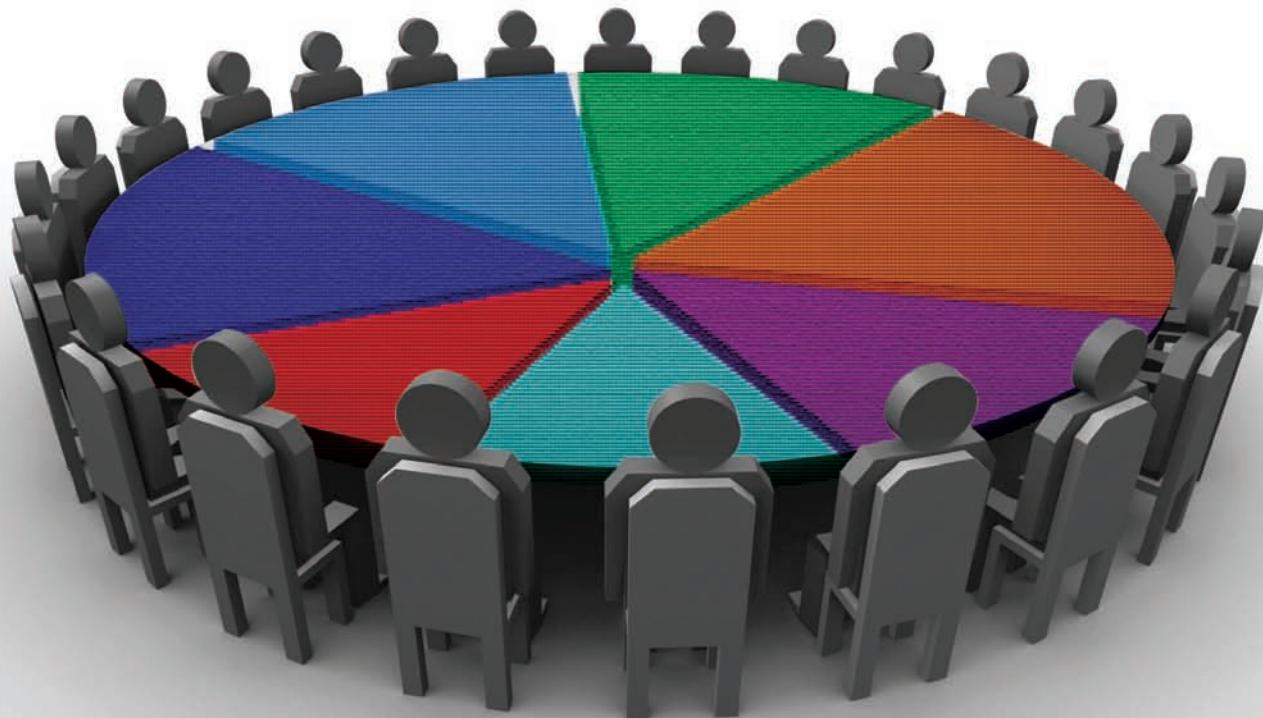


Circulation of this
edition outside the
Indian subcontinent is
UNAUTHORIZED

Seventh Edition

STATISTICS FOR MANAGEMENT

Richard I. Levin | David S. Rubin
Sanjay Rastogi | Masood Husain Siddiqui



S t a t i s t i c s
f o r
MANAGEMENT

This page is intentionally left blank.



Richard I. Levin
The University of
North Carolina at Chapel Hill

David S. Rubin
The University of
North Carolina at Chapel Hill

Sanjay Rastogi
Indian Institute of
Foreign Trade, New Delhi

Masood H. Siddiqui
Jaipuria Institute of
Management, Lucknow

PEARSON

Copyright © 2014 Dorling Kindersley (India) Pvt. Ltd.

Licensees of Pearson Education in South Asia

No part of this eBook may be used or reproduced in any manner whatsoever without the publisher's prior written consent.

This eBook may or may not include all assets that were part of the print version. The publisher reserves the right to remove any material in this eBook at any time.

ISBN 9788131774502

eISBN 9789332535626

Head Office: A-8(A), Sector 62, Knowledge Boulevard, 7th Floor, NOIDA 201 309, India

Registered Office: 11 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

Contents

Preface xiii

CHAPTER 1 Introduction 1

- 1.1 Why Should I Take This Course and Who Uses Statistics Anyhow? 2
- 1.2 History 3
- 1.3 Subdivisions Within Statistics 4
- 1.4 A Simple and Easy-to-Understand Approach 4
- 1.5 Features That Make Learning Easier 5
- 1.6 Surya Bank—Case Study 6

CHAPTER 2 Grouping and Displaying Data to Convey Meaning: Tables and Graphs 13

- 2.1 How Can We Arrange Data? 14
- 2.2 Examples of Raw Data 17
- 2.3 Arranging Data Using the Data Array and the Frequency Distribution 18
- 2.4 Constructing a Frequency Distribution 27
- 2.5 Graphing Frequency Distributions 38
 - Statistics at Work* 58
 - Chapter Review* 59
 - Flow Chart: Arranging Data to Convey Meaning* 72

CHAPTER 3 Measures of Central Tendency and Dispersion in Frequency Distributions 73

- 3.1 Summary Statistics 74
- 3.2 A Measure of Central Tendency: The Arithmetic Mean 77

3.3	A Second Measure of Central Tendency: The Weighted Mean	87
3.4	A Third Measure of Central Tendency: The Geometric Mean	92
3.5	A Fourth Measure of Central Tendency: The Median	96
3.6	A Final Measure of Central Tendency: The Mode	104
3.7	Dispersion: Why It is Important?	111
3.8	Ranges: Useful Measures of Dispersion	113
3.9	Dispersion: Average Deviation Measures	119
3.10	Relative Dispersion: The Coefficient of Variation	132
3.11	Descriptive Statistics Using Msexcel & SPSS	136
	<i>Statistics at Work</i>	140
	<i>Chapter Review</i>	141
	<i>Flow Charts: Measures of Central Tendency and Dispersion</i>	151

CHAPTER 4 Probability I: Introductory Ideas 153

4.1	Probability: The Study of Odds and Ends	154
4.2	Basic Terminology in Probability	155
4.3	Three Types of Probability	158
4.4	Probability Rules	165
4.5	Probabilities Under Conditions of Statistical Independence	171
4.6	Probabilities Under Conditions of Statistical Dependence	179
4.7	Revising Prior Estimates of Probabilities: Bayes' Theorem	189
	<i>Statistics at Work</i>	197
	<i>Chapter Review</i>	199
	<i>Flow Chart: Probability I: Introductory Ideas</i>	208

CHAPTER 5 Probability Distributions 209

5.1	What is a Probability Distribution?	210
5.2	Random Variables	214
5.3	Use of Expected Value in Decision Making	220
5.4	The Binomial Distribution	225
5.5	The Poisson Distribution	238

- 5.6 The Normal Distribution: A Distribution of a Continuous Random Variable 246
- 5.7 Choosing the Correct Probability Distribution 263
 - Statistics at Work* 263
 - Chapter Review* 265
 - Flow Chart: Probability Distribution* 274

CHAPTER 6 Sampling and Sampling Distributions 277

- 6.1 Introduction to Sampling 278
- 6.2 Random Sampling 281
- 6.3 Non-random Sampling 289
- 6.4 Design of Experiments 292
- 6.5 Introduction to Sampling Distributions 296
- 6.6 Sampling Distributions in More Detail 300
- 6.7 An Operational Consideration in Sampling: The Relationship Between Sample Size and Standard Error 313
 - Statistics at Work* 319
 - Chapter Review* 320
 - Flow Chart: Sampling and Sampling Distributions* 326

CHAPTER 7 Estimation 327

- 7.1 Introduction 328
- 7.2 Point Estimates 331
- 7.3 Interval Estimates: Basic Concepts 336
- 7.4 Interval Estimates and Confidence Intervals 341
- 7.5 Calculating Interval Estimates of the Mean from Large Samples 344
- 7.6 Calculating Interval Estimates of the Proportion from Large Samples 349
- 7.7 Interval Estimates Using the *t* Distribution 353
- 7.8 Determining the Sample Size in Estimation 364
 - Statistics at Work* 370
 - Chapter Review* 371
 - Flow Chart: Estimation* 377

CHAPTER 8 Testing Hypotheses: One-sample Tests 379

- 8.1 Introduction 380
- 8.2 Concepts Basic to the Hypothesis-testing Procedure 381
- 8.3 Testing Hypotheses 385
- 8.4 Hypothesis Testing of Means When the Population Standard Deviation is Known 393
- 8.5 Measuring the Power of a Hypothesis Test 402
- 8.6 Hypothesis Testing of Proportions: Large Samples 405
- 8.7 Hypothesis Testing of Means When the Population Standard Deviation is Not Known 411
 - Statistics at Work* 418
 - Chapter Review* 418
- Flow Chart: One-Sample Tests of Hypotheses* 424

CHAPTER 9 Testing Hypotheses: Two-sample Tests 425

- 9.1 Hypothesis Testing for Differences Between Means and Proportions 426
- 9.2 Tests for Differences Between Means: Large Sample Sizes 428
- 9.3 Tests for Differences Between Means: Small Sample Sizes 434
- 9.4 Testing Differences Between Means with Dependent Samples 445
- 9.5 Tests for Differences Between Proportions: Large Sample Sizes 455
- 9.6 Prob Values: Another Way to Look at Testing Hypotheses 464
 - Statistics at Work* 469
 - Chapter Review* 470
- Flow Chart: Two-Sample Tests of Hypotheses* 477

CHAPTER 10 Quality and Quality Control 479

- 10.1 Introduction 480
- 10.2 Statistical Process Control 482
- 10.3 \bar{x} Charts: Control Charts for Process Means 484
- 10.4 R Charts: Control Charts for Process Variability 495
- 10.5 p Charts: Control Charts for Attributes 501

10.6 Total Quality Management	508
10.7 Acceptance Sampling	514
<i>Statistics at Work</i>	522
<i>Chapter Review</i>	523
<i>Flow Chart: Quality and Quality Control</i>	529

CHAPTER 11 Chi-Square and Analysis of Variance 531

11.1 Introduction	532
11.2 Chi-Square as a Test of Independence	533
11.3 Chi-Square as a Test of Goodness of Fit: Testing the Appropriateness of a Distribution	548
11.4 Analysis of Variance	555
11.5 Inferences About a Population Variance	582
11.6 Inferences About Two Population Variances	589
<i>Statistics at Work</i>	597
<i>Chapter Review</i>	598
<i>Flow Chart: Chi-Square and Analysis of Variance</i>	608

CHAPTER 12 Simple Regression and Correlation 609

12.1 Introduction	610
12.2 Estimation Using the Regression Line	617
12.3 Correlation Analysis	643
12.4 Making Inferences About Population Parameters	657
12.5 Using Regression and Correlation Analyses: Limitations, Errors, and Caveats	664
<i>Statistics at Work</i>	667
<i>Chapter Review</i>	667
<i>Flow Chart: Regression and Correlation</i>	676

CHAPTER 13 Multiple Regression and Modeling 677

13.1 Multiple Regression and Correlation Analysis	678
13.2 Finding the Multiple-Regression Equation	679
13.3 The Computer and Multiple Regression	688

13.4 Making Inferences About Population Parameters	698
13.5 Modeling Techniques	717
<i>Statistics at Work</i>	733
<i>Chapter Review</i>	734
<i>Flow Chart: Multiple Regression and Modeling</i>	745

CHAPTER 14 Nonparametric Methods 747

14.1 Introduction to Nonparametric Statistics	748
14.2 The Sign Test for Paired Data	750
14.3 Rank Sum Tests: The Mann–Whitney <i>U</i> Test and the Kruskal–Wallis Test	758
14.4 The One-sample Runs Test	772
14.5 Rank Correlation	781
14.6 The Kolmogorov–Smirnov Test	793
<i>Statistics at Work</i>	800
<i>Chapter Review</i>	801
<i>Flow Chart: Nonparametric Methods</i>	814

CHAPTER 15 Time Series and Forecasting 817

15.1 Introduction	818
15.2 Variations in Time Series	818
15.3 Trend Analysis	820
15.4 Cyclical Variation	832
15.5 Seasonal Variation	838
15.6 Irregular Variation	847
15.7 A Problem Involving All Four Components of a Time Series	848
15.8 Time-Series Analysis in Forecasting	858
<i>Statistics at Work</i>	858
<i>Chapter Review</i>	860
<i>Flow Chart: Time Series</i>	867

CHAPTER 16 Index Numbers 869

- 16.1 Defining an Index Number 870
16.2 Unweighted Aggregates Index 874
16.3 Weighted Aggregates Index 879
16.4 Average of Relatives Methods 888
16.5 Quantity and Value Indices 895
16.6 Issues in Constructing and Using Index Numbers 900
Statistics at Work 901
Chapter Review 902
Flow Chart: Index Numbers 910

CHAPTER 17 Decision Theory 911

- 17.1 The Decision Environment 912
17.2 Expected Profit Under Uncertainty: Assigning Probability Values 913
17.3 Using Continuous Distributions: Marginal Analysis 922
17.4 Utility as a Decision Criterion 931
17.5 Helping Decision Makers Supply the Right Probabilities 935
17.6 Decision-Tree Analysis 939
Statistics at Work 952
Chapter Review 953

Appendix Tables 963**Bibliography 987****Index 991**

This page is intentionally left blank.

