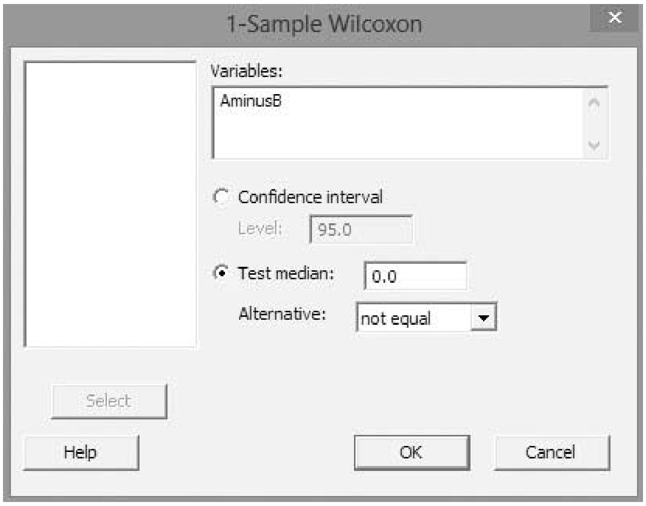



**FIGURE D.43**
Wilcoxon Signed-Ranks Dialog Box

Now select "*1-Sample Wilcoxon*" from the nonparametrics menu list (see Figure D.40). The dialog screen shown in Figure D.43 will appear. Enter the variable representing the paired differences in the "*Variables*" box. Select the "*Test median*" option and specify the hypothesized value of the median as "0." Select the form of the alternative hypothesis ("not equal," "less than," or "greater than"). Click "*OK*" to generate the MINITAB printout.

## Kruskal-Wallis Test

To run a Kruskal-Wallis test for a one-way ANOVA design, your data must be in two columns on the worksheet — one column for the dependent variable and one column representing the treatments. Select "*Kruskal-Wallis*" from the nonparametrics menu list (see Figure D.40). The dialog screen shown in Figure D.44 will appear. Specify the

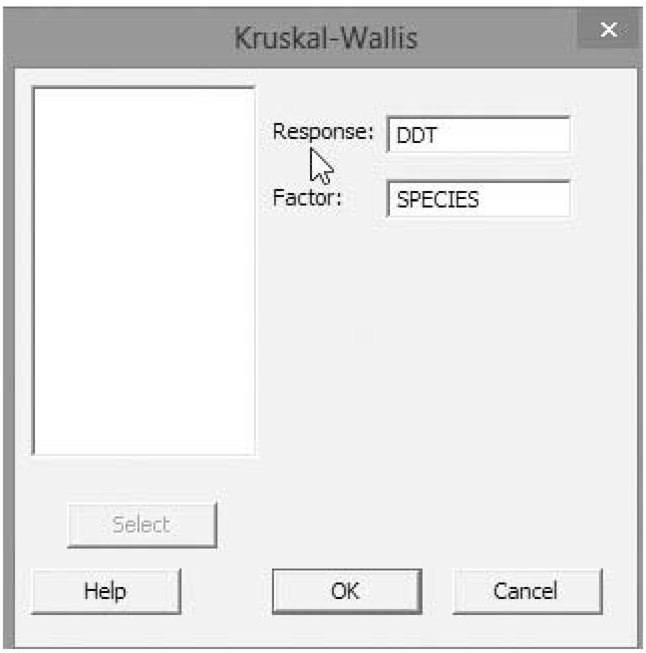



**FIGURE D.44**
Kruskal-Wallis Test Dialog Box

dependent variable in the "*Response*" box and the treatment (factor) variable in the "*Factor*" box. Click "*OK*" to view the MINITAB printout.

## Friedman Test

To run a Friedman test for a randomized block ANOVA design, your data must be in three columns on the worksheet — one column for the dependent variable, one column representing the treatments, and one column representing the blocks. Select "*Friedman*" from the nonparametrics menu list (see Figure D.40). The dialog screen shown in Figure D.45 will appear. Specify the dependent variable in the "*Response*" box, the treatment (factor) variable in the "*Treatment*" box, and the blocking variable in the "*Blocks*" box. Click "*OK*" to view the MINITAB printout.

## Rank Correlation Test

To obtain Spearman's rank correlation coefficient in MINITAB, you must first rank the values of the two quantitative variables of interest. Click the "*Calc*" button on the MINITAB menu bar and create two additional columns, one for the ranks of the $x$-variable and one for the ranks of the $y$-variable. (Use the "*Rank*" function on the MINITAB calculator as shown in Figure D.46).

After you have ranked the variables, click on the "*Stat*" button on the main menu bar, then click on "*Basic Statistics*" and "*Correlation*." On the resulting dialog box (see Figure D.47), enter the ranked variables in the "*Variables*" box and unselect the "*Display p-values*" option. Click "OK" to obtain the MINITAB printout. (You will need to look up the critical value of Spearman's rank correlation to conduct the test.)

**FIGURE D.45**
Friedman Test Dialog Box

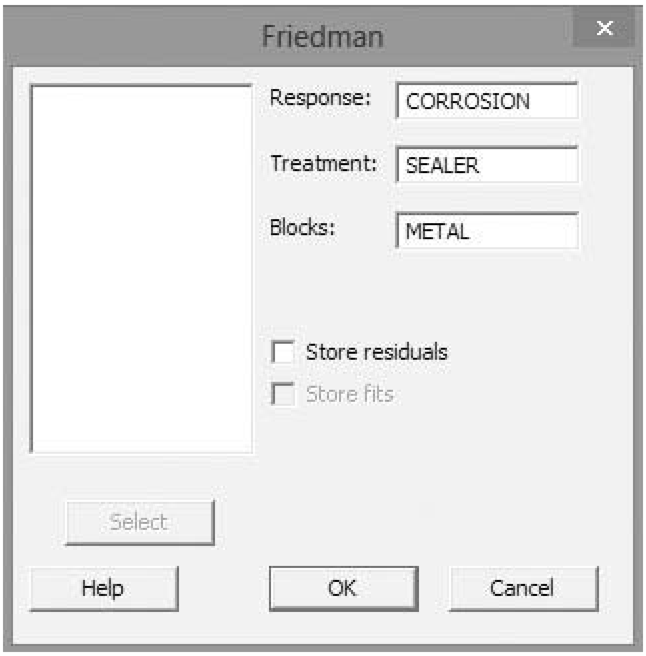


**FIGURE D.46**
MINITAB Calculator Menu Screen

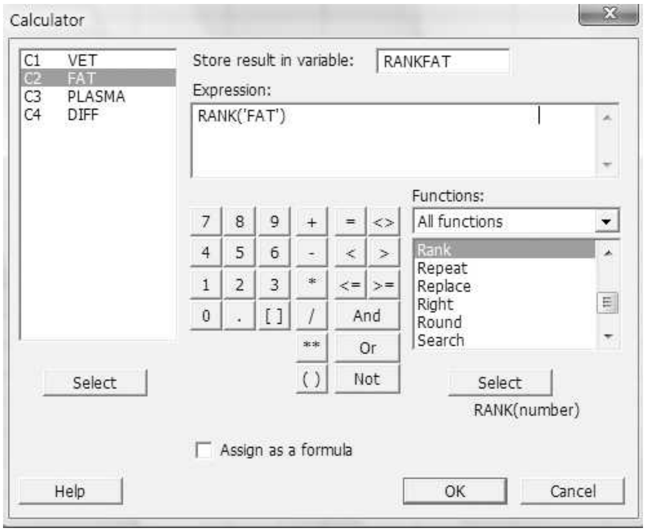
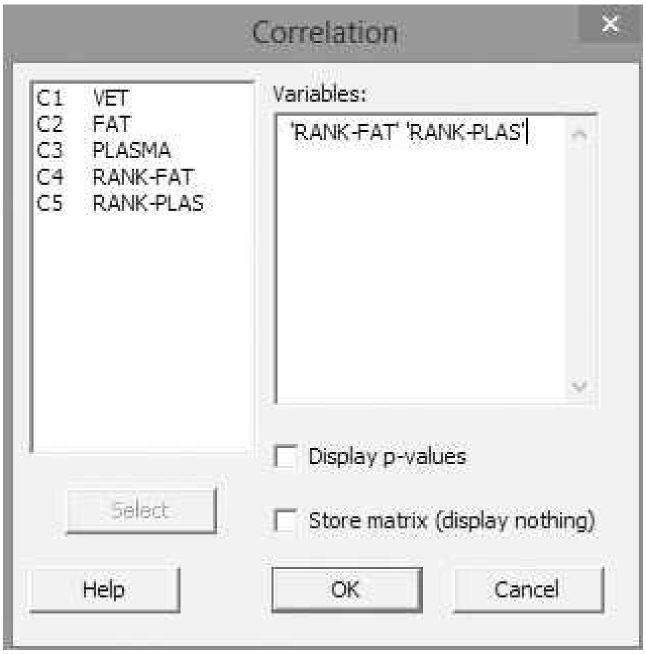



**FIGURE D.47**
MINITAB Correlation Dialog Box




## D.14 Control Charts and Capability Analysis

### Variable Control Charts

To generate an individual variable control chart using MINITAB, click on the "*Stat*" button on the main menu bar, then click on "*Control Charts*", "*Variable Charts for Individuals*", and "*Individuals*", as shown in Figure D.48. On the resulting dialog box (see Figure D.49), specify the variable to be charted in the "*Variables*" box. Click on "*I Chart Options*" and then select "*Tests*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

**FIGURE D.48**
MINITAB Menu Options for Individual Variable Control Charts

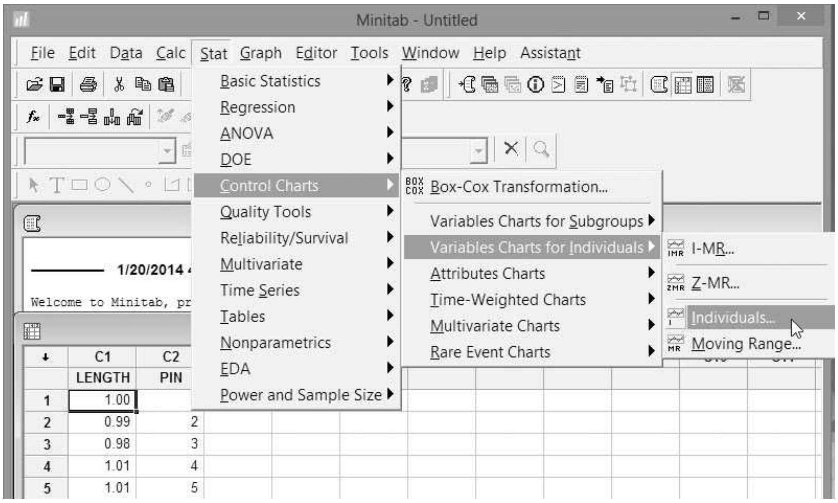
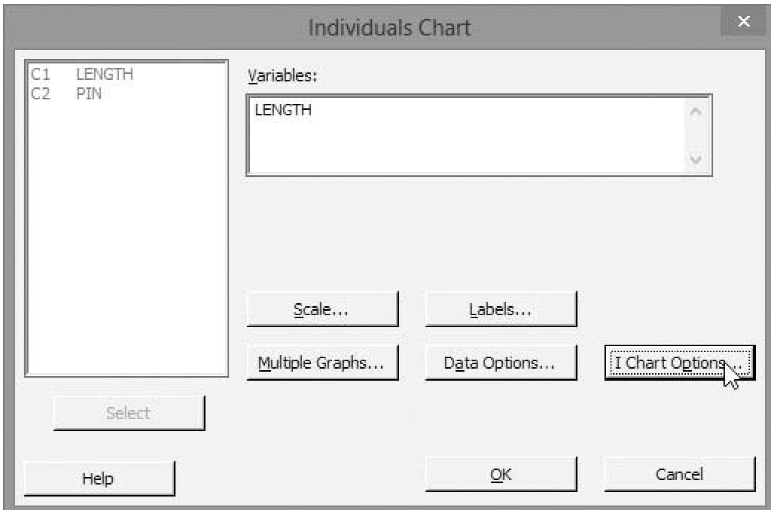



**FIGURE D.49**
Individual Variable Control Chart Dialog Box




## Mean and Range Control Charts

To generate control charts for the mean and range using MINITAB, click on the "*Stat*" button on the main menu bar, then click on "*Control Charts*", "*Variable Charts for Subgroups*", and "Xbar-R", as shown in Figure D.50. On the resulting dialog box (see Figure D.51), specify the variable to be charted in the open box on the right. Click on "*Xbar-R Options*" and then select "*Tests*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

## Attributes (Number and Percent Defectives) Control Charts

To generate a p-chart or a c-chart using MINITAB, click on the "*Stat*" button on the main menu bar, then click on "*Control Charts*" and "*Attributes Charts*", as shown in Figure D.52. On the resulting menu, select the type of chart (either "P" for a p-chart or "C" for a c-chart) you want to display. (See Figure D.52.) On the resulting dialog box, specify the variable containing the number of defects in the "Variables" box, as shown in Figure D.53. For p-charts, you will also need to specify the subgroup size. Click on

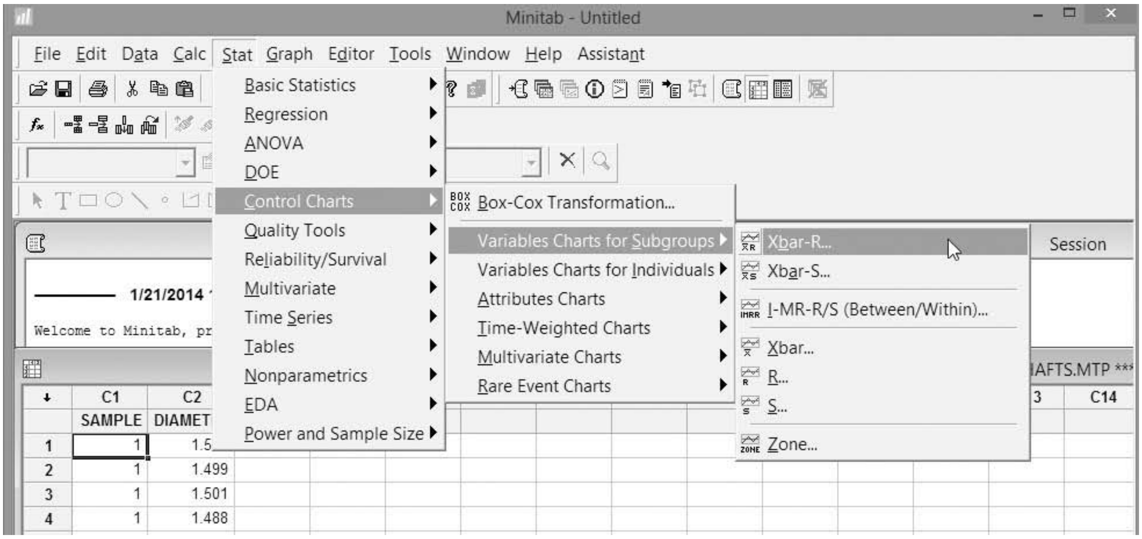



**FIGURE D.50**
MINITAB Menu Options for Mean and Range Control Charts

**FIGURE D.51**
Xbar-R Control Chart Dialog Box

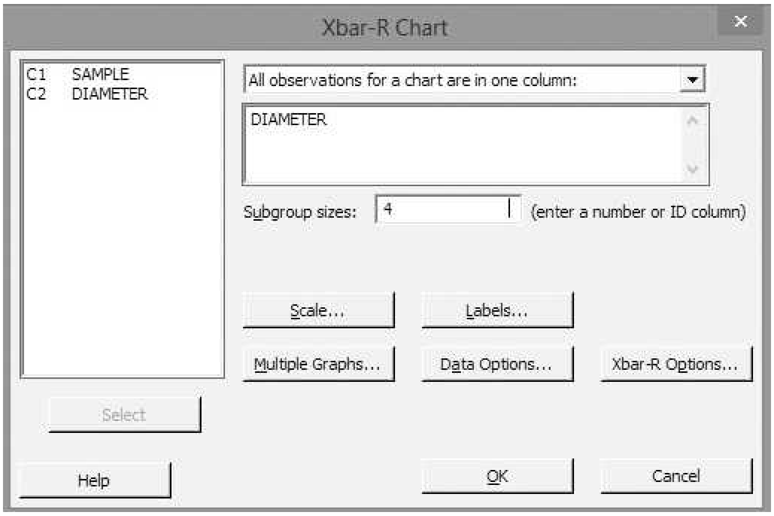



**FIGURE D.52**
MINITAB Menu Options for Mean and Range Control Charts

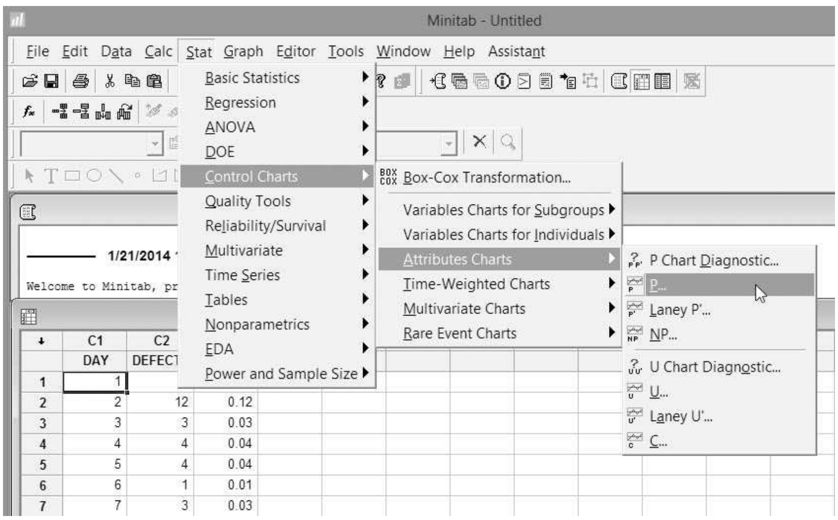
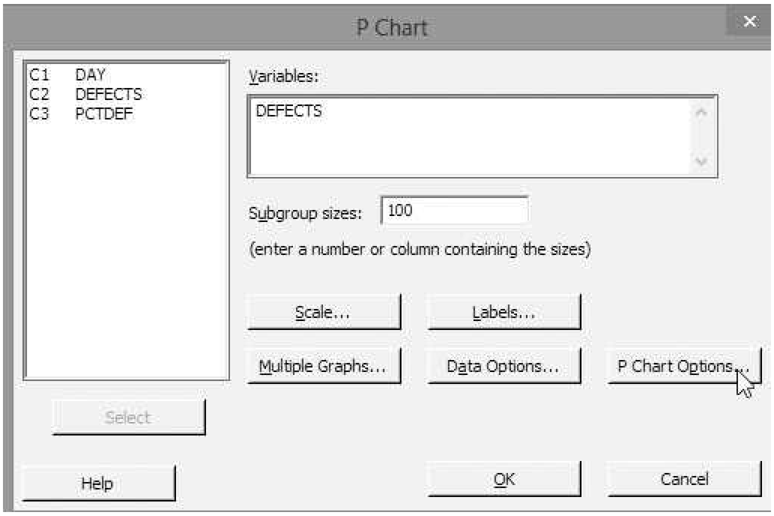


**FIGURE D.53**
P-Chart Dialog Box




"*P-Chart (or C-Chart) Options*" and then select "*Tests*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

## Capability Analysis

To conduct a capability analysis using MINITAB, click on the "*Stat*" button on the main menu bar, then click on "*Quality Tools*", "*Capability Analysis*" and "*Normal*", as shown in Figure D.54. A dialog box similar to Figure D.55 will be

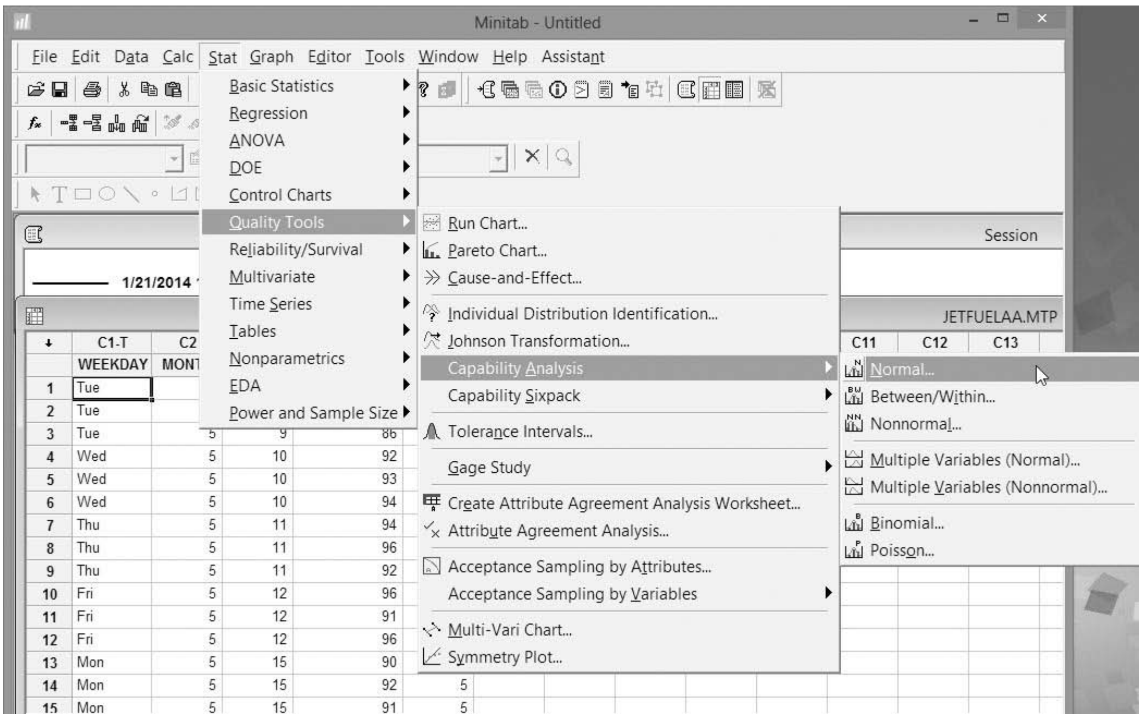
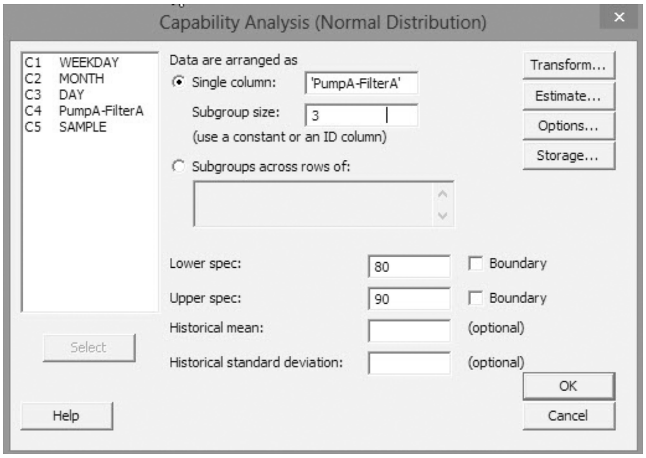



**FIGURE D.54**
MINITAB Menu Options for Process Capability Analysis

**FIGURE D.55**
Capability Analysis Dialog Box




displayed. Specify the quality variable of interest in the "*Single column*" box, subgroup size, and lower and upper specification limits on the menu screen as shown in Figure D.55. Click the "*OK*" button to produce the capability analysis graph and statistics.

## D.15 Random Samples

To generate a random sample of observations from a data set using MINITAB, click on the "*Calc*" button on the main menu bar, then click on "*Random Data*" and "*Sample from Columns*" as shown in Figure D.56. On the resulting dialog box (see Figure D.57), specify the sample size in the "*Number of rows to sample*" box, the variable to be sampled in the "*From columns*" box, and the column where the sample is to be saved in the "*Store samples in*" box. (Note: As an option, you can check "*Sample with replacement*". The default is to sample without replacement.) Click "*OK*" to generate the random sample.

**FIGURE D.56**
MINITAB Menu Options for Random Samples

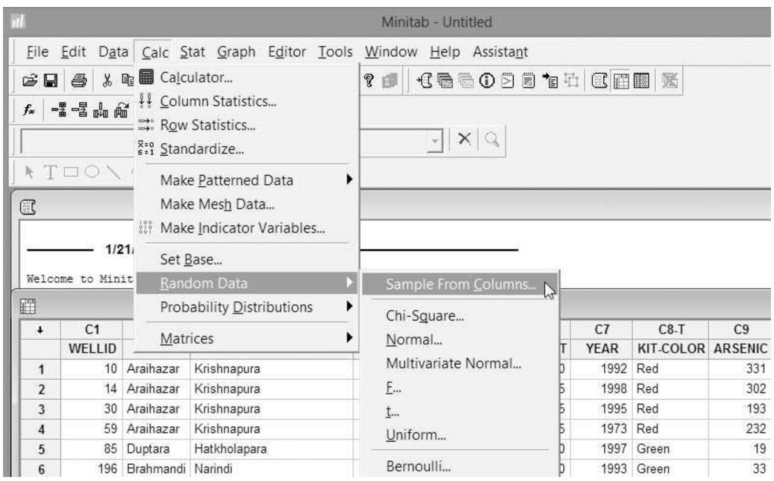
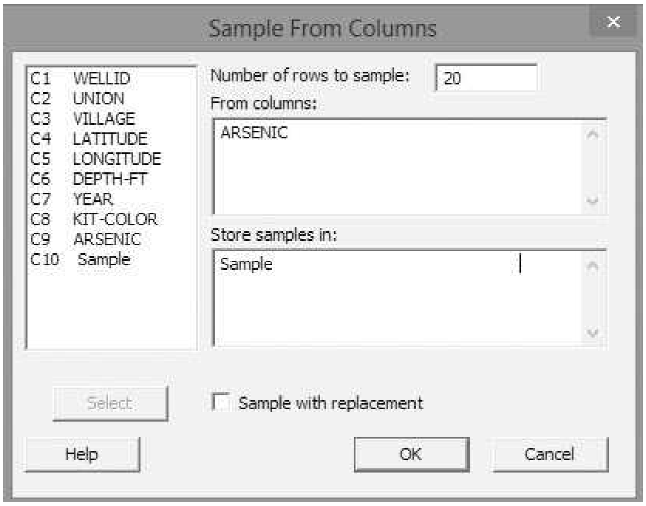


**FIGURE D.57**
Random Sample Dialog Box



# APPENDIX E SPSS for Windows Tutorial

**CONTENTS**



## E.1 SPSS Windows Environment

Upon entering into an SPSS session, you will see a screen similar to Figure E.1. The main portion of the screen is an empty spreadsheet, with columns representing variables and rows representing observations (or cases). The very top of the screen is the SPSS main menu bar, with buttons for the different functions and procedures available in SPSS. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate menu buttons. The results will appear in an SPSS viewer Output window.

## E.2 Creating/Accessing a Data Set Ready for Analysis

There are three ways you can get a data set ready for analysis in SPSS:

1. Entering data values directly into the SPSS spreadsheet
2. Accessing a previously created SPSS file
3. Accessing an external data file

## Direct Data Entry

Create an SPSS data file by entering data directly into the spreadsheet. Figure E.2 shows data entered for a variable called "RATIO." Name the variables (columns) by selecting "*Variable View*" at the bottom of the spreadsheet and typing in the name of each variable in the "*Name*" column.

## Getting an SPSS File

To access data already saved as an SPSS file, select "*File*" on the main menu bar, then "*Open*" and "*Data*", as shown in Figure E.3. In the resulting "Open Data" dialog box (see Figure E.4), select the folder where the data file resides, then select the data set (e.g., ACCIDENTS). After clicking "*Open*", the data will appear in the spreadsheet.

## Getting an External File

Finally, if the data are saved in an external text file (e.g., as a .DAT, .TXT, or Excel file), access it by selecting "*File*" on the menu bar, then "*Read Text Data*" (see Figure E.3). In the resulting "Open Data" dialog box (see Figure E.5), select the folder where the data file resides, specify the type of file (e.g., an Excel file), then select the data set name (e.g., ALLOY) and click "*Open*". Depending on the type of file, this will invoke one or more Text Import Wizard screens. Make the appropriate selections on the screen (e.g., whether or not variable names are in the first row), and click "Next" to go to the next menu screen. When done, click "Finish". The data will appear in the spreadsheet.

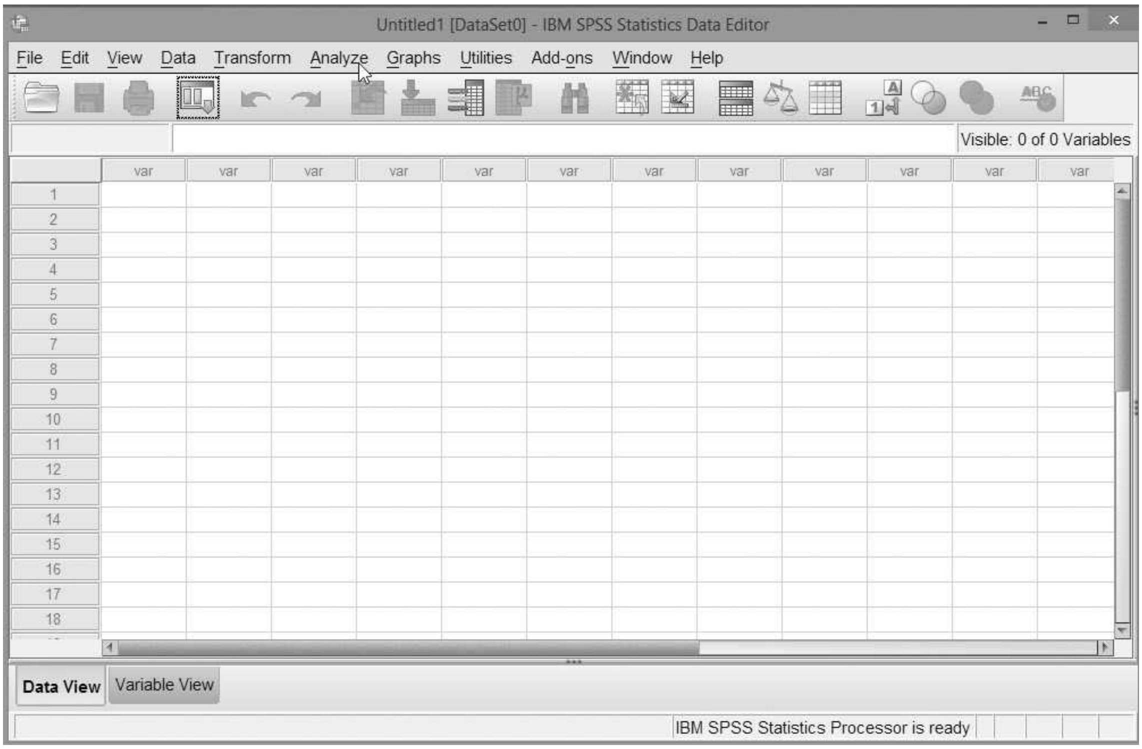
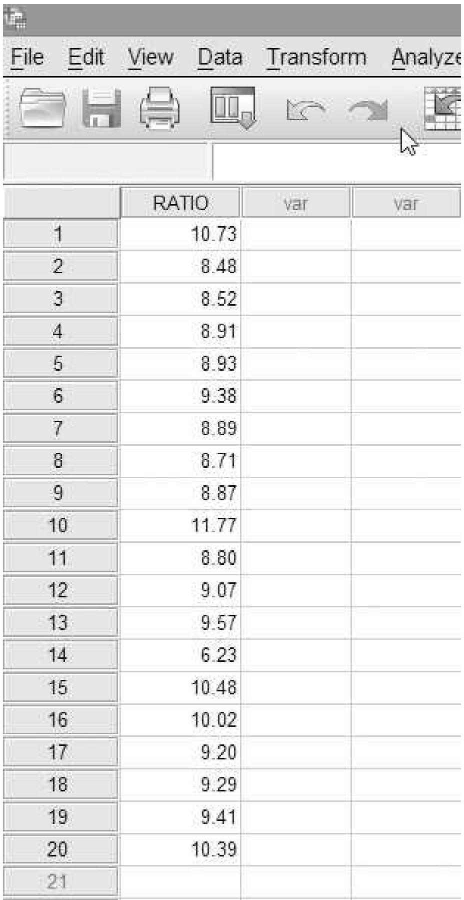



**FIGURE E.1**
Initial Screen Viewed by the SPSS User

**FIGURE E.2**
Data Entered into the SPSS Spreadsheet




## E.3 Listing Data

To access a listing (printout) of your data using SPSS, click on the "*Analyze*" button on the main menu bar, and then click on "*Reports*" and "*Report Summaries in Rows*", as shown in Figure E.6. The resulting menu, or dialog box, appears as in Figure E.7. Enter the names of the variables you want to print in the "*Data Column Variables*" box and check the "*Display cases*" box. Click "OK" to obtain a listing of the data in the SPSS output window.

## E.4 Graphing Data

### Bar Graphs, Pie Charts, Box Plots, Scatterplots, Histograms

To obtain graphical descriptions of your data using SPSS, click on the "*Graphs*" button on the main menu bar, then click on "*Legacy Dialogs*" and select the graph of your choice (Bar, Pie, Boxplot, Scatter/Dot, or Histogram), as shown in Figure E.8. On the resulting dialog box, make the appropriate variable selections and click "OK" to view the graph. (The selections for a histogram are shown in Figure E.9.)