Preface

An Opportunity for New Ideas

Writing a new edition of our textbook is an exciting time. In the two years that it takes to complete it, we get to interact with a number of adopters of our text, we benefit from the many thoughtful comments of professors who review the manuscript, our students here at the University of North Carolina at Chapel Hill always have a lot of good ideas for change, and our team at Prentice Hall organizes the whole process and provides a very high level of professional input. Even though this is the seventh edition of our book, our original goal of writing the most teacher- and student-friendly textbook in business statistics still drives our thoughts and our writing in this revision.

What Has Made This Book Different through Six Editions?

Our philosophy about what a good business statistics textbook ought to be hasn't changed since the day we started writing the first edition, twenty years ago. At that time and up through this edition, we have always strived to produce a textbook that met these four goals:

- *We think a beginning business statistics textbook ought to be intuitive and easy to learn from.* In explaining statistical concepts, we begin with what students already know from their life experience and we enlarge on this knowledge by using intuitive ideas. Common sense, real-world ideas, references, patient explanations, multiple examples, and intuitive approaches all make it easier for students to learn.
- *We believe a beginning business statistics textbook ought to cover all of the topics any teacher might wish to build into a two-semester or a two-quarter course.* Not every teacher will cover every topic in our book, but we offer the most complete set of topics for the consideration of anyone who teaches this course.
- *We do not believe that using complex mathematical notation enhances the teaching of business statistics; and our own experience suggests that it may even make learning more difficult.* Complex mathematical notation belongs in advanced courses in mathematics and statistics (and we do use it there), but not here. This is a book that will make and keep you comfortable even if you didn't get an A in college algebra.
- *We believe that a beginning business statistics textbook ought to have a strong realworld focus.* Students ought to see in the book what they see in their world every day. The approach we use, the exercises we have chosen for this edition, and the continuing focus on using statistics to solve business problems all make this book very relevant. We use a large number of real-world problems, and our

explanations tend to be anecdotal, using terms and references that students read in the newspapers, see on TV, and view on their computer monitors. As our own use of statistics in our consulting practices has increased, so have the references to how and why it works in our textbook. This book is about actual managerial situations, which many of the students who use this book will face in a few years.

New Features in This Edition to Make Teaching and Learning Easier

Each of our editions and the supplements that accompanied them contained a complete set of pedagogical aids to make teaching business statistics more effective and learning it less painful. With each revision, we added new ideas, new tools, and new helpful approaches. This edition begins its own set of new features. Here is a quick preview of the twelve major changes in the seventh edition:

- End-of-section exercises have been divided into three subsets: *Basic Concepts*, *Applications*, and *Self-Check Exercises*. The Basic Concepts are those exercises without scenarios, Applications have scenarios, and the Self-Check Exercises have worked-out solutions right in the section.
- The set of Self-Check Exercises referenced above is found at the end of each chapter section except the introductory section. Complete *Worked-Out Answers* to each of these can be found at the end of the applications exercises in that section of the chapter.
- Minitab has been adopted throughout the book as the preferred computer software package.
- *Hints and Assumptions* are short discussions that come at the end of each section in the book, just before the end-of-section exercises. These review important assumptions and tell why we made them, they give students useful hints for working the exercises that follow, and they warn students of potential pitfalls in finding and interpreting solutions.
- The number of real-world examples in the end-of-chapter *Review and Application Exercises* has been doubled, and many of the exercises from the previous edition have been updated.
- Most of the hypothesis tests in Chapters 8 and 9 are done using the standardized scale.
- The scenarios for a quarter of the exercises in this edition have been rewritten.
- Over a hundred new exercises appear in this edition.
- All of the large, multipage data sets have been moved to the data disk, which is available with this book.
- The material on *exploratory data analysis* has been significantly expanded.
- The design of this edition has been completely changed to represent the state of the art in easy-to-follow pedagogy.
- Instructions are provided to handle the data using computer software such as MS Excel and SPSS.
- A Comprehensive Case “*Surya Bank Pvt. Ltd.*” has been added along with the live data. The questions related to this case has been put at the end of each chapter in order to bring more clarity in Statistical Applications in real life scenarios.

Successful Features Retained from Previous Editions

In the time between editions, we listen and learn from teachers who are using our book. The many adopters of our sixth edition reinforced our feeling that these time-tested features should also be a part of the new edition:

- Chapter *learning objectives* are prominently displayed in the chapter opening.
- The more than 1,500 *on-page notes* highlight important material for students.

- Each chapter begins with a *real-world problem*, in which a manager must make a decision. Later in the chapter, we discuss and solve this problem as part of the teaching process.
- Each chapter has a section entitled *review of Terms Introduced* in the chapter.
- An *annotated review of all Equations Introduced* is a part of every chapter.
- Each chapter has a comprehensive *Chapter Concepts Test* using multiple pedagogies.
- A *flow chart* (with numbered page citations) in Chapters 2–16 organizes the material and makes it easier for students to develop a logical, sequential approach to problem solving.
- Our *Statistics at Work* sections in each chapter allow students to think conceptually about business statistics without getting bogged down with data. This learning aid is based on the continuing story of the “Loveland Computer Company” and the experiences of its employees as they bring more and more statistical applications to the management of their business.

Teaching Supplements to the Seventh Edition

The following supplements to the text represent the most comprehensive, classroom-tested set of supplementary teaching aids available in business statistics books today. Together they provide a powerful instructor-focused package.

- An *Instructor’s Solutions Manual* containing worked-out solutions to all of the exercises in the book.
- A comprehensive *online Test Bank Questions*.
- A complete set of *Instructor Lecture Notes*, developed in *Microsoft Powerpoint*.

It Takes a Lot of People to Make a Book

Our part in the process of creating a new edition is to present ideas that we believe work in the classroom. The Prentice Hall team takes these ideas and makes them into a book. Of course, it isn’t that easy.

The whole process starts with our editor, Tom Tucker, who rides herd on the process from his office in St. Paul. Tom is like a movie director; he makes sure everybody plays his or her part and that the entire process moves forward on schedule. Tom guides the project from the day we begin to discuss a seventh edition until the final book version appears on his desk. Without Tom, we’d be rudderless.

Then comes Kelli Rahlf, our production supervisor from Carlisle Publishers Services. In conjunction with Katherine Evancie, our Prentice Hall Production Manager, she manages the thousands of day-to-day activities that must all be completed before a book is produced. Together they move the rough manuscript pages through the editing and printing process, see that printed pages from the compositor reach us, keep us on schedule as we correct and return proofs, work with the bindery and the art folks, and do about a thousand other important things we never get to see but appreciate immensely.

A very helpful group of teachers reviewed the manuscript for the seventh edition and took the time to make very useful suggestions. We are happy to report that we incorporated most of them. This process gives the finished book a student–teacher focus we could not achieve without them; for their effort, we are grateful. The reviewers for this edition were Richard P. Behr, Broome Community College; Ronald L. Coccari, Cleveland State University; V. Reddy Dondeti, Norfolk State University; Mark Haggerty, Clarion University; Robert W. Hull, Western Illinois University; James R. Schmidt, University of Nebraska-Lincoln; and Edward J. Willies.

We use statistical tables in the book that were originally prepared by other folks, and we are grateful to the literary executor of the late Sir Ronald Fisher, F.R.S., to Dr Frank Yates, F.R.S., and to Longman

Group, Ltd., London, for permission to reprint tables from their book *Statistical Tables for Biological, Agricultural and Medical Research*, sixth edition, 1974.

Dr David O. Robinson of the Hass School of Business, Berkeley University, contributed a number of real-world exercises, produced many of the problem scenario changes, and as usual, persuaded us that it would be considerably less fun to revise a book without him.

Kevin Keyes provided a large number of new exercises, and Lisa Klein produced the index. To all of these very important, hard-working folks, we are grateful.

We are glad it is done and now we look forward to hearing from you with your comments about how well it works in your classroom. Thank you for all your help.

R.L.

D.R.

I owe a great deal to my teachers and colleagues from different management institutes for their support, encouragement, and suggestions. Sincere thanks to my student Ashish Awasthi for helping me in preparing the snapshots for SPSS and Microsoft Excel, to my assistant Kirti Yadav for helping me in preparing the manuscript. Finally, I would like to express my gratitude to my parents, special thanks to my wife Subha and my kids Sujay and Sumedha for their love, understanding, and constant support.

Sanjay Rastogi

I want to express my heartfelt and sincere gratitude towards my mother Mrs Ishrat, my wife Usma, my son Ashrat, family members, teachers, friends, colleagues, and Jaipuria Institute of Management, Lucknow, for their help and support in completion of this task.

Masood H. Siddiqui

1 Introduction

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To examine who really uses statistics and how statistics is used
 - To provide a very short history of the use of statistics
 - To present a quick review of the special features of this book that were designed to make learning statistics easier for you
-

CHAPTER CONTENTS

- | | |
|---|--|
| 1.1 Why should I Take This Course and Who Uses Statistics Anyhow? 2 | 1.4 A Simple and Easy-to-Understand Approach 4 |
| 1.2 History 3 | 1.5 Features That Make Learning Easier 5 |
| 1.3 Subdivisions within Statistics 4 | 1.6 Surya Bank—Case Study 6 |

1.1 WHY SHOULD I TAKE THIS COURSE AND WHO USES STATISTICS ANYHOW?

Every 4 years, Americans suffer through an affliction known as the presidential election. Months before the election, television, radio, and newspaper broadcasts inform us that “a poll conducted by XYZ Opinion Research shows that the Democratic (or Republican) candidate has the support of 54 percent of voters with a margin of error of plus or minus 3 percent.” What does this statement mean? What is meant by the term *margin of error*? Who has actually done the polling? How many people did they interview and how many should they have interviewed to make this assertion? Can we rely on the truth of what they reported? Polling is a big business and many companies conduct polls for political candidates, new products, and even TV shows. If you have an ambition to become president, run a company, or even star in a TV show, you need to know something about statistics and statisticians.

It’s the last play of the game and the Giants are behind by 4 points; they have the ball on the Chargers’ 20-yard line. The Chargers’ defensive coordinator calls time and goes over to the sidelines to speak to his coach. The coach knows that because a field goal won’t even tie the game, the Giants will either pass or try a running play. His statistical assistant quickly consults his computer and points out that in the last 50 similar situations, the Giants have passed the ball 35 times. He also points out to the Chargers’ coach that two-thirds of these passes have been short passes, right over center. The Chargers’ coach instructs his defensive coordinator to expect the short pass over center. The ball is snapped, the Giants’ quarterback does exactly what was predicted and there is a double-team Charger effort there to break up the pass. Statistics suggested the right defense.

The Food and Drug Administration is in final testing of a new drug that cures prostate cancer in 80 percent of clinical trials, with only a 2 percent incidence of undesirable side effects. Prostate cancer is the second largest medical killer of men and there is no present cure. The Director of Research must forward a finding on whether to release the drug for general use. She will do that only if she can be more than 99 percent certain that there won’t be any significant difference between undesirable side effects in the clinical tests and those in the general population using the drug. There are statistical methods that can provide her a basis for making this important decision.

The Community Bank has learned from hard experience that there are four factors that go a long way in determining whether a borrower will repay his loan on time or will allow it to go into default. These factors are (1) the number of years at the present address, (2) the number of years in the present job, (3) whether the applicant owns his own home, and (4) whether the applicant has a checking or savings account with the Community Bank. Unfortunately, the bank doesn’t know the individual effect of each of these four factors on the outcome of the loan experience. However, it has computer files full of information on applicants (both those who were granted a loan and those who were turned down) and knows, too, how each granted loan turned out. Sarah Smith applies for a loan. She has lived at her present address 4 years, owns her own home, has been in her current job only 3 months, and is not a Community Bank depositor. Using statistics, the bank can calculate the chance that Sarah will repay her loan on time if it is granted.

The word *statistics* means different things to different folks. To a football fan, statistics are rushing, passing, and first down numbers; to the Chargers’ coach in the second example, statistics is the chance that the Giants will throw the short pass over center. To the manager of a power station, statistics are the amounts of pollution being released into the atmosphere. To the Food and Drug Administrator in our third example, statistics is the likely percentage of undesirable effects in the general population using the new prostate drug. To the Community Bank in the fourth example, statistics is the chance that Sarah

will repay her loan on time. To the student taking this course, statistics are the grades on your quizzes and final exam in the course.

Each of these people is using the word correctly, yet each person uses it in a different way. All of them are using statistics to help them make decisions; you about your grade in this course, and the Chargers' coach about what defense to call for the final play of the game. Helping you learn why statistics is important and how to use it in your personal and professional life is the purpose of this book.

Benjamin Disraeli once said, "There are three kinds of lies: lies, damned lies, and statistics." This rather severe castigation of statistics, made so many years ago, has come to be a rather apt description of many of the statistical deceptions we encounter in our everyday lives. Darrell Huff, in an enjoyable little book, *How to Lie with Statistics*, noted that "the crooks already know these tricks; honest men must learn them in self-defense." One goal of this book is to review some of the common ways statistics are used incorrectly.

How to lie with statistics

1.2 HISTORY

The word *statistik* comes from the Italian word *statista* (meaning "statesman"). It was first used by Gottfried Achenwall (1719–1772), a professor at Marlborough and Göttingen. Dr. E. A. W. Zimmerman introduced the word *statistics* into England. Its use was popularized by Sir John Sinclair in his work *Statistical Account of Scotland 1791–1799*. Long before the eighteenth century, however, people had been recording and using data.

Origin of the word

Official government statistics are as old as recorded history. The Old Testament contains several accounts of census taking. Governments of ancient Babylonia, Egypt, and Rome gathered detailed records of populations and resources. In the Middle Ages, governments began to register the ownership of land. In A.D. 762, Charlemagne asked for detailed descriptions of church-owned properties. Early in the ninth century, he completed a statistical enumeration of the serfs attached to the land. About 1086, William the Conqueror ordered the writing of the *Domesday Book*, a record of the ownership, extent, and value of the lands of England. This work was England's first statistical abstract.

Early government records

Because of Henry VII's fear of the plague, England began to register its dead in 1532. About this same time, French law required the clergy to register baptisms, deaths, and marriages. During an outbreak of the plague in the late 1500s, the English government started publishing weekly death statistics. This practice continued, and by 1632, these *Bills of Mortality* listed births and deaths by sex. In 1662, Captain John Graunt used 30 years of these *Bills* to make predictions about the number of people who would die from various diseases and the proportions of male and female births that could be expected. Summarized in his work *Natural and Political Observations . . . Made upon the Bills of Mortality*, Graunt's study was a pioneer effort in statistical analysis. For his achievement in using past records to predict future events, Graunt was made a member of the original Royal Society.

An early prediction from statistics

The history of the development of statistical theory and practice is a lengthy one. We have only begun to list the people who have made significant contributions to this field. Later we will encounter others whose names are now attached to specific laws and methods. Many people have brought to the study of statistics refinements or innovations that, taken together, form the theoretical basis of what we will study in this book.

1.3 SUBDIVISIONS WITHIN STATISTICS

Managers apply some statistical technique to virtually every branch of public and private enterprise. These techniques are so diverse that statisticians commonly separate them into two broad categories: *descriptive statistics* and *inferential statistics*. Some examples will help us understand the difference between the two.

Suppose a professor computes an average grade for one history class. Because statistics describe the performance of that one class but do not make a generalization about several classes, we can say that the professor is using *descriptive statistics*. Graphs, tables, and charts that display data so that they are easier to understand are all examples of descriptive statistics.

Descriptive statistics

Now suppose that the history professor decides to use the average grade achieved by one history class to estimate the average grade achieved in all ten sections of the same history course. The process of estimating this average grade would be a problem in *inferential statistics*. Statisticians also refer to this category as *statistical inference*. Obviously, any conclusion the professor makes about the ten sections of the course is based on a generalization that goes far beyond the data for the original history class; the generalization may not be completely valid, so the professor must state how likely it is to be true. Similarly, statistical inference involves generalizations and statements about the *probability* of their validity.

Inferential statistics

The methods and techniques of statistical inference can also be used in a branch of statistics called *decision theory*. Knowledge of decision theory is very helpful for managers because it is used to make decisions under conditions of uncertainty when, for example, a manufacturer of stereo sets cannot specify precisely the demand for its products or when the chairperson of the English department at your school must schedule faculty teaching assignments without knowing precisely the student enrollment for next fall.

Decision theory

1.4 A SIMPLE AND EASY-TO-UNDERSTAND APPROACH

This book is designed to help you get the feel of statistics: what it is, how and when to apply statistical techniques to decision-making situations, and how to interpret the results you get. Because we are not writing for professional statisticians, our writing is tailored to the backgrounds and needs of college students, who probably accept the fact that statistics can be of considerable help to them in their future occupations but are probably apprehensive about studying the subject.

For students, not statisticians

We discard mathematical proofs in favor of intuitive ones. You will be guided through the learning process by reminders of what you already know, by examples with which you can identify, and by a step-by-step process instead of statements such as “it can be shown” or “it therefore follows.”

Symbols are simple and explained

As you thumb through this book and compare it with other basic business statistics textbooks, you will notice a minimum of mathematical notation. In the past, the complexity of the notation has intimidated many students, who got lost in the symbols even though they were motivated and intellectually capable of understanding the ideas. Each symbol and formula that is used is explained in detail, not only at the point at which it is introduced, but also in a section at the end of the chapter.

If you felt reasonably comfortable when you finished your high school algebra course, you have enough background to understand *everything* in this book. Nothing beyond basic algebra is assumed or used. Our goals are for you to be comfortable as you learn and for you to get a good intuitive grasp of statistical concepts and techniques. As a future manager, you will need to know when statistics can help your decision process and which tools to use. If you do need statistical help, you can find a statistical expert to handle the details.

No math beyond simple algebra is required

The problems used to introduce material in the chapters, the exercises at the end of each section in the chapter, and the chapter review exercises are drawn from a wide variety of situations you are already familiar with or are likely to confront quite soon. You will see problems involving all facets of the private sector of our economy: accounting, finance, individual and group behavior, marketing, and production. In addition, you will encounter managers in the public sphere coping with problems in public education, social services, the environment, consumer advocacy, and health systems.

Text problem cover a wide variety of situations

In each problem situation, a manager is trying to use statistics creatively and productively. Helping you become comfortable doing exactly that is our goal.

1.5 FEATURES THAT MAKE LEARNING EASIER

In our preface, we mentioned briefly a number of learning aids that are a part of this book. Each has a particular role in helping you study and understand statistics, and if we spend a few minutes here discussing the most effective way to use some of these aids, you will not only learn more effectively, but will gain a greater understanding of how statistics is used to make managerial decisions.

Margin Notes Each of the more than 1,500 margin notes highlights the material in a paragraph or group of paragraphs. Because the notes briefly indicate the focus of the textual material, you can avoid having to read through pages of information to find what you need. Learn to read down the margin as you work through the textbook; in that way, you will get a good sense of the flow of topics and the meaning of what the text is explaining.

Application Exercises The Chapter Review Exercises include Application Exercises that come directly from real business/economic situations. Many of these are from the business press; others come from government publications. This feature will give you practice in setting up and solving problems that are faced every day by business professionals. In this edition, the number of Application Exercises has been doubled.

Review of Terms Each chapter ends with a glossary of every new term introduced in that chapter. Having all of these new terms defined again in one convenient place can be a big help. As you work through a chapter, use the glossary to reinforce your understanding of what the terms mean. Doing this is easier than going back in the chapter trying to find the definition of a particular term. When you finish studying a chapter, use the glossary to make sure you understand what each term introduced in the chapter means.

Equation Review Every equation introduced in a chapter is found in this section. All of them are explained again, and the page on which they were first introduced is given. Using this feature of the book is a very effective way to make sure you understand what each equation means and how it is used.

Chapter Concepts Test Using these tests is a good way to see how well you understand the chapter material. As a part of your study, be sure to take these tests and then compare your answers with those in the back of the book. Doing this will point out areas in which you need more work, especially before quiz time.

Statistics at Work In this set of cases, an employee of Loveland Computers applies statistics to managerial problems. The emphasis here is not on numbers; in fact, it's hard to find any numbers in these cases. As you read each of these cases, focus on what the problem is and what statistical approach might help find a solution; forget the numbers temporarily. In this way, you will develop a good appreciation for identifying problems and matching solution methods with problems, without being bogged down by numbers.

Flow Chart The flow charts at the end of the chapters will enable you to develop a systematic approach to applying statistical methods to problems. Using them helps you understand where you begin, how you proceed, and where you wind up; if you get good at using them, you will not get lost in some of the more complex word problems instructors are fond of putting on tests.

From the Textbook to the Real World Each of these will take you no more than 2 or 3 minutes to read, but doing so will show you how the concepts developed in this book are used to solve real-world problems. As you study each chapter, be sure to review the "From the Textbook to the Real World" example; see what the problem is, how statistics solves it, and what the solution adds in value. These situations also generate good classroom discussion questions.

Classification of Exercises This feature is new with this edition of the book. The exercises at the end of each section are divided into three categories: basic concepts to get started on, application exercises to show how statistics is used, and self-check exercises with worked-out answers to allow you to test yourself.

Self-Check Exercises with Worked-Out Answers A new feature in this edition. At the beginning of most sets of exercises, there are one or two self-check exercises for you to test yourself. The worked-out answers to these self-check exercises appear at the end of the exercise set.

Hints and Assumptions New with this edition, these provide help, direction, and things to avoid before you begin work on the exercises at the end of each section. Spending a minute reading these saves lots of time, frustration, and mistakes in working the exercises.

1.6 SURYA BANK—CASE STUDY

SURYA BANK PVT. LTD. was incorporated in the first quarter of the Twentieth Century in Varanasi, by a group of ambitious and enterprising Entrepreneurs. Over the period of time, the Bank with its untiring customer services has earned a lot of trust and goodwill of its customers. The staff and the management of the bank had focused their attention on the customers from the very inceptions of the bank. It is the practice of the bank that its staff members would go out to meet the customers of various walks of life and enquire about their banking requirements on the regular basis. It was due to the bank's strong belief in the need for innovation, delivering the best service and demonstrating responsibility that had helped the bank in growing from strength to strength.

The bank had only 6 branches till 1947. Post-independence, the bank expanded and now has 198 full-fledged branches across the North, North-West and Central India, dotted across the rural, semi-urban and urban areas.

SURYA BANK PVT. LTD. concentrated on its efforts to meet the genuine requirements of the different sectors of business and was forthcoming in giving loans to the needy & weaker sections of the society. The bank also has a sound portfolio of advances consisting of wide basket of retail finance. As a matter of policy, SURYA BANK PVT. LTD. gives loans to a large spectrum of retail businessmen.

In 2011, the bank had a net-profit of ₹ 26.3 crores. The total income of the bank has been steadily increasing over the past one decade from ₹ 188.91 crores in 2000 to ₹ 610.19 crores in 2011. The financial results of the bank are given below:

SURYA BANK FINANCIAL RESULTS

Sl. No.	Financial Year	Net Profit	Total Income	Operating Expenses
1	2000	10.24	188.91	35.62
2	2001	5.37	203.28	49.03
3	2002	9.33	240.86	50.97
4	2003	14.92	258.91	97.42
5	2004	17.07	250.07	99.20
6	2005	-20.1	204.19	80.72
7	2006	10.39	237.33	86.68
8	2007	16.55	280.64	96.52
9	2008	33.01	361.51	95.23
10	2009	58.17	486.67	114.74
11	2010	23.92	625.94	194.10
12	2011	26.30	610.19	202.14

SURYA BANK PVT. LTD. is one of the first private sector banks in India to introduce a massive computerization at branch level. The bank adopted modernization and computerization as early as 1990. All its 198 branches are computerized. The bank operates around 400 ATMs across northern India. This computerization has enabled the bank to render better and efficient service to its customers.

The bank is implementing new technology in core bank on an ongoing basis so as to achieve higher customer satisfaction and better retention to the customers. The bank has embarked upon a scheme of total branch automation with centralized Data Base System to integrate all its branches. This scheme has helped the bank to implement newer banking modes like internet banking, cyber banking and mobile banking etc, which has helped the customers to access the banking account from their place of work. The bank in its endeavor to provide quality service to its customers has been constantly improvising its services for the satisfaction of its customers. To better understand the customers' needs and wants and of its customers and the level of satisfaction with respect to the services provided to its customers, Surya Bank has conducted a survey of the bank customers to understand their opinions/perceptions with respect to the services provided by the bank.

NOTE: This case is prepared for class discussion purpose only. The information provided is hypothetical, but the questionnaire and the data set are real.