### Stem-and-Leaf Plots

Select "*Analyze*" from the main SPSS menu, then "*Descriptive Statistics*," and then "*Explore*." In the "*Explore*" dialog box, select the variable to be analyzed in the

**FIGURE E.3**
Accessing an SPSS Data File

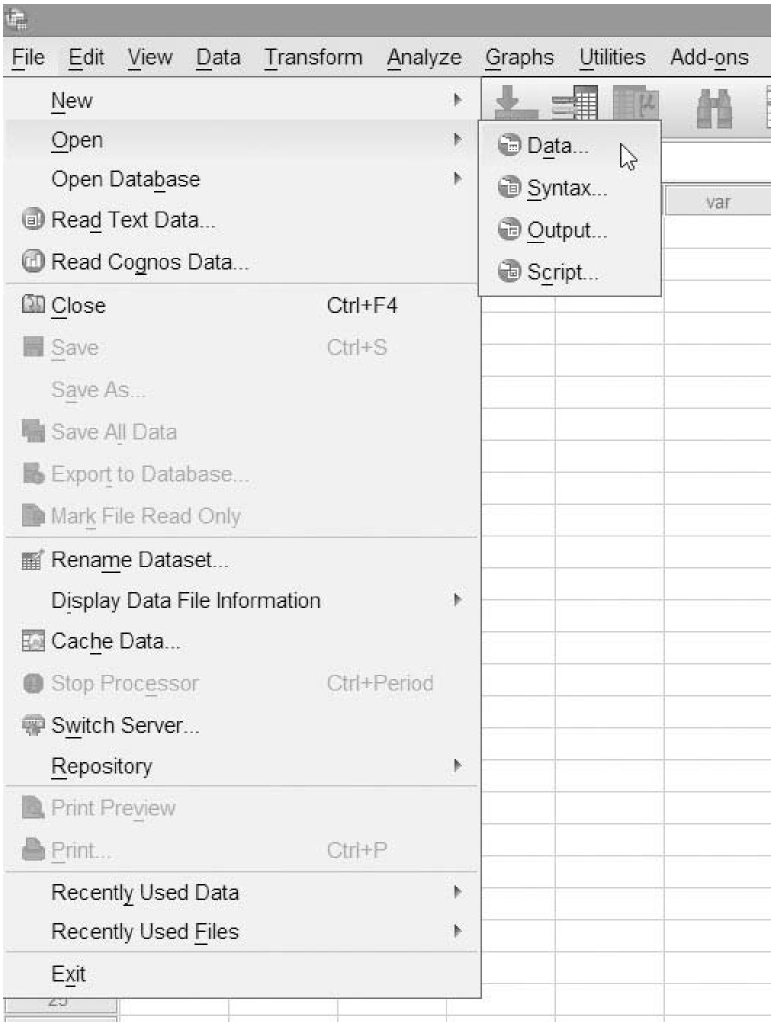



**FIGURE E.4**
SPSS Open Data Dialog Box

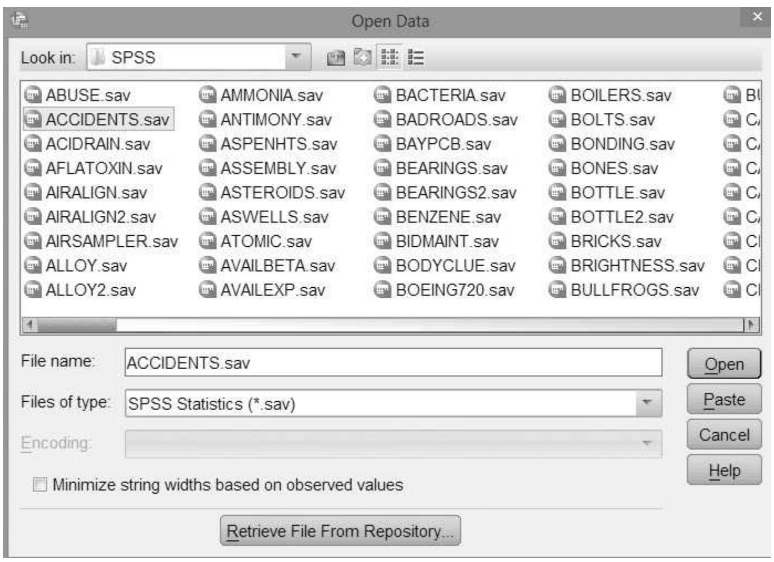
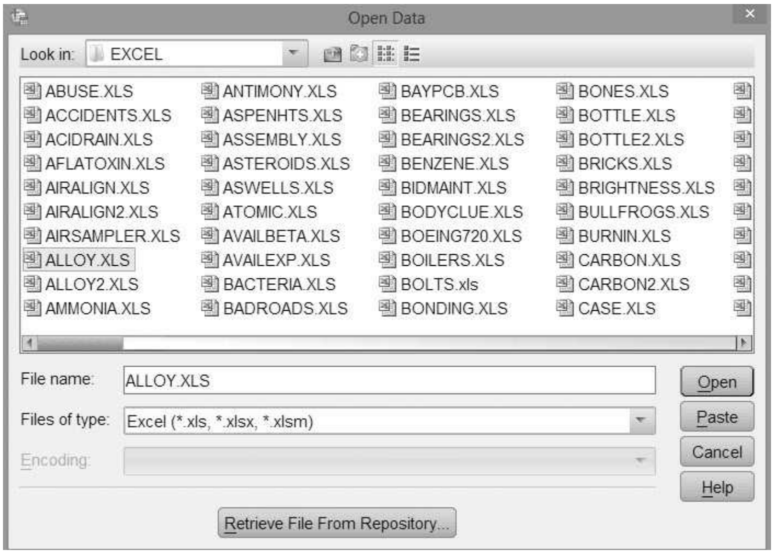


**FIGURE E.5**
Getting an External File in SPSS



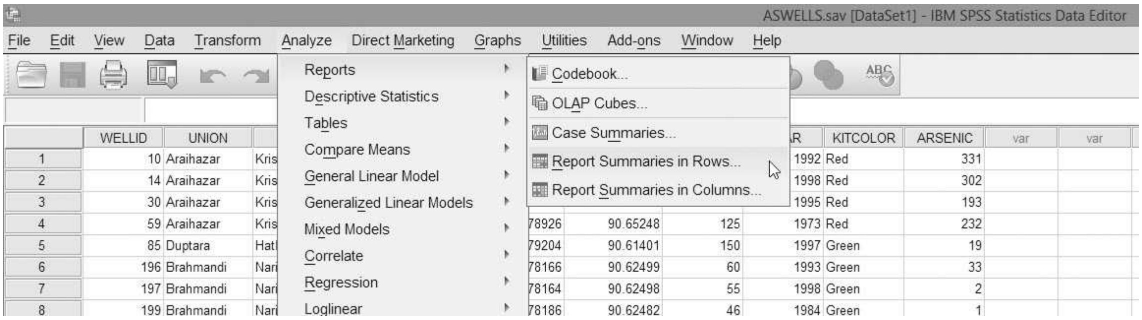




**FIGURE E.6**
SPSS Menu Options for Listing Data

**FIGURE E.7**
SPSS Report Summaries Dialog Box

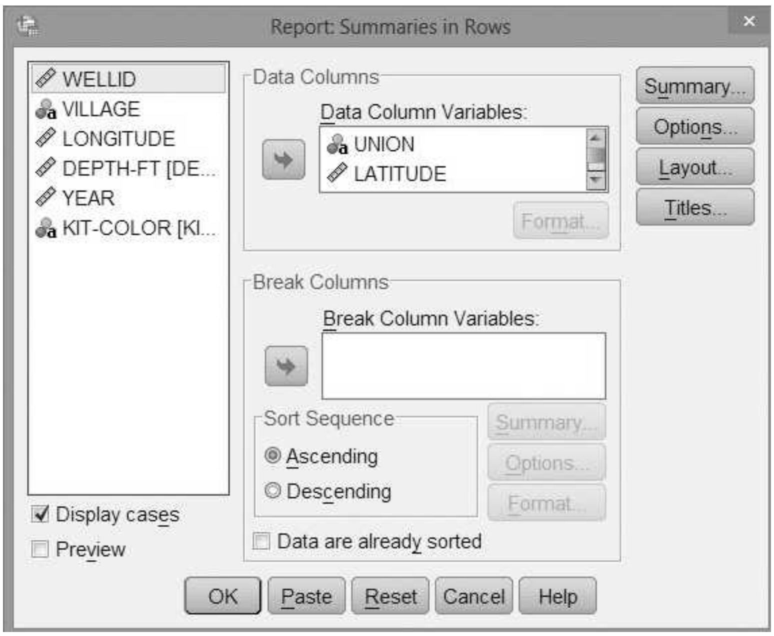
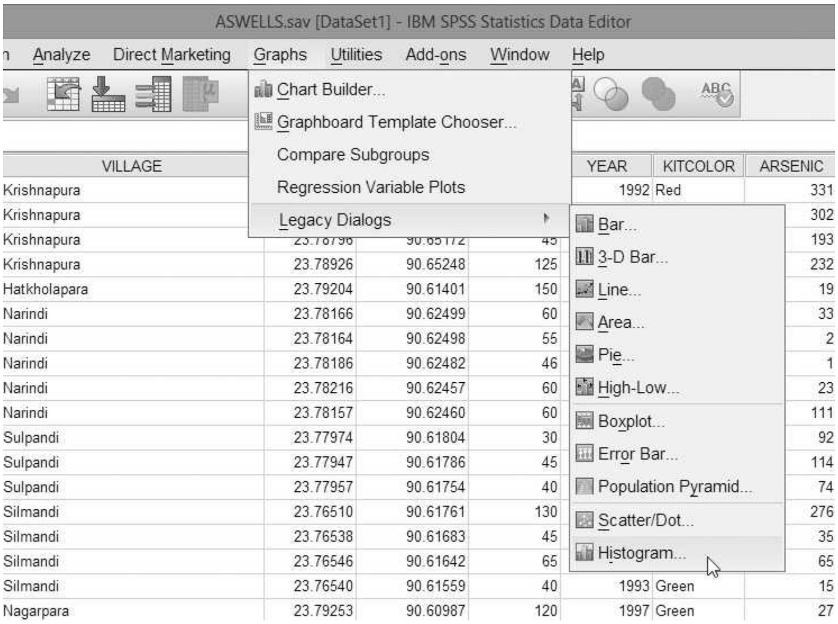


**FIGURE E.8**
SPSS Menu Options for Graphing Data




*"Dependent List"* box, as shown in Figure E.10. Click on either "*Both*" or "*Plots*" in the "*Display*" options and then click "*OK*" to display the stem-and-leaf plot.

## Pareto Diagrams

Select "*Analyze*" from the main SPSS menu, then "*Quality Control*," and then "*Pareto Charts*." Click "*Define*" on the resulting menu and then select the variable to be

**FIGURE E.9**
SPSS Histogram Dialog Box

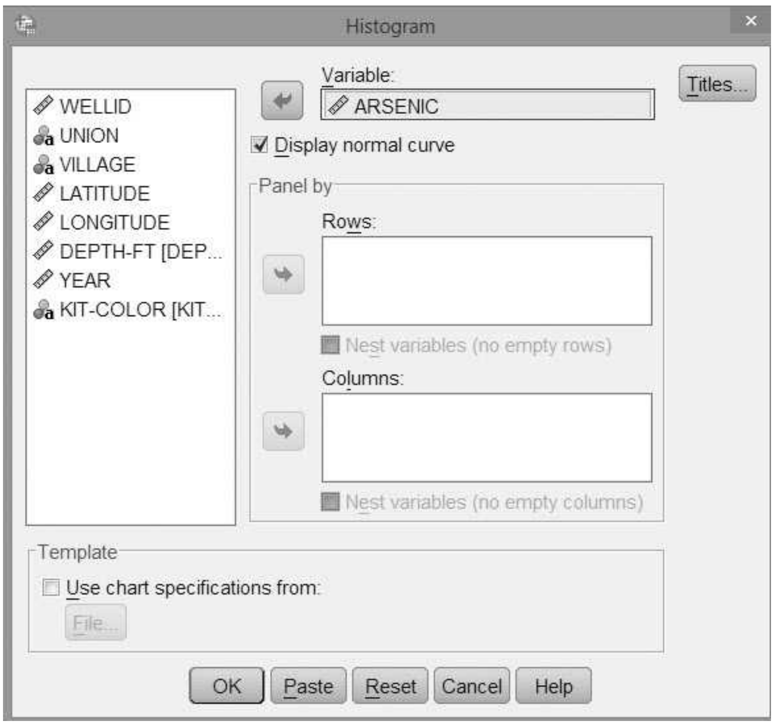



**FIGURE E.10**
SPSS Explore Dialog Box

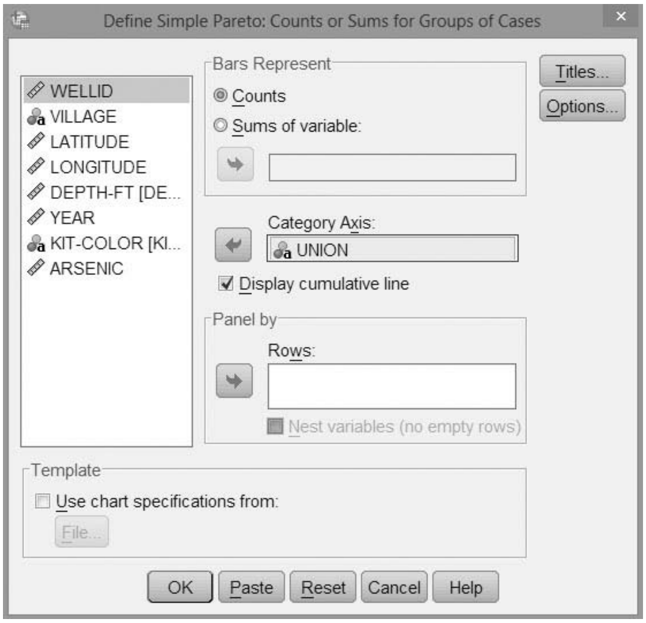



**FIGURE E.11**
SPSS Pareto Chart Dialog Box

analyzed and move it to the "*Category Axis*" box, as shown in Figure E.11. Click "*OK*" to display the Pareto diagram.

## E.5 Descriptive Statistics, Percentiles, and Correlations

### Descriptive Statistics

To obtain numerical descriptive measures for a quantitative variable (e.g., mean, median, standard deviation, etc.) using SPSS, click on the "*Analyze*" button on the main menu bar, click on "*Descriptive Statistics*," and then click on "*Descriptives*", as shown in Figure E.12. The resulting dialog box appears in Figure E.13. Select the quantitative variables you want to analyze and place them in the "*Variable(s)*" box. You can control which descriptive statistics appear by clicking the "*Options*" button on the dialog box and making your selections. Click "*OK*" to view the descriptive statistics output.

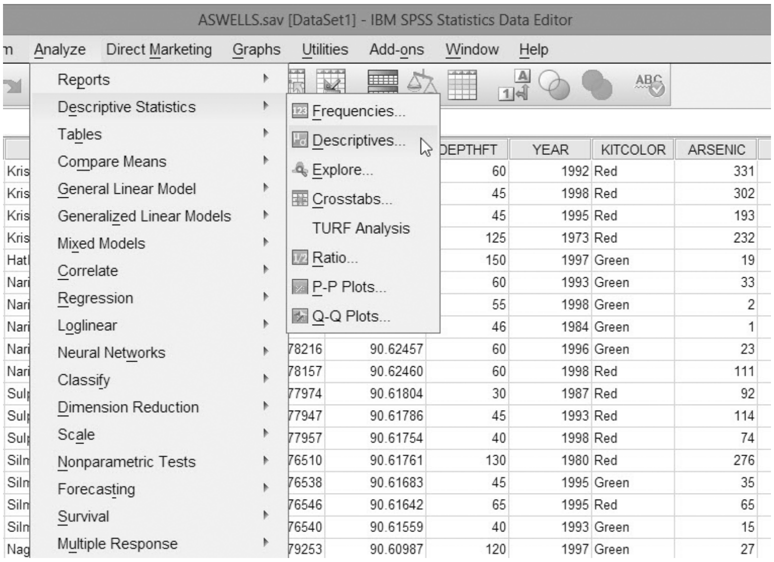



**FIGURE E.12**
SPSS Menu Options for Descriptive Statistics

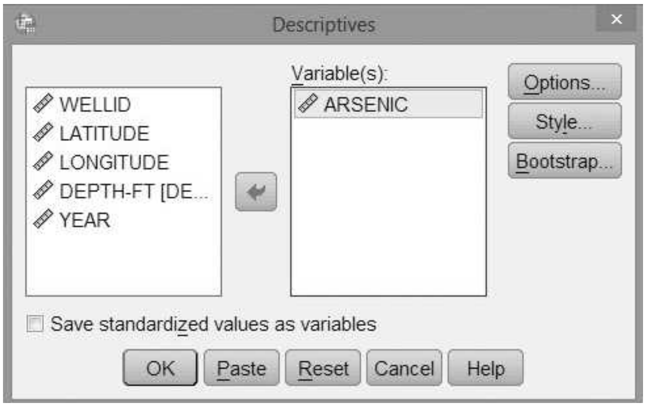



**FIGURE E.13**
Descriptive Statistics Dialog Box

## Percentiles

Select "*Analyze*" on the main menu bar, and then click on "*Descriptive Statistics*" (see Figure E.12). Select "*Explore*" from the resulting menu. In the resulting dialog box (see Figure E.14), enter the variable to be analyzed in the "Dependent List" box, select the "*Statistics*" button and check the "*Percentiles*" box on the resulting menu. Return to the "*Explore*" dialog box and click "*OK*" to obtain a list of the percentiles.

## Correlations

To obtain Pearson product moment correlations for pairs of quantitative variables, click on the "*Analyze*" button on the SPSS main menu bar, click on "*Correlate*," and then click on "*Bivariate*". On the resulting dialog box, enter the variables you want to correlate in the "*Variables*" box on the right panel. (See Figure E.15.) Make sure "*Pearson*" is checked in the "*Correlation Coefficients*" box. Click "*OK*" to obtain a printout of the correlations.

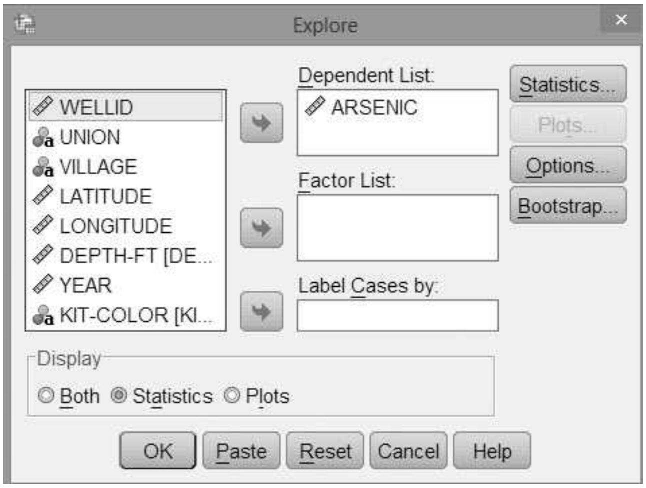



**FIGURE E.14**
SPSS Explore Dialog Box

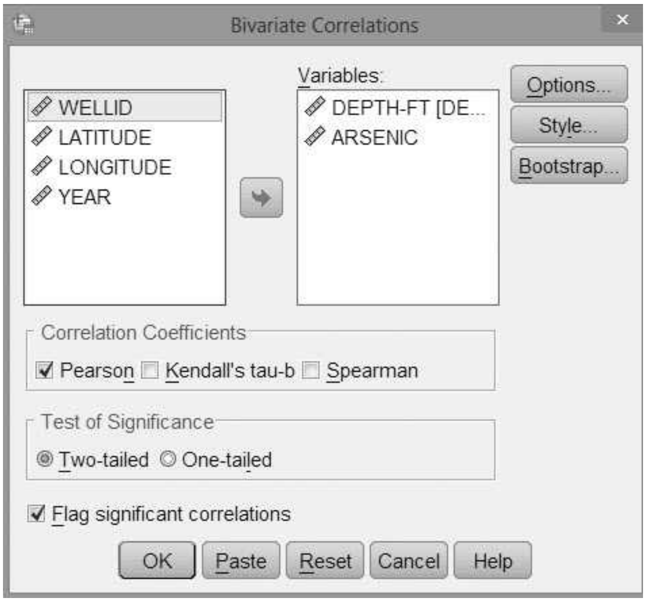


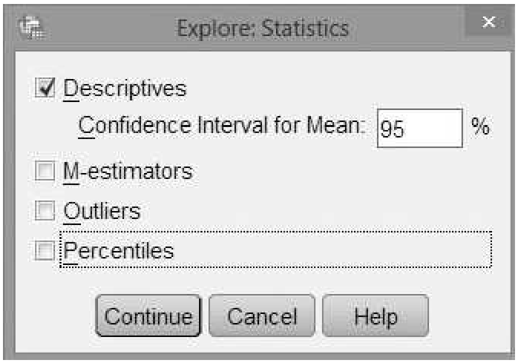

**FIGURE E.15**
SPSS Correlations Dialog Box

## E.6 Confidence Intervals and Hypothesis Tests for a Mean or Proportion

### Population Mean, Confidence Interval

To form a confidence interval for a single population mean of a quantitative variable, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Descriptive Statistics*" and "*Explore*" (See Figure E.12.) On the resulting dialog box, specify the quantitative variable in the "*Dependent List*" box and then click the "*Statistics*" button (see Figure E.14). Enter the confidence level on the resulting dialog box, as shown in Figure E.16. Click "*Continue*" then "*OK*" to obtain a printout of the results. (Note: A bootstrap confidence interval can be obtained by clicking the "Bootstrap" button on the Explore dialog box shown in Figure E.14.)

**FIGURE E.16**
SPSS Options for a Confidence Interval for the Mean



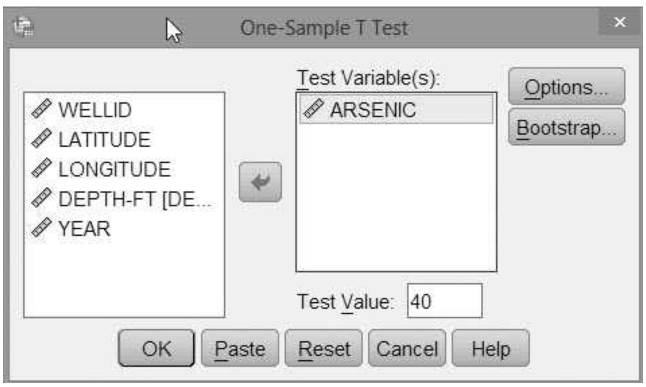

## Population Mean, Hypothesis Test

To conduct a test of hypothesis for a single population mean of a quantitative variable, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Compare Means*" and "*One- Sample T Test*". On the resulting dialog box, specify the quantitative variable in the "*Test Variable(s)*" box and enter the value of the mean in the null hypothesis in the "Test Value" box, as shown in Figure E.17. Click "*OK*." SPSS will automatically conduct a two-tailed test of hypothesis. [*Note:* The SPSS one-sample $t$-procedure uses the $t$-statistic to conduct the test of hypothesis. When the sample size $n$ is small, this is the appropriate method. When the sample size $n$ is large, the $t$-value will be approximately equal to the large-sample $z$-value and the resulting test will still be valid.]

## Population Proportion

To form a confidence interval or conduct a hypothesis test for a single population proportion for a two-level (binomial) qualitative variable, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Nonparametric Tests*" and "*1 Sample*". Click the "*Fields*" option, and then on the resulting dialog box (shown in Figure E.18, left panel), move the qualitative (categorical) variable to the "*Test Fields*" box. Click the "*Settings*" option, and then on the resulting dialog box select "*Customize Tests*" and "*Compare observed binary probability to hypothesized (Binomial test)*." Then click "*Options*." On the resulting dialog box (see Figure E.18, right panel), click "*Likelihood ratio*" in the "*Confidence Interval*" box, click "*OK*," and then click "*Run*." On

**FIGURE E.17**
SPSS Dialog Box for Testing a Mean



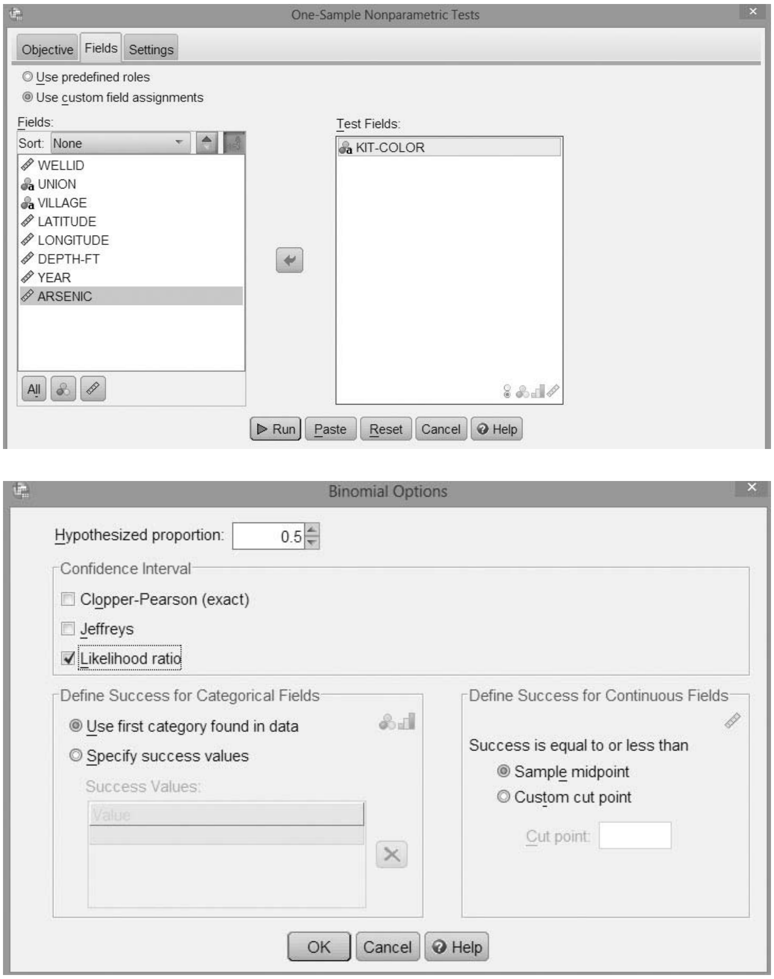



**FIGURE E.18**
SPSS Options for Binomial Proportion Test/Confidence Interval

the resulting output, double click on the "*Hypothesis Test Summary*" output to display the "*Model Viewer*" screen, as shown in Figure E.19. The results of the hypothesis test will automatically be displayed. If you want to see the results of the confidence interval, select "Confidence Interval Summary View" at the bottom of the screen (as shown in Figure E.19).

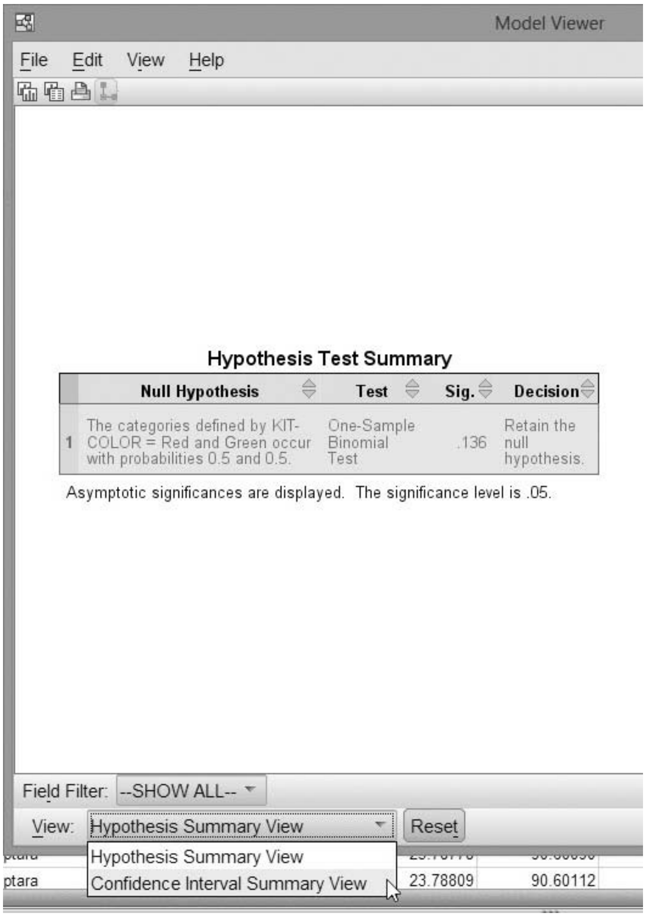
[Note: Confidence intervals or hypothesis tests for a single population variance are not available in SPSS.]

## E.7 Confidence Intervals and Hypothesis Tests for the Difference Between Means, Proportions, or Variances

### Two Means, Independent Samples

To conduct a test of hypothesis and form a confidence interval for the difference between two population means based on independent samples, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Compare Means*" and "*Independent Samples T Test*" (See Figure E.20.) On the resulting dialog box (shown in Figure E.21, left screen), specify the quantitative variable of interest in the "*Test Variable(s)*" box

**FIGURE E.19**
SPSS Model Viewer Selections for Binomial Proportion




and the qualitative variable that identifies the two populations in the "*Grouping Variable*" box. Click the "*Define Groups*" button and specify the values of the two groups in the resulting dialog box (see Figure E.21, right screen). Click "*Continue*" to return to the "*Independent-Samples T Test*" dialog screen, then click "*OK*". SPSS will automatically conduct a two-tailed test of the null hypothesis of no difference in means and produce a 95% confidence interval for the mean difference.

Note: The SPSS two-sample t-procedure uses the t-statistic to conduct the test of hypothesis. When the sample sizes are small, this is the appropriate method. When the sample sizes are large, the $t$-value will be approximately equal to the large-sample $z$-value, and the resulting test will still be valid.

## Two Means, Matched Pairs

The SPSS data file should contain two quantitative variables—one with the data values for the first group (or population) and one with the data values for the second group. (Note: The sample size should be the same for each group.) To conduct the paired difference test, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Compare Means*" and "*Paired-Samples T Test*" (see Figure E.20). On the resulting dialog box (shown in Figure E.22), specify the two quantitative variables of interest in the "Paired Variables" box. Click "*OK*" to view the results of a two-tailed test of the null hypothesis of no difference in means and a 95% confidence interval for the mean difference.

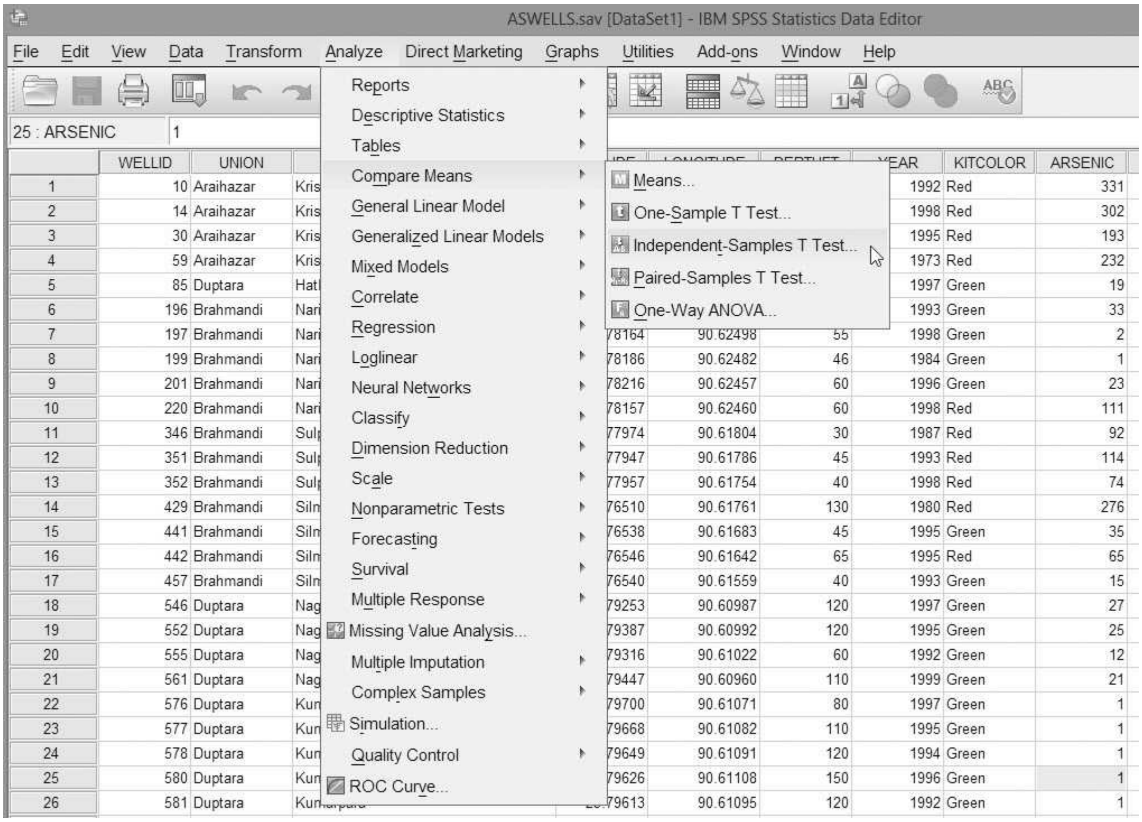



**FIGURE E.20**
SPSS Menu Options for Comparing Two Means

## Two Proportions

To conduct a test of hypothesis and form a confidence interval for the difference between two population proportions based on independent samples, you must first create an SPSS data file with three variables (columns)— (1) SAMPLE, (2) OUTCOME, and (3) NUMBER—and four rows. Each row will give the sample number, outcome

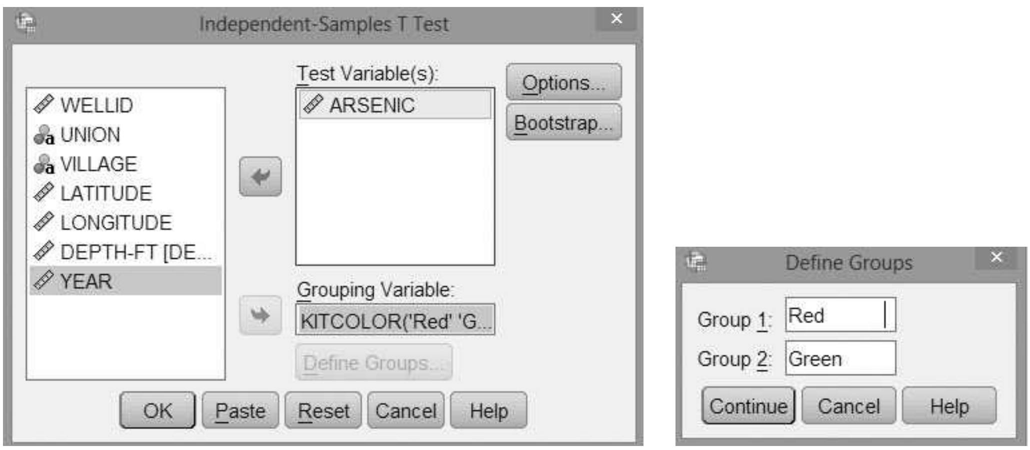



**FIGURE E.21**
Independent Samples T-Test Options

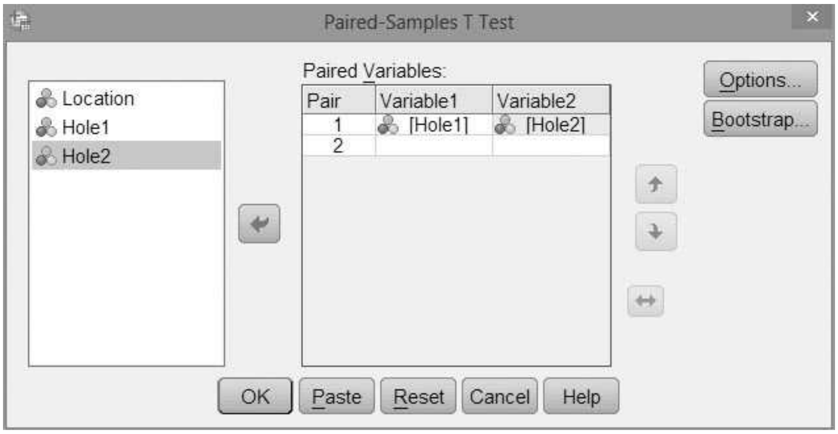



**FIGURE E.22**
SPSS Paired T-Test Options

**FIGURE E.23**
SPSS Data File for Comparing Two Proportions

| SAMPLE | OUTCOME | NUMBER |
|---|---|---|
| 1 | S | 60 |
| 1 | F | 40 |
| 2 | S | 50 |
| 2 | F | 50 |

(success or failure), and number of observations. For example, Figure E.23 shows the data file for a problem with 60 out of 100 successes for sample 1 and 50 out of 100 successes for sample 2. After creating the data file, click on the "*Data*" button on the SPSS menu bar and then click on "*Weight Cases*". Click "*Weight cases by*", then enter the "*Number*" variable into the "*Frequency Variable*" box and click "*OK*."