Questionnaire

Q. 1 Do You have an account in any bank, If yes
name of the bank

Q. 2 Which type of account do you have,
Saving
Current
Both

Q. 3 For how long have had the bank account
< 1 year
2-3 year
3-5 year
5-10 year
>10 year

Q. 4 Rank the following modes in terms of the extent to which they helped you know about e-banking services on scale 1 to 4

	Least important	Slightly important	Important	Most important
(a) Advertisement				
(b) Bank Employee				
(c) Personal enquiry				
(d) Friends or relative				

Q. 5 How frequently do you use e-banking
Daily
2-3 times in a week
Every week
Fort nightly
Monthly
Once in a six month
Never

Q. 6 Rate the add-on services which are available in your e-banking account on scale 1 to 5

	Highly unavailable	Available	Moderate	Available	Highly available
(a) Seeking product & rate information					
(b) Calculate loan payment information					
(c) Balance inquiry					
(d) Inter account transfers					
(e) Lodge complaints					
(f) To get general information					
(g) Pay bills					
(h) Get in touch with bank					

Q. 7 Rate the importance of the following e-banking facilities while selecting a bank on the scale 1 to 4

	Least important	Slightly important	If important	Most important
(a) Speed of transaction				
(b) Reliability				
(c) Ease of use				
(d) Transparency				
(e) 24×7 any time banking				
(f) Congestion				
(g) Lower amount transactions are not possible				
(h) Add on services and schemes				
(i) Information retrieval				
(j) Ease of contact				
(k) Safety				
(l) Privacy				
(m) Accessibility				

Q. 8 Rate the level of satisfaction of the following e-banking facilities of your bank on the scale 1 to 4

	Highly dissatisfied	Dissatisfied	Satisfied	Highly satisfied
(a) Speed of transaction				
(b) Reliability				
(c) Ease of use				
(d) Transparency				
(e) 24×7 any time banking				
(f) Congestion				
(g) Lower amount transactions are not possible				
(h) Add on services and schemes				
(i) Information retrieval				
(j) Ease of contact				
(k) Safety				
(l) Privacy				
(m) Accessibility				

Q. 9	Rate the level of satisfaction with e-services provided by your bank	Highly dissatisfied Dissatisfied Satisfied Highly satisfied	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Q. 10	How frequently you find problem in using the e-banking	Daily Monthly 2–3 times in a week Once in a six month Every week Nightly never	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Q. 11 Rate the following problems you have faced frequently using e-banking.

	Least faced	Slightly faced	Faced	Frequently faced
(a) Feel it is unsecured mode of transaction				
(b) Misuse of information				
(c) Slow transaction				
(d) No availability of server				
(e) Not a techno savvy				
(f) Increasingly expensive and time consuming				
(g) Low direct customer connection				

Q. 12 How promptly your problems have been solved	Instantly Within a week 10–15 days Within a month	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
--	--	--

Q. 13 Rate the following statements for e-banking facility according to your agreement level on scale 1–5

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
(a) It saves a person's time					
(b) Private banks are better than public banks					
(c) This facility was first initiated by private banks so they have an edge over the public banks					

(Continued)

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
(d) Information provided by us is misused					
(e) It is good because we can access our bank account from anywhere in the world					
(f) It makes money transfer easy and quick					
(g) It is an important criterion to choose a bank to open an account					
(h) Limited use of this is due to lack of awareness					
(i) Complaint handling through e-banking is better by private banks than public banks					
(j) Banks provides incentives to use it					
(k) This leads to lack of personal touch					
(l) People do not use e-banking because of extra charge					

- Q. 14** Age in years 20–30 years 31–45 years 45–60 years >60 years
- Q. 15** Gender Male Female
- Q. 16** Marital Status Single Married
- Q. 17** Education Intermediate Graduate Postgraduate Professional course
- Q. 18** Profession Student Employed in private sector Employed in Govt sector Professional Self employed House wife

Q. 19	Monthly Personal Income in INR	<10,000	<input type="checkbox"/>
		10,000–20,000	<input type="checkbox"/>
		20,001–35,000	<input type="checkbox"/>
		35,001–50,000	<input type="checkbox"/>
		>50,000	<input type="checkbox"/>

Our own work experience has brought us into contact with thousands of situations where statistics helped decision makers. We participated personally in formulating and applying many of those solutions. It was stimulating, challenging, and, in the end, very rewarding as we saw sensible application of these ideas produce value for organizations. Although very few of you will likely end up as statistical analysts, we believe very strongly that you can learn, develop, and have fun studying statistics, and that's why we wrote this book. Good luck!

The authors' goals

Grouping and Displaying Data to Convey Meaning: Tables and Graphs

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To show the difference between samples and populations
 - To convert raw data to useful information
 - To construct and use data arrays
 - To construct and use frequency distributions
 - To graph frequency distributions with histograms, polygons, and ogives
 - To use frequency distributions to make decisions
-

CHAPTER CONTENTS

- 2.1 How Can We Arrange Data? 14
- 2.2 Examples of Raw Data 17
- 2.3 Arranging Data Using the Data Array and the Frequency Distribution 18
- 2.4 Constructing a Frequency Distribution 27
- 2.5 Graphing Frequency Distributions 38

- Statistics at Work 58
- Terms Introduced in Chapter 2 59
- Equations Introduced in Chapter 2 60
- Review and Application Exercises 60
- Flow Chart: Arranging Data to Convey Meaning 72

The production manager of the Dalmon Carpet Company is responsible for the output of over 500 carpet looms. So that he does not have to measure the daily output (in yards) of each loom, he samples the output from 30 looms each day and draws a conclusion as to the average carpet production of the entire 500 looms. The table below shows the yards produced by each of the 30 looms in yesterday's sample. These production amounts are the raw data from which the production manager can draw conclusions about the entire population of looms yesterday.

YARDS PRODUCED YESTERDAY BY EACH OF 30 CARPET LOOMS

16.2	15.4	16.0	16.6	15.9	15.8	16.0	16.8	16.9	16.8
15.7	16.4	15.2	15.8	15.9	16.1	15.6	15.9	15.6	16.0
16.4	15.8	15.7	16.2	15.6	15.9	16.3	16.3	16.0	16.3

Using the methods introduced in this chapter, we can help the production manager draw the right conclusion. ■

Data are collections of any number of related observations. We can collect the number of telephones that several workers install on a given day or that one worker installs per day over a period of several days, and we can call the results our data. A collection of data is called a *data set*, and a single observation a *data point*.

Some definitions

2.1 HOW CAN WE ARRANGE DATA?

For data to be useful, our observations must be organized so that we can pick out patterns and come to logical conclusions. This chapter introduces the techniques of arranging data in tabular and graphical forms. Chapter 3 shows how to use numbers to describe data.

Collecting Data

Statisticians select their observations so that all relevant groups are represented in the data. To determine the potential market for a new product, for example, analysts might study 100 consumers in a certain geographical area. Analysts must be certain that this group contains people representing variables such as income level, race, education, and neighborhood.

Represent all groups

Data can come from actual observations or from records that are kept for normal purposes. For billing purposes and doctors' reports, a hospital, for example, will record the number of patients using the X-ray facilities. But this information can also be organized to produce data that statisticians can describe and interpret.

Find data by observation or from records

Data can assist decision makers in educated guesses about the causes and therefore the probable effects of certain characteristics in given situations. Also, knowledge of trends from past experience can enable concerned citizens to be aware of potential outcomes and to plan in advance. Our marketing survey may reveal that the product is preferred by African-American homemakers of suburban communities, average incomes, and average education. This product's advertising copy should address this target audience. If hospital records show that more patients used the X-ray facilities in June than in January, the hospital personnel

Use data about the past to make decisions about the future

division should determine whether this was accidental to this year or an indication of a trend, and perhaps it should adjust its hiring and vacation practices accordingly.

When data are arranged in compact, usable forms, decision makers can take reliable information from the environment and use it to make intelligent decisions. Today, computers allow statisticians to collect enormous volumes of observations and compress them instantly into tables, graphs, and numbers. These are all compact, usable forms, but are they reliable? Remember that the data that come out of a computer are only as accurate as the data that go in. As computer programmers say, “GIGO,” or “Garbage In, Garbage Out.” Managers must be very careful to be sure that the data they are using are based on correct assumptions and interpretations. Before relying on any interpreted data, from a computer or not, test the data by asking these questions:

1. Where did the data come from? Is the source biased—that is, is it likely to have an interest in supplying data points that will lead to one conclusion rather than another?
2. Do the data support or contradict other evidence we have?
3. Is evidence missing that might cause us to come to a different conclusion?
4. How many observations do we have? Do they represent all the groups we wish to study?
5. Is the conclusion logical? Have we made conclusions that the data do not support?

Tests for data

Study your answers to these questions. Are the data worth using? Or should we wait and collect more information before acting? If the hospital was caught short-handed because it hired too few nurses to staff the X-ray room, its administration relied on insufficient data. If the advertising agency targeted its copy only toward African-American suburban home makers when it could have tripled its sales by appealing to white suburban homemakers, too, it also relied on insufficient data. In both cases, testing available data would have helped managers make better decisions.

The effect of incomplete or biased data can be illustrated with this example. A national association of truck lines claimed in an advertisement that “75 percent of everything you use travels by truck.” This might lead us to believe that cars, railroads, airplanes, ships, and other forms of transportation carry only 25 percent of what we use. Reaching such a conclusion is easy but not enlightening. Missing from the trucking assertion is the question of double counting. What did they do when something was carried to your city by rail and delivered to your house by truck? How were packages treated if they went by airmail and then by truck? When the double-counting issue (a very complex one to treat) is resolved, it turns out that trucks carry a much lower proportion of the goods you use than truckers claimed. Although trucks are involved in *delivering* a relatively high proportion of what you use, railroads and ships still carry more goods for more total miles.

Double-counting example

Difference between Samples and Populations

Statisticians gather data from a sample. They use this information to make inferences about the population that the sample represents. Thus, a population is a whole, and a sample is a fraction or segment of that whole.

Sample and population defined

We will study samples in order to be able to describe populations. Our hospital may study a small, representative group of X-ray records rather than examining each record for the last 50 years. The Gallup Poll may interview a sample of only 2,500 adult Americans in order to predict the opinion of all adults living in the United States.

Function of samples

Studying samples is easier than studying the whole population; it costs less and takes less time. Often, testing an airplane part for strength destroys the part; thus, testing fewer parts is desirable. Sometimes testing involves human risk; thus, use of sampling reduces that risk to an acceptable level. Finally, it has been proven that examining an entire population still allows defective items to be accepted; thus, sampling, in some instances, can *raise* the quality level. If you're wondering how that can be so, think of how tired and inattentive you might get if you had to look at thousands and thousands of items passing before you.

Advantages of samples

A *population* is a collection of all the elements we are studying and about which we are trying to draw conclusions. We must define this population so that it is clear whether an element is a member of the population. The population for our marketing study may be all women within a 15-mile radius of center-city Cincinnati who have annual family incomes between \$20,000 and \$45,000 and have completed at least 11 years of school. A woman living in downtown Cincinnati with a family income of \$25,000 and a college degree would be a part of this population. A woman living in San Francisco, or with a family income of \$7,000, or with 5 years of schooling would not qualify as a member of this population.

Function of populations

A *sample* is a collection of some, but not all, of the elements of the population. The population of our marketing survey is *all* women who meet the qualifications listed above. Any group of women who meet these qualifications can be a sample, as long as the group is only a fraction of the whole population. A large helping of cherry filling with only a few crumbs of crust is a sample of pie, but it is not a representative sample because the proportions of the ingredients are not the same in the sample as they are in the whole.

Need for a representative sample

A *representative sample* contains the relevant characteristics of the population *in the same proportions* as they are included in that population. If our population of women is one-third African-American, then a sample of the population that is representative in terms of race will also be one-third African-American. Specific methods for sampling are covered in detail in Chapter 6.

Finding a Meaningful Pattern in the Data

There are many ways to sort data. We can simply collect them and keep them in order. Or if the observations are measured in numbers, we can list the data points from lowest to highest in numerical value. But if the data are skilled workers (such as carpenters, masons, and ironworkers) at construction sites, or the different types of automobiles manufactured by all automakers, or the various colors of sweaters manufactured by a given firm, we must organize them differently. We must present the data points in alphabetical order or by some other organizing principle. One useful way to organize data is to divide them into similar categories or classes and then count the number of observations that fall into each category. This method produces a *frequency distribution* and is discussed later in this chapter.

Data come in a variety of forms

The purpose of organizing data is to enable us to see quickly some of the characteristics of the data we have collected. We look for things such as the range (the largest and smallest values), apparent patterns, what values the data may tend to group around, what values appear most often, and so on. The more information of this kind that we can learn from our sample, the better we can understand the population from which it came, and the better we can make decisions.

Why should we arrange data?

EXERCISES 2.1

Applications

- 2-1** When asked what they would use if they were marooned on an island with only one choice for a pain reliever, more doctors chose Bayer than Tylenol, Bufferin, or Advil. Is this conclusion drawn from a sample or a population?
- 2-2** Twenty-five percent of the cars sold in the United States in 1996 were manufactured in Japan. Is this conclusion drawn from a sample or a population?
- 2-3** An electronics firm recently introduced a new amplifier, and warranty cards indicate that 10,000 of these have been sold so far. The president of the firm, very upset after reading three letters of complaint about the new amplifiers, informed the production manager that costly control measures would be implemented immediately to ensure that the defects would not appear again. Comment on the president's reaction from the standpoint of the five tests for data given on page 15.
- 2-4** "Germany will remain ever divided" stated Walter Ulbricht after construction of the Berlin Wall in 1961. However, toward the end of 1969, the communists of East Germany began allowing free travel between the east and west, and twenty years after that, the wall was completely destroyed. Give some reasons for Ulbricht's incorrect prediction.
- 2-5** Discuss the data given in the chapter-opening problem in terms of the five tests for data given on page 15.

2.2 EXAMPLES OF RAW DATA

Information before it is arranged and analyzed is called *raw data*. It is "raw" because it is unprocessed by statistical methods.

The carpet-loom data in the chapter-opening problem was one example of raw data. Consider a second. Suppose that the admissions staff of a university, concerned with the success of the students it selects for admission, wishes to compare the students' college performances with other achievements, such as high school grades, test scores, and extracurricular activities. Rather than study every student from every year, the staff can draw a sample of the population of all the students in a given time period and study only that group to conclude what characteristics appear to predict success. For example, the staff can compare high school grades with college grade-point averages (GPAs) for students in the sample. The staff can assign each grade a numerical value. Then it can add the grades and divide by the total number of grades to get an average for each student. Table 2-1 shows a sample of these raw data in tabular form: 20 pairs of average grades in high school and college.

Problem facing admissions staff

TABLE 2-1 HIGH SCHOOL AND COLLEGE GRADE-POINT AVERAGES OF 20 COLLEGE SENIORS

H.S.	College	H.S.	College	H.S.	College	H.S.	College
3.6	2.5	3.5	3.6	3.4	3.6	2.2	2.8
2.6	2.7	3.5	3.8	2.9	3.0	3.4	3.4
2.7	2.2	2.2	3.5	3.9	4.0	3.6	3.0
3.7	3.2	3.9	3.7	3.2	3.5	2.6	1.9
4.0	3.8	4.0	3.9	2.1	2.5	2.4	3.2

TABLE 2-2 POUNDS OF PRESSURE PER SQUARE INCH THAT CONCRETE CAN WITHSTAND

2500.2	2497.8	2496.9	2500.8	2491.6	2503.7	2501.3	2500.0
2500.8	2502.5	2503.2	2496.9	2495.3	2497.1	2499.7	2505.0
2490.5	2504.1	2508.2	2500.8	2502.2	2508.1	2493.8	2497.8
2499.2	2498.3	2496.7	2490.4	2493.4	2500.7	2502.0	2502.5
2506.4	2499.9	2508.4	2502.3	2491.3	2509.5	2498.4	2498.1

When designing a bridge, engineers are concerned with the stress that a given material, such as concrete, will withstand. Rather than test every cubic inch of concrete to determine its stress capacity, engineers take a sample of the concrete, test it, and conclude how much stress, on the average, that kind of concrete can withstand. Table 2-2 summarizes the raw data gathered from a sample of 40 batches of concrete to be used in constructing a bridge.

Bridge-building problem**HINTS & ASSUMPTIONS**

Data are *not* necessarily information, and having more data doesn't necessarily produce better decisions. The goal is to summarize and present data in useful ways to support prompt and effective decisions. The reason we have to organize data is to see whether there are patterns in them, patterns such as the largest and smallest values, and what value the data seem to cluster around. If the data are from a sample, we assume that they fairly represent the population from which they were drawn. All good statisticians (and users of data) recognize that using biased or incomplete data leads to poor decisions.

EXERCISES 2.2**Applications**

- 2-6 Look at the data in Table 2-1. Why do these data need further arranging? Can you form any conclusions from the data as they exist now?
- 2-7 The marketing manager of a large company receives a report each month on the sales activity of one of the company's products. The report is a listing of the sales of the product by state during the previous month. Is this an example of raw data?
- 2-8 The production manager in a large company receives a report each month from the quality control section. The report gives the reject rate for the production line (the number of rejects per 100 units produced), the machine causing the greatest number of rejects, and the average cost of repairing the rejected units. Is this an example of raw data?

2.3 ARRANGING DATA USING THE DATA ARRAY AND THE FREQUENCY DISTRIBUTION

The *data array* is one of the simplest ways to present data. It arranges values in ascending or descending order. Table 2-3 repeats the carpet data from our chapter-opening problem, and Table 2-4 rearranges these numbers in a data array in ascending order.

Data array defined

TABLE 2-3 SAMPLE OF DAILY PRODUCTION IN YARDS OF 30 CARPET LOOMS

16.2	15.8	15.8	15.8	16.3	15.6
15.7	16.0	16.2	16.1	16.8	16.0
16.4	15.2	15.9	15.9	15.9	16.8
15.4	15.7	15.9	16.0	16.3	16.0
16.4	16.6	15.6	15.6	16.9	16.3

TABLE 2-4 DATA ARRAY OF DAILY PRODUCTION IN YARDS OF 30 CARPET LOOMS

15.2	15.7	15.9	16.0	16.2	16.4
15.4	15.7	15.9	16.0	16.3	16.6
15.6	15.8	15.9	16.0	16.3	16.8
15.6	15.8	15.9	16.1	16.3	16.8
15.6	15.8	16.0	16.2	16.4	16.9

Data arrays offer several advantages over raw data:

Advantages of data arrays

- We can quickly notice the lowest and highest values in the data.** In our carpet example, the range is from 15.2 to 16.9 yards.
- We can easily divide the data into sections.** In Table 2-4, the first 15 values (the lower half of the data) are between 15.2 and 16.0 yards, and the last 15 values (the upper half) are between 16.0 and 16.9 yards. Similarly, the lowest third of the values range from 15.2 to 15.8 yards, the middle third from 15.9 to 16.2 yards, and the upper third from 16.2 to 16.9 yards.
- We can see whether any values appear more than once in the array.** Equal values appear together. Table 2-4 shows that nine levels occurred more than once when the sample of 30 looms was taken.
- We can observe the distance between succeeding values in the data.** In Table 2-4, 16.6 and 16.8 are succeeding values. The distance between them is 0.2 yards (16.8–16.6).

In spite of these advantages, sometimes a data array isn't helpful. Because it lists every observation, it is a cumbersome form for displaying large quantities of data. We need to compress the information and still be able to use it for interpretation and decision making. How can we do this?

Disadvantages of data arrays

A Better Way to Arrange Data: The Frequency Distribution

One way we can compress data is to use a *frequency table* or a *frequency distribution*. To understand the difference between this and an array, take as an example the average inventory (in days) for 20 convenience stores:

Frequency distributions handle more data

In Tables 2-5 and 2-6, we have taken identical data concerning the average inventory and displayed them first as an array in ascending order and then as a frequency distribution. To obtain Table 2-6, we had to divide the data in groups of similar values. Then we recorded the number of data points that fell into each group. Notice that we lose some information in constructing the frequency distribution. We no longer know, for example, that the value 5.5 appears four times or that the value 5.1 does not appear at all. Yet we gain information concerning the *pattern* of average inventories. We can see from Table 2-6 that average inventory falls most often in the range from 3.8 to 4.3 days. It is unusual to find an average inventory in the range from 2.0

They lose some information

But they gain other information

TABLE 2-5 DATA ARRAY OF AVERAGE INVENTORY (IN DAYS) FOR 20 CONVENIENCE STORES

2.0	3.8	4.1	4.7	5.5
3.4	4.0	4.2	4.8	5.5
3.4	4.1	4.3	4.9	5.5
3.8	4.1	4.7	4.9	5.5

TABLE 2-6 FREQUENCY DISTRIBUTION OF AVERAGE INVENTORY (IN DAYS) FOR 20 CONVENIENCE STORES (6 CLASSES)

Class (Group of Similar Values of Data Points)	Frequency (Number of Observations in Each Class)
2.0 to 2.5	1
2.6 to 3.1	0
3.2 to 3.7	2
3.8 to 4.3	8
4.4 to 4.9	5
5.0 to 5.5	4

to 2.5 days or from 2.6 to 3.1 days. Inventories in the ranges of 4.4 to 4.9 days and 5.0 to 5.5 days are not prevalent but occur more frequently than some others. Thus, frequency distributions sacrifice some detail but offer us new insights into patterns of data.

A frequency distribution is a table that organizes data into classes, that is, into groups of values describing one characteristic of the data. The average inventory is one characteristic of the 20 convenience stores. In Table 2-5, this characteristic has 11 different values. Table 2-6, for example, uses 6. We could compress the data even further and use only 2 classes: less than 3.8 and greater than or equal to 3.8. Or we could increase the number of classes by using smaller intervals, as we have done in Table 2-7.

A frequency distribution shows **the number of observations from the data set that fall into each of the classes**. If you can determine the frequency with which values occur in each class of a data set, you can construct a frequency distribution.

Function of classes in a frequency distribution

values. But these same data could be divided into any number of classes. Table 2-6, for example, uses 6. We could compress the data even further and use only 2 classes: less than 3.8 and greater than or equal to 3.8. Or we could increase the number of classes by using smaller intervals, as we have done in Table 2-7.

Why is it called a frequency distribution?

Characteristics of Relative Frequency Distributions

So far, we have expressed the frequency with which values occur in each class as the total number of data points that fall within that class. We can also express the frequency of each value as a *fraction* or a *percentage* of the total number of observations. The frequency of an average inventory of 4.4 to 4.9 days, for example, is 5 in Table 2-6 but 0.25 in Table 2-8. To get this value of 0.25, we divided the frequency for that class (5) by the total number of observations in the data set (20). The answer can be expressed as a fraction ($\frac{5}{20}$) a decimal (0.25), or a percentage (25 percent). A *relative frequency distribution* presents frequencies in terms of fractions or percentages.

Relative frequency distribution defined

Notice in Table 2-8 that the sum of all the relative frequencies equals 1.00, or 100 percent. This is true because a relative frequency distribution pairs each class with its appropriate fraction or percentage of the total data. Therefore, the classes in any relative or simple frequency distribution are *all-inclusive*. All the data fit into one category or another. Also notice

*Classes are all-inclusive
They are mutually exclusive*

TABLE 2-8 RELATIVE FREQUENCY DISTRIBUTION OF AVERAGE INVENTORY (IN DAYS) FOR 20 CONVENIENCE STORES

Class	Frequency	Relative Frequency: Fraction of Observations in Each Class
2.0 to 2.5	1	0.05
2.6 to 3.1	0	0.00
3.2 to 3.7	2	0.10
3.8 to 4.3	8	0.40
4.4 to 4.9	5	0.25
5.0 to 5.5	<u>4</u>	<u>0.20</u>
	20	1.00
		(sum of the relative frequencies of all classes)

TABLE 2-9 MUTUALLY EXCLUSIVE AND OVERLAPPING CLASSES

Mutually exclusive	1 to 4	5 to 8	9 to 12	13 to 16
Not mutually exclusive	1 to 4	3 to 6	5 to 8	7 to 10

that the classes in Table 2-8 are *mutually exclusive*; that is, no data point falls into more than one category. Table 2-9 illustrates this concept by comparing mutually exclusive classes with ones that overlap. In frequency distributions, there are no overlapping classes.

Up to this point, our classes have consisted of numbers and have described some quantitative attribute of the items sampled. We can also classify information according to qualitative characteristics, such as race, religion, and gender, which do not fall naturally into numerical categories. Like classes of quantitative attributes, these classes must be all-inclusive and mutually exclusive. Table 2-10 shows how to construct both simple and relative frequency distributions using the qualitative attribute of occupations.

Classes of qualitative data

TABLE 2-10 OCCUPATIONS OF SAMPLE OF 100 GRADUATES OF CENTRAL COLLEGE

Occupational Class	Frequency Distribution (1)	Relative Frequency Distribution (1) ÷ 100
Actor	5	0.05
Banker	8	0.08
Businessperson	22	0.22
Chemist	7	0.07
Doctor	10	0.10
Insurance representative	6	0.06
Journalist	2	0.02
Lawyer	14	0.14
Teacher	9	0.09
Other	<u>17</u>	<u>0.17</u>
	100	1.00

Although Table 2-10 does not list every occupation held by the graduates of Central College, it is still all-inclusive. Why? The class “other” covers all the observations that fail to fit one of the enumerated categories. We will use a word like this whenever our list does not specifically list all the possibilities. For example, if our characteristic can occur in any month of the year, a complete list would include 12 categories. But if we wish to list only the 8 months from January through August, we can use the term *other* to account for our observations during the 4 months of September, October, November, and December. Although our list does not specifically list all the possibilities, it is all-inclusive. This “other” is called an *open-ended class* when it allows either the upper or the lower end of a quantitative classification scheme to be limitless. The last class in Table 2-11 (“72 and older”) is open-ended.

Classification schemes can be either quantitative or qualitative *and* either discrete or continuous. *Discrete classes* are separate entities that do not progress from one class to the next without a break. Such classes as the number of children in each family, the number of trucks owned by moving companies, and the occupations of Central College graduates are discrete. Discrete data are data that can take on only a limited number of values. Central College graduates can be classified as either doctors or chemists but not something in between. The closing price of AT&T stock can be $39\frac{1}{2}$ or $39\frac{7}{8}$ (but not 39.43), or your basketball team can win by 5 or 27 points (but not by 17.6 points).

Continuous data do progress from one class to the next without a break. They involve numerical measurement such as the weights of cans of tomatoes, the pounds of pressure on concrete, or the high school GPAs of college seniors. Continuous data can be expressed in either fractions or whole numbers.