Now, click on the "*Analyze*" button on the SPSS menu bar and then click on "*Descriptive Statistics*" and "*Crosstabs*." (See Figure E.24.) On the resulting menu, specify "SAMPLE" in the "*Row(s)*" box and "OUTCOME" in the "*Column(s)*" box as shown in Figure E.25. Also, click the "*Statistics*" option button and select "*Chi-square*," then click the "*Cells*" option button and select "*Observed Counts*" and "*Row Percentages*." Click "*Continue*," then click "*OK*." On the resulting SPSS printout, look

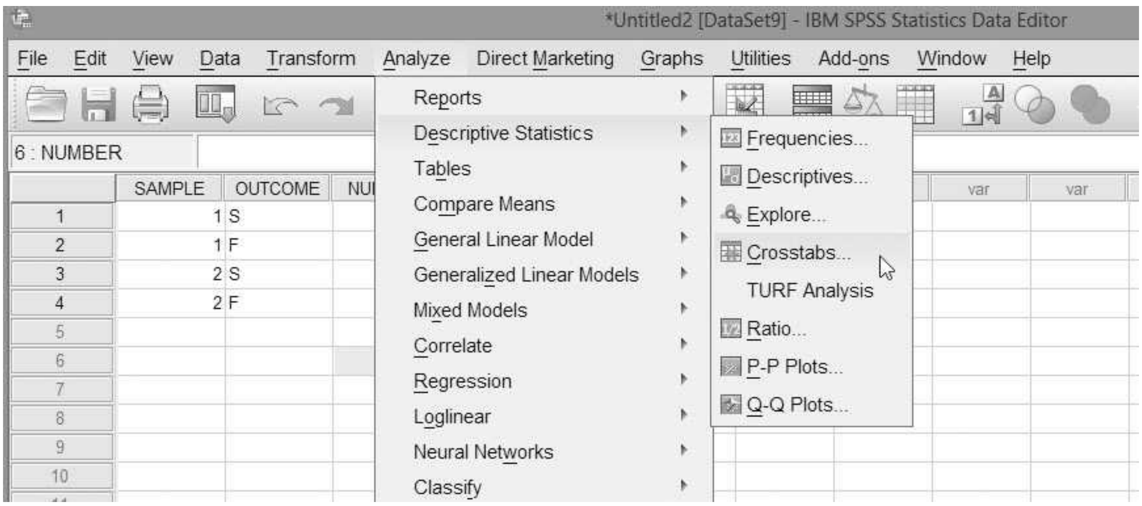
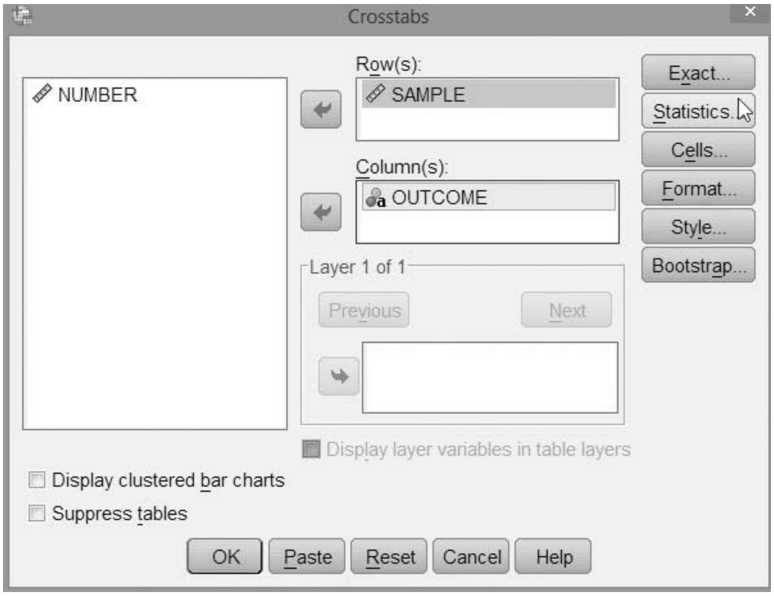



**FIGURE E.24**
SPSS Menu Options for Comparing Two Proportions

**FIGURE E.25**
SPSS Menu Crosstabs Dialog Box




for the *p*-value associated with the "Likelihood Ratio" test (this is equivalent to the large-sample *z*-test).

### Two Variances

Follow the steps outlined for "Two Means, Independent Samples" above. On the resulting SPSS printout, there will be an F-test for comparing population variances. (This test, called Levene's Test, is a nonparametric test that is similar to the F-test presented in the text.)

## E.8 Categorical Data Analysis

SPSS can produce a frequency table for a single qualitative variable (i.e., a one-way table) and can conduct a chi-square test for independence of two qualitative variables in a two-way (contingency) table.

### One-Way Table

Open an SPSS spreadsheet file that contains the variable with category values for each of the *n* observations in the data set. (*Note:* SPSS requires that these categories be specified numerically, e.g., 1, 2, 3.) Click on the "*Analyze*" button on the SPSS menu bar and then click on "*Nonparametric Tests*," "*Legacy Dialogs*," and "*Chi-square*," as shown in Figure E.26. The resulting dialog box appears as shown in Figure E.27.

Specify the qualitative variable of interest in the "*Test Variable List*" box. If you want to test for equal cell probabilities in the null hypothesis, then select the "*All categories equal*" option under the "*Expected Values*" box (as shown in Figure E.27). If the null hypothesis specifies unequal cell probabilities, then select the "*Values*" option under the "*Expected Values*" box. Enter the hypothesized cell probabilities in the adjacent box, one at a time, clicking "*Add*" after each specification. Click "*OK*" to generate the SPSS printout.

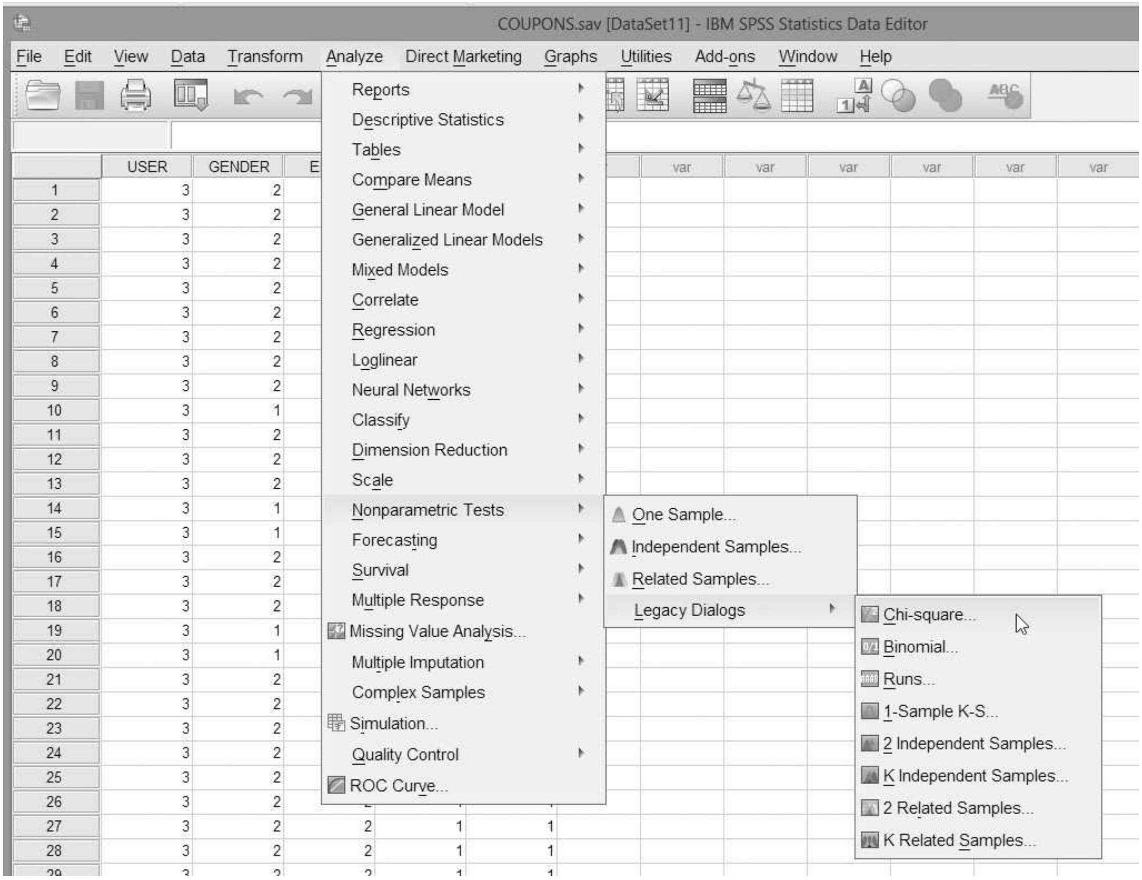


**FIGURE E.26**
SPSS Menu Options for a One-Way Frequency Table Analysis

### Two-Way Table

Open an SPSS spreadsheet file that contains the two qualitative variables with category values for each of the $n$ observations in the data set. Click on the "*Analyze*" button on the SPSS menu bar and then click on "*Descriptive Statistics*" and "*Crosstabs*," as shown in Figure E.24. The resulting dialog box appears as shown in Figure E.28. Specify one qualitative variable in the "*Row(s)*" box and the other qualitative variable in the "*Column(s)*" box. Click the "*Statistics*" button and select the "*Chi-square*" option.

Click "*Continue*" to return to the "*Crosstabs*" dialog box. If you want the contingency table to include expected values, row percentages, and/or column percentages, click the "*Cells*" button and make the appropriate menu selections. When you return to the "*Crosstabs*" menu screen, click "*OK*" to generate the SPSS printout.

Note: If your SPSS spreadsheet contains summary information (i.e., the cell counts for the contingency table) rather than the actual categorical data values for each observation, you must weight each observation in your data file by the cell count for that observation prior to running the chi-square analysis. Do this by selecting the "*Data*" button on the SPSS menu bar and then click on "*Weight Cases*" and specify the variable that contains the cell counts.

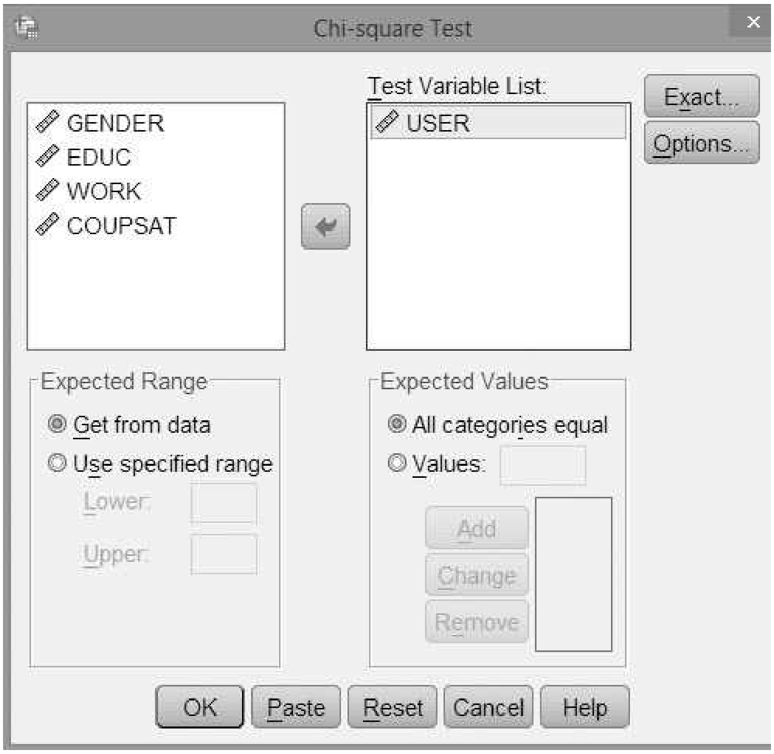



**FIGURE E.27**
One-Way Frequency Table Dialog Box

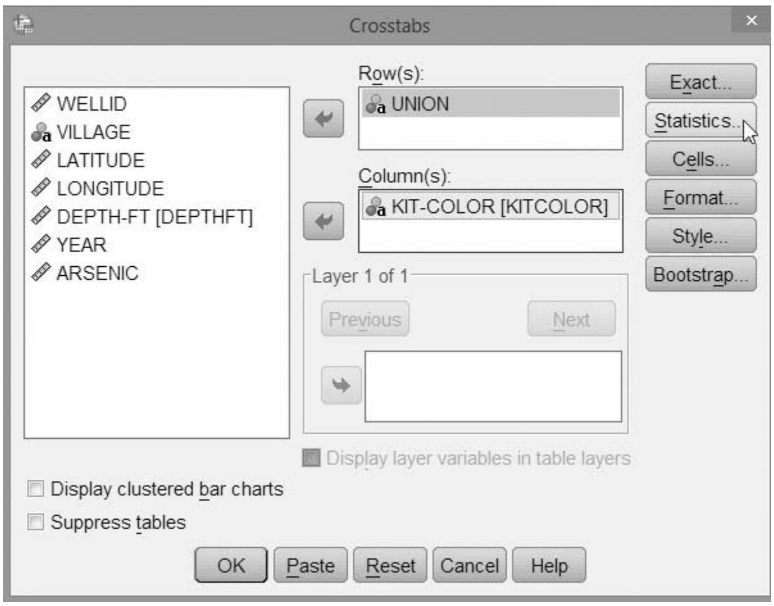



**FIGURE E.28**
SPSS Crosstabs Dialog Box

## E.9 Simple Linear Regression

To conduct a simple linear regression analysis, click on the "*Analyze*" button on the SPSS menu bar, then click on "*Regression*" and "*Linear*". (See Figure E.29.) On the resulting dialog box, specify the quantitative dependent variable in the "*Dependent*" box and the quantitative independent variable in the "*Independent(s)*" box, as shown in Figure E.30. Be sure to select "*Enter*" in the "*Method*" box.

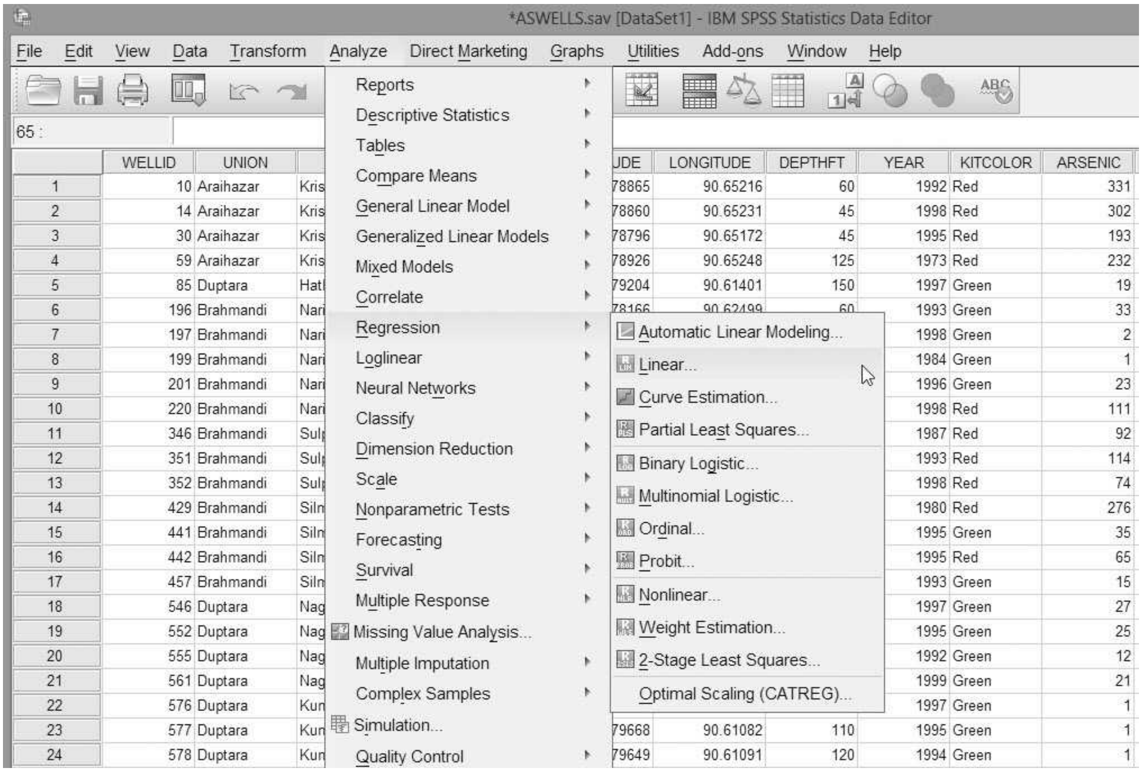
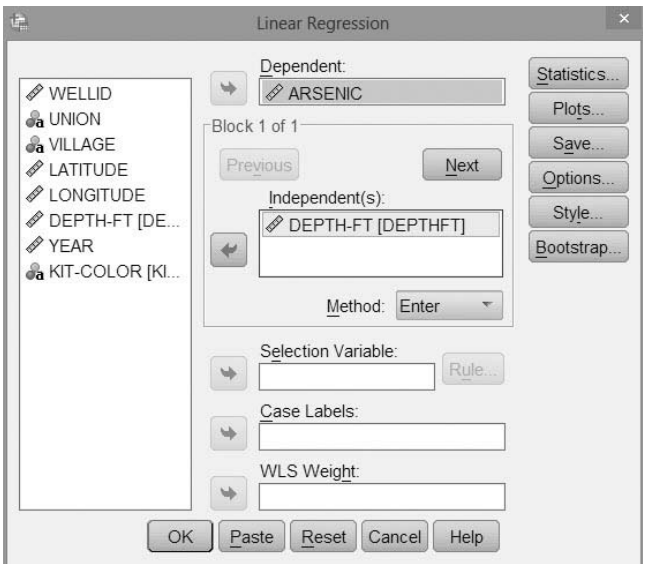



**FIGURE E.29**
SPSS Menu Options for Simple Linear Regression

**FIGURE E.30**
SPSS Linear Regression Dialog Box




To produce confidence intervals for the model parameters, click the "*Statistics*" button and check the appropriate menu items in the resulting menu list. To obtain prediction intervals for $y$ and confidence intervals for $E(y)$, click the "*Save*" button and check the appropriate items in the resulting menu list, as shown in Figure E.31. (The prediction intervals will be added as new columns to the SPSS data spreadsheet.) From this screen, you can also save residuals for plotting. To return to the main

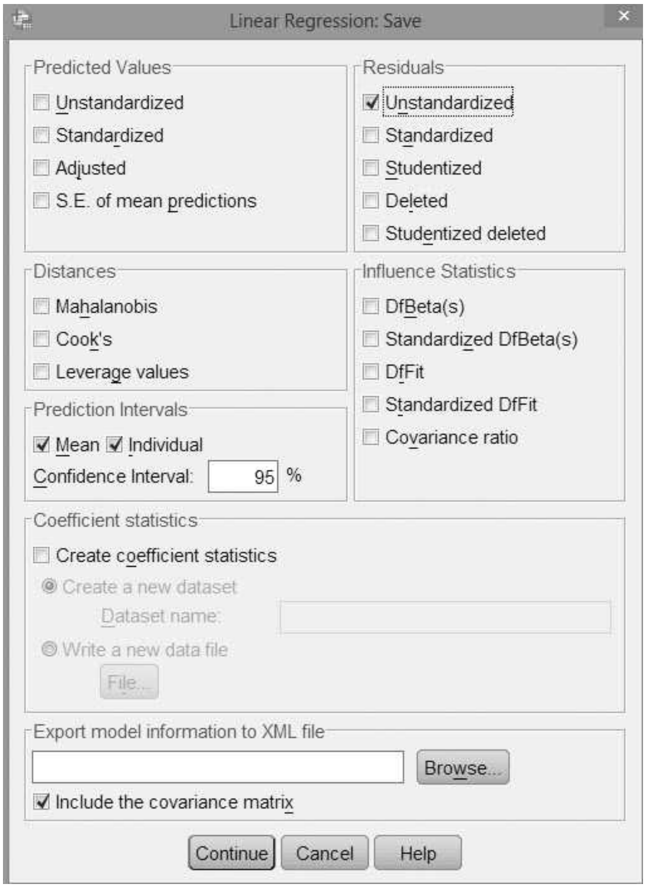



**FIGURE E.31**
SPSS Simple Linear Regression Options

Regression dialog box from any of these optional screens, click "*Continue*." Click "OK" on the Regression dialog box to view the linear regression results.

## E.10 Multiple Regression

To conduct a multiple regression analysis, click on the "*Analyze*" button on the SPSS menu bar, then click on "*Regression*" and "*Linear*". (See Figure E.29.) On the resulting dialog box, specify the quantitative dependent variable in the "*Dependent*" box and the independent variables in the "*Independent(s)*" box, as shown in Figure E.32.

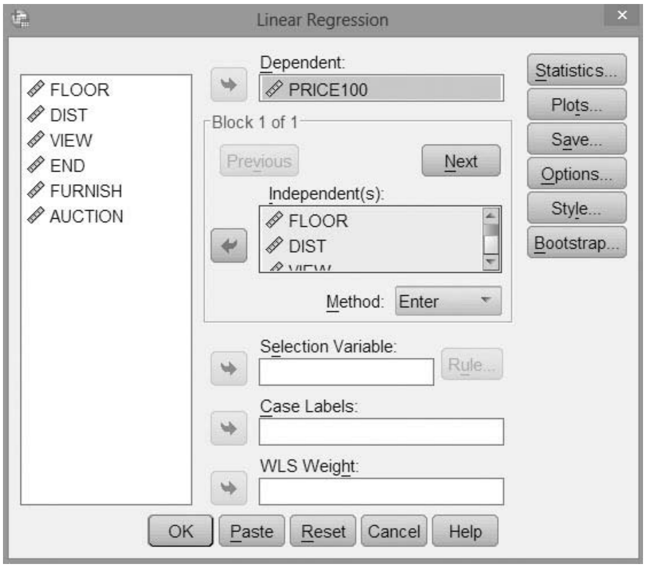
[*Note:* If your model includes dummy variables, interactions and/or squared terms, you must create and add these variables to the SPSS spreadsheet *prior* to running a regression analysis. You can do this by clicking the "*Transform*" button on the SPSS main menu and selecting the "*Compute*" option (for squared terms and interactions) or the "*Create Dummy Variables*" option.]

To perform a standard regression analysis, select "*Enter*" in the "*Method*" box. To perform a stepwise regression analysis, select "*Stepwise*" in the "*Method*" box.

To perform a nested model *F*-test for additional model terms, click the "*Next*" button and enter the terms you want to test in the "*Independent(s)*" box. [*Note:* These terms, plus the terms you entered initially, form the complete model for the nested *F*-test.] Next, click the "*Statistics*" button and select "*R squared change*." Click "*Continue*" to return to the main SPSS linear regression dialog box.

To produce confidence intervals for the model parameters, click the "*Statistics*" button and check the appropriate menu items in the resulting menu list. To

**FIGURE E.32**
SPSS Linear Regression Dialog Box

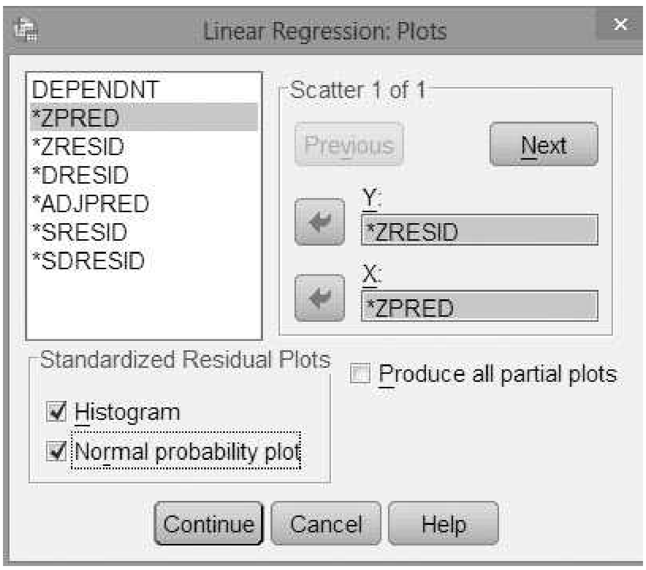



obtain prediction intervals for $y$ and confidence intervals for $E(y)$, click the "*Save*" button and check the appropriate items in the resulting menu list, as shown in Figure E.31. (The prediction intervals will be added as new columns to the SPSS data spreadsheet.) From this screen, you can also save residuals for plotting. To return to the main Regression dialog box from any of these optional screens, click "*Continue*."

Residual plots are obtained by clicking the "*Plots*" button and making the appropriate selections on the resulting menu (see Figure E.33). Influence diagnostics (e.g., studentized deleted residuals, leverage values, Cook's distances) are obtained by clicking the "*Save*" option and checking the diagnostics on the resulting menu screen (see Figure E.31).

After making all your menu selections, click "*OK*" on the Regression dialog box to view the multiple regression results.

**FIGURE E.33**
Menu Selections for Residual Plots



## E.11 Analysis of Variance

To conduct an ANOVA using SPSS, click on the "*Analyze*" button on the main menu bar, then click on "*General Linear Model*", and "*Univariate*". (See Figure E.34.) On the resulting dialog screen (Figure E.35), specify the response variable in the "*Dependent Variable*" box and the factor variables in the "*Fixed Factor(s)*" box. Click "*Model*" and specify the effects in the ANOVA model, as shown in Figure E.36. Click "Continue" to return to the ANOVA Variables dialog box.

To perform multiple comparisons of treatment means, click the "*Post Hoc*" button to obtain the dialog box shown in Figure E.37. On this box, check the comparison method (e.g., "*Bonferroni*") and specify the factor(s) to be analyzed in the "*Post Hoc Tests for*" box. Click "Continue" to return to the ANOVA Variables dialog box.

To perform a test of equality of variances, click on the "*Options*" button and check "*Homogeneity tests*" on the resulting menu screen. Click "*Continue*", then click "*OK*" to view the ANOVA results.

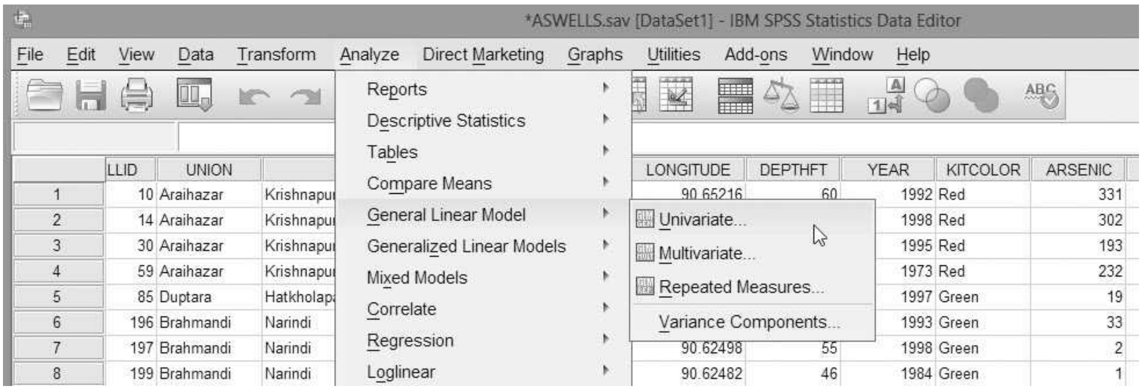



**FIGURE E.34**
SPSS Menu Options for ANOVA

**FIGURE E.35**
ANOVA Variables Dialog Box

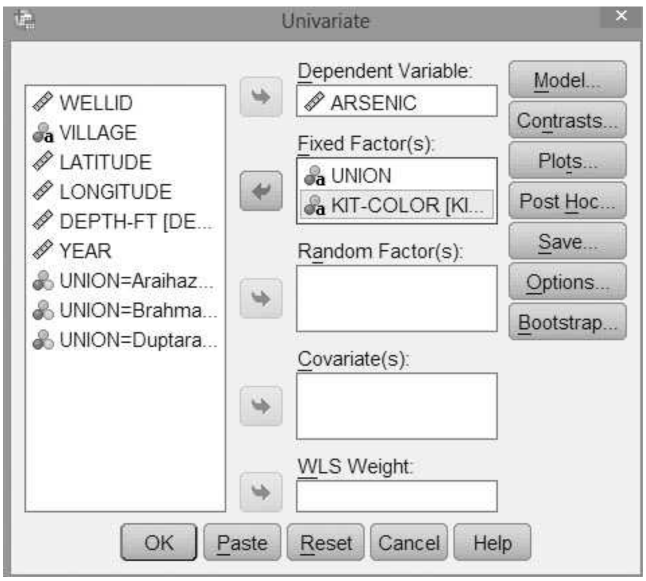


**FIGURE E.36**
ANOVA Model Dialog Box

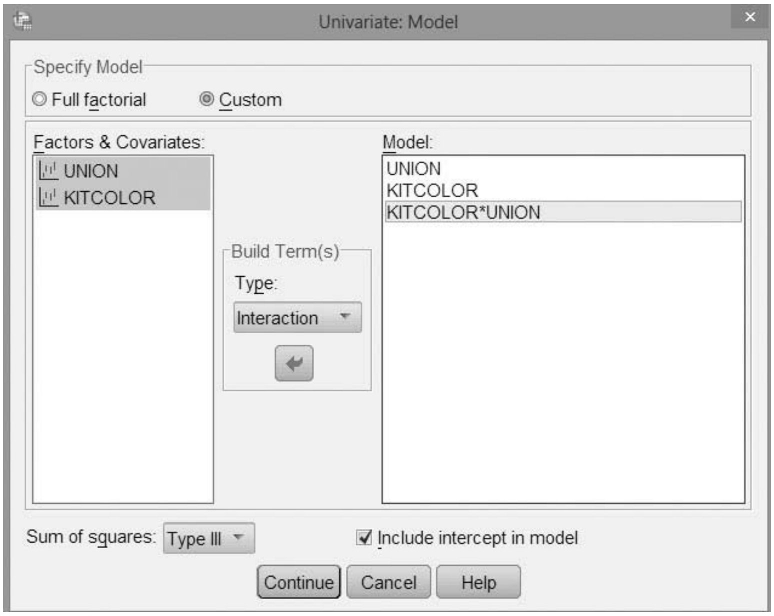
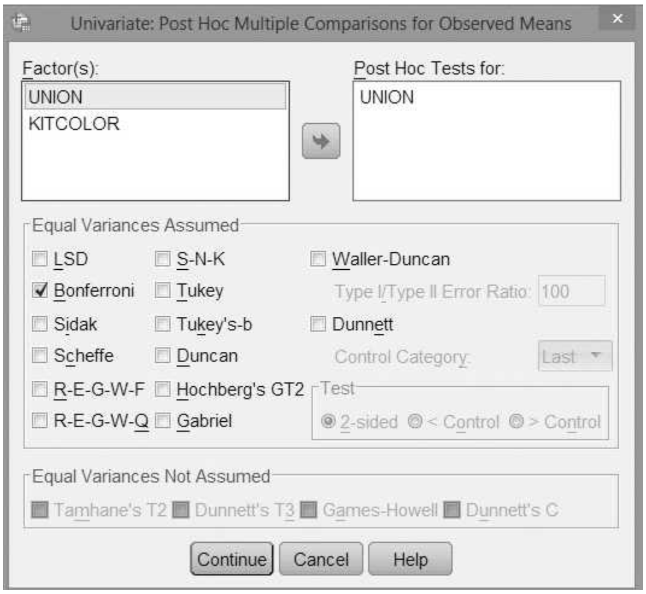



**FIGURE E.37**
Multiple Comparisons Dialog Box for ANOVA




## E.12 Nonparametric Tests

SPSS can perform the following nonparametric tests: sign test, Wilcoxon rank sum test, Wilcoxon signed-ranks test, Kruskal-Wallis test, Friedman test and Spearman's rank correlation test. All but Spearman's test are produced by making the following menu selections: Click on the "*Analyze*" button on the SPSS main menu bar, then click on "*Nonparametric Tests*", and "*Legacy Dialogs*". The resulting menu list is shown in Figure E.38. Select the type of nonparametric analysis you want to run (e.g., "*2 Independent Samples*"). The menu options for each of the different nonparametric tests are described below.

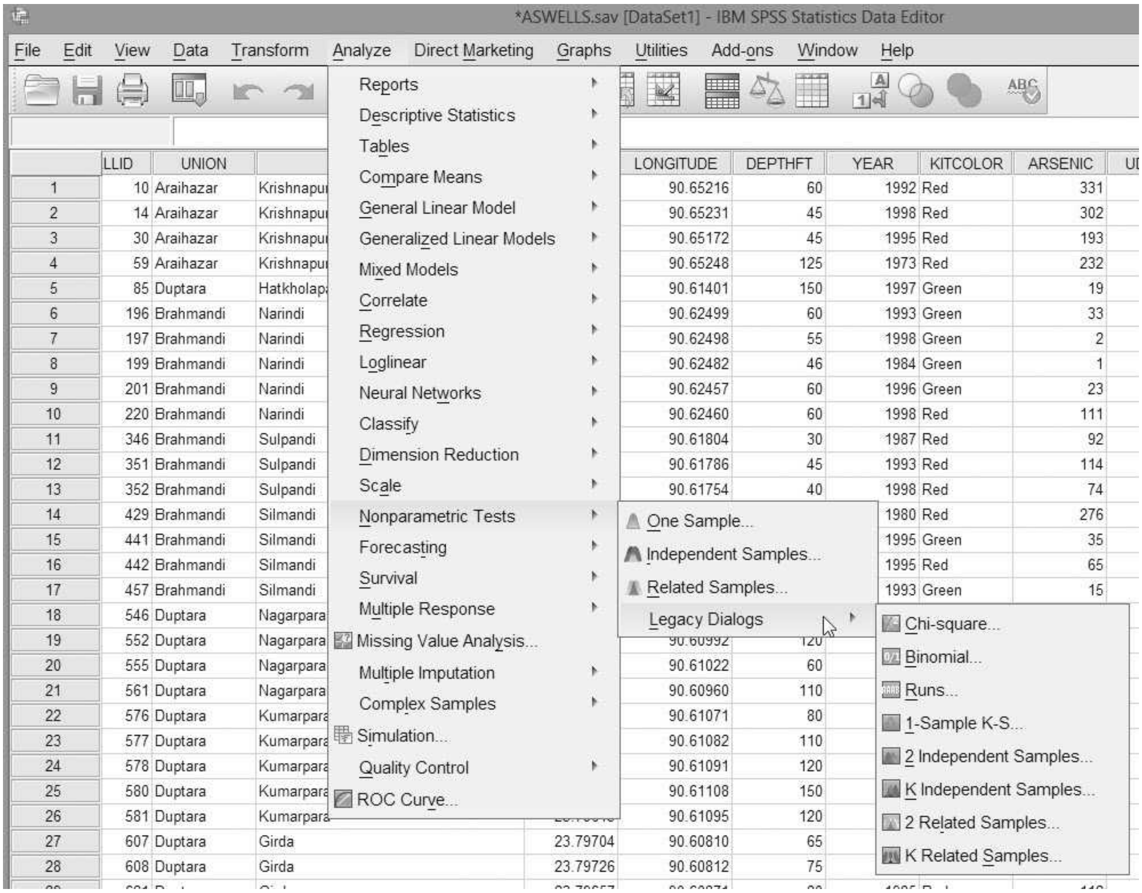



**FIGURE E.38**
SPSS Menu Options for Nonparametric Tests

## Sign Test or Signed-Ranks Test

Note: For the sign test, one variable on the SPSS file is the variable to be analyzed and the other variable will have the value of the hypothesized median for all cases. For the signed rank test, the two variables represent the two variables in the paired difference.

Select "*2 Related Samples*" from the nonparametrics menu list (see Figure E.38). On the resulting dialog box (see Figure E.39), select the two quantitative variables of interest for "*Variable 1*" and "*Variable 2*" in the "*Test Pairs*" box. Under "*Test Type*," select the "*Sign*" option for a sign test or the "*Wilcoxon*" option for a signed rank test. Click "*OK*" to generate the SPSS printout.

## Rank Sum Test

Note: The SPSS data file should contain two variables, one that represents the quantitative variable of interest and the other with two numerical coded values (e.g., 1 and 2). These two values represent the two groups or populations to be compared.