Open-ended classes for lists that are not exhaustive

TABLE 2-11 AGES OF BUNDER COUNTY RESIDENTS

Class: Age (1)	Frequency (2)	Relative Frequency (2) ÷ 89,592
Birth to 7	8,873	0.0990
8 to 15	9,246	0.1032
16 to 23	12,060	0.1346
24 to 31	11,949	0.1334
32 to 39	9,853	0.1100
40 to 47	8,439	0.0942
48 to 55	8,267	0.0923
56 to 63	7,430	0.0829
64 to 71	7,283	0.0813
72 and older	<u>6,192</u>	<u>0.0691</u>
	89,592	1.0000

Discrete classes

Continuous classes

HINTS & ASSUMPTIONS

There are many ways to present data. Constructing a data array in either descending or ascending order is a good place to start. Showing how many times a value appears by using a frequency distribution is even more effective, and converting these frequencies to decimals (which we call relative frequencies) can help even more. Hint: We should remember that discrete variables are things that can be counted but continuous variables are things that appear at some point on a scale.

EXERCISES 2.3

Self-Check Exercises

SC 2-1 Here are the ages of 50 members of a country social service program:

83	51	66	61	82	65	54	56	92	60
65	87	68	64	51	70	75	66	74	68
44	55	78	69	98	67	82	77	79	62
38	88	76	99	84	47	60	42	66	74
91	71	83	80	68	65	51	56	73	55

Use these data to construct relative frequency distributions using 7 equal intervals and 13 equal intervals. State policies on social service programs require that approximately 50 percent of the program participants be older than 50.

- (a) Is the program in compliance with the policy?
- (b) Does your 13-interval relative frequency distribution help you answer part (a) better than your 7-interval distribution?
- (c) Suppose the Director of Social Services wanted to know the proportion of program participants between 45 and 50 years old. Could you estimate the answer for her better with a 7- or a 13-interval relative frequency distribution?

SC 2-2 Using the data in Table 2-1 on page 12, arrange the data in an array from highest to lowest high school GPA. Now arrange the data in an array from highest to lowest college GPA. What can you conclude from the two arrays that you could not from the original data?

Applications

2-9 Transmission Fix-It stores recorded the number of service tickets submitted by each of its 20 stores last month as follows:

823	648	321	634	752
669	427	555	904	586
722	360	468	847	641
217	588	349	308	766

The company believes that a store cannot really hope to break even financially with fewer than 475 service actions a month. It is also company policy to give a financial bonus to any store manager who generates more than 725 service actions a month. Arrange these data in a data array and indicate how many stores are not breaking even and how many are to get bonuses.

2-10 Use the data from Transmission Fix-It in Exercise 2-9. The company financial VP has set up what she calls a “store watch list,” that is, a list of the stores whose service activity is low enough to warrant additional attention from the home office. This category includes stores whose service activity is between 550 and 650 service actions a month. How many stores should be on that list based on last month’s activity?

2-11 The number of hours taken by transmission mechanics to remove, repair, and replace transmissions in one of the Transmission Fix-It stores one day last week is recorded as follows:

4.3	2.7	3.8	2.2	3.4
3.1	4.5	2.6	5.5	3.2
6.6	2.0	4.4	2.1	3.3
6.3	6.7	5.9	4.1	3.7

Construct a frequency distribution with intervals of 1.0 hour from these data. What conclusions can you reach about the productivity of mechanics from this distribution? If Transmission Fix-It management believes that more than 6.0 hours is evidence of unsatisfactory performance, does it have a major or minor problem with performance in this particular store?

- 2-12** The Orange County Transportation Commission is concerned about the speed motorists are driving on a section of the main highway. Here are the speeds of 45 motorists:

15	32	45	46	42	39	68	47	18
31	48	49	56	52	39	48	69	61
44	42	38	52	55	58	62	58	48
56	58	48	47	52	37	64	29	55
38	29	62	49	69	18	61	55	49

Use these data to construct relative frequency distributions using 5 equal intervals and 11 equal intervals. The U.S. Department of Transportation reports that, nationally, no more than 10 percent of the motorists exceed 55 mph.

- (a) Do Orange County motorists follow the U.S. DOT's report about national driving patterns?
- (b) Which distribution did you use to answer part (a)?
- (c) The U.S. DOT has determined that the safest speed for this highway is more than 36 but less than 59 mph. What proportion of the motorists drive within this range? Which distribution helped you answer this question?

- 2-13** Arrange the data in Table 2-2 on page 12 in an array from highest to lowest.
- (a) Suppose that state law requires bridge concrete to withstand at least 2,500 lb/sq in. How many samples would fail this test?
 - (b) How many samples could withstand a pressure of at least 2,497 lb/sq in. but could not withstand a pressure greater than 2,504 lb/sq in.?
 - (c) As you examine the array, you should notice that some samples can withstand identical amounts of pressure. List these pressures and the number of samples that can withstand each amount.
- 2-14** A recent study concerning the habits of U.S. cable television consumers produced the following data:

Number of Channels Purchased	Number of Hours Spent Watching Television per Week
25	14
18	16
42	12
96	6
28	13
43	16
39	9
29	7
17	19
84	4
76	8
22	13
104	6

Arrange the data in an array. What conclusion(s) can you draw from these data?

- 2-15** The Environmental Protection Agency took water samples from 12 different rivers and streams that feed into Lake Erie. These samples were tested in the EPA laboratory and rated as to the amount of solid pollution suspended in each sample. The results of the testing are given in the following table:

Sample	1	2	3	4	5	6
Pollution Rating (ppm)	37.2	51.7	68.4	54.2	49.9	33.4
Sample	7	8	9	10	11	12
Pollution Rating (ppm)	39.8	52.7	60.0	46.1	38.5	49.1

- (a) Arrange the data into an array from highest to lowest.
- (b) Determine the number of samples having a pollution content between 30.0 and 39.9, 40.0 and 49.9, 50.0 and 59.9, and 60.0 and 69.9.
- (c) If 45.0 is the number used by the EPA to indicate excessive pollution, how many samples would be rated as having excessive pollution?
- (d) What is the largest distance between any two consecutive samples?

- 2-16** Suppose that the admissions staff mentioned in the discussion of Table 2-1 on page 12 wishes to examine the relationship between a student's differential on the college SAT examination (the difference between actual and expected score based on the student's high school GPA) and the spread between the student's high school and college GPA (the difference between the college and high school GPA). The admissions staff will use the following data:

H.S. GPA	College GPA	SAT Score	H.S. GPA	College GPA	SAT Score
3.6	2.5	1,100	3.4	3.6	1,180
2.6	2.7	940	2.9	3.0	1,010
2.7	2.2	950	3.9	4.0	1,330
3.7	3.2	1,160	3.2	3.5	1,150
4.0	3.8	1,340	2.1	2.5	940
3.5	3.6	1,180	2.2	2.8	960
3.5	3.8	1,250	3.4	3.4	1,170
2.2	3.5	1,040	3.6	3.0	1,100
3.9	3.7	1,310	2.6	1.9	860
4.0	3.9	1,330	2.4	3.2	1,070

In addition, the admissions staff has received the following information from the Educational Testing Service:

H.S. GPA	Avg. SAT Score	H.S. GPA	Avg. SAT Score
4.0	1,340	2.9	1,020
3.9	1,310	2.8	1,000
3.8	1,280	2.7	980
3.7	1,250	2.6	960
3.6	1,220	2.5	940
3.5	1,190	2.4	920
3.4	1,160	2.3	910
3.3	1,130	2.2	900
3.2	1,100	2.1	880
3.1	1,070	2.0	860
3.0	1,040		

- (a) Arrange these data into an array of spreads from highest to lowest. (Consider an increase in college GPA over high school GPA as positive and a decrease in college GPA below high school GPA as negative.) Include with each spread the appropriate SAT differential. (Consider an SAT score below expected as negative and above expected as positive.)
- (b) What is the most common spread?
- (c) For this spread in part (b), what is the most common SAT differential?
- (d) From the analysis you have done, what do you conclude?

Worked-Out Answers to Self-Check Exercises

SC 2-1

7 Intervals		13 Intervals			
Class	Relative Frequency	Class	Relative Frequency	Class	Relative Frequency
30–39	0.02	35–39	0.02	70–74	0.10
40–49	0.06	40–44	0.04	75–79	0.10
50–59	0.16	45–49	0.02	80–84	0.12
60–69	0.32	50–54	0.08	85–89	0.04
70–79	0.20	55–59	0.08	90–94	0.04
80–89	0.16	60–64	0.10	95–99	0.04
90–99	0.08	65–69	0.22		1.00
1.00					

- (a) As can be seen from either distribution, about 90 percent of the participants are older than 50, so the program is not in compliance.
- (b) In this case, both are equally easy to use.
- (c) The 13-interval distribution gives a better estimate because it has a class for 45–49, whereas the 7-interval distribution lumps together all observations between 40 and 49.

SC 2-2 Data array by high school GPA:

High School GPA	College GPA	High School GPA	College GPA
4.0	3.9		3.4
4.0	3.8		3.2
3.9	4.0		2.9
3.9	3.7		2.7
3.7	3.2		2.6
3.6	3.0		2.6
3.6	2.5		2.4
3.5	3.8		2.2
3.5	3.6		2.2
3.4	3.6		2.1

Data array by college GPA:

College GPA	High School GPA	College GPA	High School GPA
4.0	3.9	3.2	3.7
3.9	4.0	3.2	2.4
3.8	4.0	3.0	3.6
3.8	3.5	3.0	2.9
3.7	3.9	2.8	2.2
3.6	3.5	2.7	2.6
3.6	3.4	2.5	3.6
3.5	3.2	2.5	2.1
3.5	2.2	2.2	2.7
3.4	3.4	1.9	2.6

From these arrays we can see that high GPAs at one level tend to go with high GPAs at the other, although there are some exceptions.

2.4 CONSTRUCTING A FREQUENCY DISTRIBUTION

Now that we have learned how to divide a sample into classes, we can take raw data and actually construct a frequency distribution. To solve the carpet-loom problem on the first page of the chapter, follow these three steps:

1. Decide on the type and number of classes for dividing the data. In this case, we have already chosen to classify the data by the quantitative measure of the number of yards produced rather than by a qualitative attribute such as color or pattern. Next, we need to decide how many different classes to use and the range each class should cover. The range must be divided by *equal* classes; that is, the width of the interval from the beginning of one class to the beginning of the next class must be the same for every class. If we choose a width of 0.5 yard for each class in our distribution, the classes will be those shown in Table 2-12.

Classify the data

Divide the range by equal classes

If the classes were unequal and the width of the intervals differed among the classes, then we would have a distribution that is much more difficult to interpret than one with equal intervals. Imagine how hard it would be to interpret the data presented in Table 2-13!

Problems with unequal classes

The number of classes depends on the number of data points and the range of the data collected. The more data points or the wider the range of the data, the more classes it takes to divide the data. Of course, if we have only 10 data points, it is senseless to have as many as 10 classes. As a rule, statisticians rarely use fewer than 6 or more than 15 classes.

Use 6 to 15 classes

Because we need to make the class intervals of equal size, the number of classes determines the width of each class. To find the intervals, we can use this equation:

Determine the width of the class intervals

TABLE 2-12 DAILY PRODUCTION IN A SAMPLE OF 30 CARPET LOOMS WITH 0.5-YARD CLASS INTERVALS

Class in Yards	Frequency
15.1–15.5	2
15.6–16.0	16
16.1–16.5	8
16.6–17.0	4
	30

TABLE 2-13 DAILY PRODUCTION IN A SAMPLE OF 30 CARPET LOOMS USING UNEQUAL CLASS INTERVALS

Class	Width of Class Intervals	Frequency
15.1–15.5	$15.6 - 15.1 = 0.5$	2
15.6–15.8	$15.9 - 15.6 = 0.3$	8
15.9–16.1	$16.2 - 15.9 = 0.3$	9
16.2–16.5	$16.6 - 16.2 = 0.4$	7
16.6–16.9	$17.0 - 16.6 = 0.4$	4
		30

Width of a Class Interval

$$\text{Width of class intervals} = \frac{\text{Next unit value after largest value in data} - \text{Smallest value in data}}{\text{Total number of class intervals}} \quad [2-1]$$

We must use the *next value of the same units* because we are measuring the *interval* between the first value of one class and the first value of the next class. In our carpet-loom study, the last value is 16.9, so 17.0 is the next value. We shall use six classes in this example, so the width of each class will be:

$$\begin{aligned} & \frac{\text{Next unit value after largest value in data} - \text{Smallest value in data}}{\text{Total number of class intervals}} \\ &= \frac{17.0 - 15.2}{6} \\ &= \frac{1.8}{6} \\ &= 0.3 \text{ yd} \leftarrow \text{width of class intervals} \end{aligned} \quad [2-1]$$

Step 1 is now complete. We have decided to classify the data by the quantitative measure of how many yards of carpet were produced. We have chosen 6 classes to cover the range of 15.2 to 16.9 and, as a result, will use 0.3 yard as the width of our class intervals.

Examine the results

2. **Sort the data points into classes and count the number of points in each class.** This we have done in Table 2-14. Every data point fits into at least one class, and no data point fits into more than one class. Therefore, our classes are all-inclusive and mutually exclusive. Notice that the lower boundary of the first class corresponds with the smallest data point in our sample, and the upper boundary of the last class corresponds with the largest data point.

Create the classes and count the frequencies

3. **Illustrate the data in a chart.** (See Figure 2-1.)