Select "2 *Independent Samples*" from the nonparametrics menu list (see Figure E.38). On the resulting dialog box (see Figure E.40), specify the quantitative variable of interest in the "*Test Variable List*" box and the coded variable in the "*Grouping Variable*" box. Click the "*Define Groups*" button and specify the values of the two groups in the resulting dialog box. Then click "*Continue*" to return to the

**FIGURE E.39**
Nonparametric Two-Related Samples Dialog Box

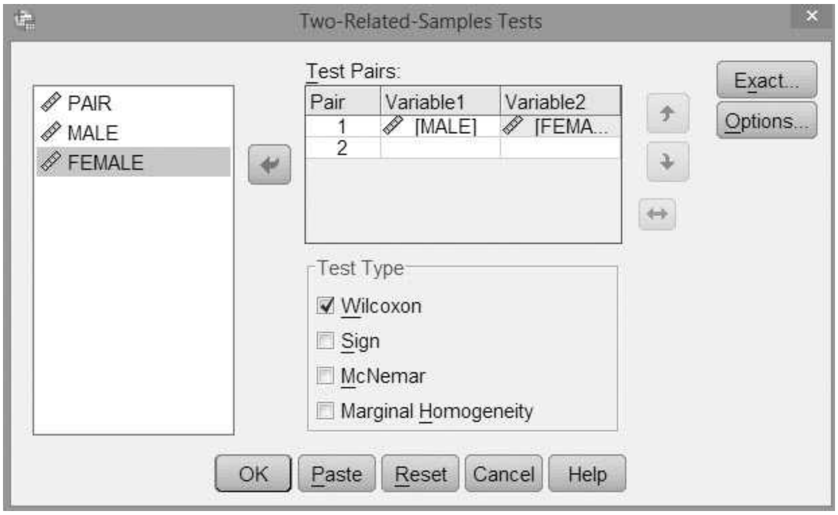
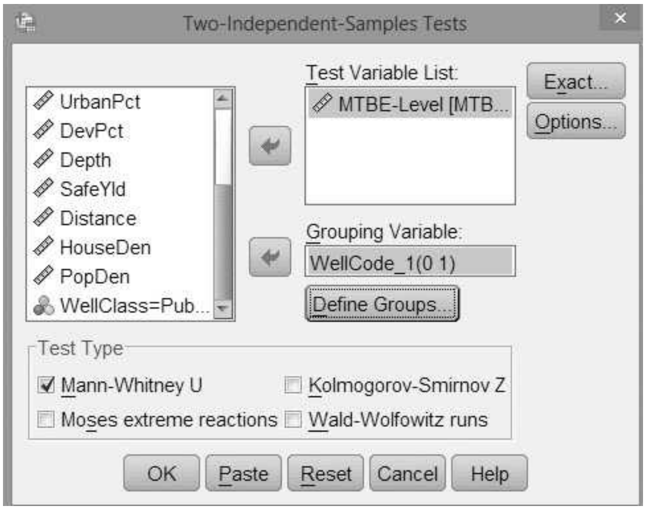



**FIGURE E.40**
Nonparametric Two Independent Samples Dialog Box



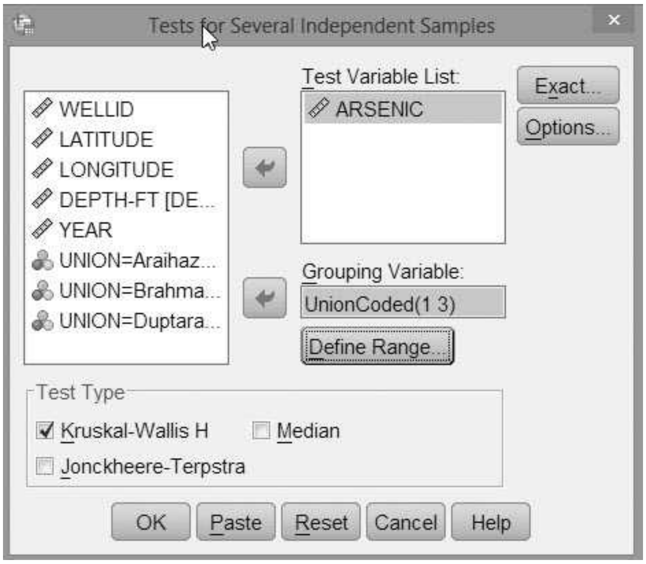

"*Two-Independent-Samples*" dialog screen. Select the "*Mann-Whitney U*" option under "*Test Type*." Click "*OK*" to generate the SPSS printout.

## Kruskal-Wallis Test

Note: The SPSS data file should contain one quantitative variable (the response, or dependent, variable) and one factor variable with at least two levels. (These values must be numbers, e.g., 1, 2, 3, etc.)

Select "*K Independent Samples*" from the nonparametrics menu list (see Figure E.38). On the resulting dialog box (see Figure E.41), specify the response variable in the "*Test Variable List*" box and the factor variable in the "*Grouping Variable*" box. Click the "*Define Range*" button and specify the values of the grouping factor in the resulting dialog box. Then click "*Continue*" to return to the "*K Independent Samples*" dialog screen. Select the "*Kruskal-Wallis*" option under "*Test Type*." Click "*OK*" to generate the SPSS printout.

**FIGURE E.41**
Nonparametric K Independent Samples Dialog Box




## Friedman Test

Note: The SPSS data file should contain k quantitative variables, representing the *k* treatments to be compared. The cases in the rows represent the blocks.

Select "*K Related Samples*" from the nonparametrics menu list (see Figure E.38). On the resulting dialog box (see Figure E.42), specify the treatment variables in the "*Test Variables*" box. Select the "Friedman" option under "Test Type." Click "OK" to generate the SPSS printout.

## Spearman's Rank Correlation Test

To obtain Spearman's rank correlation coefficient for the two quantitative variables of interest, click on the "*Analyze*" button on the main menu bar, then click on "*Correlate*" and "*Bivariate*." (See Figure E.15.) The resulting dialog box appears in Figure E.43. Enter the variables of interest in the "*Variables*" box. Check the "*Spearman*" option under "*Correlation Coefficients*." Click "*OK*" to obtain the SPSS printout.

**FIGURE E.42**
Nonparametric K Related Samples Dialog Box

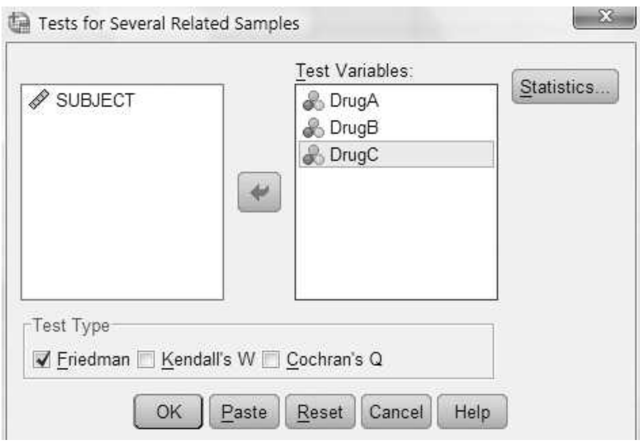
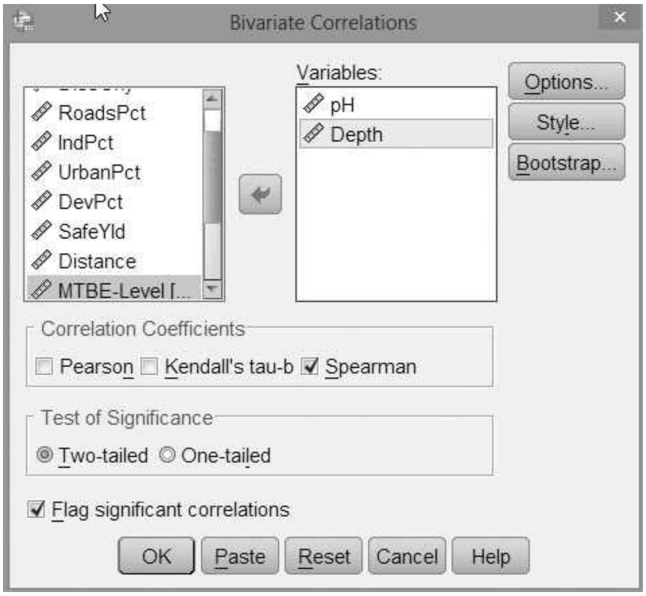


**FIGURE E.43**
SPSS Correlation Dialog Box




## E.13 Control Charts and Capability Analysis

### Variable Control Charts

To generate an individual variable control chart using SPSS, click on the "*Analyze*" button on the main menu bar and then click on "*Quality Control*" and "*Control Charts*," as shown in Figure E.44. On the resulting dialog box (shown in Figure E.45), select "*Individuals, Moving Range*" and "*Cases are units*", then click the "*Define*" button. On the resulting dialog box (see Figure E.46), specify the variable you want to graph in the "*Process Measurement*" box and the variable that identifies the individual measurements in the "*Identify points by*" box. Optionally, select "*Control Rules*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

### Mean and Range Control Charts

To generate both a mean and range control chart using SPSS, click on the "*Analyze*" button on the main menu bar and then click on "*Quality Control*" and "*Control Charts*," as shown in Figure E.44. On the resulting dialog box (shown in Figure E.45), select "*Xbar, R*, s" and "*Cases are units*", then click the "*Define*" button. On the resulting dialog box (see Figure E.47), specify the variable you want to graph in the "*Process Measurement*" box and the variable that identifies the subgroups in the "*Subgroups Defined by*" box. In the "*Charts*" box, select the type of Xbar chart you want and check "*Display R chart*". Optionally, select "*Control Rules*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

### Attributes (Number and Percent Defectives) Control Charts

To generate either a p-chart (for percent defectives) or a c-chart (for number with an attribute) using SPSS, click on the "*Analyze*" button on the main menu bar and then click on "*Quality Control*" and "*Control Charts*," as shown in Figure E.44. On the

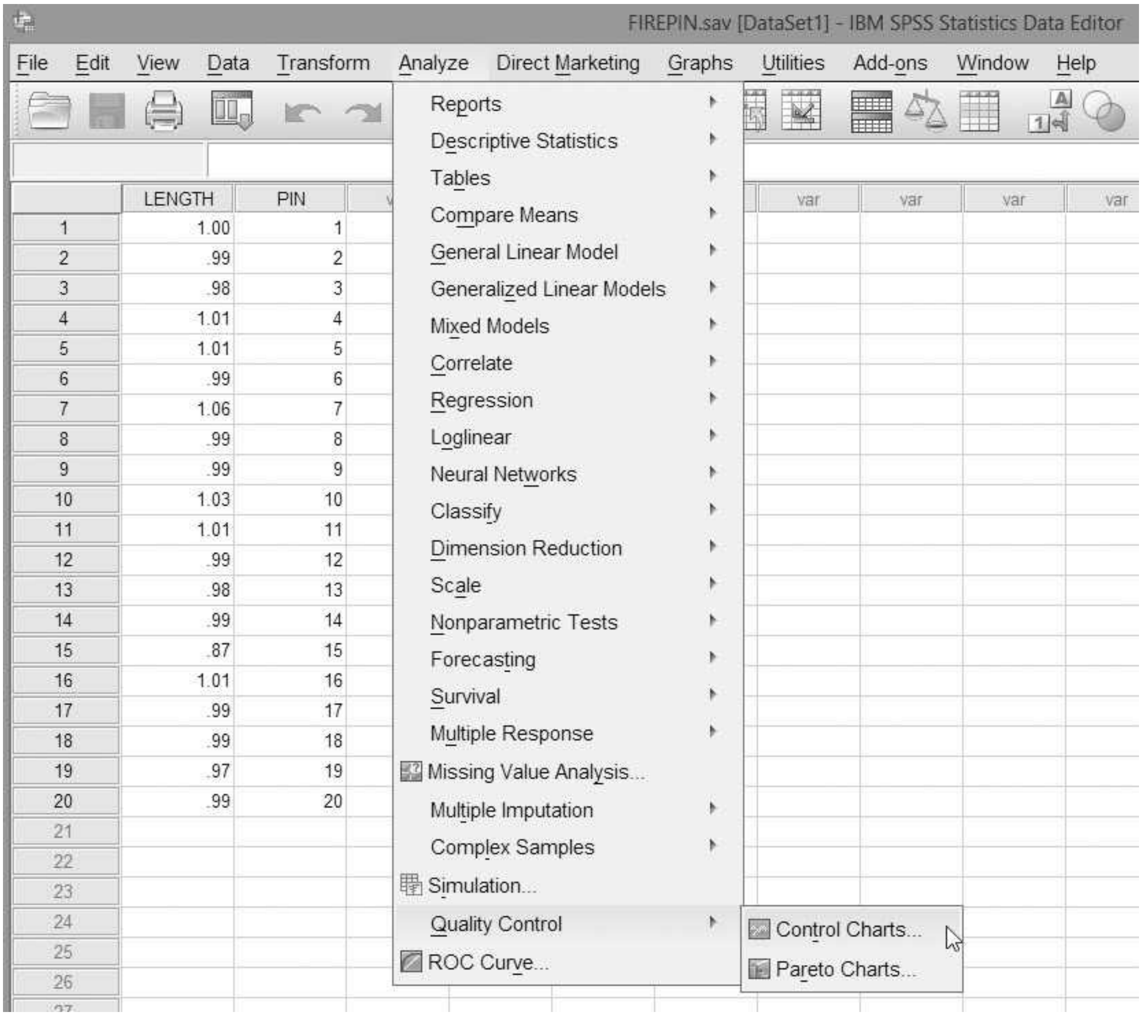



**FIGURE E.44**
SPSS Menu Options for Control Charts

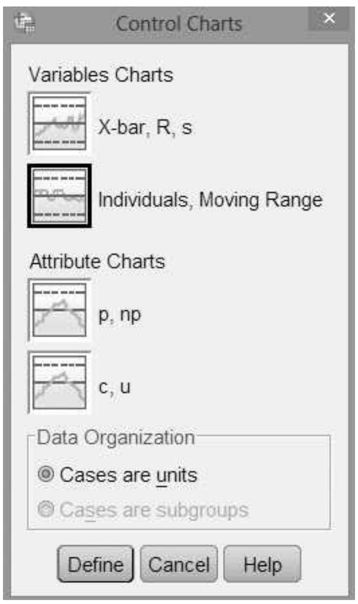



**FIGURE E.45**
Control Charts Selection Box

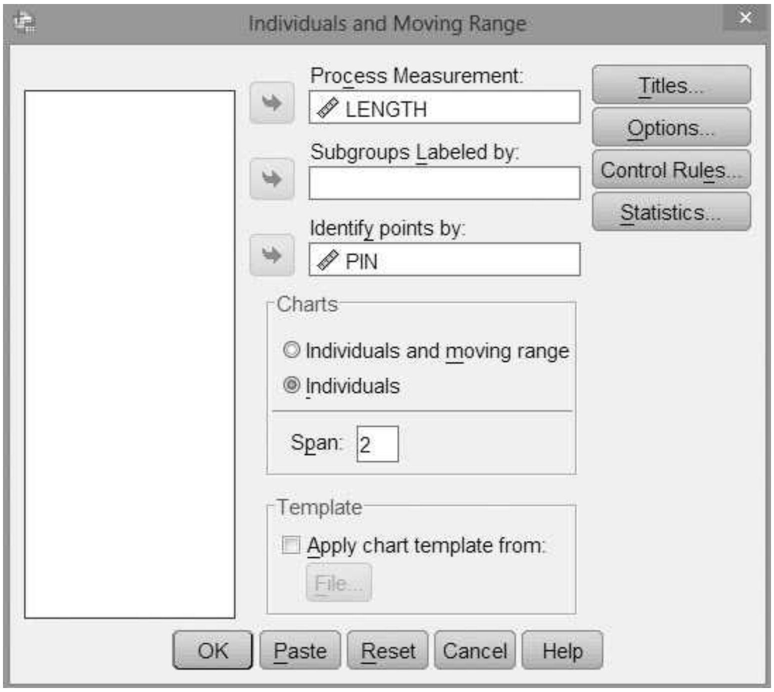



**FIGURE E.46**
Individual Variable Control Chart Dialog Box

resulting dialog box (shown in Figure E.45), select "p, np" for a p-chart or "c, u" for a c-chart. Select "*Cases are subgroups*", then click the "*Define*" button. On the resulting dialog box (see Figure E.48), specify the variable that represents the number of defects in the "*Number Nonconforming*" box and the variable that identifies the subgroups in the "*Subgroups Labeled by*" box. Enter the sample size for each subgroup in the designated box. (For a c-chart, enter "1" for the sample size.) Optionally, select "*Control Rules*" to specify any pattern-analysis rules you want to apply. Click "*OK*" to generate the control chart.

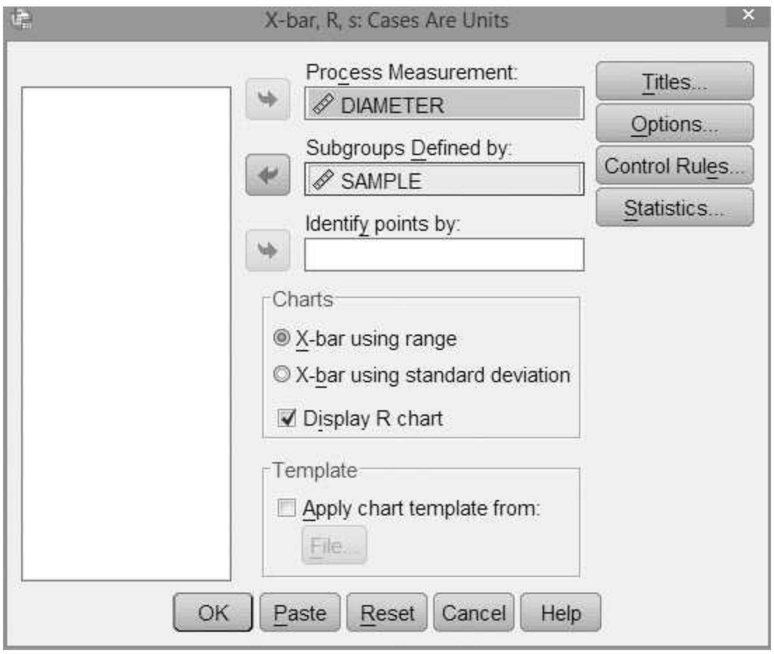
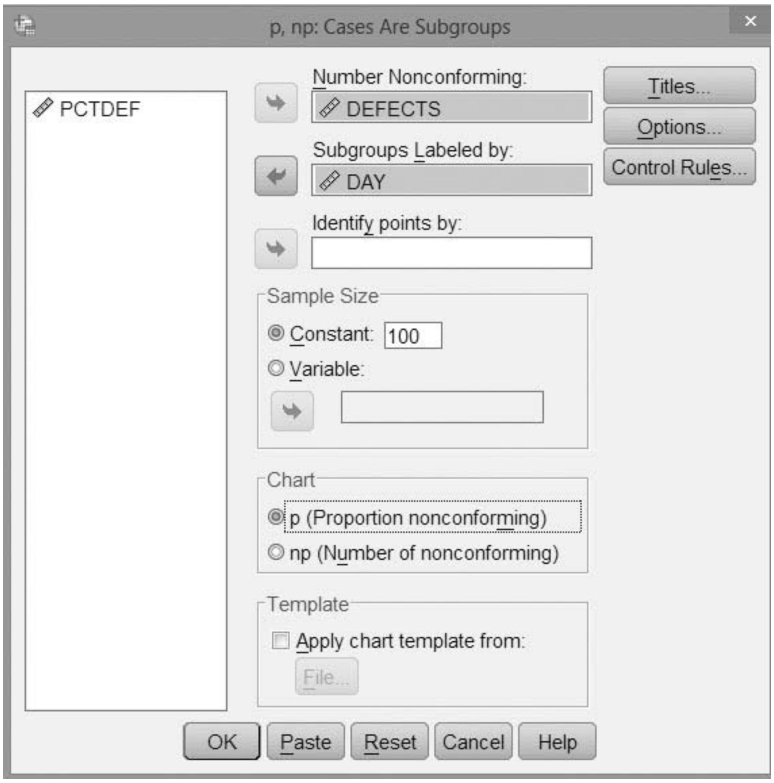



**FIGURE E.47**
Xbar-R Control Chart Dialog Box

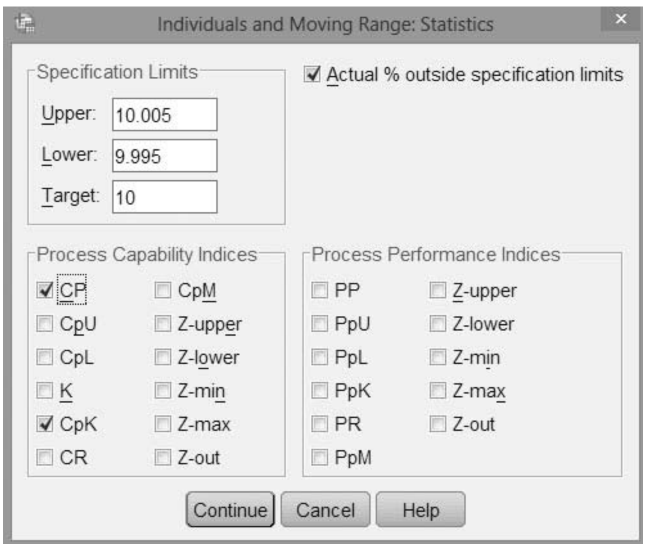
**FIGURE E.48**
P-Chart Dialog Box




## Capability Analysis

To conduct a capability analysis for a process, click on the "*Analyze*" button on the main menu bar and then click on "*Quality Control*" and "*Control Charts*," as shown in Figure E.44. On the resulting dialog box (shown in Figure E.45), select "*Individuals, Moving Range*" and "*Cases are units*", then click the "*Define*" button. On the resulting dialog box (see Figure E.46), specify the variable you want to analyze in the "*Process Measurement*" box and the variable that identifies the individual measurements in the "*Identify points by*" box. Select "*Statistics*" to view the Capability Analysis dialog box, as shown in Figure E.49. Enter upper, lower and target specification limits, and

**FIGURE E.49**
Capability Analysis Dialog Box



check the statistics (e.g., CpK) you want to compute. Click "Continue" then "*OK*" to produce the capability analysis graph and statistics.

## E.14 Random Samples

To generate a random sample of observations from a data set using SPSS, click on the "*Data*" button on the main menu bar, then click on "*Select Cases*" as shown in Figure E.50. On the resulting dialog box, select "*Random sample of cases*" from the list and then click on the "*Sample*" button, as shown on the left panel of Figure E.51. On the next dialog box (right panel of Figure E.51), specify the sample size there as a percentage of cases or a raw number by making the appropriate menu selections. Click "*Continue*" and then click "*OK*." The SPSS spreadsheet will reappear with the selected (sampled) cases.

**FIGURE E.50**
SPSS Menu Options for Random Samples

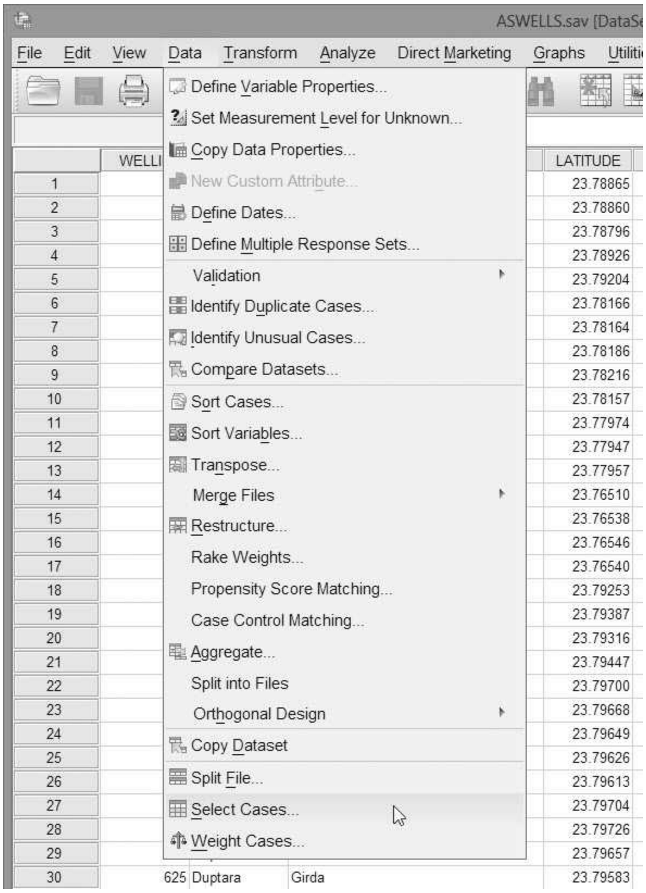
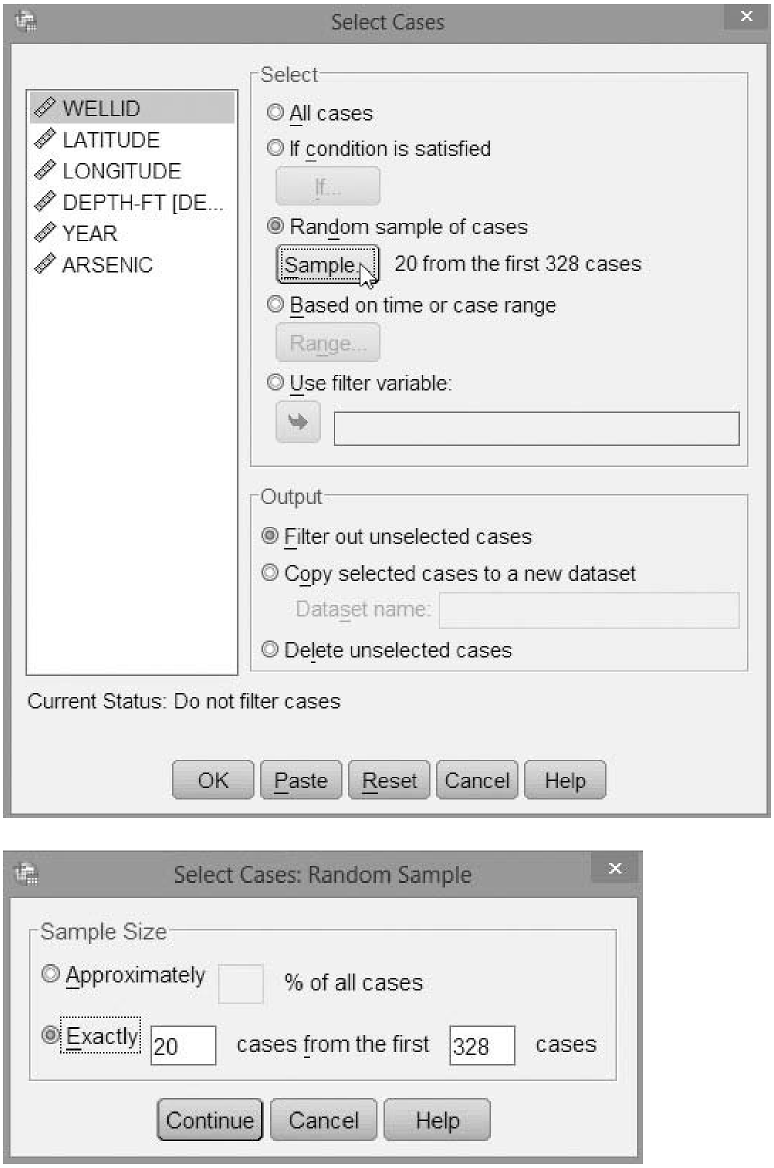


**FIGURE E.51**
SPSS Options for Selecting a Random Sample



# References


## Chapter 1

*Brochures about Survey Research*, Section on Survey Research Methods, American Statistical Association, 2004. (*www.amstat.org*)

*Careers in Statistics*, American Statistical Association, Biometric Society, Institute of Mathematical Statistics and Statistical Society of Canada, 2004. (*www.amstat.org*)

Peck, R., Casella, G., Cobb, G. W., Hoerl, R., Nolan, D., Starbuck, R., and Stern, H. *Statistics: A Guide to the Unknown*, 4th ed. Boston: Thomson/Brooks/Cole, 2006.

## Chapter 2

Freedman, D., Pisani, R., and Purves, R. *Statistics*. New York: W. W. Norton and Co., 1978.

Huff, D. *How to Lie with Statistics*. New York: Norton, 1954.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. *Introduction to Probability and Statistics*, 10th ed. North Scituate, MA: Duxbury, 1999.

Tufte, E. R. *Envisioning Information*. Cheshire, CT.: Graphics Press, 1990.

———. *Visual Explanations*. Cheshire, CT.: Graphics Press, 1997.

———. *Visual Display of Quantitative Information*. Cheshire, CT.: Graphics Press, 1983.

Sincich, T., Levine, D., and Stephan, D. *Practical Statistics by Example*. Upper Saddle River, NJ: Prentice Hall, 2002.

Tukey, J. W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

## Chapter 3

Bennett, D. J. *Randomness*. Cambridge, MA: Harvard University Press, 1998.

Epstein, R. A. *The Theory of Gambling and Statistical Logic*, rev. ed. New York: Academic Press, 1977.

Feller, W. *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. 1. New York: Wiley, 1968.

Lindley, D. V. *Making Decisions*, 2nd ed. London: Wiley, 1985.

Parzen, E. *Modern Probability Theory and Its Applications*. New York: Wiley, 1960.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. *Mathematical Statistics with Applications*, 6th ed. Boston: Duxbury, 2002.

Williams, B. *A Sampler on Sampling*. New York: Wiley, 1978.

Winkler, R. L. *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart and Winston, 1972.

Wright, G., and Ayton, P., eds. *Subjective Probability*. New York: Wiley, 1994.

## Chapter 4

Feller, W. *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed. New York: Wiley, 1968.

Hogg, R. V., and Craig, A. *Introduction to Mathematical Statistics*, 5th ed. Upper Saddle River, NJ: Prentice Hall, 1995.

Mendenhall, W. *Introduction to Mathematical Statistics*, 8th ed. Boston: Duxbury, 1991.

Mood, A. M., Graybill, F. A., and Boes, D. C. *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, 1963.

Mosteller, F., Rourke, R. E. K., and Thomas, G. B. *Probability with Statistical Applications*, 2nd ed. Reading, MA: Addison-Wesley, 1970.

Parzen, E. *Modern Probability Theory and Its Applications*. New York: Wiley, 1964.

Parzen, E. *Stochastic Processes*. San Francisco: Holden-Day, 1962.

*Standard Mathematical Tables*, 17th ed. Cleveland: Chemical Rubber Company, 1969.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. *Mathematical Statistics with Applications*, 6th ed. North Scituate, MA: Duxbury, 2002.

## Chapter 5

Hogg, R. V., and Craig, A. T. *Introduction to Mathematical Statistics*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 1995.

Lindgren, B. W. *Statistical Theory*, 3rd ed. New York: Macmillan, 1976.

Mood, A. M., Graybill, F. A., and Boes, D.C. *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, 1974.

Parzen, E. *Modern Probability Theory and Its Applications*. New York: Wiley, 1964.

Pearson, K. *Tables of the Incomplete Beta Function*. New York: Cambridge University Press, 1956.

Pearson, K. *Tables of the Incomplete Gamma Function*. New York: Cambridge University Press, 1956.

Ramsey, P. P. and Ramsey, P. H. “Simple tests of normality in small samples.” *Journal of Quality Technology*, Vol. 22, 1990.

Ross, S. M. *Stochastic Processes*, 2nd ed. New York: Wiley, 1996.

*Standard Mathematical Tables*, 17th ed. Cleveland: Chemical Rubber Company, 1969.

*Tables of the Binomial Probability Distribution*. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 6, 1950.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. *Mathematical Statistics with Applications*, 6th ed. North Scituate, MA: Duxbury, 2002.

Weibull, W. “A Statistical Distribution Function of Wide Applicability.” *Journal of Applied Mechanics*, Vol. 18 (1951), pp. 293–297.

Winkler, R. L., and Hays, W. *Statistics: Probability, Inference, and Decision*, 2nd ed. New York: Holt, Rinehart and Winston, 1975.

## Chapter 6

Freedman, D., Pisani, R., and Purves, R. *Statistics*, New York: Norton, 1978.

Hoel, P. G. *Introduction to Mathematical Statistics*, 4th ed. New York: Wiley, 1971.

Hogg, R. V., and Craig, A. T. *Introduction to Mathematical Statistics*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 1995.

Larsen, R. J., and Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2001.

Lindgren, B. W. *Statistical Theory*, 3rd ed. New York: Macmillan, 1976.

Mood, A. M., Grabill, F. A., and Boes, D. *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, 1974.

Snedecor, G. W., and Cochran, W. G. *Statistical Methods*, 7th ed. Ames, IA: lowa State University Press, 1980.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. *Mathematical Statistics with Applications*, 6th ed. North Scituate, MA: Duxbury, 2002.

## Chapter 7

Carlin, B., and Louis, T. “Bayes and empirical Bayes methods for data analysis.” *Statistics and Computing*, Vol. 7, No. 2, 1997.

Davison, A., and Hinkley, D. *Bootstrap Methods and Their Applications*. Cambridge, MA: Cambridge University Press, 1997.

Efron, B., and Tibshirani, R. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

Freedman, D., Pisani, R., and Purves, R. *Statistics*, 3rd ed. New York: Norton, 1998.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian Data Analysis*, 2nd ed. New York: Chapman and Hall, 2004.

Hoel, P. G. *Introduction to Mathematical Statistics*, 5th ed. New York: Wiley, 1984.

Hogg, R., McKean, J., and Craig, A. *Introduction to Mathematical Statistics*, 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2005.

Hogg, R., and Tanis, E. *Probability and Statistical Inference*, 7th ed. Upper Saddle River, NJ: Prentice-Hall, 2006.

Lehmann, E., and Casella, G. *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag, 1998.

Mendenhall, W., Beaver, R. J., and Beaver, B. *Introduction to Probability and Statistics*, 12th ed. Belmont, CA: Thomson, 2006.

Mood, A., Graybill, F., and Boes, D. *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, 1974.

Mosteller, F., and Tukey, J. *Data Analysis and Regression*. Reading, MA: Addison-Wesley, 1977.

Robert, C., and Casella, G. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.

Satterthwaite, F. W. "An approximate distribution of estimates of variance components." *Biometrics Bulletin*, Vol. 2, 1946, pp. 110–114.

Snedecor, G. W., and Cochran, W. *Statistical Methods*, 7th ed. Ames, IA: Iowa State University Press, 1980.

Steel, R. G. D., and Torrie, J. H. *Principles and Procedures of Statistics*, 2nd ed. New York: McGraw-Hill, 1980.

Tukey, J. W. "Bias and confidence in not-quite large samples." *Annals of Mathematical Statistics*, Vol. 29, 1958.

Wackerly, D., Mendenhall, W., and Scheaffer, R. *Mathematical Statistics with Applications*, 6th ed. Boston: Duxbury, 1996.

## Chapter 8

Carlin, B., and Louis, T. "Bayes and empirical Bayes methods for data analysis." *Statistics and Computing*, Vol. 7, No. 2, 1997.

Davison, A., and Hinkley, D. *Bootstrap Methods and Their Applications*. Cambridge, MA: Cambridge University Press, 1997.

Efron, B., and Tibshirani, R. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

Freedman, D., Pisani, R., and Purves, R. *Statistics*, 3rd ed. New York: Norton, 1998.

Hoel, P. G. *Introduction to Mathematical Statistics*, 6th ed. New York: Wiley, 1987.

Hogg, R., McKean, J., and Craig, A. *Introduction to Mathematical Statistics*, 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2005.

Hogg, R., and Tanis, E. *Probability and Statistical Inference*, 7th ed. Upper Saddle River, NJ: Prentice-Hall, 2006.

Mendenhall, W., Beaver, R. J., and Beaver, B. *Introduction to Probability and Statistics*, 12th ed. Belmont, CA: Thomson, 2006.

Satterthwaite, F. W. "An approximate distribution of estimates of variance components." *Biometrics Bulletin*, Vol. 2, 1946, pp. 110–114.

Steel, R. G. D., and Torrie, J. H. *Principles and Procedures of Statistics*, 2nd ed. New York: McGraw-Hill, 1980.

Wackerly, D., Mendenhall, W., and Scheaffer, R. *Mathematical Statistics with Applications*, 6th ed. Boston: Duxbury, 1996.

## Chapter 9


Agresti, A. *Categorical Data Analysis*. New York: Wiley, 1990.

Cochran, W. G. "The $\chi^2$ test of goodness of fit." *Annals of Mathematical Statistics*, Vol. 23, 1952.

Cochran, W. G. "Some methods for strengthening the common $\chi^2$ tests." *Biometrics*, Vol. 10, 1954.

Conover, W. J. *Practical Nonparametric Statistics*, 2nd ed. New York: Wiley, 1980.

Fisher, R. A. "The logic of inductive inference (with discussion)." *Journal of the Royal Statistical Society*, Vol. 98, 1935, pp. 39–82.

Hollander, M., and Wolfe, D. A. *Nonparametric Statistical Methods*. New York: Wiley, 1973.

Savage, I. R. "Bibliography of nonparametric statistics and related topics." *Journal of the American Statistical Association*, 1953, p. 48.


## Chapter 10


Chatterjee, S., and Price, B. *Regression Analysis by Example*, 2nd ed. New York: Wiley, 1991.

Draper, N., and Smith, H. *Applied Regression Analysis*, 3rd ed. New York: Wiley, 1987.

Graybill, F. *Theory and Application of the Linear Model*. North Scituate, MA: Duxbury, 1976.

Kleinbaum, D., and Kupper, L. *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. North Scituate, MA: Duxbury, 1997.

Mendenhall, W. *Introduction to Linear Models and the Design and Analysis of Experiments*, Belmont, CA: Wadsworth, 1968.

Mendenhall, W., and Sincich, T. *A Second Course in Statistics: Regression Analysis*, 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2003.

Montgomery, D., Peck, E., and Vining, G. *Introduction to Linear Regression Analysis*, 3rd ed. New York: Wiley, 2001.

Mosteller, F., and Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley, 1977.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. *Applied Linear Statistical Models*, 4th ed. Homewood, IL: Richard D. Irwin, 1996.

Rousseeuw, P. J., and Leroy, A. M. *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

Weisburg, S. *Applied Linear Regression*, 2nd ed. New York: Wiley, 1985.


## Chapter 11


Barnett, V., and Lewis, T. *Outliers in Statistical Data*. New York: Wiley, 1978.

Belsley, D. A., Kuh, E., and Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 1980.

Box, G. E. P., and Jenkins, G. M. *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day, Inc., 1970.

Chatterjee, S., and Price, B. *Regression Analysis by Example*, 2nd ed. New York: Wiley, 1991.

Draper, N. R., and Smith, H. *Applied Regression Analysis*, 2nd ed. New York: Wiley, 1981.

Fuller, W. *Introduction to Statistical Time Series*. New York: Wiley, 1976.

Graybill, F. A. *Theory and Application of the Linear Model*. North Scituate, MA: Duxbury, 1976.

Mendenhall, W. *Introduction to Linear Models and the Design and Analysis of Experiments*. Belmont, CA: Wadsworth, 1968.

Mendenhall, W., and Sincich T. *A Second Course in Statistics: Regression Analysis*, 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2003.

Mosteller, F., and Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley, 1977.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. *Applied Linear Statistical Models*, 4th ed. Homewood, IL: Richard Irwin, 1996.

Rousseeuw, P. J., and Leroy, A. M. *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

Weisberg, S. *Applied Linear Regression*, 2nd ed. New York: Wiley, 1985.

## Chapter 12

Daniel, C., and Wood, F. *Fitting Equations to Data*, 2nd ed. New York: Wiley, 1980.

Draper, N., and Smith, H. *Applied Regression Analysis*, 3rd ed. New York: Wiley, 1998.

Graybill, F. A. *Theory and Application of the Linear Model*. North Scituate, MA: Duxbury, 1976.

Geisser, S. "The predictive sample reuse method with applications." *Journal of the American Statistical Association*, Vol. 70, 1975.

Mendenhall, W. *Introduction to Linear Models and the Design and Analysis of Experiments*. Belmont, CA: Wadsworth, 1968.

Mendenhall, W., and Sincich, T. *A Second Course in Statistics: Regression Analysis*, 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2003.

Montgomery, D., Peck, E., and Vining, G. *Introduction to Linear Regression Analysis*, 3rd ed. New York: Wiley, 2001.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. *Applied Linear Statistical Models*, 4th ed. Homewood, IL: Richard D. Irwin, 1996.

Snee, R., "Validation of regression models: Methods and examples." *Technometrics*, Vol. 19, 1977.

## Chapter 13

Box G. E. P., Hunter, W. G., and Hunter, J. S. *Statistics for Experimenters*. New York: Wiley, 1957.

Cochran, W. G., and Cox, G. M. *Experimental Designs*, 2nd ed. New York: Wiley, 1957.

Davies, O. L. *The Design and Analysis of Industrial Experiments*. New York: Hafner, 1956.

Kirk, R. E. *Experimental Design*, 2nd ed. Belmont, CA: Brooks/Cole, 1982.

Mendenhall, W. *Introduction to Linear Models and the Design and Analysis of Experiments*. Belmont, CA: Wadsworth, 1968.

Neter, J., Kutner, M. Nachtsheim, C. and Wasserman, W. *Applied Linear Statistical Models*, 4th ed. Homewood, IL: Richard D. Irwin, 1996.

Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

## Chapter 14

Box, G. E. P., Hunter, W. G., and Hunter, J. S. *Statistics for Experimenters*. New York: Wiley, 1978.

Cochran, W. G., and Cox, G. M. *Experimental Designs*, 2nd ed. New York: Wiley, 1957.

Hicks, C. R. *Fundamental Concepts in the Design of Experiments*, 3rd ed. New York: CBC College Publishing, 1982.

Hochberg, Y., and Tamhane, A. C. *Multiple Comparison Procedures*. New York: Wiley, 1987.

Hsu, J. C. *Multiple Comparisons, Theory and Methods*. New York: Chapman & Hall, 1996.

Johnson, R., and Wichern, D. *Applied Multivariate Statistical Methods*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1992.

Kirk, R. E. *Experimental Design*, 2nd ed. Belmont, CA: Brooks/Cole, 1982.

Kramer, C. Y. "Extension of multiple range tests to group means with unequal number of replications." *Biometrics*, Vol. 12, 1956, pp. 307–310.

Levene, H. *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press, 1960, pp. 278–292.

Mendenhall, W. *Introduction to Linear Models and the Design and Analysis of Experiments*. Belmont, CA: Wadsworth, 1968.

Miller, R. G. *Simultaneous Statistical Inference*, 2nd ed. New York: Springer-Verlag, 1981.

Montgomery, D. C. *Design and Analysis of Experiments*, 3rd ed. New York: John Wiley & Sons, 1991.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. *Applied Linear Statistical Models*, 4th ed. Homewood, IL: Richard D. Irwin, 1996.

Scheffe, H. "A method for judging all contrasts in the analysis of variance." *Biometrika*, Vol. 40, 1953, pp. 87–104.

Scheffe, H. *The Analysis of Variance*. New York: Wiley, 1959.

Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*. New York: Wiley, 1992.

Tukey, J. W. "Comparing individual means in the analysis of variance." *Biometrics*, Vol. 5, 1949, pp. 99–114.

Uusipaikka, E. "Exact simultaneous confidence intervals for multiple comparisons among three or four mean values." *Journal of the American Statistical Association*, Vol. 80, 1985, pp. 196–201.

Winer, B. J. *Statistical Principals in Experimental Design*, 2nd ed. New York: McGraw-Hill, 1971.

## Chapter 15

Agresti, A., and Agresti, B. F. *Statistical Methods for the Social Sciences*, 2nd ed. San Francisco: Dellen, 1986.

Conover, W. J. *Practical Nonparametric Statistics*, 2nd ed. New York: Wiley, 1980.

Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. Boston: PWS-Kent, 1990.

Dunn, O. J. "Multiple comparisons using rank sums." *Technometrics*, Vol. 6, 1964.

Friedman, M. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *Journal of the American Statistical Association*, Vol. 32, 1937.

Gibbons, J. D. *Nonparametric Statistical Inference*, 2nd ed. New York: McGraw-Hill, 1985.

Hollander, M., and Wolfe, D. A. *Nonparametric Statistical Methods*. New York: Wiley, 1973.

Kruskal, W. H., and Wallis, W. A. "Use of ranks in one-criterion variance analysis." *Journal of the American Statistical Association*, Vol. 47, 1952.

Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, 1975.

Marascuilo, L. A., and McSweeney, M. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Monterey, CA: Brooks/Cole, 1977.

Wilcoxon, F., and Wilcox, R. A. "Some rapid approximate statistical procedures." The American Cyanamid Co., 1964.

## Chapter 16

Alwan, L. C., and Roberts, H. V. "Time-series modeling for statistical process control." *Journal of Business and Economic Statistics*, 1988, Vol. 6, pp. 87–95.

Banks, J. *Principles of Quality Control*. New York: Wiley, 1989.

Box, G. E. P. "Evolutionary operation: A method for increasing industrial productivity." *Applied Statistics*, Vol. 6, 1957, pp. 3–23.

Box, G. E. P., and Hunter, J. S. "Condensed calculations for evolutionary operation programs." *Technometrics*, Vol. 1, 1959, pp. 77–95.

Checkland, P. *Systems Thinking, Systems Practice*. New York: Wiley, 1981.

Deming, W. E. *Quality, Productivity, and Competitive Position*. Cambridge, MA: MIT Press, 1982.

DeVor, R. E., Chang, T., and Southerland, J. W. *Statistical Quality Design and Control*. New York: Macmillan, 1992.

Duncan, A. J. *Quality Control and Industrial Statistics*. Homewood, IL: Irwin, 1986.

Feigenbaum, A. V. *Total Quality Control*, 3rd ed. New York: McGraw-Hill, 1983.

Garvin, D. A. *Managing Quality*. New York: Free Press/Macmillan, 1988.

Gitlow, H., Gitlow, S., Oppenheim, A., and Oppenheim, R. *Tools and Methods for the Improvement of Quality*. Homewood, IL: Irwin, 1989.

Grant, E. L., and Leavenworth, R. S. *Statistical Quality Control*, 6th ed. New York: McGraw-Hill, 1988.

Hald, A. *Statistical Theory of Sampling Inspection of Attributes*. New York: Academic Press, 1981.

Hart, Marilyn K. "Quality tools for improvement." *Production and Inventory Management Journal*, First Quarter 1992, Vol. 33, No. 1, p. 59.

Ishikawa, K. *Guide to Quality Control*, 2nd ed. White Plains, NY: Kraus International Publications, 1986.

Joiner, B. L., and Goudard, M. A. "Variation, management, and W. Edwards Deming." *Quality Process*, Dec. 1990, pp. 29–37.

Juran, J. M., and Gryna, F. M., Jr. *Quality Planning Analysis*, 2nd ed. New York: McGraw-Hill, 1980.

Kane, V. E. *Defect Prevention*. New York: Marcel Dekker, 1989.

*Military Standard 105D*. Washington, DC: U.S. Government Printing Office, 1963.

Moen, R. D., Nolan, T. W., and Provost, L. P. *Improving Quality through Planned Experimentation*. New York: McGraw-Hill, 1991.

Montgomery, D. C. *Introduction to Statistical Quality Control*, 2nd ed. New York: Wiley, 1991.

National Bureau of Standards, *Tables of the Binomial Distribution*. Washington, DC: U.S. Government Printing Office, 1950.

Nelson, L. L. "The Shewhart control chart—Tests for special causes." *Journal of Quality Technology*, Oct. 1984, Vol. 16, No. 4, pp. 237–239.

Ott, E. R. *Process Quality Control: Trouble-shooting and Interpretation of Data*. New York: McGraw-Hill, 1975.

Romig, H. G. 50–100 *Binomial Tables*. New York: Wiley, 1953.

Ryan, T. P. *Statistical Methods for Quality Improvement*. New York: Wiley, 1989.

Shewhart, W. A. *Economic Control of Quality of Manufactured Product*. Princeton, NJ: Van Nostrand Reinhold, 1931.

*Statistical Quality Control Handbook*. Indianapolis, IN: AT&T Technologies, Select Code 700-444 (inquiries: 800-432-6600); originally published by Western Electric Company, 1956.

Wadsworth, H. M., Stephens, K.S., and Godfrey, A. B. *Modern Methods for Quality Control and Improvement*. New York: Wiley, 1986.

Wheeler, D. J., and Chambers, D. S. *Understanding Statistical Process Control*. Knoxville, TN: Statistical Process Controls, Inc., 1986.

## Chapter 17

Allison, P. D. *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute, 1998.

Barlow, R. E., and Proschan, F. *The Mathematical Theory of Reliability*, New York: Wiley, 1965.

Box, G. E. P. "Problems in the analysis of growth and wear curves." *Biometrics*, Vol. 6, 1950.

Cohen, A. C., Jr. "On estimating the mean and standard deviation of truncated normal distribution." *Journal of the American Statistical Association*, Vol. 44, 1949, pp. 518–525.

———. "A note on truncated distributions." Industrial Quality Control, Vol. 6, 1949, p. 22.

Cox, D. R. "Regression models and life tables (with discussion)." *Journal of the Royal Statistical Society, Series B*, Vol. 34, 1972.

Davis, D. J. "An analysis of some failure data." *Journal of the American Statistical Association*, Vol. 47, 1952, pp. 113–150.

Epstein, B. "Statistical problems in life testing." *Seventh Annual Quality Control Conference Papers*, 1953, pp. 385–398.

Epstein, B., and Sobel, M. "Life testing." *Journal of the American Statistical Association*, Vol. 48, 1953, pp. 486–502.

Kalbfleisch, J. D., and Prentice, R. L. *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.

Miller, I., and Freund, J. E. *Probability and Statistics for Engineers*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1977.

Therneau, T. M, and Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000.

Weibull, W. "A statistical distribution function of wide applicability." *Journal of Applied Mechanics*, Vol. 18, 1951, pp. 293–297.

Zelen, M. *Statistical Theory of Reliability*. Madison, WI: University of Wisconsin Press, 1963.

# Selected Short Answers

## Chapter 1

**1.1** **a.** all young women who recently participated in a STEM program **b.** 159 surveyed women **c.** 27% feel STEM program increased their interest in science.
**1.3** populations: (1) male students who are video game players, (2) male students who are not video game players
**1.5** **a.** earthquakes **b.** sample.
**1.7** **a.** level of carbon monoxide gas; a week at the weather station **b.** population
**1.9** **a.** qualitative **b.** quantitative **c.** quantitative
**1.11** **a.** qualitative **b.** qualitative **c.** quantitative **d.** quantitative **e.** quantitative **f.** quantitative **g.** quantitative **h.** qualitative
**1.13** **a.** smokers **b.** screening method and age of tumor detection **c.** qualitative; quantitative **d.** which screening method is more effective in pinpointing small tumors.
**1.19** **a.** all computer security personnel at U.S. corporations and government agencies **b.** survey; nonresponse bias **c.** unauthorized use or not; qualitative **d.** 41% of all firms had unauthorized use of their computer systems
**1.23** **a.** all hardware components **b.** 100 components tested for life length **c.** quantitative **d.** estimate mean life length of all components
**1.25** **a.** 2-milliliter cleaning solutions **b.** amount of hydrochloric acid necessary to neutralize solution **c.** all possible 2-milliliter cleaning solutions **d.** five 2-milliliter solutions prepared by the chemist
**1.27** **a.** undergraduate engineering students **b.** population: all undergraduate engineering students at Penn State; sample: 21 undergraduate engineering students selected for the study **c.** quantitative **d.** estimate average Perry score for all undergraduate students at Penn State to be 3.27
**1.29** **a.** structural status of a bridge **b.** qualitative **c.** population **d.** observational study

## Chapter 2

**2.1** **a.** bar chart **b.** type of robotic limbs **c.** legs only **d.** None: .1415; Both: .0755; Legs only: .5943; Wheels only: .1887
**2.5** **a.** hotspot: qualitative; beach condition: qualitative; bar condition: qualitative; erosion rate: quantitative
**2.7** Most LEO satellites owned by government (45.6%); most GEO satellites owned by commercial sector (65.0%)
**2.9** **b.** no
**2.13** **a.** 10–20 **b.** .68 **c.** .175 **d.** .12
**2.15** **e.** .444
**2.19** **b.** .941 **c.** stem-and-leaf display
**2.21** **a.** .26 **b.** .086
**2.23** increases value of zeta potential
**2.25** **a.** 6; 5; all **b.** 6; 5.5; 4 and 6
**2.27** **a.** 16.5; increase **b.** 16.16; no change **c.** no mode
**2.29** mean $=$ 9.72, median $=$ 10.94
**2.31** **a.** 1.81 **b.** 1.35 **c.** 4 **d.** 2.85 **e.** .45 **f.** dioxide level less with crude oil
**2.33** **e.** Group B
**2.35** **a.** mean $= -1.09$, median $= -0.655$ **b.** $-8.11$ **c.** mean $= -.52$, median $= -.52$; mean
**2.37** **a.** no, skewed to right **b.** $\bar{y} = 3.21$, $s = 1.37$ **c.** (.47, 5.96) **d.** at least .75 **e.** $\approx$.95 **f.** .93; yes
**2.39** **a.** .18 **b.** .0041 **c.** .064 **d.** morning
**2.41** **a.** 67.2 **b.** 14.48 **c.** Chebyshev's rule: at least 88.8% of measurements for Group A fall between 30.18 and 117.06
**2.43** **a.** yes **b.** no
**2.45** **a.** $(-0.900, 2.900)$ **b.** $(-16.220, 25.340)$
**2.47** (204.815, 264.665)
**2.49** **a.** 10% **b.** 90%
**2.51** **a.** \$141,417 **b.** \$96,417 **c.** $-1.76$
**2.53** **a.** 1.57 **b.** $-3.36$
**2.55** **a.** $z = -3.83$ **b.** $z = -1.58$
**2.57** **a.** $z = 15.83$ **b.** $z = 1.26$ **c.** calcium/gypsum
**2.59** **a.** 50% of clinkers have Barium values below 170 mg/kg **b.** 25% of clinkers have Barium values below 115 mg/kg **c.** 75% of clinkers have Barium values below 260 mg/kg **d.** 145 **e.** $(-102.5. 477.5)$ **f.** no outliers
**2.61** **a.** no, $z = 1.57$ **b.** yes, $z = -3.36$
**2.63** yes, $z = -2.5$
**2.65** **a.** 117.3, 118.5, and 122.4 **b.** 50.4 **c.** none
**2.67** **a.** appears like BP was collecting more barrels of oil on each successive day
**2.69** **a.** fate **b.** burned; recycled; exported; land disposed **c.** .517, .32, .023, .14
**2.73** (0.833, 2.929)
**2.75** 1.06
**2.79** **d.** $\bar{y} = 62.96$, $s = .61$ **e.** 96.97%; yes **f.** 62.57, 63.01, 63.36, 63.71
**2.81** no, $z = 2.3$
**2.83** **a.** seabirds & length: quantitative; oil: qualitative **b.** transect **c.** oiled: 38%; unoiled: 62% **e.** distributions similar **f.** (0, 16.67) **g.** (0, 15.43) **h.** unoiled
**2.85** **a.** grounding **b.** $\bar{y} = 59.82$, $s = 53.36$; (0, 166.54)
**2.87** **a.** quantitative **b.** frequency distribution **c.** .28 **d.** yes

## Chapter 3

**3.1** **a.** legs only, wheels only, both legs and wheels, and neither legs nor wheels **b.** P(Legs only) = .594, P(Wheels only) = .189, P(both) = .076, P(neither) = .141 **c.** .265 **d.** .670
**3.3** passing ship
**3.5** **a.** SL, IT, CP, NP, and 0 **b.** .06, .26, .21, .35, .12 **c.** .06
**3.7** **a.** Pu/B/BL, Pu/B/D, Pu/U/BL, Pu/U/D, Pr/B/BL, Pr/B/D, Pr/U/BL, Pr/U/D **b.** .256, .184, .067, .031, .363, .099, 0, 0 **c.** .314
**3.9** **a.** .261 **b.** Trunk: .85; Leaves: .10; Branch: .05
**3.11** **a.** 12 **b.** no
**3.13** **a.** .56 **b.** .94
**3.15** **a.** .271 **b.** .706 **c.** .088
**3.17** **a.** 32 simple events: FFFFF, FFFFW, . . . , WWWWW **b.** .97
**3.19** **a.** AC, AW, AF, IC, IW, and IF **b.** .148, .066, .426, .176, .052, .132 **c.** .64 **d.** .118 **e.** .176 **f.** .427 **g.** .676
**3.21** .984
**3.23** .286
**3.25** $P(A \mid B) = 0$
**3.27** **a.** .7628 **b.** .1445
**3.29** **a.** .667 **b.** .458
**3.31** .559
**3.33** **a.** .531 **b.** .531
**3.35** .35
**3.37** .09
**3.39** **a.** $P(\text{A} \mid \text{I}) = .9$, $P(\text{B} \mid \text{I}) = .95$, $P(\text{A} \mid \text{N}) = .2$, $P(\text{B} \mid \text{N}) = .1$ **b.** .855 **c.** .02 **d.** .995
**3.41** .04
**3.43** **a.** .116 **b.** .728
**3.45** $(.5 + \alpha - \alpha\beta)\beta$
**3.49** Door, since $P(\text{D} \mid \text{J}) = .6122$
**3.51** Novice, since $P(\text{Novice} \mid \text{Fail}) = .764$
**3.53** **a.** .158 **b.** .316 **c.** .526 **d.** #3
**3.55** no, since $P(\text{D} \mid \text{G}) = .108$
**3.59** **a.** 18 **b.** 4/18
**3.61** **a.** 24 **b.** 100
**3.63** **a.** 16 **b.** 24
**3.65** **a.** 729 **b.** 120
**3.67** **a.** 168 **b.** 8/168 **c.** 2/7
**3.69** 63,063,000
**3.71** **a.** 63/1,326 **b.** .0465
**3.73** *P*(at least one defective) = .039 if claim true
**3.75** **a.** No; *P*(3 misses) = .166 if $p = .45$
**b.** Yes; *P*(lO misses) = .0025 if $p = .45$
**3.77** **a.** BB, TG, GG, S, G **b.** .28, .11, .11, .26, .24 **c.** .52 **d.** .48
**3.79** **a.** .974 **b.** .12
**3.81** **a.** .92 **b.** 1
**3.83** **a.** 1,440 **b.** 240
**3.85** **a.** .12 **b.** .473
**3.87** **a.** .06 **b.** .94
**3.89** .2362, .1942, .5696
**3.91** 26
**3.95** **a.** 60 **b.** 3/5 **c.** 3/10
**3.97** **a.** .0019808 **b.** .00394 **c.** .0000154
**3.99** *P*(at least 1 division error in 1 billion divisions) = .105

## Chapter 4

**4.1** **b.** $p(0) = .116$, $p(1) = .312$, $p(2) = .336$, $p(3) = .181$, $p(4) = .049$, $p(5) = .005$ **d.** .054
**4.3** **a.** $p(1) = .4$, $p(2) = .54$, $p(3) = .02$, $p(4) = .04$ **b.** .06
**4.5** **a.** .23 **b.** .081 **c.** .77
**4.7** **a.** 0, 1, 2 **b.** $p(0) = 5/8$, $p(1) = 2/8$, $p(2) = 1/8$
**4.9** **b.** $p(30) = .0086$, $p(40) = .1441$, $p(50) = .3026$, $p(60) = .5447$ **c.** .8473
**4.11** $p(1) = 3/5$, $p(2) = 3/10$, $p(3) = 1/10$
**4.13** **a.** 1.8 **b.** .99 **c.** .96
**4.15** .29
**4.17** **a.** 2.9; 3 **b.** 3; 4 **c.** 3; 3
**4.19** $\mu = \$3{,}600$, $\sigma^2 = 3{,}920{,}000$; (\$0, \$7,559.80)
**4.21** 5.9938
**4.27** **b.** 40 **c.** 24 **d.** (30.2, 49.8)
**4.29** **a.** $\binom{5}{y}.25^y(.75)^{5-y}$ **b.** .2637 **c.** .6328
**4.31** .1394
**4.33** .049
**4.35** $\binom{4}{y}.5^y(.5)^{4-y}$, binomial
**4.37** **a.** .001 **b.** yes
**4.43** **a.** 12.5, 5, 32.5 **b.** .002 **c.** probabilities suspect
**4.45** **a.** .0001139 **b.** .0355
**4.47** **a.** .0319 **b.** .0337 **c.** 5.2
**4.49** $n(4p_1 + p_2)$
**4.53** **a.** $\binom{y-1}{9}.2^{10}.80^{y-10}$ **b.** 50 **c.** .0047
**4.55** **a.** Geometric: $(.4)(.6)^{y-1}$ **b.** 2.5 **c.** 3.75 **d.** (0, 6.7)
**4.57** **a.** 3.73 **b.** .26788 **c.** .04125
**4.59** **a.** 63 **b.** 62.5 **c.** (0, 188)
**4.61** **a.** .657 **b.** $\mu = 3.33$, $\sigma = 2.79$ **c.** no **d.** .671
**4.63** **c.** $\mu = 1.41$, $\sigma = 1.05$ **d.** .28
**4.65** hypergeometric: $\binom{4}{y}\binom{6}{3-y}/\binom{10}{3}$
**4.67** .2693
**4.69** **a.** .0883 **b.** .1585
**4.71** **a.** .197 **b.** .112 **c.** .038
**4.73** .0144
**4.75** .551
**4.77** **a.** .202 **b.** .323 **c.** $\mu = 1.6$, $\sigma = 1.26$
**4.79** **a.** 2 **b.** no, $P(y > 10) = .0028$
**4.81** **a.** .731 **b.** .03 **c.** 4.24; (9.5, 26.5)
**4.83** **a.** .333 **b.** .1465 **c.** .2519 **d.** .1014
**4.89** $\binom{3}{y}.32^y.68^{3-y}$
**4.91** **a.** $p(1) = .48$, $p(2) = .2496$, $p(3) = .1298$, etc.
**b.** $(.48)(.52)^{y-1}$ **c.** $\mu = 2.08$, $\sigma = 1.50$ **d.** (1, 5.08)
**4.93** **a.** .25, .25, .25, .25 **b.** .0001
**4.95** **a.** .10 **b.** .70
**4.97** **a.** .30 **b.** 1

**4.99** **a.** .08 **b.** yes
**4.101** **a.** $\mu = 1.57, \sigma = 1.25$ **b.** .209
**4.103** **a.** .986 **b.** 0 **c.** 5
**4.105** **a.** .0995 **b.** .0738
**4.107** **a.** .96 **b.** .713 **c.** .00088
**4.109** **a.** 11/5 **b.** 14/25
**4.111** **a.** $e^{\lambda(t-1)}$ **b.**

## Chapter 5

**5.1** **a.** 3/8 **b.** $F(y) = y^{3/8}$ **c.** 1/8 **d.** .0156 **e.** .2969
**5.3** **a.** 1 **b.** $F(y) = \begin{cases} \frac{1}{2} + y + \frac{y^2}{2} & -1 \le y < 0 \\ \frac{1}{2} + y - \frac{y^2}{2} & 0 \le y \le 1 \end{cases}$ **c.** 0.125 **d.** 0.375
**5.5** **a.** 3 **b.** $F(y) = \frac{75y - y^3}{500} + \frac{1}{2}$ **c.** 0.896
**5.7** **b.** $F(y) = 1 - e^{-0.04y}$ **c.** 0.8187
**5.9** **a.** $F(y) = y^2/4, 0 < y < 2$ **b.** NBU
**5.11** **a.** 1/2 **b.** 0.05 **c.** $\approx$ 0.95 **d.** 0.9838
**5.13** **a.** 25 **b.** 625 **c.** $\approx$ 0.95 **d.** 0.9502
**5.17** **a.** 2 **b.** 0.25 **c.** 0.375
**5.19** 113.5
**5.21** $\mu = .5, \sigma = .289, P_{10} = .10, Q_L = .25, Q_U = .75$
**5.23** .4444
**5.25** **b.** $a + (b - a)y$
**5.29** **a.** 0.0329 **b.** 0.4678 **c.** 99.94
**5.31** **a.** .8413 **b.** .7528
**5.33** .448
**5.35** **a.** 0.8185 **b.** 0.9082
**5.37** 0.0548
**5.39** **a.** .5 **b.** $8 **c.** $20
**5.43** No
**5.45** **a.** IQR/s = 1.52, app. normal
**5.47** IQR/s = 1.34
**5.49** nonnormal
**5.51** no
**5.53** **a.** .449 **b.** .865
**5.55** **a.** 0.753403 **b.** 0.6667 **c.** 0.809861, 0.8111
**5.57** 0.693147$\beta$
**5.59** **a.** .3679 **b.** .6065 **c.** .1353 **d.** .0041
**5.61** **a.** $\exp(-t/25{,}000)$ **b.** .7044 **c.** $2\exp(-t/25{,}000) - \exp(-t/12{,}500)$ **d.** .9126
**5.67** 1/16
**5.69** **a.** 3.232, 0.42097 **b.** $\approx$ 0.95 **c.** 0.9631
**5.71** .3935
**5.73** **a.** $(.88623)\sqrt{\beta}$ **b.** $(.2146)\beta$ **c.** $\exp(-C^2/\beta)$
**5.75** 1.75 months
**5.79** 0.31254
**5.83** **a.** $\mu = .0385, \sigma^2 = .00137$ **b.** .778
**5.85** **a.** .834 **b.** .006
**5.87** 168
**5.95** **a.** 0.4207 **b.** 0
**5.97** **a.** $\mu = 7, \sigma = .29$ **b.** .3
**5.99** **a.** .9406 **b.** .0068
**5.101** **a.** 20 **b.** .2231 **c.** .0498; .9502
**5.103** **a.** .5507 **b.** .2636 **c.** $\mu = 60, \sigma^2 = 1{,}800$ **d.** .0916
**5.105** **a.** .321 **b.** .105
**5.107** 109.02
**5.109** **a.** Y less variable than a normal distribution **b.** more
**5.111** 1/6
**5.113** **a.** $\alpha = 9, \beta = 2$ **b.** $\mu = .818, \sigma^2 = .0124$ **c.** .624
**5.115** **a.** .9671 **b.** .2611 **c.** no
**5.117** **a.** 1 **b.** $F(y) = 1 - e^{-y}$ **c.** .9257 **e.** .3611

## Chapter 6

**6.1** **b.** .3, .1, .025,.3, .125, .15 **c.** .1, .55, .35 **d.** $y = 0$: 0, .5, .25, 0, .25, 0; $y = 1$: .364, .091, .545, 0, 0; $y = 2$: .286, 0, 0, 0, .286, .429 **e.** $x = 0$: 0, .667, .333; $x = 1$: .5, .5, 0; $x = 2$: 1, 0, 0; $x = 3$: 0, 1, 0; $x = 4$: .2, 0, .8; $x = 5$: 0, 0, 1
**6.3.** **a.** $p(0, 0) = 6/45, p(0, 1) = 4/45, p(0, 2) = 0, p(1, 0) = 16/45, p(1, 1) = 8/45, p(1, 2) = 1/45, p(2, 0) = 6/45, p(2, 1) = 4/45, p(2, 2) = 0$
**b.** $p_1(0) = 10/45, p_1(1) = 25/45, p_1(2) = 10/45$
**c.** $p_2(0) = 28/45, p_2(1) = 16/45, p_2(0) = 1/45$
**d.** 1/45
**6.5** **a.** $p_2(y \mid x)$
**b.** $p_1(1) = p_1(2) = p_1(3) = 1/3$
**c.** $p(1, 30) = 0.02, p(1, 40) = 0.08, p(1, 50) = 0.08, p(1, 60) = 0.1533, p(2, 30) = 0.0333, p(2, 40) = 0.08, p(2, 50) = 0.12, p(2, 60) = 0.10, p(3,30) = 0.05, p(3, 40) = 0.06, p(3, 50) = 0.10, p(3, 60) = 0.1233$
**6.7** **a.** .11, .25, .40, .24 **b.** .175, .25, .375, .20
**6.9** **a.** $p(1, 1) = 0, p(1, 2) = 1/3, p(1, 3) = 0, p(2, 1) = 1/3, p(2, 2) = 0, p(2, 3) = 0, p(3, 1) = 0, p(3, 2) = 0, p(3, 3) = 1/3$
**6.13** **b.** $f_1(x) = e^{-x}$, exponential **c.** $f_2(y) = 1/40$, uniform
**6.15** **b.** .4624
**6.17** **a.** $-1$ **b.** $f_2(y) = \left(\frac{3}{2} - y\right)$ **c.** $f_1(x \mid y) = (x - y)/\left(\frac{3}{2} - y\right)$
**6.21** **a.** 2 **b.** 49.495
**6.23** **a.** 0 **b.** 30
**6.25** **a.** 5/4 **b.** $-1/12$ **c.** 2 **d.** 2/3
**6.29** no
**6.31** no
**6.33** $p(1, 0) = 0.005, p(1, 12) = 0.01, p(1, 24) = 0.01, p(1, 36) = 0.475, p(2, 0) = 0.001, p(2, 35) = 0.001, p(2, 70) = 0.498$
**6.35** **a.** $f(x, y) = (1/25)\exp\{-(x + y)/5\}$ **b.** 10
**6.39** no
**6.41** .375

**6.43** $-.0854$
**6.45** $-1/5$
**6.47** **a.** 0 **b.** 0
**6.53** $\mu = 7, \sigma^2 = 5.83$
**6.57** $E(\hat{p}) = p, V(\hat{p}) = pq/n$
**6.59** $f(c) = (1/15)\exp\{-(c - 2)/15\}, c \geq 2$
**6.61** $f(w) = (\mu/2)\exp\{-w/(2\mu)\}$; exponential with $\beta = 2\mu$
**6.63** $y = \sqrt{w}$, where $w$ is uniform (0, 1)
**6.65** $E(\ell) = 11, V(\ell) = 54.5$
**6.69** **a.** $f(w) = 1, 0 < w < 1$ **b.** $f(w) = (w + 1)/2, -1 \leq w < 1$ **c.** $f(w) = 2/w^3, w \geq 1$
**6.73** **a.** 0.4, 0.0476 **b.** Approximately normal **c.** 0 **d.** Yes
**6.75** **a.** $\mu_{\bar{y}} = 293, \sigma_{\bar{y}} = 119.8$ **c.** .0158
**6.77** **a.** .3264 **b.** 1.881 **c.** not valid
**6.79** no
**6.81** **a.** 60; 36 **b.** normal **c.** $\approx 0$
**6.83** .0034
**6.87** **a.** no **b.** yes **c.** yes
**6.89** $\approx 0$
**6.91** possibly; $P(Y > 20) = .1762$
**6.93** **a.** 0.92 **b.** 0.0084
**6.95** **a.** .109 **b.** .0025 **c.** .04
**6.97** **a.** Student's $T$ with 9 df **b.** $\chi^2$ with 9 df
**6.99** **a.** Student T distribution with 15 df **b.** 0.999958
**6.105** **a.** $\approx$ normal, $\mu_{\bar{y}} = 43, \sigma_{\bar{y}} = 1.11$
**b.** $\approx$ normal, $\mu_{\bar{y}} = 1{,}050, \sigma_{\bar{y}} = 59.45$
**c.** $\approx$ normal, $\mu_{\bar{y}} = 24, \sigma_{\bar{y}} = 15.5$
**6.107** **a.** $f_1(x) = \left(x + \frac{1}{2}\right); f_2(y) = \left(y + \frac{1}{2}\right)$
**c.** $f_1(x \mid y) = (x + y)/\left(y + \frac{1}{2}\right)$;
$f_2(y \mid x) = (x + y)/\left(x + \frac{1}{2}\right)$
**e.** yes; no
**f.** $E(d) = 5/12, V(d) = 5/144$; $.42 \pm .56$
**6.109** **a.** $\approx 0$ **b.** .0094
**6.111** $f(w) = \{(w + 2)/200 \text{ if } -2 < w < 8, 1/20 \text{ if } 8 < w < 23\}$
**6.113** **a.** $\approx$ normal, $\mu_{\bar{y}} = 121.74, \sigma_{\bar{y}} = 4.86$ **b.** .7348
**6.115** $P(\bar{y} \leq 400.8) = .315$; 2nd operator
**6.117** .008
**6.119** .9332
**6.125** **a.** 113 **b.** no
**6.129** $f(w) = (1/\beta)e^{-w/\beta}$; exponential
**6.133** $y = \sqrt{-\ln(1 - w)}$