These three steps enable us to arrange the data in both tabular and graphic form. In this case, our information is displayed in Table 2-14 and in Figure 2-1. These two frequency distributions omit some

TABLE 2-14 DAILY PRODUCTION IN A SAMPLE OF 30 CARPET LOOMS WITH 0.3 YARD CLASS INTERVALS

Class	Frequency
15.2–15.4	2
15.5–15.7	5
15.8–16.0	11
16.1–16.3	6
16.4–16.6	3
16.7–16.9	3
	30

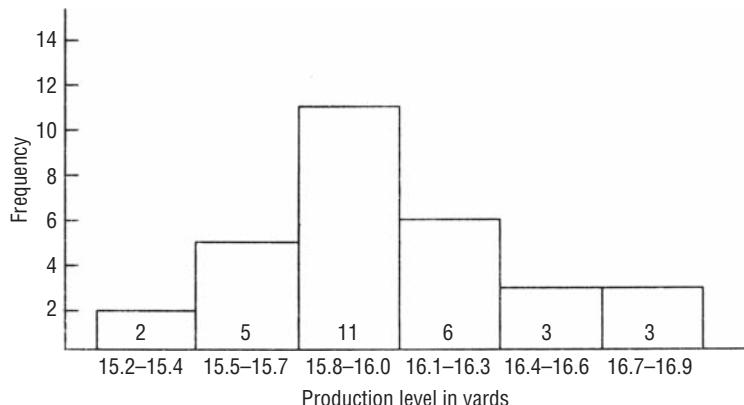


FIGURE 2-1 FREQUENCY DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS USING 0.3-YARD CLASS INTERVALS

of the detail contained in the raw data of Table 2-3, but they make it easier for us to notice patterns in the data. One obvious characteristic, for example, is that the class 15.8–16.0 contains the most elements; class 15.2–15.4, the fewest.

Notice in Figure 2-1 that the frequencies in the classes of 0.3-yard widths follow a regular progression: The number of data points begins with 2 for the first class, builds to 5, reaches 11 in the third class, falls to 6, and tumbles to 3 in the fifth and sixth classes. We will find that the larger the width of the class intervals, the smoother

Notice any trends

this progression will be. However, if the classes are too wide, we lose so much information that the chart is almost meaningless. For example, if we collapse Figure 2-1 into only two categories, we obscure the pattern. This is evident in Figure 2-2.

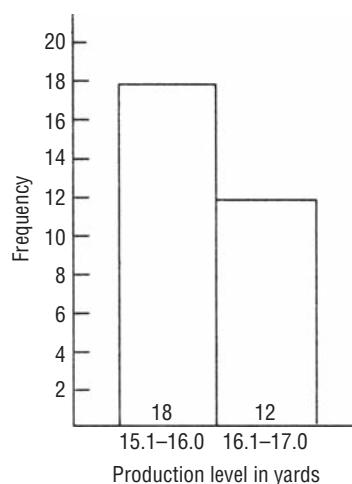


FIGURE 2-2 FREQUENCY DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS USING 1-YARD CLASS INTERVALS

Using the Computer to Construct Frequency Distributions

Throughout this text, we will be using simple examples to illustrate how to do many different kinds of statistical analyses.

Hand calculations are cumbersome

With such examples, you can learn what sort of calculations have to be done. We hope you will also be able to understand the concepts behind the calculations, so you will appreciate why these particular calculations are appropriate. However, the fact of the matter remains that hand calculations are cumbersome, tiresome, and error-prone. Many real problems have so much data that doing the calculations by hand is not really feasible.

For this reason, most real-world statistical analysis is done on computers. You

Software packages for statistical analysis

Summary Statistics for Discrete Variables					
TOTBY10	Count	Percent	Cumcnt	Cumpct	
25	1	0.50	1	0.50	
35	1	0.50	2	1.01	
45	9	4.52	11	5.53	
55	27	13.57	38	19.10	
65	68	34.17	106	53.27	
75	65	32.66	171	85.93	
85	26	13.07	197	98.99	
95	2	1.01	199	100.00	
N=	199				

FIGURE 2-3 MINITAB FREQUENCY DISTRIBUTION OF RAW TOTAL SCORES

prepare the input data and interpret the results of the analysis and take appropriate actions, but the machine does all the “number crunching.” There are many widely used software packages for statistical analyses, including Minitab, SAS, SPSS, and SYSTAT.* It is not our intention to teach you the details of how to use any of these to do your analyses, but we will be using primarily Minitab and occasionally the SAS System to illustrate typical sorts of outputs these packages produce.

Appendix Table 10 contains grade data for the 199 students who used this text in our course. In Figure 2-3, we have used Minitab to create a frequency distribution of the students’ raw total scores in the course. The TOTBY10 column values are the midpoints of the classes. Often, you will also be interested in *bivariate frequency distributions*, in which the data are classified with respect to two different attributes. In Figure 2-4, we have such a distribution showing the letter grades in each of the six sections of the class. The variable NUMGRADE has values 0 to 9, which correspond to letter grades F, D, C–, C, C+, B–, B, B+, A–, and A.

Using the grade data

Appendix Table 11 contains earnings data for 224 companies whose 1989 last-quarter earnings were published in *The Wall Street Journal* during the week of February 12, 1990. In Figure 2-5, we have used Minitab to create a frequency distribution of those last-quarter earnings. The variable Q489 is the 1989 last-quarter earnings, rounded to the nearest dollar.

Because companies listed on the New York Stock Exchange (3) tend to have different financial characteristics from those listed on the American Stock Exchange (2), and because those, in turn, are different from companies listed “over-the-counter” (1), we also used Minitab to produce the bivariate distribution of the same earnings data in Figure 2-6.

HINTS & ASSUMPTIONS

When we construct a frequency distribution we need to carefully choose the classes into which we divide data. This is true even when we use a computer program to set up the classes. For example, a computer program might divide the ages of respondents to a marketing research survey into the consistent classes: 15–19, 20–24, 25–29, and so on. But if the product being researched is intended for college students, it may make more sense to set up the classes as 18, 19–22, and 23 and above. Be aware that using a computer in statistics doesn’t substitute for common sense.

*Minitab is a registered trademark of Minitab, Inc., University Park, PA. SAS is a registered trademark of SAS Institute, Inc., Cary, N.C. SPSS is a registered trademark of SPSS, Inc., Chicago, IL. SYSTAT is a registered trademark of SYSTAT, Inc., Evanston, IL.

Tabulated Statistics								
ROWS: NUMGRADE		COLUMNS: SECTION						
		1	2	3	4	5	6	ALL
0	2	3	0	1	3	2	11	
	1.01	1.51	--	0.50	1.51	1.01		5.53
1	3	6	5	2	4	6	26	
	1.51	3.02	2.51	1.01	2.01	3.02	13.07	
2	2	2	1	2	7	4	18	
	1.01	1.01	0.50	1.01	3.52	2.01	9.05	
3	9	11	3	9	6	6	44	
	4.52	5.53	1.51	4.52	3.02	3.02	22.11	
4	3	6	10	6	7	2	34	
	1.51	3.02	5.03	3.02	3.52	1.01	17.09	
5	1	5	5	1	0	3	15	
	0.50	2.51	2.51	0.50	--	1.51	7.54	
6	2	5	3	2	2	3	17	
	1.01	2.51	1.51	1.01	1.01	1.51	8.54	
7	1	1	1	2	1	1	7	
	0.50	0.50	0.50	1.01	0.50	0.50	3.52	
8	2	2	8	1	3	0	16	
	1.01	1.01	4.02	0.50	1.51	--	8.04	
9	2	5	1	0	3	0	11	
	1.01	2.51	0.50	--	1.51	--	5.53	
ALL	27	46	37	26	36	27	199	
	13.57	23.12	18.59	13.07	18.09	13.57	100.00	
CELL CONTENTS --								
COUNT								
% OP TBL								

FIGURE 2-4 MINITAB BIVARIATE FREQUENCY DISTRIBUTION SHOWING GRADES IN EACH SECTION

Summary Statistics for Discrete Variables				
Q489	Count	Percent	Cumcnt	Cumpct
-5	1	0.45	1	0.45
-4	2	0.89	3	1.34
-2	1	0.45	4	1.79
-1	9	4.02	13	5.80
0	164	73.21	177	79.02
1	43	19.20	220	98.21
2	2	0.89	222	99.11
5	2	0.89	224	100.00
N=	224			

FIGURE 2-5
MINITAB FREQUENCY
DISTRIBUTION OF
1989 LAST-QUARTER
EARNINGS

Tabulated Statistics				
ROWS: Q489		COLUMNS: EXCHANGE		
	1	2	3	ALL
-5	0	0	1	1
	--	--	100.00	100.00
	--	--	1.33	0.45
	--	--	0.45	0.45
-4	1	0	1	2
	50.00	--	50.00	100.00
	0.90	--	1.33	0.89
	0.45	--	0.45	0.89
-2	1	0	0	1
	100.00	--	--	100.00
	0.90	--	--	0.45
	0.45	--	--	0.45
-1	5	2	2	9
	55.56	22.22	22.22	100.00
	4.50	5.26	2.67	4.02
	2.23	0.89	0.89	4.02
0	97	31	36	164
	59.15	18.90	21.95	100.00
	87.39	81.58	48.00	73.21
	43.30	13.84	16.07	73.21
1	7	4	32	43
	16.28	9.30	74.42	100.00
	6.31	10.53	42.67	19.20
	3.12	1.79	14.29	19.20
2	0	0	2	2
	--	--	100.00	100.00
	--	--	2.67	0.89
	--	--	0.89	0.89
5	0	1	1	2
	--	50.00	50.00	100.00
	--	2.63	1.33	0.89
	--	0.45	0.45	0.89
ALL	111	38	75	224
	49.55	16.96	33.48	100.00
	100.00	100.00	100.00	100.00
	49.55	16.96	33.48	100.00
CELL CONTENTS --				
COUNT				
% OF ROW				
% OF COL				
% OF TBL				

FIGURE 2-6 MINITAB BIVARIATE FREQUENCY DISTRIBUTION SHOWING EARNINGS ON EACH EXCHANGE

EXERCISES 2.4

Self-Check Exercises

- SC 2-3** High Performance Bicycle Products Company in Chapel Hill, North Carolina, sampled its shipping records for a certain day with these results:

Time from Receipt of Order to Delivery (in Days)									
4	12	8	14	11	6	7	13	13	11
11	20	5	19	10	15	24	7	19	6

Construct a frequency distribution for these data and a relative frequency distribution. Use intervals of 6 days.

- (a) What statement can you make about the effectiveness of order processing from the frequency distribution?
- (b) If the company wants to ensure that half of its deliveries are made in 10 or fewer days, can you determine from the frequency distribution whether they have reached this goal?
- (c) What does having a relative frequency distribution permit you to do with the data that is difficult to do with only a frequency distribution?

- SC 2-4** Mr. Frank, a safety engineer for the Mars Point Nuclear Power Generating Station, has charted the peak reactor temperature each day for the past year and has prepared the following frequency distribution:

Temperatures in °C	Frequency
Below 500	4
501–510	7
511–520	32
521–530	59
530–540	82
550–560	65
561–570	33
571–580	28
580–590	27
591–600	23
Total	360

List and explain any errors you can find in Mr. Franks's distribution.

Applications

- 2-17** Universal Burger is concerned about product waste, so they sampled their burger waste record from the past year with the following results:

Number of Burgers Discarded During a Shift									
2	16	4	12	19	29	24	7	19	
22	14	8	24	31	18	20	16	6	

Construct a frequency distribution for these data and a relative frequency distribution. Use intervals of 5 burgers.

- One of Universal Burger's goals is for at least 75 percent of shifts to have no more than 16 burgers wasted. Can you determine from the frequency distribution whether this goal has been achieved?
- What percentage of shifts have waste of 21 or fewer burgers? Which distribution did you use to determine your answer?

2-18 Refer to Table 2-2 on page 18 and construct a relative frequency distribution using intervals of 4.0 lb/sq in. What do you conclude from this distribution?

2-19 The Bureau of Labor Statistics has sampled 30 communities nationwide and compiled prices in each community at the beginning and end of August in order to find out approximately how the Consumer Price Index (CPI) has changed during August. The percentage changes in prices for the 30 communities are as follows:

0.7	0.4	-0.3	0.2	-0.1	0.1	0.3	0.7	0.0	-0.4
0.1	0.5	0.2	0.3	1.0	-0.3	0.0	0.2	0.5	0.1
-0.5	-0.3	0.1	0.5	0.4	0.0	0.2	0.3	0.5	0.4

- Arrange the data in an array from lowest to highest.
- Using the following four equal-sized classes, create a frequency distribution: -0.5 to -0.2, -0.1 to 0.2, 0.3 to 0.6, and 0.7 to 1.0.
- How many communities had prices that either did not change or that increased less than 1.0 percent?
- Are these data discrete or continuous?