## Chapter 7

**7.1** **b.** $\bar{y}$
**7.3** **b.** $pq/n$
**7.9** **a.** $\bar{y}$ **b.** yes
**7.11** **a.** $\bar{y}/2$ **b.** $E(\hat{\beta}) = \beta, V(\hat{\beta}) = \beta^2/(2n)$
**7.13** **a.** $\bar{y}$ **b.** yes **c.** $\beta^2/n$
**7.17** $\bar{y} \pm z_{\alpha/2}\sqrt{\frac{\bar{y}}{n}}$
**7.19** $(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
**7.23** $(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
**7.25** (196.19, 283.81), Assume $\approx$ Normal
**7.27** **a.** (16.529, 19.471) **b.** Yes
**7.29** **a.** 97.17 **b.** (−4.83, 199.16) **c.** distribution of the MTBE levels are $\approx$ normal; no
**7.31** **a.** all lichen specimens in Alaska **b.** (.0053, .0128) **d.** $\approx$ normal
**7.33** **a.** (2.497, 3.932) **c.** 99%
**7.35** **a.** (.83, 1.32) **c.** $\approx$ normal
**7.37** **a.** (0.24281, 0.28135) **b.** No
**7.39** No
**7.41** (0.1205, 0.2395), Yes
**7.43** **a.** $48.3 \pm 36.77$ **b.** yes
**7.45** (−5.10, 3.70), No
**7.47** **a.** $436.5 \pm 47.6$ **b.** $-1.09 \pm .51$
**7.49** **a.** Twin holes at same location not independent **c.** 0.140, 1.264 **d.** (−0.425, 0.715)
**7.51** (−0.0881, 0.4614), No difference
**7.53** **a.** $-10 \pm 10.99$ **b.** $-9 \pm 20.38$ **c.** $-8 \pm 9.77$
**7.55** 95% confidence interval for $(\mu_{\text{meter}} - \mu_{\text{stat}})$: $0.000523 \pm 0.0004$
**7.57** **a.** .60 **b.** $.6 \pm .021$
**7.59** $.5427 \pm .0452$
**7.61** **a.** (0.4715, 0.7172) **b.** No
**7.63** (0.7042, 0.9321), Not accurate
**7.65** **a.** $.644 \pm .099$ **b.** yes
**7.67** **a.** $p_1 - p_2$ **b.** (−0.1351, −0.0032) **c.** Proportions are different
**7.69** **a.** .153 **b.** .215 **c.** $-.061 \pm .069$ **d.** no
**7.71** **a.** (0.0127, 0.2853), Supports Theory 1 **b.** (−0.161, 0.127), Supports Theory 2
**7.73** **a.** 14.0671 **b.** 23.5418 **c.** 23.2093 **d.** 17.5346 **e.** 16.7496
**7.75** (0.0069, 0.0270)
**7.77** (3179, 7618)
**7.79** For $\sigma^2$ (6.3, 18.2)
**7.81** **c.** $\sigma^2 = 8348.0257$
**7.83** **a.** 2.40 **b.** 3.35 **c.** 1.65 **d.** 5.86
**7.85** **a.** (1.462, 2.149) **b.** Yes
**7.87** (.0071, .1806)
**7.89** **a.** .95 **b.** .001 **c.** 97
**7.91** 35
**7.93** 116
**7.95** 450
**7.97** $n_1 = n_2 = 534$
**7.103** $\hat{p}_B = [1/(n + 3)](.80n + 1)$
**7.105** **a.** normal, with $\mu = (n + 1)(\bar{y} + 5/n)/n$
**7.107** 97
**7.109** **a.** (.44, 17.81) **b.** no evidence of a difference
**7.111** **a.** $.23 \pm .017$ **b.** $.20 \pm .016$
**7.113** **a.** (4.73, 9.44) **b.** possibly
**7.115** 1,729
**7.117** (33.64, 392.78)
**7.119** **a.** 1,083 **b.** wider **c.** 38%
**7.121** **a.** $-.35 \pm .09$ **b.** $p_C < p_T$
**7.123** 14,735
**7.125** $.2 \pm .066$
**7.127** **a.** bias $= \frac{1}{2}$ **b.** $1/(12n)$ **c.** $\bar{y} - \frac{1}{2}$
**7.129** **c.** $2y/\chi^2_{\alpha/2} < \beta < 2y/\chi^2_{1-\alpha/2}$

## Chapter 8

**8.1** $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true})$; $\beta = P(\text{Accept } H_0 \mid H_0 \text{ false})$
**8.3** **a.** Type II **b.** Type I
**8.5** **a.** .033 **b.** .617 **c.** .029
**8.11** $H_0$: $\mu = 20$, $H_a$: $\mu > 20$
**8.13** $H_0$: $\mu = 22$, $H_a$: $\mu < 22$,
**8.15** $H_0$: $(\mu_1 - \mu_2) = 0$, $H_a$: $(\mu_1 - \mu_2) > 0$
**8.17** $H_0$: $(\mu_1 - \mu_2) = 0$, $H_a$: $(\mu_1 - \mu_2) \neq 0$
**8.19** **a.** .3124 **b.** .0178 **c.** $\approx 0$ **d.** .1470
**8.21** **a.** fail to reject $H_0$ **b.** fail to reject $H_0$ **c.** reject $H_0$ **d.** fail to reject $H_0$
**8.23** **a.** $H_0$: $\mu = 1$, $H_a$: $\mu \neq 1$ **b.** $\bar{y}$ is sample statistics, need variability
**c.** $t = -47.09$, $p = 0.000$ **d.** $\alpha =$ probability of concluding ratio differs from 1 when ratio equals 1
**e.** Reject $H_0$ **f.** population $\approx$ normal
**8.25** **a.** $H_0$: $\mu = 1.4$, $H_a$: $\mu > 1.4$
**b.** Probability of concluding mean daily amount of distilled water collected is greater than 1.4 when it is equal to 1.4 is 0.10.
**c.** $\bar{y} = 5.243$, $s = 0.192$ **d.** $t = 34.64$ **e.** $p = 0.000$ **f.** Reject $H_0$
**8.27** Yes, reject $H_0$
**8.29** $z = 5.47$, reject $H_0$
**8.31** **a.** $p$-value $= .8396$, do not reject $H_0$
**8.35** **a.** Yes, $z = 1.85$ **b.** Yes, $z = -1.85$ **c.** No, CLT
**8.37** $z = -1.55$, do not reject $H_0$
**8.39** **a.** no, $t = -1.22$ **b.** yes, $t = -4.20$
**8.41** **a.** Yes, $t = 11.87$ **b.** $t = 2.94$
**8.43** $t = 2.83$, reject $H_0$
**8.45** **a.** $t = 2.68$, means are different **b.** $t = 6.34$, means are different **c.** $t = 1.64$, means are not different
**8.47** **a.** $t = .43$, do not reject $H_0$ **b.** Yes
**8.49** **a.** $t = -2.97$, do not reject $H_0$ **b.** $-.4197$; no **c.** $t = .57$, do not reject $H_0$; $-.2274$, no **d.** $t = 3.23$, do not reject $H_0$; .1923, no
**8.51** no, $t = -.713$
**8.53** no, $t = -3.16$, reject $H_0$: $(\mu_1 - \mu_2) = 0$
**8.55** **a.** $H_0$: $p = .10$, $H_a$: $p < .10$ **b.** $z < -2.326$ **c.** $z = -2.11$ **d.** do not reject $H_0$
**8.57** No, $z = 0.69$
**8.59** $z = 1.33$, reject $H_0$
**8.61** Yes, $z = 3.05$
**8.63** $\alpha = 0.01$, $z = 2.67$, reject $H_0$
**8.65** **a.** $z = 0.10$, no difference **b.** $z = 1.18$, no difference
**8.67** $z = 8.34$, proportions are different
**8.69** yes, $z = 11.04$
**8.71** **a.** no, $z = 1.80$ **b.** yes, $z = -4.01$
**8.73** **a.** $\chi^2 = 3{,}031.4$, reject $H_0$
**8.75** $\chi^2 = 10.94$, do not reject $H_0$
**8.77** **a.** $H_0$: $\sigma^2 = .54$, $H_a$: $\sigma^2 > .54$ **b.** .7425 **c.** $\chi^2 = 40.8$, do not reject $H_0$
**8.79** no, $\chi^2 = 6.91$
**8.81** **a.** $F = 17.79$, reject $H_0$ **b.** Yes
**8.83** **a.** $F = 2.26$, no **b.** $p$-value $= 0.096$
**8.85** $F = 2.47$, do not reject $H_0$
**8.87** **a.** no, $F = 1.09$
**8.95** $P(p > .5 \mid x = 29) = .004$, $P(p < .5 \mid x = 29) = .996$; reject $H_0$
**8.97** reject $H_0$ if $P(\mu < \mu_0) > P(\mu > \mu_0)$, using posterior normal distribution with mean $= (n + 1)(\bar{y} + 5/n)/n$ and variance $= 1/(n + 1)$
**8.99** **a.** $H_0$: $\sigma_1^2/\sigma_2^2 = 1$, $H_a$: $\sigma_1^2/\sigma_2^2 \neq 1$ **b.** $F = 1.37$ **c.** $F > 7.39$ **d.** $p$-value $= .726$ **e.** do not reject $H_0$
**8.101** **a.** 0.0654 **b.** $\beta = 0.9413$, power $= 0.0587$ **c.** $\beta = 0.3222$, power $= 0.6778$
**8.103** **a.** $t = -.019$, do not reject $H_0$ **b.** $t = -.019$, do not reject $H_0$
**8.105** **a.** no, $t = -2.20$ **c.** $.1 < \beta < .5$ **d.** $.01 < p$-value $< .025$
**8.109** yes, $z = -2.40$, $p$-value $= .0166$
**8.111** yes, $F = 1.75$, $p$-value $= .0189$
**8.113** **a.** $H_0$: $\mu = 10$, $H_a$: $\mu < 10$ **c.** $z = -2.33$, reject $H_0$
**8.115** **a.** $H_0$: $(\mu_1 - \mu_2) = 0$, $H_a$: $(\mu_1 - \mu_2) > 0$ **b.** $z > 1.645$ **c.** reject $H_0$

## Chapter 9

**9.1** **a.** jaw habits; grinding, clenching, both, and neither **c.** $.50 \pm .127$ **d.** $.23 \pm .214$
**9.3** **a.** $.44 \pm .031$ **b.** $-.23 \pm .04$
**9.5** **a.** $.175 \pm .028$ **b.** $-.262 \pm .054$
**9.7** **a.** $.678 \pm .039$ **b.** $.356 \pm .078$
**9.11** yes; $\chi^2 = 963.4$, $p$-value $= 0$
**9.13** $\chi^2 = 2.39$, do not reject $H_0$
**9.15** yes; $\chi^2 = 8.04$, $p$-value $= .045$
**9.17** yes; $\chi^2 = 3.61$, $p$-value $= .307$, do not reject $H_0$
**9.19** $\chi^2 = 4.84$, $p$-value $= .089$, do not reject $H_0$
**9.21** **a.** $H_0$: Nappe and FIA are independent, $H_a$: Nappe and FIA are dependent **b.** yes **c.** $\chi^2 > 5.99147$ **d.** do not reject $H_0$
**9.23** yes; $\chi^2 = 37.53$, $p$-value $= 0$
**9.25** **a.** Below/Private = 81, Below/Public = 72, Detect/Private = 22, Detect/Public = 48 **b.** $\chi^2 = 8.84$, $p$-value $= .0028$, reject $H_0$
**c.** Below/Bedrock = 138, Below/Uncon = 15, Detect/Bedrock = 63, Detect/Uncon = 7 **d.** $\chi^2 \approx 0$, $p$-value $= .9637$, do not reject $H_0$
**9.27** **a.** no **b.** no **c.** yes **d.** $\chi^2 = 1.03$, do not reject $H_0$
**9.29** **a.** expected cell count for true/yes is less than 5; chi-square test of independence not valid
**9.31** yes; $\chi^2 = 64.24$, $p$-value $= 0$
**9.33** **a.** 10 teeth bonded for each adhesive type
**b.** $\chi^2 = 5.03$, $p$-value $= .17$, do not reject $H_0$ **c.** no
**9.35** $\chi^2 = 31.87$, reject $H_0$
**9.37** **a.** expected cell counts are less than 5
**b.** $p$-value $= .2616$, do not reject $H_0$
**9.39** $p$-value $= 0$, reject $H_0$
**9.41** yes, $\chi^2 = 508.74$
**9.43** $\chi^2 = .32$, do not reject $H_0$
**9.45** $\chi^2 = 4.39$, do not reject $H_0$
**9.47** no, $\chi^2 = 4.4$
**9.49** **a.** yes, $\chi^2 = 14.67$ **b.** $-.169 \pm .161$
**9.51** **a.** yes, $\chi^2 = 313.15$ **b.** $.181 \pm .069$
**9.53** **a.** $.275 \pm .182$ **b.** $.125 \pm .261$ **c.** $\chi^2 = 2.6$, do not reject $H_0$
**9.55** yes, $\chi^2 = 39.77$

## Chapter 10

**10.1** $\beta_0 = 1, \beta_1 = 1; y = 1 + x$
**10.3** **a.** $y$-intercept = 3, slope = 2 **b.** $y$-intercept = 1, slope = 1 **c.** $y$-intercept = −2, slope = 3 **d.** $y$-intercept = 0, slope = 5
**e.** $y$-intercept = 4, slope. = −2
**10.5** **a.** $y = \beta_0 + \beta_1 x + \varepsilon$; negative **b.** yes **c.** no
**10.7** **a.** $y = \beta_0 + \beta_1 x + \varepsilon$ **b.** positive **c.** $\hat{\beta}_0 = 1469$, $\hat{\beta}_1 = 210.77$
**10.9** **a.** $\hat{y} = 6.313 + 0.9665x$ **d.** 15.98%
**10.11** **a.** $y = \beta_0 + \beta_1 x + \varepsilon$; **b.** $\hat{y} = -.607 + 1.062x$ **c.** positive **e.** $\hat{y} = -.148 + 1.022x$; positive
**10.13** decrease 102 units
**10.15** **a.** $\hat{y} = -.146 + 1.553x$ **b.** increase 1.553 units
**10.25** **b.** $\hat{y} = 1.265 + 0.589x$ **c.** SSE $= 4.695$, $s^2 = 0.204$ **d.** .452
**10.27** **a.** $\hat{y} = -632 + 212.1x$ **b.** 11,283; 106.2 **c.** estimate of $\sigma$
**10.31** yes, $t = 14.87$, reject $H_0$; (.0041, .0055)
**10.33** **a.** $-.114 \pm .018$ **b.** $t = -11.05$, reject $H_0$
**10.35** $-.0023 \pm .0016$
**10.37** **a.** positive **b.** $\hat{y} = -.30 + .1845x$ **c.** yes, $t = 3.77$, $p$-value $= .0005$
**10.39** **a.** possibly **b.** yes **c.** $\hat{y} = -11.03 + 1.627x$ **e.** yes, $t = 17.99$ **f.** $1.627 \pm 0.182$
**10.43** **b.** both positive **c.** .706
**10.45** **a.** $t = 17.75$, reject $H_0$ **b.** no
**10.47** **a.** $y = \beta_0 + \beta_1 x + \varepsilon$
**10.49** **c.** reject $H_0$ at $\alpha = .01$
**10.53** **a.** $t = 32.8$, reject $H_0$, $r^2 = .901$ **b.** (41.86, 77.86) **c.** narrower
**10.55** **a.** $15.98 \pm 9.66$ **b.** $15.98 \pm 3.66$
**10.57** $2.92 \pm 2.55$
**10.59** $4.95 \pm 0.16$

**10.61** **a.** $\hat{y} = 6.62 - 0.073x$ **b.** $\hat{y} = 9.31 - 0.108x$ **c.** Brand A: (2.76, 3.94); Brand B: (4.17, 4.76) **d.** Brand A: (1.12, 5.57); Brand B: (3.35, 5.58) **e.** (−4.25, 2.96)
**10.65** **a.** misspecified model **b.** unequal error variances **c.** unequal error variances **d.** nonnormal errors
**10.67** **b.** yes, curvilinear **c.** mean error of 0 **d.** add curvature to the model
**10.69** no
**10.71** not valid; nonnormal errors, missspecified model
**10.73** **a.** yes **b.** .612 **c.** $t = 4.89$, reject $H_0$ **d.** $r = .309$; $t = 1.81$, do not reject $H_0$ **e.** all data: $r = -.880$, $t = -11.72$, reject $H_0$; omit duck chow: $r = -.646$, $t = -4.71$, reject $H_0$
**10.75** model statistically useful: $t = 7.43$, $r^2 = .81$; assumptions reasonably satisfied
**10.77** **b.** $\hat{y} = 308.14 + 41.7x$ **c.** $t = 3.02$, do not reject $H_0$ **d.** $\hat{y} = 302.59 + 64.1x$; $t = 4.79$, reject $H_0$
**10.79** **a.** yes, $t = -3.79$ **b.** no, possible heteroscedastic and non-normal errors
**10.81** **a.** yes **b.** $\hat{\beta}_0 = 1.192$, $\hat{\beta}_1 = .987$ **d.** $t = 6.91$, reject $H_0$
**10.83** **a.** $\hat{y} = -.1124 + .0944x$ **b.** yes; $t = 11.39$ **c.** $.926 \pm .197$
**10.85** $\hat{y} = 2.55 + 2.76x$

## Chapter 11

**11.1** **b.** $X'X = \begin{bmatrix} 1 & 62 \\ 62 & 720.52 \end{bmatrix}$; $X'Y = \begin{bmatrix} 97.8 \\ 1087.78 \end{bmatrix}$
**d.** $\hat{\beta} = \begin{bmatrix} 6.3126 \\ 108.9665 \end{bmatrix}$ **e.** SSE = 41.1

**11.3** **a.** $Y = \begin{bmatrix} 18.3 \\ 11.6 \\ 32.2 \\ 30.9 \\ 12.5 \\ 9.1 \\ 11.8 \\ 11.0 \\ 19.7 \\ 12.0 \end{bmatrix}$ $X = \begin{bmatrix} 1 & 2.48 \\ 1 & 2.48 \\ 1 & 2.39 \\ 1 & 2.44 \\ 1 & 2.50 \\ 1 & 2.58 \\ 1 & 2.59 \\ 1 & 2.59 \\ 1 & 2.51 \\ 1 & 2.49 \end{bmatrix}$ **b.** $X'X = \begin{bmatrix} 10 & 25.05 \\ 25.05 & 62.7893 \end{bmatrix}$; $X'Y = \begin{bmatrix} 169.1 \\ 419.613 \end{bmatrix}$ **c.** $\hat{\beta} = \begin{bmatrix} 272.3815 \\ -101.9846 \end{bmatrix}$
**d.** SSE = 226.8552, $s^2 = 28.3569$ **e.** $t = -3.78$, $p = 0.0027$, reject $H_0$ **f.** $R^2 = 0.6416$ **g.** (4.5371, 30.3028)
**11.5** **a.** $\hat{\beta}_0 = 21.1424$, $\hat{\beta}_1 = -0.6067$ **b.** $F = 8.16$, reject $H_0$ **c.** (15.00, 16.95)
**11.7** **a.** yes; $F = 17.8$ **b.** $t = -3.50$, reject $H_0$ **c.** $-6.38 \pm 4.72$
**11.9** **a.** 93,002 **b.** 98,774
**11.19** **a.** $F = 4.38$, reject $H_0$ **b.** $R_a^2 = 0.629$ **c.** $s = 11.2206$ **d.** (−0.2181, 1.0841) **e.** $t = -0.74$, do not reject $H_0$
**11.21** **b.** $F = 3.72$, reject $H_0$ **c.** $t = 2.52$, do not reject $H_0$
**11.23** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ **b.** $\hat{y} = 86.9 - 02099x_1 + 0.1515x_2 + 0.0733x_3$ **c.** $F = 2.66$, no **d.** $R_a^2 = 0.0379$, $2s = 5.9309$ **e.** (82.6017, 95.5656)
**11.25** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
**b.** $\hat{y} = 13{,}614.4 + 0.089x_1 - 9.201x_2 + 14.394x_3 + 0.352x_4 - 0.848x_5$ **c.** 458.83
**11.27** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$
**b.** $\hat{y} = .998 - .022x_1 + .156x_2 - .017x_3 - .0095x_4 + .421x_5 + .417x_6 - .155x_7$
**d.** $F = 5.29$, reject $H_0$; $R_a^2 = .625$, $s = 0.437$ **e.** (−1.233, 1.038)
**11.29** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ **b.** $\hat{y} = -63{,}238 + 18.8x_1 + 445{,}486x_2 - 139.8x_1x_2$
**c.** $F = 110.44$, reject $H_0$; $R_9^2 = 0.9376$, $2s = 48{,}720.6$ **d.** $t = -2.47$, reject $H_0$ **e.** decrease by 51.1
**11.31** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ **b.** interaction important
**11.33** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2 x_5 + \beta_7 x_3 x_5$
**b.** $\hat{y} = 13{,}645.9 + 0.046x_1 - 12.68x_2 + 23.003x_3 - 3.023x_4 + 1.288x_5 + 0.016x_2x_5 - 0.041x_3x_5$
**c.** $t = 4.40$, reject $H_0$ **d.** $t = -3.77$, reject $H_0$
**11.35** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ **b.** $\hat{y} = 1.0077 - 0.00718x_1 + 0.51715x_2 - 0.00599x_1x_2$
**c.** $t = -1.79$, reject $H_0$ **d.** −0.03114
**11.37** **a.** $\hat{y} = 6.266 + 0.0079145x - 0.000004x^2$ **b.** $F = 210.56$, reject $H_0$ **c.** $R_a^2 = 0.972$
**d.** $H_0$: $\beta_2 = 0$, $H_a$: $\beta_2 < 0$ **e.** $t = -3.23$, reject $H_0$ **f.** (7.5947, 8.3624)
**11.39** **a.** downward curvature **b.** 6.25 **c.** 10.25 **d.** 200
**11.41** $H_0$: $\beta_2 = 0$, $H_a$: $\beta_2 > 0$, $t = 0.08$, do not reject $H_0$

**11.43** **a.** $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ **b.** $\hat{y} = 0.334 - 0.810x + 0.941x^2$ **c.** $F = 62.17$, reject $H_0$
**d.** $t = 8.36$, reject $H_0$: $\beta_2 = 0$ **e.** $R_a^2 = 0.196$, $s = 0.088$
**11.45** **a.** curvilinear trend **b.** $\hat{y} = 1.007 - 1.167x + 0.290x^2$ **c.** yes, $t = 7.36$
**11.47** **a.** $\hat{y} = 85.014 + 0.04045x$ **b.** yes, $F = 21.77$ **c.** no; unusual value for $\hat{y} = 777.0$ **d.** yes
**e.** model is useful ($p$-value $= .0001$); Boston Harbor residual has $z = -2.78$
**11.49** yes; influential observations are #11, #32, #36, and #47
**11.51** May not be normal
**11.53** **a.** yes; assumption of equal variances violated **b.** use transformation $y^* = \sqrt{y}$
**11.55** No
**11.57** $x_1$ and $x_2$ could be correlated
**11.59** unable to test model adequacy; df(Error) $= 0$
**11.61** **a.** $\hat{y} = 2.743 + .801x_1$; yes, $t = 15.92$ **b.** $\hat{y} = 1.665 + 12.395x_2$; yes, $t = 11.76$
**c.** $\hat{y} = -11.795 + 25.068x_3$; yes, $t = 2.51$
**11.63** **b.** $\boldsymbol{X'X} = \begin{bmatrix} 12 & 11{,}280 & 8.12 \\ 11{,}280 & 11{,}043{,}750 & 7{,}632.8 \\ 8.12 & 7{,}632.8 & 6.762 \end{bmatrix}$; $\boldsymbol{X'Y} = \begin{bmatrix} 8.019 \\ 9{,}131.205 \\ 6.627 \end{bmatrix}$ **c.** $= \begin{bmatrix} 2.45026 & -.00213 & -.53387 \\ -.0023 & .00000227 & -4.88 \times 10^{-18} \\ -.53387 & -4.88 \times 10^{-18} & .78897 \end{bmatrix}$
**d.** $\hat{\beta} = \begin{bmatrix} -3.3727 \\ .00362 \\ .94760 \end{bmatrix}$; $\hat{y} = -3.3727 + .00362x_1 + .94760x_2$ **f.** .784 **g.** $F = 20.90$, reject $H_0$
**h.** $.0036 \pm .0011$ **i.** $.948 \pm .654$ **j.** $(-.126, 1.427)$
**11.65** **a.** $\hat{y} = 10.625 + 2.4125x_1 + .2325x_2 - .04225x_1x_2$ **b.** $F = 74.57$, reject $H_0$ **c.** yes, $t = -6.91$
**11.67** **a.** $\hat{y} = 0.132 - 9.307x_1 + 1.558x_2$ **b.** $F = 35.84$, reject $H_0$ ($p$-value $= .0005$) **c.** no; $t = -1.84$ **d.** yes; $t = 8.47$
**e.** .923 **f.** .152 **g.** $(-0.296, 0.564)$
**11.69** **a.** $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ **b.** plot supports theory **c.** $\hat{y} = 438.31 - 1684.27x + 2502.28x^2$ **d.** $t = 5.32$, reject $H_0$
**11.71** yes, $t = 4.20$, $p$-value $= .004$
**11.73** assumptions reasonably satisfied
**11.75** **a.** possible curvilinear **b.** $\hat{y} = 42.25 - .0114x + .000000608x^2$ **c.** no, $t = 1.66$
**11.77** assumptions reasonably satisfied; one outlier detected
**11.79** **a.** $E(S_v) = \beta_0 + \beta_1 x + \beta_2 x^2$ **b.** $E(V_v) = \beta_0 + \beta_1 x + \beta_2 x^2$ **c.** variance-stabilizing transformation
**d.** $\hat{y} = 96.55 + .00823x - .00000532x^2$; 97.67 **e.** $\log(\hat{y}) = 6.47 - .002373x$ **f.** yes, $t = -5.36$ **g.** (138.38, 1469.97)

## Chapter 12

**12.1** **a.** nitrate concentration **b.** water source; qualitative
**12.3** **a.** qualitative **b.** qualitative **c.** quantitative **d.** quantitative
**12.5** **a.** quantitative **b.** quantitative **c.** qualitative **d.** qualitative **e.** qualitative **f.** quantitative **g.** qualitative
**12.7** **a.** quantitative **b.** quantitative **c.** qualitative
**12.9** **a.** 1st-order **b.** 2nd-order **c.** 3rd-order **d.** 2nd-order **e.** 1st-order
**12.11** **a.** $E(y) = \beta_0 + \beta_1 x_2 + \beta_2 (x_2)^2$ **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1)^2 + \beta_3 (x_1)^3$
**12.13** **a.** 73% of sample variation in lane utilization is explained by the model
**b.** $F = 2699.6$, reject $H_0$ **d.** $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$
**12.15** **b.** $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$
**12.17** **a.** $\hat{y} = .0670 + .3158x$ **b.** experiment 4 **c.** curvilinear **d.** $t = 4.99$, no evidence of curvature
**12.21** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ **b.** increase in swimming speed for every one body length per second increase in body wave speed, holding both tail amplitude deviation and tail velocity deviation constant. **c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$
**d.** $\beta_3$ **e.** $\beta_2 + \beta_4$
**12.23** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ **b.** $\beta_3$ **c.** add six 2-variable interaction terms **d.** $\beta_3 + 50\beta_6 + 30\beta_8 + 2\beta_{10}$
**12.25** **a.** both quantitative **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ **c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
**d.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 (x_1)^2 + \beta_5 (x_2)^2$
**12.27** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 (x_1)^2 + \beta_5 (x_2)^2$ **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
**c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
**12.29** **a.** $r = .824$ **b.** $u = (x - 83.36)/24.05$ **c.** $r = .119$ **d.** $E(\hat{y}) = .0489 + .00827u + .00674u^2$
**12.31** **a.** $r = .974$ **b.** $u = (x - 15.10)/8.14$ **c.** $r = -.046$ **d.** $E(\hat{y}) = .0983 - .1641u + .1108u^2$
**12.33** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $x_1 = \{1$ if groundwater, 0 if not$\}$, $x_2 = \{1$ if sub-surface, 0 if not$\}$
**b.** $\beta_0 = \mu_{\text{over}}$; $\beta_1 = \mu_{\text{ground}} - \mu_{\text{over}}$; $\beta_2 = \mu_{\text{sub}} - \mu_{\text{over}}$
**12.35** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ $x_1 = \{1$ if Benzene, 0 if not$\}$, $x_2 = \{1$ if Toluene, 0 if not$\}$, $x_3 = \{1$ if Chloroform, 0 if not$\}$, $x_4 = \{1$ if Methanol, 0 if not$\}$ **b.** $\beta_0 = \mu_A$; $\beta_1 = \mu_B - \mu_A$; $\beta_2 = \mu_T - \mu_A$; $\beta_3 = \mu_C - \mu_A$; $\beta_4 = \mu_M - \mu_A$

**12.37** **a.** $\beta_0$ **b.** $\mu_{\text{Set}} - \mu_{\text{Gill}}$ **c.** $H_0$: $\beta_1 = \beta_2 = 0$
**12.39** **a.** Group **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $x_1 = \{1 \text{ if group 2, 0 if not}\}$, $x_2 = \{1 \text{ if group 3, 0 if not}\}$
**c.** $\beta_0 = \mu_1$; $\beta_1 = \mu_2 - \mu_1$; $\beta_2 = \mu_3 - \mu_1$
**12.41** **a.** $E(y) = \beta_0 + \beta_1 x$, $x = \{1 \text{ if flightless, 0 if not}\}$ **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $x_1 = \{1 \text{ if vertebrates, 0 if not}\}$, $x_2 = \{1 \text{ if vegetables, 0 if not}\}$, $x_3 = \{1 \text{ if invertebrates, 0 if not}\}$ **c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $x_1 = \{1 \text{ if within ground, 0 if not}\}$, $x_2 = \{1 \text{ if trees, 0 if not}\}$, $x_3 = \{1 \text{ if above ground, 0 if not}\}$ **d.** $\hat{y} = 641 + 30{,}647x$
**e.** $t = 5.75$, reject $H_0$ **f.** $\hat{y} = 903 + 2{,}997x_1 + 26{,}206x_2 - 660x_3$ **g.** $F = 8.43$, reject $H_0$
**h.** $\hat{y} = 73.732 - 9.132x_1 - 45.01x_2 - 39.51x_3$ **i.** $F = 8.07$, reject $H_0$
**12.43** **a.** $x_1 = \{1 \text{ large/public, 0 if not}\}$, $x_2 = \{1 \text{ if large/private, 0 if not}\}$, $x_3 = \{1 \text{ if small/public, 0 if not}\}$
**b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ **c.** evidence of differences in mean likelihoods for the 4 size/type categories
**d.** $x_1 = \{1 \text{ large, 0 if small}\}$, $x_2 = \{1 \text{ if public, 0 if private}\}$ **e.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
**f.** large/public: $\beta_0 + \beta_1 + \beta_2$; large/private: $\beta_0 + \beta_1$; small/public: $\beta_0 + \beta_2$; small/private: $\beta_0$
**g.** $\mu_{\text{large}} - \mu_{\text{small}} = \beta_1$ holding type fixed **h.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
**i.** large/public: $\beta_0 + \beta_1 + \beta_2 + \beta_3$; large/private: $\beta_0 + \beta_1$; small/public: $\beta_0 + \beta_2$; small/private: $\beta_0$
**12.45** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$, $x_3 - x_6$ are dummy variables for compound
**c.** $E(y) = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_6 + \beta_6 x_2 x_3 + \beta_7 x_2 x_4 + \beta_8 x_2 x_5 + \beta_9 x_2 x_6$
**d.** $\beta_2$; $(\beta_2 + \beta_6)$; $(\beta_2 + \beta_7)$; $(\beta_2 + \beta_8)$; $(\beta_2 + \beta_9)$
**12.47** **a.** $E(y) = \beta_0 + \beta_1 x_3 + \beta_2 x_7$ **b.** $\mu_{\text{Timberjack}} - \mu_{\text{Valmet}}$ holding dominant hand power level fixed
**c.** $E(y) = \beta_0 + \beta_1 x_3 + \beta_2 x_7 + \beta_3 x_3 x_7$ **d.** $\beta_1$ **e.** $\beta_2 + 75\beta_3$
**f.** $E(y) = \beta_0 + \beta_1 x_3 + \beta_2 x_7 + \beta_3 x_3 x_7 + \beta_4 x_3^2 + \beta_5 x_3^2 x_7$ **g.** $\beta_4$
**12.49** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $x_2 = \{1 \text{ if method } G\text{, 0 if not}\}$, $x_3 = \{1 \text{ if method } R_1\text{, 0 if not}\}$
**c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$ **d.** G: $\beta_1 + \beta_4$; $R_1$: $\beta_1 + \beta_5$; $R_2$: $\beta_1$
**12.51** $t = 3.27$, $p$-value $= .003$, evidence of interaction; without: 6.719; with: 9.757
**12.53** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$
**c.** TDS: $\beta_1 + \beta_4$; FE: $\beta_1 + \beta_5$; AL: $\beta_1$
**12.55** **a.** $H_0$: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ **b.** $E(y) = \beta_0 + \beta_1 x_1$ **c.** difference in mean lengths for 3 gear types **d.** $H_0$: $\beta_4 = \beta_5 = 0$
**e.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ **f.** no evidence of interaction
**12.57** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{11} x_{11}$
**b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{11} x_{11} + \beta_{12} x_1 x_9 + \beta_{13} x_1 x_{10} +$
$\beta_{14} x_1 x_{11} + \beta_{15} x_2 x_9 + \cdots + \beta_{18} x_3 x_9 + \cdots + \beta_{21} x_4 x_9 + \cdots + \beta_{33} x_8 x_9 + \beta_{34} x_8 x_{10} + \beta_{35} x_8 x_{11}$
**c.** $H_0$: $\beta_{12} = \beta_{13} = \cdots = \beta_{35} = 0$
**12.59** $F = 24.19$; complete 2nd-order more useful
**12.61** **a.** $H_0$: $\beta_4 = \beta_5 = 0$ **b.** $H_0$: $\beta_3 = \beta_4 = \beta_5 = 0$ **c.** no, $F = .93$
**12.63** **a.** $H_0$: $\beta_2 = \beta_3 = 0$ **b.** $F = 6.99$, reject $H_0$ **c.** test $H_0$: $\beta_4 = \beta_5 = 0$ in the model
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$
**d.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1)^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 (x_1)^2 x_2 + \beta_8 (x_1)^2 x_3$
**e.** test $H_0$: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
**12.65** **a.** 6; $E(y) = \beta_0 + \beta_1 x_j$ **b.** $x_2$; $t = -90$ is largest in absolute value **c.** 5; $E(y) = \beta_0 + \beta_1 x_2 + \beta_2 x_j$
**e.** inflated P(at least 1 Type I error); no higher-order or interaction terms
**12.67** **a.** 11 **b.** 10 **c.** model statistically useful **d.** inflated P(at least 1 Type I error); no higher-order or interaction terms
**e.** $E(y) = \beta_0 + \beta_1 x_6 + \beta_2 x_{11} + \beta_3 x_6 x_{11} + \beta_4 (x_6)^2 + \beta_5 (x_{11})^2$ **f.** test $H_0$: $\beta_4 = \beta_5 = 0$
**12.69** **a.** 11 **b.** 10 **c.** 1 **d.** $E(y) = \beta_0 + \beta_1 x_{11} + \beta_2 x_4 + \beta_3 x_2 + \beta_4 x_7 + \beta_5 x_{10} + \beta_6 x_1 + \beta_7 x_9 + \beta_8 x_3$
**12.71** **a.** 8 **b.** largest $|t|$ value **c.** 7 **e.** inflated P(at least 1 Type I error); no higher-order or interaction terms
**12.73** $E(y) = \beta_0 + \beta_1 x$
**12.75** $t = -6.60$, reject $H_0$: $\beta_2 = 0$ in favor of $H_a$: $\beta_2 < 0$
**12.77** no estimate of $\sigma^2$
**12.79** **a.** $-.782 + .0399x_1 - .021x_2 - .0033x_1 x_2$ **b.** $-.782 + .0399(1) - .021(10) - .0033(1)(10) = -.9851$
**12.81** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1)^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 (x_1)^2 x_2$ **b.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_3 x_2$
**12.83** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $x_2 = \{1 \text{ if program } B\text{, 0 if not}\}$, $x_3 = \{1 \text{ if program } C\text{, 0 if not}\}$ **b.** $H_0$: $\beta_2 = \beta_3 = 0$
**c.** $F = 2.60$, do not reject $H_0$
**12.85** **a.** $u = (x - 4.5)/2.45$ **b.** $-1.429, -1.021, -.612, -.204, 204, .612, 1.021, 1.429$ **c.** .976
**d.** 0 **e.** $\hat{y} = -0.656 + 105.07u + 90.61u^2$; $F = 26.66$, model useful
**12.87** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 (x_1)^2 + \beta_4 (x_2)^2$
**12.89** $\hat{y} = 2.095 + 1.643x_1 + .029x_2 + .0212x_1 x_2 - .00000595(x_1)^2 - .00000469(x_2)^2$

## Chapter 13

**13.1** **a.** noise (variability) and volume (sample size) **b.** remove extraneous source of variation
**13.3** **a.** pipe location **b.** randomized block; treatments: instant-off & instant-on; blocks: 19 pipe locations **c.** accuracy
**d.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{19} x_{19}$, where $x_1 = \{1$ if instant-off, 0 if instant-on$\}$, $x_2 - x_{19}$ = dummy variables for blocks
**13.5** **a.** cockatiel **b.** yes **c.** experimental group **d.** 1, 2, 3 **e.** 3 **f.** total consumption
**g.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $x_1 = \{1$ if group 1, 0 if not$\}$, $x_2 = \{1$ if group 2, 0 if not$\}$
**13.7** **a.** account for month-to-month variation **b.** California, Utah, and Alaska **c.** Nov. 2000, Oct. 2001, and Nov. 2001
**13.9** **a.** $y_{B,1} = \beta_0 + \beta_2 + \beta_4 + \varepsilon_{B,1}$; $y_{B,2} = \beta_0 + \beta_2 + \beta_5 + \varepsilon_{B,2}$; . . . ;
$y_{B,10} = \beta_0 + \beta_2 + \varepsilon_{B,10}$; $\bar{y}_B = \beta_0 + \beta_2 + (\beta_4 + \beta_5 + \cdots + \beta_{12})/10 + \bar{\varepsilon}_B$
**b.** $y_{D,1} = \beta_0 + \beta_4 + \varepsilon_{D,1}$; $y_{D,2} = \beta_0 + \beta_5 + \varepsilon_{D,2}$; . . . ; $y_{D,10} = \beta_0 + \varepsilon_{D,10}$; $\bar{y}_D = \beta_0 + (\beta_4 + \beta_5 + \cdots + \beta_{12})/10 + \bar{\varepsilon}_D$
**13.11** **a.** factorial design **b.** Factor 1: Level of coagulant (5, 10, 20, 50, 100, and 200 mg/liter); Factor 2: pH level (4.0, 5.0, 6.0, 7.0, 8.0 and 9.0); treatments: (5/4.0), (5/5.0), (5/6.0), . . ., (200/9.0)
**13.13** **a.** quality **b.** temperature (QN), pressure (QL) **c.** (1100/500), (1100/550), (1100/600), . . ., (1200/600) **d.** steel ingots
**13.15** **a.** no
**13.17** df(Error) $> 0$
**13.19** 18
**13.25** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \cdots$
$+ \beta_9 x_1 x_5 + \beta_{10} x_2 x_3 + \beta_{11} x_2 x_4 + \beta_{12} x_2 x_5 + \beta_{13} x_1 x_2 x_3 + \beta_{14} x_1 x_2 x_4 + \beta_{15} x_1 x_2 x_5$; 0 df
**13.27** **a.** flextime, staggered hours, fixed hours **b.** collect independent random samples of workers from each work schedule
**c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $x_1 = \{1$ if flextime, 0 if not$\}$, $x_2 = \{1$ if staggered hours, 0 if not$\}$
**13.29** **a.** $3 \times 3$ factorial **b.** pay rate (QN), workday length (QN)
**13.31** **a.** Sex and Weight **b.** both qualitative **c.** 4: (ML), (MH), (FL), and (FH)

## Chapter 14

**14.1** **a.** boxes in each size are different **b.** yes **c.** possibly not, due to large variations
**14.3** **a.** extracted teeth; bonding times (1, 24, or 48 hours); breaking strength **b.** $H_0$: $\mu_1 = \mu_{24} = \mu_{48}$ **c.** $F > 5.49$
**d.** reject $H_0$ **e.** breaking strengths normally distributed for each treatment, with equal variances
**14.5** **a.** $H_0$: $\mu_{\text{Set}} = \mu_{\text{pot}} = \mu_{\text{gill}}$ **b.** evidence of differences in mean body lengths
**14.7** **a.** MS(Exposure) = .003333, MSE = .000207, $F$ = 16.1 **b.** yes
**14.9** **a.** $E(y) = \beta_0 + \beta_1 x$, $x = \{1$ if current alloy, 0 if new RAA alloy$\}$ **b.** $\hat{y} = 641 - 48x$; df$(T)$ = 1, df$(E)$ = 4, SST = 3456, SSE = 1040, MST = 3456, MSE = 260, $F$ = 13.29 **c.** 3,456 **d.** 260 **e.** 1 **f.** 4 **g.** $F$ = 13.29 **h.** $F > 6.61$
**i.** reject $H_0$ **j.** $t = -3.65$, reject $H_0$ **l.** two-tailed
**14.11** yes, $F$ = 7.25, $p$-value = .0008
**14.13** **b.** treatments: scopolamine, glycopyrrolate, and no drug; response: number of pairs recalled **c.** 6.167, 9.375, 10.625; no, no measure of reliability **d.** yes, $F$ = 27.07, $p$-value $\approx$ 0
**14.15** **a.** CM = 1,691,387.13; SST = 357,986.87 **b.** 1,151,602 **c.** 1,509,588.9 **d.** MST = 89,496.72, MSE = 15,775.37, $F$ = 5.67 **e.** yes
**14.17** **a.** randomized block **b.** df(Method) = 2, df(Error) = 6, SS(Method) = .39, SS(Month) = 32.34, F(Month) = 156.23
**c.** $F$ = 2.83, $p$-value = .08, fail to reject $H_0$: $\mu_{\text{ANN}} = \mu_{\text{TSR}} = \mu_{\text{Actual}}$
**14.19** $F$ = 57.99, reject $H_0$
**14.21** **a.** dependent: skin factor; treatments: 4 products; blocks: 10 wells

**c.**

| Source | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Product | 3 | 911 | 304 | 6.45 | 0.002 |
| Well | 9 | 340469 | 37830 | 803.79 | 0.000 |
| Error | 27 | 1271 | 47 | | |
| Total | 39 | 342651 | | | |

**d.** yes, $F$ = 6.45, $p$-value = .002

**14.23** Reject $H_0$: $\mu_{\text{Standard}} = \mu_{\text{Supervent}} = \mu_{\text{Ecopack}}$, $F$ = 7.90, $p$-value = .064
**14.25** **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_{104} x_{104}$, $x_1 = \{1$ if full-dark, 0 if not$\}$, $x_2 = \{1$ if transient light, 0 if not$\}$, $x_3$, $x_4$, . . ., $x_{104}$ are dummy variables for genes (blocks) **b.** $H_0$: $\beta_1 = \beta_2 = 0$ **c.** $F$ = 5.33, $p$-value = .0056, reject $H_0$
**14.27** **a.** $2 \times 2$ factorial **b.** Age (younger, older) and Diet (fine, coarse) **c.** hen **d.** shell thickness
**e.** effect of diet on shell thickness is not dependent on age
**f.** no significant difference between $\mu_{\text{older}}$ and $\mu_{\text{younger}}$
**g.** significant difference between $\mu_{\text{fine}}$ and $\mu_{\text{coarse}}$
**14.29** **c.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$, $x_1 = \{1$ if Baker's, 0 if Brewer's$\}$, $x_2 = \{1$ if 45°, 0 if not$\}$, $x_3 = \{1$ if 48°, 0 if not$\}$, $x_4 = \{1$ if 51°, 0 if not$\}$ **d.** $H_0$: $\beta_5 = \beta_6 = \beta_7 = 0$; partial F-test on interaction terms **e.** reject $H_0$ **f.** yeast: conduct t-test $H_0$: $\beta_1 = 0$; temperature: conduct partial F-test $H_0$: $\beta_2 = \beta_3 = \beta_4 = 0$
**g.** interaction present

**14.31** **a.** 5 × 3 factorial; factors: cutting tool (5 levels) and speed (3 levels); treatments: 15 combinations of cutting tool and speed; experimental unit: run; dependent variable : feed force
**b.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_1x_5 + \cdots + \beta_{14}x_4x_6$, where $x_1 - x_4$ are dummy variables for cutting tool, $x_5 - x_6$ are dummy variables for speed **c.** $H_0$: $\beta_7 = \beta_8 = \cdots = \beta_{14} = 0$ **d.** $F = 21.96$, reject $H_0$ **f.** no
**14.33** Trap *X* Color interaction: $F = .26$, $p$-value $= .618$; Trap main effect: $F = 2.30$, $p$-value $= .143$; Color main effect: $F = 54.86$, $p$-value $= 0$
**14.35** Effect of mowing frequency on vegetation height depends on mowing height ($F = 10.18$, $p$-value $= 0$)
**14.37** **b.** 1st-order **c.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$ **e.** no, $t = .55$, $p$-value $= .591$
**f.** $\hat{y} = -2.09528 + 0.003684x_1 - 0.238x_2 + 0.000733x_1x_2$ **g.** 2.71 **h.** (2.5772, 2.8352)
**14.39** **a.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1x_4 + \cdots + \beta_{15}x_1x_2x_3x_4$ **b.** df(error) = 0
**c.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1x_4 + \beta_8x_2x_3 + \beta_9x_2x_4 + \beta_{10}x_3x_4$
**d.** $\hat{y} = 5.95 + .388x_1 + .755x_2 + .403x_3 + 1.088x_4 - .038x_1x_2 + .138x_1x_3 - .183x_1x_4 + .043x_2x_3 - .428x_2x_4 - .283x_3x_4$
**e.** only $C \times F$ interaction is significant **f.** yes, Agent and Liquid main effects; both significant at $\alpha = .10$
**14.41** **a.** 81 **b.** 80 terms: 8 main effect terms, 24 two-variable interactions, 32 three-variable interactions, and 16 four-variable interactions
**c.** df(IC) = 2, df(CC) = 2, df(RC) = 2, df(RT) = 2, df(any 2-way interaction) = 4, df(any 3-way interaction) = 8, df(4-way interaction) = 16, df(Error) = 81, df(Total) = 161 **d.** no **e.** only CC
**14.43** **a.** yes, $F = 74.16$ **b.** Alloy *X* Time, Alloy *X* Material
**14.45** **a.** $F = 304.6$, reject $H_0$: $\beta_3 = \beta_4 = \cdots = \beta_{11} = 0$
**14.47** **a.** 3 **b.** 5 **c.** 15 **d.** $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ ($i = 1, 2, 3; j = 1, 2, \ldots, 5$) **e.** $\hat{\sigma}_\alpha^2 = .00129$, $\hat{\sigma}^2 = .0558$
**f.** $F = .02$, $p$-value $= .977$, do not reject $H_0$
**14.49** Source (df): Production Lot (9), Batch within Lot (40), Shipping lot within Batch (950), Total (999)
**14.51** **b.** $\hat{\sigma}_B^2 = .038333$, $\hat{\sigma}_W^2 = .057464$ **c.** yes; $F = 6.34$, $p$-value $= .0006$
**14.53** **a.** 6 **b.** $\mu_{12} < (\mu_3, \mu_6, \mu_9)$
**14.55** $\mu_1 > \mu_2 > \mu_3$
**14.57** **a.** 6 **b.** Highest: Sourdough; Lowest: Control and Yeast
**14.59** $(\mu_{\text{UMRB2}}, \mu_{\text{UMRB3}}) > (\mu_{\text{SD}}, \mu_{\text{SWRA}})$
**14.61** only means for 6 and 14 weeks are not significantly different
**14.63** $\mu_{10} < (\mu_5, \mu_3, \mu_0)$
**14.65** unequal variances
**14.67** unequal variances
**14.69** assumptions reasonably satisfied
**14.71** **a.** safety score **b.** 3 **c.** $H_0$: $\mu_{\text{Scientist}} = \mu_{\text{Journalist}} = \mu_{\text{Official}}$ **d.** 7.065 **e.** less than .01
**14.73** **a.** completely randomized **b.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2$, $x_1 = \{1$ if touch-tone, 0 if not$\}$, $x_2 = \{1$ if human operator, 0 if not$\}$
**c.** $H_0$: $\mu_T = \mu_H = \mu_S$ **d.** $H_0$: $\beta_1 = \beta_2 = 0$ **e.** large within-sample variance
**14.75** **a.** 3 agents (nickel, iron, copper) **b.** 7 ingots **c.** yes, $F = 6.36$
**14.77** **a.** no, $F = .39$ **b.** $F = 19.17$, reject $H_0$
**14.79** no evidence of interaction; no evidence of room order main effect; evidence of aid-type main effect
**14.81** yes, $F = 9.50$, $p$-value $= .0061$; $\mu_{\text{PD}-1} > \mu_{\text{IADC517}}$
**14.83** **b.** yes, $F = 5.39$ **c.** $\hat{y} = -12.306 - 0.1875\text{E} + 0.10\text{T} + 0.01125\text{ET} + 0.01708\text{E}^2 + 0.00146\text{T}^2$ **d.** 50.78 **e.** (85.22, 89.34)
**14.85** **a.** no, $F = 2.32$ **b.** no, $F = 4.68$
**14.87** **a.** yes, $F = 40.78$ **b.** $x_1, x_3, x_4$ **c.** $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1x_4 + \beta_8x_2x_3 + \beta_9x_2x_4 + \beta_{10}x_3x_4 + \beta_{11}x_1x_2x_3 + \beta_{12}x_1x_2x_4 + \beta_{13}x_1x_3x_4 + \beta_{14}x_2x_3x_4 + \beta_{15}x_1x_2x_3x_4$ **d.** 16
**14.89** **a.** CM = 8,912,304,025; SST = 92,833,225 **b.** 75,145,616 **c.** 167,978,841 **d.** MST = 92,833,225, MSE = 766,792, $F = 121.07$
**f.** $\hat{y} = 7{,}834.67 + 91.76x$ **g.** 9,669.87 ± 176.43 **h.** .553

## Chapter 15

**15.1** **a.** $H_0$: $\tau = 300$, $H_a$: $\tau > 300$ **b.** 4 **c.** .3438 **d.** do not reject $H_0$
**15.3** **a.** test unreliable **b.** Sign test **c.** $S = 9$ **d.** $p = 0.5$ **e.** Do not reject $H_0$
**15.5** $S = 11$, $p$-value $= .4119$; do not reject $H_0$
**15.7** **a.** $H_0$: $\tau = 1.5$, $H_a$: $\tau > 1.5$ **b.** 3 **c.** .855 **d.** do not reject $H_0$
**15.9** $S = 9$, $p = 0.0730$, reject $H_0$, no
**15.13** **b.** 104 **c.** 86 **d.** 49 **e.** do not reject $H_0$
**15.15** no; $T_{70} = 66$
**15.17** $T_A = 18$, do not reject $H_0$
**15.19** $z = -8.617$, yes

**15.23** **a.** Differences may not be normal **b.** Wilcoxon signed ranks test
**c & d.**

| Difference | −0.2 | −0.2 | −0.1 | 2.6 | 0.7 | 0.9 | 1.7 | −1.6 | 1.0 | 1.1 | −1.7 | 0.5 | 0.1 | −1.5 | −1.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank | 3.5 | 3.5 | 1.5 | 15 | 6 | 7 | 13.5 | 12 | 8 | 9 | 13.5 | 5 | 1.5 | 11 | 10 |

**e.** $T_+ = 65, T_- = 55$
**f.** $T = 55$, do not reject $H_0$, yes

**15.25** **a.** $H_0$: Driver and Passenger injury rating distributions are identical **b.** $T_+ = 23$ **c.** $T_+ \leq 19$
**d.** do not reject $H_0$, $p$-value $= 0.0214$

**15.27** no, $T_+ = 23$

**15.29** $T_+ = 1$, reject $H_0$

**15.35** **b.** 84 **c.** 145 **d.** 177 **e.** $H = 18.40$ **f.** reject $H_0$ **g.** $z = 3.38$, reject $H_0$

**15.37** **a.** $H_0$: 5 population probability distributions are identical, $H_a$: At least 2 population probability distributions differ in location
**c.** Reject $H_0$

**15.39** $H = 5.16$, no difference

**15.41** $H = 16.27$, reject $H_0$

**15.45** **a.** $F_r > 9.21034$ **b.** reject $H_0$ **c.** do not reject $H_0$ **d.** reject $H_0$

**15.47** $F_r = 1.00$, do not reject $H_0$

**15.49** $F_r = 6.78$, do not reject $H_0$

**15.51** **a.**

| Length | 22.5 | 16 | 13.5 | 14 | 13.75 | 12.5 |
|---|---|---|---|---|---|---|
| Rank | 6 | 5 | 2 | 4 | 3 | 1 |

**b.**

| Concentration | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| Rank | 1 | 2 | 3 | 4 | 5 | 6 |

**c.** $r_s = -0.829$ **d.** Yes

**15.53** **b.** reject $H_0$ for Transactions and Locatability

**15.55** **a.** $C = 5, p = 0.235$, do not reject $H_0$ **b.** $C = 15, p = 0.001$, reject $H_0$

**15.57** **a.** $r_{s1} = 0.643$ **b.** $r_{s2} = 0.524$ **c.** $r_{s3} = 1.000$
**b.** $Y_1$: C $= 12, p = 0.089$, do not reject $H_0$ ; $Y_2$: C $= 12, p = 0.089$, do not reject $H_0$;
$Y_3$: C $= 28, p \approx 0.000$, reject $H_0$

**15.59** **a.** $S = 14, p = 0.0577$, do not reject $H_0$ **b.** $S = 12, p = 0.1796$, do not reject $H_0$
**c.** $T = 50$, do not reject $H_0$ **d.** $r_s = 0.774$, reject $H_0$

**15.63** **a.** $H_0$: $\tau = 5$ **b.** $z = 6.07$, $p$-value $= 0$ **c.** reject $H_0$

**15.65** $T_+ = 3$, do not reject $H_0$

**15.67** **a.** nonnormal distributions **c.** $H = 2.97$, do not reject $H_0$

**15.69** yes; $F_r = 7.85$

**15.71** $F_r = 9.10$, $p$-value $= .011$, reject $H_0$

**15.73** yes; $z = 3.91$

**15.75** **a.** .90 **b.** reject $H_0$

**15.77** no; $H = 2.03$

## Chapter 16

**16.1** **a.** $\bar{x} = 228.67$ **b.** LCL $= -170.84$, UCL $= 628.18$ **c.** yes

**16.3** Site 1: $\bar{x} = 89.548$, LCL $= 83.4142$, UCL $= 95.6818$, out of control
Site 2: $\bar{x} = 89.0332$, LCL $= 82.3556$, UCL $= 95.7106$, out of control

**16.5** **a.** $\bar{x} = 5.895$, LCL $= 4.836$, UCL $= 6.954$ **b.** yes

**16.7** **a.** $\bar{x} = .13974$, LCL $= .1313$, UCL $= .1482$ **b.** yes

**16.9** **a.** $\bar{\bar{x}} = 70.00$ **b.** $\bar{R} = 32.5$ **c.** LCL $= 59.99$, UCL $= 80.01$ **d.** Yes **e.** No

**16.11** **a.** $\bar{x} = .9958$ **b.** LCL $= .9531$, UCL $= 1.0385$ **d.** yes

**16.13** **a.** 0.0013 **b.** 0.000097

**16.15** **a.** $\bar{x} = .14065$, LCL $= .13565$, UCL $= .14565$ **c.** yes

**16.19** **a.** LCL $= 0$, UCL $= 31.45$ **b.** $R = 22.5$, in control

**16.23** **a.** $\overline{R} = 0.02375$, LCL $= 0$, UCL $= 0.0778$ **b.** $\overline{\overline{x}} = 0.222563$, LCL $= 0.177913$, UCL $= 0.267213$ **c.** in control
**16.25** **a.** $\overline{R} = 0.8065$, LCL $= 0$, UCL $= 2.0767$, yes **b.** $\overline{R} = 0.75$, LCL $= 0$, UCL $= 1.93215$, out of control
**16.27** trend in parts d and f
**16.29** no trends
**16.33** no trends
**16.35** **a.** LCL $= .0008$, UCL $= .0202$ **b.** in control
**16.37** $\overline{p} = 0.06046$, LCL $= 0.02848$, UCL $= 0.09244$, out of control
**16.39** **b.** $\overline{p} = .075$ **c.** LCL $= 0$, UCL $= .25169$; yes
**16.41** **b.** $\overline{p} = .2571$ **c.** LCL $= .0717$, UCL $= .4426$ **d.** no; $\overline{p} = .247$, LCL $= .064$, UCL $= .430$ **e.** no trends
**16.43** **b.** $\overline{c} = 6.5$ **c.** LCL $= 0$, UCL $= 14.15$; yes **d.** no trends
**16.45** **a.** no **b.** no trends
**16.47** **a.** $54.26125 \pm 3.92272$ **b.** no **c.** 93; $n$ not large enough
**16.49** **a.** $26 \pm 29.11$; $(1 - \alpha) = 1$ **b.** specific cause
**16.55** **b.** 15.2% **c.** .5045; process not capable
**16.57** **b.** 51% **c.** 3.997 **d.** yes
**16.59** **a.** .5490, .1671, .0353, .0052,.000488 **b.** .1710 **c.** .1671
**16.61** **a.** .0246 **b.** .5443
**16.65** $\overline{x} = 1.4985$, LCL $= 1.4731$, UCL $= 1.5239$
**16.67** $\overline{x} = 5.89667$, LCL $= 5.36471$, UCL $= 6.42863$, in control; $\overline{R} = .52$, LCL $= 0$, UCL $= 1.339$, out of control
**16.69** **b.** producer's risk: plan 1 $= .0815$, plan 2 $= .0334$; prefer plan 2 **c.** consumer's risk: plan 1 $= .5282$, plan 2 $= .1935$; prefer plan 2
**16.71** **a.** $n = 32$, $a = 7$ **b.** $n = 50$, $a = 10$
**16.73** **a.** .0755 **b.** .6769 **d.** $n = 125$, $a = 10$ **e.** .041 **f.** .564
**16.75** **a.** $\overline{p} = .0614$, LCL $= .0105$, UCL $= .1124$ **b.** no
**16.77** **b.** no; $C_p < 1$

## Chapter 17

**17.1** $\lambda$
**17.3** **a.** $f(0) = .0044$, $F(0) = .0013$, $z(0) = .0044$; $f(1) = .054$, $F(1) = .0228$, $z(1) = .0553$; $f(2) = .242$, $F(2) = .1587$, $z(2) = .2876$; $f(3) = .3989$, $F(3) = .5$, $z(3) = .7979$; $f(4) = .242$, $F(4) = .8413$, $z(4) = 1.5247$; $f(5) = .054$, $F(5) = .9772$, $z(5) = 2.368$; $f(6) = .0044$, $F(6) = .9987$, $z(6) = 3.4091$
**17.5** **a.** 1/460 **b.** 1/2880 **c.** 1/395
**17.7** **a.** $F(t) = 1 - \exp(-t^2/100)$ **b.** $R(t) = \exp(-t^2/100)$, $z(t) = t/50$ **c.** $R(8) = .5273$; $z(8) = .16$
**17.9** **a.** $9.4057 \times 10^{11}$ **b.** $(3.7211x\ 10^{12})t^{2.5}$ **c.** .006578
**17.11** **a.** (384.73, 847.57) **b.** (.00118, .00260)
**17.13** **a.** (93.8, 173.5) **b.** (143.9, 292.1)
**17.15** (18,041.27, 294,622.44)
**17.17** (211.4, 394.5)
**17.19** **a.** $\hat{\alpha} = 1.9879$, $\hat{\beta} = 16.029$ **b.** (1.9405, 2.0353) **c.** (14.9423, 17.1947)
**17.21** **a.** $z(t) = .124t^{.9879}$ **b.** $z(4) = .4877$
**17.23** **a.** 9, 4, 2 **b.** $\hat{\alpha} = 1.6938$, $\hat{\beta} = 3.326$ **c.** $\alpha$: $(-.7335, 4.1211)$; $\beta$: (.5383, 20.5250) **d.** .6219
**17.25** .60192
**17.27** **a.** series **b.** $(1 - p_1)(1 - p_2)\cdots(1 - p_k)$
**17.29** .99507
**17.31** **a.** .983 **b.** .632
**17.33** **a.** .8671 **b.** .1424 **c.** .000195
**17.35** **a.** $F(t) = t/\beta$, $R(t) = 1 - t/\beta$, $z(t) = 1/(\beta - t)$ **c.** .6
**17.37** **a.** $\hat{\alpha} = 2.0312$, $\hat{\beta} = 7.3942$ **b.** $z(t) = (.2747)t^{1.0312}$, $R(t) = \exp(-t^{2.0312}/7.3942)$ **c.** .8735
**17.39** .0608
**17.41** **a.** (2,153.3, 17,322.3) **b.** .4284; (.1560, .7938) **c.** .000212; (.0000577, .0004644)
**17.43** .9613

## Appendix A

**A.1** **a.** $\begin{bmatrix} 6 & 3 \\ -2 & -5 \end{bmatrix}$ **b.** $\begin{bmatrix} 3 & 0 & 9 \\ -9 & 4 & 5 \end{bmatrix}$ **c.** $\begin{bmatrix} 5 & 4 \\ 1 & -4 \end{bmatrix}$

**A.3** **a.** $3 \times 4$ **b.** No

**A.5** **a.** $\begin{bmatrix} 2 & 3 \\ -9 & 0 \\ 8 & -2 \end{bmatrix}$ **b.** $[3 \quad 0 \quad 4]$ **c.** $[14 \quad 7]$

**A.7** **a.** $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ **c.** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

**A.9** **a.** $\mathbf{A} = \begin{bmatrix} 10 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 68 \end{bmatrix}$; $\mathbf{V} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{bmatrix}$; $\mathbf{G} = \begin{bmatrix} 60 \\ 60 \\ 176 \end{bmatrix}$ **c.** $\mathbf{V} = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}$

# Credits

## CHAPTER 2

App. Ex. 2.2, "Engineering jobs related to studies." *Mechanical Engineering*, Vol. 126, No. 11, November 2004; App. Ex. 2.3, "Identification and Characterization of Erosional Hotspots," William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March, 18, 2002; App. Ex. 2.4, Blair, A.S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board," *Process Safety Progress*, Vol. 23, No. 4, Dec. 2004; App. Ex. 2.5, Adapted from the 2001 CSI/FBI Computer Crime and Security Survey, *Computer Security Issues & Trends*, Vol. 7, No. 1, Spring 2001, p. 16; App. Ex. 2.7, Hill, T.P. "The First Digit Phenomenon." *American Scientist*, Vol. 86, No. 4, July-Aug 1998, p. 363; App. Ex. 2.10, Reprinted with permission. © 2005 American Chemical Society; App. Ex. 2.11, Lichen Radionuclide Baseline Research project, 2003; Fig. 2.1, American Society for Engineering Education, *Prism*, October, 2004; Fig. 2.2, New Orders-Units © 2005. Robotic Industries Association; Fig. 2.12, Binzel, R. P., and Xu, S. "Chips off of Asteroid 4 Vesta: Evidence for the parent body of basaltic achondrite meteorites." *Science*, Vol. 260, Apr. 3, 1993, p. 187; Fig. 2.63, Reprinted with permission. © 1985 American Chemical Society; Fig. 2.64, Chin, Jih-Hua et al. "The computer simulation and experimental analysis of chip monitoring for deep hole drilling." *Journal of Engineering for Industry*, Transactions of the ASME, Vol. 115, May 1993, p. 187; Table 2.6, "Competitive PWB manufacturing: What is needed to maintain a viable industry in Europe?" Philip Britton, *Circuit World* (2000, Vol26, 3 – Table 1 pg 18) © MCB University Press. Republished with permission, Emerald Group Publishing Limited.

## CHAPTER 3

App. Ex. 3.8, Hill, T.P. "The First Digit Phenomenon." *American Scientist*, Vol. 86, No. 4, July-Aug 1998, p. 363; Fig. 3.1, Chen, J. R., et al. "Emergency response of toxic chemicals in Taiwan: The system and case studies," *Process Safety Progress*, Vol. 23, No. 3, Sept. 2004; Fig. 3.2, "Identification and Characterization of Erosional Hotspots," William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March, 18, 2002; Fig. 3.3 Extracted from *The Orange County (Calif.) Reporter*, Aug. 7, 1990; Fig. 3.4, Blair, A. S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board," *Process Safety Progress*, Vol. 23, No. 3, Sept. 2004; Fig. 3.20, Chandler, H. E. "Materials trends at Mazda Motor Corporation," *Metal Progress*, Vol. 129, No. 6, May 1986, p. 57; Fig. 3.24, Kaneda, K., et al. "An unmanned watching system using video cameras." *JEEE Computer Applications in Power*, Apr. 1990, p. 24; Fig. 3.73, Ennis, R. L., et al., "Acontinuous real-time expert system for computer operations." *IBM Journal of Research and Development*, Vol. 30, No. 1, Jan. 1986, p. 19. Copyright 1986 by International Business Machines Corporation; reprinted with permission; Fig. 3.74, Meagher, J. J., and Seazzero, J. A. "Measuring Inspector Variability." *39th Annual Quality congress Transactions*, May 1985, pp. 75–81. American Society for Quality Control; Table 3.2, Adapted from Cook, M., Simon, P., and Hoffman, R. "Unintentional carbon monoxide poisoning in Colorado," *American Journal of Public Health*, Vol. 85, No. 7, July 1995; Table 3.35, Polus, A., and Livneh, M. "Vehicle flow characteristics on acceleration lanes," *Journal of Transportation Engineering*, Vol. III, No. 6, Nov. 1985, pp. 600–601; Table 3.75, Nature by Prendergast, J. R. Copyright 1993 by Nature Publishing Group. Reproduced with permission of Nature Publishing Group in the format Textbook via Copyright Clearance Center; Table 3.83, Kinchen, A. L. "Projected outcomes of exploration programs based on current program status and the impact of prospects under consideration." *Journal of Petroleum Technology*, Vol. 38, No. 4, Apr. 1986, p. 462. (Table 1). © 1986 Society of Pettoleum Engineers.

## CHAPTER 4

App. Ex. 4.11, Kinchen A. L. "Projected outcomes of exploration programs based on current program status and the impact of prospects under consideration." *Journal of Petroleum Technology*, Vol. 38, No. 4, Apr. 1986, p. 462. (Table 1). © 1986 Society of Petroleum Engineers; Fig. 4.1, Annals of the Entomological Society of America by SOSA, A. J. Copyright 2005 by Entomological Soc. of America. Reproduced with permission of Entomological Soc. of America in the format Textbook via Copyright Clearance Center; Fig. 4.4, "Identification and Characterization of Erosional Hotspots," William & Mary Virginia Institute of Marine Science, U.S. Army Corps of Engineers Project Report, March, 18, 2002; Fig. 4.43, Adapted from the 2001 CSI/FBI Computer Crime and Security Survey, *Computer Security Issues & Trends*, Vol. 7, No. 1, Spring 2001, p. 16; Fig. 4.51, Gonzalez, J. and Valdes, J. B. "Bivariate drought recurrence analysis using tree ring reconstructions," *Journal of Hydrologic Engineering*, Vol. 8, No. 5, Sep/Oct 2003; Fig. 4.83, Chandler, H. E. "Materials trends at Mazda Motor Corporation." *Metal Progress*, Vol. 129, No. 6, May 1986, p. 57 (Figure 3); Table 4.42, *Mechanical Engineering*, Vol. 126, No. 11, November 2004; Table SIA 4.1 Department of Defense Reliability Analysis Center, START: Analysis of "One-Shot" Devices, Vol. 7, No. 4, 2000 (Table 2).

## CHAPTER 5

App. Ex. 5.36, Binzel, R. P., and Xu, S. "Chips off of Asteroid 4 Vesta: Evidence for the parent body of basaltic achondrite meteorites." *Science*, Vol. 260, Apr. 3, p. 187 (Table 1); App. Ex. 5.89, Chin, Jih-Hua et al. "The computer simulation and experimental analysis of chip monitoring for deep hole drilling." *Journal of Engineering for Industry*, Transactions of the ASME, Vol. 115, May 1993, p. 187 (Figure 12); Fig. 5.35, Cogley, J. G., and Jung-Rothenhausler, F. "Uncertainly in digital elvation models of Axel Heiberg Island, Arctic Canada," *Arctic, Antarctic, and Alpine Research*, Vol. 36, No. 2, May, 2004 (Figure 3); Fig. 5.37, Scholz, H. "Fish Creek Community Forest: Exploratory statistical analysis of selected data," working paper. Northern Lights College, British Columbia, Canada; Fig. 5.38, Good, T. P., Hamms, T. K., and Ruckelshaus, M.H. "Misuse of checklist assessments in endangered species recovery efforts," *Conservation Ecology*, Vol. 7, No. 2, Dec. 2003 (Figure 3); Table 5.86, Bozkurt, E., et al. "Geochemistry and the tectonic significance of augen gneisses from the southern

Menderes Massif (West Turkey)," *Geological Magazine*, Vol. 132, No. 3, May 1995, p. 291 (Table 1).

## CHAPTER 6

Table 6.67, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*.