2-20 Sarah Anne Rapp, the president of Baggit, Inc., has just obtained some raw data from a marketing survey that her company recently conducted. The survey was taken to determine the effectiveness of the new company slogan, "When you've given up on the rest, Baggit!" To determine the effect of the slogan on the sales of Luncheon Baggits, 20 people were asked how many boxes of Luncheon Baggits per month they bought before and after the slogan was used in the advertising campaign. The results were as follows:

Before/After	Before/After	Before/After	Before/After
4	3	2	1
4	6	6	9
1	5	6	7
3	7	5	8
5	5	3	6
		3	5

- Create both frequency and relative frequency distributions for the "Before" responses, using as classes 1–2, 3–4, 5–6, 7–8, and 9–10.
- Work part (a) for the "After" responses.
- Give the most basic reason why it makes sense to use the same classes for both the "Before" and "After" responses.
- For each pair of "Before/After" responses, subtract the "Before" response from the "After" response to get the number that we will call "Change" (example: $3 - 4 = -1$), and

create frequency and relative frequency distributions for “Change” using classes –5 to –4, –3 to –2, –1 to 0, 1 to 2, 3 to 4, and 5 to 6.

- (e) Based on your analysis, state whether the new slogan has helped sales, and give one or two reasons to support your conclusion.

- 2-21** Here are the ages of 30 people who bought video recorders at Symphony Music Shop last week:

26	37	40	18	14	45	32	68	31	37
20	32	15	27	46	44	62	58	30	42
22	26	44	41	34	55	50	63	29	22

- (a) From looking at the data just as they are, what conclusions can you come to quickly about Symphony’s market?
 (b) Construct a 6-category closed classification. Does having this enable you to conclude anything more about Symphony’s market?

- 2-22** Use the data from Exercise 2-21.

- (a) Construct a 5-category open-ended classification. Does having this enable you to conclude anything more about Symphony’s market?
 (b) Now construct a relative frequency distribution to go with the 5-category open-ended classification. Does having this provide Symphony with additional information useful in its marketing? Why?

- 2-23** John Lyon, owner of Fowler’s Food Store in Chapel Hill, North Carolina, has arranged his customers’ purchase amounts last week into this frequency distribution:

\$ Spent	Frequency	\$ Spent	Frequency	\$ Spent	Frequency
0.00–0.99	50	16.00–18.99	1,150	34.00–36.99	610
1.00–3.99	240	19.00–21.99	980	37.00–39.99	420
4.00–6.99	300	22.00–24.99	830	40.00–42.99	280
7.00–9.99	460	25.00–27.99	780	43.00–45.99	100
10.00–12.99	900	28.00–30.99	760	46.00–48.99	90
13.00–15.99	1,050	31.00–33.99	720		

John says that having 17 intervals each defined by 2 numbers is cumbersome. Can you help him simplify the data he has without losing too much of their value?

- 2-24** Here are the midpoints of the intervals for a distribution representing minutes it took the members of a university track team to complete a 5-mile cross-country run.

25 35 45

- (a) Would you say that the team coach can get enough information from these midpoints to help the team?
 (b) If your answer to part (a) is “no,” how many intervals do seem appropriate?

- 2-25** Barney Mason has been examining the amount of daily french fry waste (in pounds) for the past 6 months at Universal Burger and has created the following frequency distribution:

French Fry Waste in Pounds	Frequency
0.0–3.9	37
4.0–7.9	46
8.0–11.9	23
12.0–16.9	27
17.0–25.9	7
26.0–40.9	0
	180

List and explain any errors you can find in Barney's distribution.

- 2-26** Construct a discrete, closed classification for the possible responses to the "marital status" portion of an employment application. Also, construct a 3-category, discrete, open-ended classification for the same responses.
- 2-27** Stock exchange listings usually contain the company name, the high and low bids, the closing price, and the change from the previous day's closing price. Here's an example:

Name	High Bid	Low Bid	Closing	Change
System Associates	11½	10⅞	11¼	+½

Is a distribution of all (a) stocks on the New York Stock Exchange by industry, (b) closing prices on a given day, and (c) changes in prices from the previous day

- (1) Quantitative or qualitative?
- (2) Continuous or discrete?
- (3) Open-ended or closed?

Would your answer to part (c) be different if the change were expressed simply as "higher," "lower," or "unchanged"?

- 2-28** The noise level in decibels of aircraft departing Westchester County Airport was rounded to the nearest decibel and grouped in a frequency distribution having intervals with midpoints at 100 and 130. Under 100 decibels is not considered loud at all, and anything over 140 decibels is almost deafening. If Residents for a Quieter Neighborhood is gathering data for its lawsuit against the airport, is this distribution adequate for its purpose?
- 2-29** Use the data from Exercise 2-28. If the lawyer defending the airport is collecting data preparatory to going to trial, would she approve of the midpoints of the intervals in Exercise 2-28 for her purposes?
- 2-30** The president of Ocean Airlines is trying to estimate when the Federal Aviation Administration (FAA) is most likely to rule on the company's application for a new route between Charlotte and Nashville. Assistants to the president have assembled the following waiting times for applications filed during the past year. The data are given in days from the date of application until an FAA ruling.

34	40	23	28	31	40	25	33	47	32
44	34	38	31	33	42	26	35	27	31
29	40	31	30	34	31	38	35	37	33
24	44	37	39	32	36	34	36	41	39
29	22	28	44	51	31	44	28	47	31

- (a) Construct a frequency distribution using 10 closed intervals, equally spaced. Which interval contains the most data points?
- (b) Construct a frequency distribution using 5 closed intervals, equally spaced. Which interval contains the most data points?
- (c) If the president of Ocean Airlines had a relative frequency distribution for either (a) or (b), would that help him estimate the answer he needs?

2-31 For the purpose of performance evaluation and quota adjustment, Ralph Williams monitored the auto sales of his 40 salespeople. Over a 1-month period, they sold the following number of cars:

7	8	5	10	9	10	5	12	8	6
10	11	6	5	10	11	10	5	9	13
8	12	8	8	10	15	7	6	8	8
5	6	9	7	14	8	7	5	5	14

- (a) Based on frequency, what would be the desired class marks (midpoints of the intervals)?
- (b) Construct a frequency and relative frequency distribution having as many of these marks as possible. Make your intervals evenly spaced and at least two cars wide.
- (c) If sales fewer than seven cars a month is considered unacceptable performance, which of the two answers, (a) or (b), helps you more in identifying the unsatisfactory group of salespeople?

2-32 Kessler's Ice Cream Delight attempts to keep all of its 55 flavors of ice cream in stock at each of its stores. Their marketing-research director suggests that keeping better records for each store is the key to preventing stockouts. Don Martin, director of store operations, collects data to the nearest half gallon on the daily amount of each flavor of ice cream that is sold. No more than 20 gallons of any flavor are ever used on one day.

- (a) Is the flavor classification discrete or continuous? Open or closed?
- (b) Is the "amount of ice cream" classification discrete or continuous? Open or closed?
- (c) Are the data qualitative or quantitative?

(d) What would you suggest Martin do to generate better data for market-research purposes?

2-33 Doug Atkinson is the owner and ticket collector for a ferry that transports people and cars from Long Island to Connecticut. Doug has data indicating the number of people, as well as the number of cars, that have ridden the ferry during the past 2 months. For example,

JULY 3 NUMBER OF PEOPLE, 173 NUMBER OF CARS, 32

might be a typical daily entry for Doug. Doug has set up six equally spaced classes to record the daily number of people, and the class marks are 84.5, 104.5, 124.5, 144.5, 164.5, and 184.5. Doug's six equally spaced classes for the daily number of cars have class marks of 26.5, 34.5, 42.5, 50.5, 58.5, and 66.5. (The class marks are the midpoints of the intervals.)

- (a) What are the upper and lower boundaries of the classes for the number of people?
- (b) What are the upper and lower boundaries of the classes for the number of cars?

Worked-Out Answers to Self-Check Exercises

SC 2-3	Class	1–6	7–12	13–18	19–24	25–30
Frequency		4	8	4	3	1
Relative Frequency		0.20	0.40	0.20	0.15	0.05

- (a) Assuming that the shop is open 6 days a week, we see that fully 80 percent of the orders are filled in 3 weeks or less.
- (b) We can tell only that between 20 percent and 60 percent of the deliveries are made in 10 or fewer days, so the distribution does not generate enough information to determine whether the goal has been met.
- (c) A relative frequency distribution lets us present frequencies as fractions or percentages.

SC 2-4 The distribution is not all-inclusive. The data point 500°C is left out, along with the points between 541°C and 549°C, inclusive. In addition, the distribution is closed on the high end, which eliminates all data points above 600°C. These omissions might explain the fact that the total number of observations is only 360, rather than 365 as might be expected for a data set compiled over one year. (Note: It is not absolutely necessary that the distribution be open-ended on the high end, especially if no data points were recorded above 600°C. However, for completeness, the distribution should be continuous over the range selected, even though no data points may fall in some of the intervals.) Finally, the classifications are not mutually exclusive. Two points, 530°C and 580°C, are contained in more than one interval. When creating a set of continuous classifications, care must be taken to avoid this error.

2.5 GRAPHING FREQUENCY DISTRIBUTIONS

Figures 2-1 and 2-2 (on page 29) are previews of what we are going to discuss now: how to present frequency distributions graphically.

Identifying the horizontal and vertical axes

Graphs give data in a two-dimensional picture. On the *horizontal* axis, we can show the values of the variable (the characteristic we are measuring), such as the carpet output in yards. On the *vertical* axis, we mark the frequencies of the classes shown on the horizontal axis. Thus, the height of the boxes in Figure 2-1 measures the number of observations in each of the classes marked on the horizontal axis. Graphs of frequency distributions and relative frequency distributions are useful because they emphasize and clarify patterns that are not so readily discernible in tables. They attract a reader's attention to patterns in the data. Graphs can also help us do problems concerning frequency distributions. They will enable us to estimate some values at a glance and will provide us with a pictorial check on the accuracy of our solutions.

Function of graphs

Histograms

Figures 2-1 and 2-2 (page 29) are two examples of histograms. A *histogram* is a series of rectangles, each proportional in width to the range of values within a class and proportional in height to the number of items falling in the class. If the classes we use in the frequency distribution are of equal width, then the vertical bars in the histogram are also of equal width. The height of the bar for each class corresponds to the number of items in the class. As a result, the area contained in each rectangle (width times height) is the same percentage of the area of all the rectangles as the frequency of that class is to all the observations made.

Histograms described

A histogram that uses the relative frequency of data points in each of the classes rather than the actual number of points is called a *relative frequency histogram*. The relative frequency histogram has the same shape as an absolute frequency histogram made from the same data set. This is true because in both, the relative size of each rectangle is the frequency of that class compared to the total number of observations.

Function of a relative frequency histogram

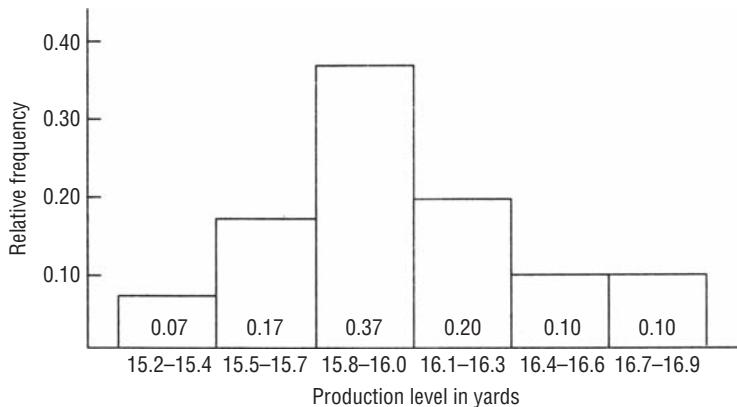


FIGURE 2-7 RELATIVE FREQUENCY DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS USING 0.3-YARD CLASS INTERVALS

Recall that the relative frequency of any class is the number of observations in that class divided by the total number of observations made. The sum of all the relative frequencies for any data set is equal to 1.0. With this in mind, we can convert the histogram of Figure 2-1 into a relative frequency histogram, such as we find in Figure 2-7. Notice that the only difference between these two is the left-hand vertical scale. Whereas the scale in Figure 2-1 is the *absolute* number of observations in each class, the scale in Figure 2-7 is the number of observations in each class as a *fraction* of the total number of observations.

Being able to present data in terms of the relative rather than the absolute frequency of observations in each class is useful because, while the absolute numbers may change (as we test more looms, for example), the relationship among the classes may remain stable. Twenty percent of all the looms may fall in the class “16.1–16.3 yards” whether we test 30 or 300 looms. It is easy to compare the data from different sizes of samples when we use relative frequency histograms.

Advantage of the relative frequency histogram

Frequency Polygons

Although less widely used, *frequency polygons* are another way to portray graphically both simple and relative frequency distributions. To construct a frequency polygon, we mark the frequencies on the vertical axis and the values of the variable we are measuring on the horizontal axis, as we did with histograms. Next, we plot each class frequency by drawing a dot above its midpoint, and connect the successive dots with straight lines to form a polygon (a many-sided figure).

Use midpoints on the horizontal axis

Figure 2-8 is a frequency polygon constructed from the data in Table 2-14 on page 29. If you compare this figure with Figure 2-1, you will notice that classes have been added at *each end* of the scale of observed values. These two new classes contain zero observations but