## CHAPTER 7

App. Ex. 7.29, Lichen Radionuclide Baseline Research project, 2003; App. Ex. 7.33, Ewing, R. "Roadway Levels of Service in an Era of Growth Management." In Transportation Research Record 1364, Transportation Research Board, National Research Council, Washington, D.C., 1992, Table 4, page 69, Reproduced with permission of TRB; App. Ex. 7.78, Reprinted with permission from *Environmental Science & Technology*. Copyright © 1993 American Chemical Society; Fig. 7.24, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; App. Ex. 7.108, Loncarevic, B. D., Fenniger, T., and Lefebvre, D. "The Sept-lies layered mafic intrusion: Geophysical Expression," *Canadian Journal of Earth Science*, Vol. 27, Aug. 1990, p. 505; Fig. 7.31, Hillsborough County Water Department Environmental Laboratory, Tampa, Florida; Fig. 7.36, Reprinted with permission from the *Journal of Agricultural, Biological, and Environmental Statistics*. Copyright © 2005 by the American Statistical Association. All rights reserved; Fig. 7.39, Ushio, H., and Watabe, S. "Ultrastructural and biochemical analysis of the sarcoplasmic reticulum from crayfish fast and slow striated muscles," *The Journal of Experimental Zoology*, Vol. 267, Sept. 1993, p. 16 (Table 1); Fig. 7.41, Avent, R. R. "Design criteria for epoxy repair of timber structures," *Journal of Structural Engineering*, Vol. 112, No. 2, Feb. 1986, pp. 232; Fig. 7.46, IEICE Transactions on Information and Systems by Ichihara, H., Shintani, M., & Inoue, T. Copyright 2005 by Oxford Univ Press Inc (US). Reproduced with permission of Oxford Univ Press Inc (US) in the format Textbook via Copyright Clearance Center; Fig. 7.47, Reprinted with permission from *Environmental Science & Technology*. Copyright © 1993 American Chemical Society; Fig. 7.48, Wall, D. J., and Peterson, C. "Model for winter heat loss in uncovered clarifiers." *Journal of Environmental Engineering*, Vol. 112, No. 1, Feb. 1986, p. 128; Fig. 7.49, Qibai, C. Y. H. and Shi, H. "An investigation on the physiological and psychological effects of infrasound on persons," *Journal of Low Frequency Noise, Vibration and Active Control*, Vol. 23, No. 1, Mar. 2004 (Table V); Fig. 7.77, Reproduced with permission from Strelow, D. and Singh, S. "Motion estimation form image inertial measurements," *The International Journal of Robotics Research*, Vol. 23, No. 12, Dec. 2004 (Table 4). © Sage Productions, 2004, by permission of Sage Publications Ltd.; Fig. 7.79, Avent, R. R. "Design criteria for epoxy repair of timber structures." *Journal of Structural Engineering*, Vol. 112, No. 2, Feb. 1986, pp. 232; Fig. 7.92, Adapted from the *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2); Fig. 7.106, Butcher, B. T., Reed, M. A., and O'Neil, C. E. "Biochemical and immunologic characterization of cotton bract extraxt and its effect on in vitro cyclic AMP production" *Environmental Research*, Vol. 39, No. 1, Feb. 1986, p. 119. With permission from Elsevier; Table 7.42, Reprinted, with permission, from the *Journal of Testing and Evaluation*, Vol. 9, No. 4, July 1981, pp. 175–181., copyright ASTM International, 100 Barr Harbor Drive, West Conshohocken, PA 19428; Table 7.42, Martin, A. M., et al. "Estimation of the serviceability of forest access roads," *International Journal of Forest Engineering*, Vol. 10, No. 2, July 1999 (adapted from Table 3); Table 7.76, Martin, A. M., et al. "Estimation of the serviceability of forest access roads", *International Journal of Forest Engineering*, Vol. 10, No. 2, July 1999 (adapted from Table 3); Table 7.104, Applied Spectroscopy by Wopenka, B. Copyright 1986 by Soc For Applied Spectroscopy. Reproduced with permission of Soc for Applied Spectroscopy in the format Textbook via Copyright Clearance Center.

## CHAPTER 8

App. Ex. 8.45, Thomas E. Bradstreet, Merck Research Labs, BI. 3–2, West Point, Penn. 19486. Used with permission; App. Ex. 8.84, Pfeiffer, M., et al. "Community organization and species richness of ants in Mongolia along an ecological gradient from steppe to Gobi desert," *Journal of Biogeography*, Vol. 30, No. 12, Dec. 2003 (Tables 1 and 2); Fig. 8.9, Zararis, P. D., and Penelis, G. Jr. "Reinforced concrete T-beams in torsion and bending." *Journal of the American Concrete Institute*, Vol. 83, No. 1, Jan.–Feb. 1986, p. 153; Fig. 8.16, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; Fig. 8.41, Adapted from Gill, R. T., et al. "Genome-wide dynamic transcriptional profiling of the light to dark transition in Synechocystis Sp. PCC6803," *Journal of Bacteriology*, Vol. 184, No. 13, July 2002; Fig. 8.46, Reprinted from *Chemosphere*, Vol. 15, No. 2, Feb. 1986, p. 125. © 1986, with permission from Elsevier; Fig. 8.102, Wall, D. J., and Peterson, C. "Model for winter heat loss in uncovered clarifiers." *Journal of Environmental Engineering*, Vol. 112, No. 1, Feb. 1986, p. 128; Fig. 8.74, Pinchin, M. J. "A study of the trace organics profiles of raw and potable water systems," *Journal of the Institute of Water Engineering & Scientists*, Vol. 40, No. 1, Feb. 1986, p. 87; Fig. 8.93, Goodman, J. R., Vanderbilt, M. D., and Criswell, M. E. "Reliability-based design of wood transmission line structures." *Journal of Structural Engineering*, Vol. 109, No. 3, 1983, pp. 690–704; Fig. 8.98, Reprinted with permission from *Environmental Science & Technology*. Copyright © 1993 American Chemical Society; Fig. 8.101, Reichman, O. J. "Desert granivore foraging and its impact on seed densities and distributions." *Ecology*, Dec. 1979, Vol. 60, pp. 1085–1092; Table 8.2, Moore, H. E., and Gussow, D. G. "Radium and radon in Dade County ground water and soil samples." *Florida Scientist*, Vol. 54, No. 3/4, Summer/Autumn, 1991, p. 1555 (Portion of Table 3); Table 8.31, Reprinted with permission from the *Journal of Agricultural, Biological, and Environmental Statistics.* Copyright © 2005 by the American Statistical Association. All rights reserved; Table 8.35, Reprinted from *Ecological Engineering*, Vol. 22, No. 1, Feb. 2004, (Table 5) © 2004, with permission from Elsevier; Table 8.37, Reprinted with permission from *Environmental Science & Technology*. Copyright © 1993 American Chemical Society; Table 8.44, Yih, Y., Liang, T., and Moskowitz, H. "Robot scheduling in a circuit board production line: A hybrid OR/ANN approach." *IEEE Transactions*, Vol. 25, No. 2, March 1993, p. 31 (Table 1). © 1993 IEEE; Table 8.42, IEICE Transactions on Information and Systems by Ichihara, H., Shintani, M., & Inoue, T. Copyright 2005 by Oxford Univ Press Inc (US) Reproduced with permision of Oxford Univ Press INc (US) in the format Textbook via Copyright Clearance Center; Table 8.47, Kerkhof, P. and Geboers, M. "Toward a unified theory of isotropic molecular transport phenomena," *AIChE Journal*, Vol. 51, No. 1, January 2005 (Table 2); Table 8.54, Basu, A., and McKay, D. S. "Lunar soil evolution processes and Apollo 16 core 60013/60014." *Meteoritics*, Vol. 30, No. 2, Mar. 1995, p. 166 (Table 2); Table 8.78, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; Table 8.1, Copyright © American Statistical Association and Society for Industrial and Applied Mathematics. Reprinted with permission.

## CHAPTER 9

App. Ex. 9.21, Wileyto, E. P. et al. "Self-marking recapture models for estimating closed insect populations," *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 5, No. 4, December 2000 (Table 5A); App. Ex. 9.26, Johns, C., Holman, B., Niemeier, A., and Shumway, R. "Nonlinear regression for modeling censored one-dimensional concentration profiles of fugitive dust plumes," *Journal of Agricultural, Biological, and Environmental Sciences*, Vol. 6, No. 1, March 2001 (from data file provided by co-author Brit Holmen); App. Ex. 9.29, Sunny, J. and Vallathan, A. "A comparative in vitro study with new generation ethyl cyanacrylate (Smartboard) and a composite bonding agent," *Trends in Biomaterials & Artificial Organs*, vol. 16, No. 2, Jan. 2003 (Table 6); App. Ex. 9.3, Menard, H. W. "Time, chance, and the origin of manganese nodules." *American Scientist*, Sept.–Oct. 1976; App. Ex. 9.33, Mosley, L., Sharp, D., and Singh, S. "Effects of a tropical cyclone on the drinking-water quality of a remote Pacific island," *Disasters*, Vol. 28, No. 4, 2004 (from Table 3); App. Ex. 9.35, Sunny, J. and Vallathan, A. "A comparative in vitro study with new generation ethyl cyanacrylate (Smartboard) and a composite bonding agent," *Trends in Biomaterials & Artificial Organs*, vol. 16, No. 2, Jan. 2003 (Table 6); Fig. 9.4, Brunn, S., et al. "Final report survey of Three Mile Island area residents." Department of Geography, Michigan State University, Aug. 1979. Used with permission; Fig. 9.42, Jaeger, R. G. "Dear enemy recognition and the costs of aggression between salamanders." *The American Naturalist*, June 1981, Vol. 117, pp. 962–973. Reprinted by permission of the University of Chicago Press © 1981 The University of Chicago; Table 9.47, Johnson, R. W. "Testing colour proportions of M & M's." *Teaching Statistics*, Vol. 15. No. 1, Spring 1993, p. 2 (Table 1); Table 9.4, Reprinted with permission from *Environmental Science & Technology*. Copyright © 1993 American Chemical Society; Table 9.7, Copyright © 1986, American Water Works Association. Adapted by permission; App. Ex. 9.39, Zeighami, E. A., and Morris, M. D. "Thyroid cancer risk in the population around the Nevada test site." *Health Physics*, Vol. 50, No. 1, Jan. 0986, p. 26 (Table 2); Table 9.19, Yeh, W. and Bell, T. "Significance of dextral reactivation of an E-W transfer fault in the formation of the Pennsylvania orocline, central Appalachians," *Tectonics*, Vol. 23, No. 5, October 2004 (Table 2); Table 9.8, Gilbert, P. "Developing an AIDS vaccine by sieving." *Chance*, Vol. 13, No. 4, Fall 2000, pp. 16–21; Table 9.13, Nature by Prendergast, J. R., et al. Copyright 1993 by Nature Pubg Group. Reproduced with permission of Nature Pubg Group in the format Textbook via Copyright Clearance Center; Table 9.12, Blair, A.S. "Management system failures identified in incidents investigated by the U.S. Chemical Safety and Hazard Investigation Board," *Process Safety Progress*, Vol. 23, No. 4, Dec. 2004 (Table 1).

## CHAPTER 10

App. Ex. 10.6, Pandit, R., and U.S. Palekar. "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry, Transactions of the ASME*, Vol. 115, Aug. 1993, p. 326 (Table 2); App. Ex. 10.7, American Ceramic Society Bulletin by Bonadia, P. Copyright 2005 by Am Ceramic Soc Inc. Reproduced with permission of Am Ceramic Soc Inc in the format Textbook via Copyright Clearance Center; App. Ex. 10.8, Barry, J. "Estimating rates of spreading and evaporation of volatile liquids," *Chemical Engineering Progress*, Vol. 101, No. 1, Jan., 2005; App. Ex. 10.12, Marto, P. J., et al. "An experimental study of R-113 film condensation on horizontal integral-fin tubes." *Journal of Heat Transfer*, Vol. 112, Aug. 1990, p. 763 (Table 2); App. Ex. 10.21, Hoffman, G. and Tsuge, O. "ITmk3-Application of a new ironmaking technology for the iron ore mining industry," *Mining Engineering*, Vol. 56, No. 9, October 2004 (Figure 8); App. Ex. 10.22, Copyright © 2002 from *Drug Development and Industrial Pharmacy* by Reynolds, T., Mitchell, S., and Balwinski, K. Reproduced by permission of Taylor & Francis Group, LLC, *http://www.taylorandfrancis.com*; App. Ex. 10.61, Bennett, W. S. "An error analysis of the FCC site-attenuation approximation." *IEEE Transactions on Electromagnetic Compatibility*, Vol. EMC-27, No. 3, Aug. 1985, p. 113 (Table IV). © 1985 IEEE; App. Ex. 10.62, Reprinted from *Corrosion Science*, Vol. 49, No. 9, Chattoraj, I., et al. "Polarization and resistivity measurements of post-crystallization changes in amorphous Fe-B-Si alloys." p. 712 (Table a), Copyright (1993), with permission from Elsevier; App. Ex. 10.68, Abou El Naga, H. H., and Salem, A. E. M. "Base oils thermooxidation," *Lubrication Engineering*, Vol. 24, No. 4, Apr. 1986, p. 213. Reprinted by permission of the American Society of Lubrication Engineers. All rights reserved; App. Ex. 10.32, Hageseth, G. T., and Cody, A. L. "Energy-level model for isothermal seed germination." *Journal of Experimental Botany*. Vol. 44, No. 258, Jan. 1993, p. 123 (Figure 9); App. Ex. 10.65, Park, J.Y., Ruther, W. E., Kassner, T. F., and Shack, W. J. "Stress corrosian crack growth rates in Type 304 stainless steel in simulated BWR environments," *Transactions of the American Society of Mechanical Engineers*, Vol. 108, No. 1, Jan. 1986, p. 23 (Table 4); App. Ex. 10.33, Wade, T. G., K. H. Riitters, J. D. Wickham, and K. B. Jones. 2003. Distribution and causes of global forest fragmentation. Conservation Ecology 7(2): 7. [online] URL: *http://www.consecol.org/vol7/iss2/art7/*; App. Ex. 10.52, Scholz, H. "Fish Creek Community Forest: Exploratory statistical analysis of selected data," working paper. Northern Lights College, British Columbia, Canada; App. Ex. 10.6, Porco, C. C., et al. "Cassini imaging science: Initial results on Phoebe and Iapetus," *Science*, Vol. 307, No. 5713, Feb. 25, 2005 (Figure 8); App. Ex. 10.66, Heger, F. J., and McGrath, T. J. "Radial tension strength of pipe and other curved flexural members," *Journal of the American Concrete Institute*, Vol. 80, No. 1, 1983, pp. 33–39; App. Ex. 10.71, Fairley, W., B., et al. "Bricks, buildings, and the Bronx: Estimating masonry deterioration." *Chance*, Vol. 7, No. 3, Summer 1994, p. 36; Table 10.11, Penner, R., and Watts, D. G. "Mining information." *The American Statistician*, Vol. 45, No. 1, Feb. 1991, p. 6 (Table 1); Table SIA 10.1, Enright, J. T. "Testing dowsing: The failure of the Munich Experiments." *Skeptical Inquirer*, Jan./Feb. 1999, p. 45 (Figure 6a).

## CHAPTER 11

App. Ex. 11.1, American Ceramic Society Bulletin by Bonadia, P. Copyright 2005 by Am Ceramic Soc Inc. Reproduced with permission of Am Ceramic Soc Inc in the format Textbook via Copyright Clearance Center; App. Ex. 11.2, Copyright © 2002 from *Drug Development and Industrial Pharmacy* by Reynolds, T., Mitchell, S., and Balwinski, K. Reproduced by permission of Taylor & Francis Group, LLC, *http://www.taylorandfrancis.com*; App. Ex. 11.3, Pandit, R., and U.S. Palekar. "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry, Transactions of the ASME*, Vol. 115, Aug. 1993, p. 326 (Table 2); App. Ex. 11.5, Pacansky, J., England, C. D., and Waltman, R. "Infrared spectroscopic studies of poly (perfluoropropylencoxide) on gold substrates: A classical dispension analysis for the refractive index." *Applied Spectroscopy*, Vol. 40, No. 1, Jan. 1986, p. 9 (Table 1); App. Ex. 11.21, Bhargava, R. and Meher-Homji, C. B. "Parametric analysis of existing gas turbines with inlet evaporative abd overspray fogging," *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005; App. Ex. 11.24, Schmidt, M., Schneider, D. P., and Gunn, J. E. "Spectroscoptic CCD surveys for quasars at large redshift." *The Astronomical Journal*, Vol. 110, No. 1, July 1995, p. 70 (Table 1); App. Ex. 11.31, Reprinted from the *Journal of Colloid and Interface Science*, Vol. 173, No. 2, Aug. 1995, Fordedal, H., "A multivariate analysis of W/O emulsions in high external electric fields as studies by means of dielectric time domain spectroscopy," p. 398 (Table 2)., © 1995, with permission from Elsevier; App. Ex. 11.36, Hayes, B. "How to avoid yourself," *American Scientist*, Vol. 86, No. 4, July–Aug. 1998, p. 317 (Figure 5); App. Ex. 11.38, Wade, T. G., K. H. Riitters, J. D. Wickham, and K. B. Jones, 2003. Distribution and causes of global forest fragmentation. *Conservation Ecology* 7(2): 7. [online] URL: *http://www.consecol.org/vol7/iss2/art7/*; App. Ex. 11.39, Takizawa, K., et al. "Characteristics of $C_3$ radicals in high- density $C_4F_8$ plasmas studied by laser-induced fluorescence spectroscopy," *Journal of Applied Physics*, Vol. 88, No. 11, Dec. 1, 2000 (Figure 7); App. Ex. 11.42, Bassett, W. A., Weathers, M. S., and Wu, T. C. "Compressibility of SiC up to 68.4 Dpa." *Journal of Applied Physics*, Vol. 74, No. 6, Sept. 15, 1993, p. 3825 (Table 1); App. Ex. 11.43, Copyright © 1993. American Chemical Company. Reprinted with permission; App. Ex. 11.55, Vuorinen, J. "Applications of diffusion theory to permeability tests on concrete, Part II: Pressure-saturation test on concrete and coefficient of permeability," *Magazine of Concrete Research*, Vol. 37, No. 132, Sept. 1985, p. 156. (Table II.I); App. Ex. 11.59, Caswell R. H., and Trak, B. "Some geotechnical characteristics of fragmented Queenston Shale," *Canadian Geotechnical Journal*, Vol. 22, No. 3, Aug. 1985, pp. 403–408; App. Ex. 11.63, Reprinted from *Geoderma*, Vol. 67, No. 1–2, Sharpley, A. N., Robinson, J. S., and Smith, S. J., "Bioavailable phosphorus dynamics in agricultural soils and effects on water quality," p. 11 (Table 4). Copyright © 1995, with permission from Elsevier; Fig. 11.5, Grimes, P. & Kentor, J. "Exporting the greenhouse: Foreign capital penetration and $CO_2$ emissions 1980–1996," *Journal of World-Systems Research*, Vol. IX, No. 2, Summer 2003 (Appendix B). Used with permission; Fig. 11.51, Grimes, P. & Kentor, J. "Exporting the greenhouse: Foreign capital penetration and $CO_2$ emissions 1980–1996," *Journal of World-Systems Research*, Vol. IX, No. 2, Summer 2003 (Appendix B). Used with permission; Fig. 11.54, Hamilton, D. "Sometimes $R^2$ correlated variables are not always reduntant," *The American Statistician*, Vol. 41, No. 2, May 1987, pp. 129–132; Fig. 11.69, Pallardy, S. G., and Kozlowski, T. T. "Water relations of Populus clones." *Ecology*, Feb. 1981, Vol. 62, pp. 159–169. Copyright 1981, the Ecological Society of America; Table 11.6, Reprinted from the *Journal of Urban Economics*, Vol. 21, Rolleston, B. S., "Determinants of restrictive suburban zoning: An empirical analysis," p. 15 (Table 4). © 1987, with permission from Elsevier; Table 11.17, Grimes, P. & Kentor, J. "Exporting the greenhouse: Foreign capital penetration and $CO_2$ emissions 1980–1996," *Journal of World-Systems Research*, Vol. IX, No. 2, Summer 2003 (Appendix B). Used with permission.

## CHAPTER 12

App. Ex. 12.2, Data from "Achieving Uniformity in a Semiconductor Fabrication Process Using Spatial Modeling" by Hughes-Oliver et al, *JASA*, March 1998, Vol. 93; App. Ex. 12.27, Hayes, B. "How to avoid yourself," *American Scientist*, Vol. 86, No. 4, July–Aug. 1998, p. 317 (Figure 5); App. Ex. 12.28, Pacansky, J., England, C. D., and Waltman, R. "Infrared spectroscopic studies of poly (perfluoropropylencoxide) on gold substrates: A classical dispension analysis for the refractive index." *Applied Spectroscopy*, Vol. 40, No. 1, Jan. 1986, p. 9 (Table 1); App. Ex. 12.47, Leigh, L. E. "Contestability in deregulated airline markets: Some empirical tests." *Transportation Journal*, Winter 1990, p. 55 (Table 4). Reprinted from the American Society of Transportation and Logistics, Inc., for educational purposes only; App. Ex. 12.48, Litzinger, T. A., and Buzza, T. G. "Performance and emissions of a diesel engine using a coal-derived fuel." *Journal of Energy Resources Technology*, Vol. 112, Mar. 1990, p. 32, Table 3; App. Ex. 12.66, Wang, G. C. "Microscopic investigation of CO-2 flooding process." *Journal of Petroleum Technology*, Vol. 34, No. 8, Aug. 1982, pp. 1789–1797. © 1982 Society of Petroleum Engineers; Table 12.3, Petric, D., et al. "Dependence of $CO_2$ baited suction trap captures on temperature variations." *Journal of the American Mosquito Control Association*, Vol. 11, No. 1, Mar. 1995, p. 8.

## CHAPTER 13

Table SIA13.1, 13.2, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*.

## CHAPTER 14

Fig. 14.1, Reprinted, with permission, from the *Journal of Testing and Evaluation*, Vol. 20, No. 4, July 1992, p. 319 (Figure 3), copyright ASTM International, 100 Barr Harbor Drive, West Conshohocken, PA 19428; App. Ex. 14.62, Copyright © 1985 American Chemical Society. Reprinted with permission; App. Ex. 14.1, Reprinted from the *Journal of Hazardous Materials*, Vol. 42, No. 2, J. D. Ortego et al., "A review of polymeric geosynthetics used in hazardous waste facilities," p. 142 (Table 9), Copyright © 1995, with permission from Elsevier; App. Ex. 14.12, Rogers, W. H. & Moeller, G. (1984). "Comparison of abbreviation methods: Measures of preference and decoding performance." *Human Factors*, 26(1), 49–59; App. Ex. 14.13, Khoury, G. A., Grainger, B. N., and Sullivan, P. J. E. "Strain of concrete during first heating to 600°C under load." *Magazine of Concrete Research*, Vol. 37, No. 133, Dec. 1985, p. 198 (Table 2); App. Ex. 14.17, Cox, B.G., and Keisall, K. J. "Construction of Cape Peron Ocean Outlet Perth, Western Australia." *Proceedings of the Institute of Civil Engineers*, Part 1, Vol. 80, Apr. 1986, p. 479 (Table 1); App. Ex. 14.18, Qibai, C. Y. H., and Shi, H. "An Investigation on the physiological and psychological effects of infrasound on persons," *Journal of Low Frequency Noise, Vibration and Active Control*, Vol. 23, No. 1, March 2004 (Tables I-IV); App. Ex. 14.2, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; App. Ex. 14.21, Adapted from Gill, R. T., et al. "Genome-wide dynamic transcriptional profilling of the light to dark transition in Synechocystis Sp. PCC6803," *Journal of Bacteriology*, Vol. 184, No. 13, July 2002; App. Ex. 14.22, Boggs, J. J. "The eradication of leisure." *New Technology, Work, and Employment*, Vol. 16, No. 2, July 2001 (Table 3); App. Ex. 14.28, Tomlinson, W. J., and Cooper, G. A. "Fracture mechanism of brass/Sn-Pb-Sb solder joints and the effect of production variables on the joint strength." *Journal of Materials Science*, Vol. 21, No. 5, May 1986, p. 1731 (Table II). Copyright 1986 Chapman and Hall; App. Ex. 14.31, Reprinted from *Combustion and Flame*, Vol. 50, Matsui, K., Tsuji, H., and Makino, A., "The effects of water vapor concentration on the rate of combustion of an artificial graphite in humid air flow," pp. 107–118. © 1983, with permission from Elsevier; App. Ex. 14.39, Reprinted from *Engineering Geology*, Vol. 22, No. 2, Seedmen, R. W., and Emerson, W. W., "The formation of planes of weakness in the highwall at Goonyella Mine, Queensland, Australia," p. 164 (Table I). © 1985, with permission from Elsevier; App. Ex. 14.52, Casali, S. P., Williges, B. H., and Dryden, R. D. "Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance." *Human Factors*, Vol. 32, No. 2, April 1990, p. 190 (Figure 2); App. Ex. 14.64, Rawlins, S. C., and Oh Hing Wan, J. "Resistance in some Caribbean population of Aedes aegypti to several insecticides." *Journal of American Mosquito Control Associations*, Vol. 11, No. 1, Mar. 1995 (Table 1); App. Ex. 14.66, Forsberg, C. W., et al. "The release of fermentable carbohydrate from peat by steam explosion and its use in the microbial production of solvents." *Biotechnology and Bioengineering*, Vol. 28, No. 2, Feb. 1986, p. 179 (Table 1). Copyright 1986; Fig. 14.74, Aroni, S., and Fletcher, G. "Observations on mortar lining of steel pipelines." *Journal of Transportation Engineering*, Nov. 1979; Table 14.25, Butler, D. L., Acquino, A. L., Hissong, A. A., & Scott, P. A. (1993). "Wayfinding by newcomers in a complex building." *Human Factors*, 35(1), 159–174; Table 14.27, Reprinted with permission of APICS The Association for Operations Management, Production and Inventory Management Journal, 3rd quarter, 1999.

## CHAPTER 15

App. Ex. 15.6, Reprinted with permission of the Institute of Industrial Engineers, 3577 Parkway Lane, Suite 200, Norcross, GA 30092, 770-449-0461. Copyright © 2005; App. Ex. 15.2, Lichen Radionuclide Baseline Research project, 2003; App. Ex. 15.3, Farshad, F. & Pesacreta, T.

"Coated pipe interior surface roughness as measured by three scanning probe instruments," *Anti-corrosion Methods and Materials*, Vol. 50, No. 1, 2003 (Table III); App. Ex. 15.7, Moore, H. E., and Gussow, D. G. "Radium and radon in Dade County ground water and soil samples." *Florida Scientist*, Vol. 54, No. 3/4, Summer/Autumn, 1991, p. 1555 (Portion of Table 3); App. Ex. 15.15, Copyright © 1993. American Chemical Society; App. Ex. 15.16, Gastwith, J. L., and Mahmoud, H. "An efficient robust nonparametric test for scale change for data from a gamma distribution." *Technometrics*, Vol. 28, No. 1, Feb. 1986, p. 83 (Table 2); App. Ex. 15.19, Reprinted from *Chemosphere*, Vol. 15, No. 2, Feb. 1986, Badsha, K., and Eduljee, G. "PCB in the U.K. environment-A preliminary survey," p. 213, © 1986, with permission from Elsevier; App. Ex. 15.27, Yih, Y., Liang, T., and Moskowitz, H. "Robot scheduling in a circuit board production line: A hybrid OR/ANN approach." *IEEE Transactions*, Vol. 25, No. 2, March 1993, p. 31 (Table 1). © 1993 IEEE; App. Ex. 15.24, IEICE Transactions on Information and Systems by Ichihara, H., Shintani, M., & Inoue, T. Copyright 2005 by OXFORD UNIV PRESS INC (US) (J). Reproduced with permission of OXFORD UNIV PRESS INC (US) (J) in the format Textbook via Copyright Clearance Center; App. Ex. 15.26, Kerkhof, P. and Geboers, M. "Toward a unified theory of isotropic molecular transport phenomena," *AIChE Journal*, Vol. 51, No. 1, January 2005 (Table 2); App. Ex. 15.61, Butcher, B. T., Reed, M. A., and O'Neil, C. E. "Biochemical and immunologic characterization of cotton bract extract and its effect on in vitro cyclic AMP production." *Environmental Research*, Vol. 39, No. 1, Feb. 1986, p. 119. With permission from Elsevier; App. Ex. 15.29, Wall, D. J., and Peterson, C. "Model for winter heat loss in uncovered clarifiers." *Journal of Environmental Engineering*, Vol. 112, No. 1, Feb. 1986, p. 128; App. Ex. 15.4, Copyright 1985. American Chemical Society, Reprinted with permission; App. Ex. 15.45, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; App. Ex. 15.47, Boggs, J. J. "The eradication of leisure," *New Technology, Work, and Employment*, Vol. 16, No. 2, July 2001 (Table 3); App. Ex. 15.48, Saleh, S. D., and Desai, K. "Occupational stress for engineers." *IEEE Transactions on Engineering Management*, Vol. EM-33, No. 1, Feb. 1986, p. 8 (Table II). © 1986 IEEE; App. Ex. 15.49, American Ceramic Society Bulletin by Bonadia, P. Copyright 2005 by Am Ceramic Soc Inc. Reproduced with permission of Am Ceramic Soc Inc in the format Textbook via Copyright Clearance Center; App. Ex. 15.52, Pandit, R., and U. S. Palekar. "Response time considerations for optimal warehouse layout design." *Journal of Engineering for Industry. Transactions of the ASME*, Vol. 115, Aug. 1993, p. 326 (Table 2); App. Ex. 15.53, Reprinted from *Corrosion Science*, Vol. 49, No. 9, Sept. 1993, Chattoraj, I. "Polarization and resistivity measurements of post-crystallization changes in amorphous Fe-B-Si alloys," p. 712, © 1993, with permission from Elsevier; App. Ex. 15.71, Bennett, W. S. "An error analysis of the FCC site-attenuation approximation." *IEEE Transactions on Electromagnetic Compatibility*, Vol. EMC-27, No. 3, Aug. 1985, p. 113 (Table IV). © 1985 IEEE; App. Ex. 15.56, Riddington, J. R., and Ghazali, M. Z. "Hypothesis for shear failure in masonary joints." *Proceedings of the Institute of Civil Engineers*, Part 2, Mar. 1990, Vol. 89, p. 96 (Figure 7); App. Ex. 15.55, Marto, P. J., et al. "An experimental study of R-113 film condensation on horizontal integral-fin tubes." *Journal of Heat Transfer*, Vol. 112, Aug. 1990, p. 763 (Table 2); App. Ex. 15.6, Strickman, D., et al. "Meteorological effects on the biting activity of Leptoconops americanus (Diptera: Ceratopogonidae)." *Journal of the American Mosquito Control Association*, Vol. II; App. Ex. 15.66, Abou El Naga, H. H., and Salem, A. E. M. "Base oils thermooxidation," *Lubrication Engineering*, Vol. 24, No. 4, Apr. 1986, p. 213. Reprinted by permission of the American Society of Lubrication Engineers. All rights reserved; App. Ex. 15.69, Reprinted from *Microelectronics Reliability*, Vol. 26, No. 1, Hollander, M., Park, D. H., and Proschan, F., "Testing whether F is 'more NBU' than G," p. 43 (Table I), Copyright © 1986, with permission from Elsevier; Table 15.51, Republished with permission, Emerald Group Publishing Limited. *www.emeraldinsight.com/acmm.htm*; Table SIA 15:1, Reprinted from *Chemosphere*, Vol. 20, Nos. 7–9, Schecter, A. et al. "Partitioning of 2,3,7,8-chlorinated dibenzo-p-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans." 954–955 (Tables I and II), Copyright © 1990, with permission from Elsevier.

## CHAPTER 16

App. Ex. 16.1, Holmes, J. S. "Software measurement using SCM," *Software Quality Professional*, Vol. 7, No. 1, Nov. 2004 (Figure 5); App. Ex. 16.1, Jerry Kinard, Western Carolina University; and Brian Kinard, Mississippi State University; App. Ex. 16.5, Grant, E. L., Leavenworth, R. S. *Statistical Quality Control*, 5th ed. New York, McGraw-Hill, 1980 (Table 1–1). Reprinted with permission; App. Ex. 16.11, Grant, E. L., and Leavenworth, R. S. *Statistical Quality Control*, 5th ed. New York, McGraw-Hill, 1980 (Table 1–2). Reprinted with permission; App. Ex. 16.62, Kolarik, W. *Creating Quality Concepts, Systems, Strategies, and Tools*. New York:McGraw-Hill, 1995; Table 16.33, Grant, E. L., and Leavenworth, R. S. *Statistical Quality Control*, 5th ed. New York, McGraw-Hill, 1980 (Table 8–1). Reprinted with permission.

## CHAPTER 17

App. Ex. 17.11, Schmee, J., Gladstein, D., and Nelson, W. "Confidence limits for parameters of a normal distribution from singly-censored samples, using maximum likelihood." *Technometrics*, Vol. 27, No. 2, May 1985, p. 119; App. Ex. 17.18, Nelson, W. "Weibull analysis of reliability data with few or no failures." *Journal of Quality Technology*, Vol. 17, no. 3, July 1985, p. 141 (Table I). © 1985 American Society for Quality Control. Reprinted by permission; Appendix B; Table 5, Abridged from Table 1 of A. Hald, *Statistical Tables and Formulas* (New York Wiley), 1952. Reproduced by permission of A. Hald and the publisher, John Wiley & Sons, Inc.; Table 6, Abridged from W. H. Beyer (ed.) *CRC Standard Mathematical Tables*, 24th edition. (Cleveland: The Chemical Rubber Company) 1976. Reproduced with permission; Table 7, This table is reproduced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.), *The Biometrika Tables for Statisticians*, Vol. 1, 3rd. Ed., Biometrika, 1996; Table 8, Thompson, C. M., "Tables of the Percentage Points of the $x^2$-Distribution," *Biometrika*, 1941, 32, 188–189. Reproduced by permission of the Biometrika Trustees; Table 9, From Merrington, and Thompson, C. M., "Tables of Percentage Points of the Inverted Beta (F)-Distribution." *Biometrika*, 1943, 33, 73–88; Table 13, *Biometrika Tables for Statisticians*, Vol. I, 3rd. Ed., edited by E. S. Pearson and H. O. Hartley (Cambridge University Press, 1966). Reproduced by permission of Professor E. S. Pearson and the Biometrika Trustees; Table 14, *Biometrika Tables for Statisticians*, Vol. I., 3rd. Ed., edited by E. S. Pearson and H. O. Hartley (Cambridge University Press, 1966). Reproduced by permission of Professor E. S. Pearson and the Biometrika Trustees; Table 17, From E. G. Olds, "Distribution of Sums of Squares of Rank Differences for Small Samples." *Annals of Mathematical Statistics*, 1938, 9; Table 19, Reprinted, with permission, from the *ASTM Manual on Quality Control of Materials*, American Society for Testing Materials, copyright ASTM International, 100 Barr Harbor Drive, West Conshohocken, PA 19428; Table 20, From *Techniques of Statistical Analysis* by C. Eisenhart, M. W. Hastay, and W. A. Wallis. Copyright 1947, McGraw-Hill Book Company, Inc. Reproduced with permission of McGraw-Hill; Table 21, Wilfrid J. Dixon and Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3rd ed., McGraw-Hill Book Company, New York, 1969. Used with permission of McGraw-Hill.

STATISTICS

SIXTH EDITION

# STATISTICS

## for Engineering and the Sciences

William M. Mendenhall ▪ Terry L. Sincich

**Statistics for Engineering and the Sciences, Sixth Edition** is designed for a two-semester introductory course on statistics for students majoring in engineering or any of the physical sciences. This popular text continues to teach you the basic concepts of data description and statistical inference as well as the statistical methods necessary for real-world applications. You will understand how to collect and analyze data and think critically about the results.

**New to the Sixth Edition**

- Many new and updated exercises based on contemporary engineering and scientific-related studies and real data
- More statistical software printouts and corresponding instructions for use that reflect the latest versions of the SAS, SPSS, and MINITAB software
- Introduction of the case studies at the beginning of each chapter
- Streamlined material on all basic sampling concepts, such as random sampling and sample survey designs, which gives you an earlier introduction to key sampling issues
- New examples on comparing matched pairs versus independent samples, selecting the sample size for a designed experiment, and analyzing a two-factor experiment with quantitative factors
- New section on using regression residuals to check the assumptions required in a simple linear regression analysis

The first several chapters of the book identify the objectives of statistics, explain how to describe data, and present the basic concepts of probability. The text then introduces the two methods for making inferences about population parameters: estimation with confidence intervals and hypothesis testing. The remaining chapters extend these concepts to cover other topics useful in analyzing engineering and scientific data, including the analysis of categorical data, regression analysis, model building, analysis of variance for designed experiments, nonparametric statistics, statistical quality control, and product and system reliability.

K25936
ISBN: 978-1-4987-2885-0
90000
9 781498 728850
www.crcpress.com

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK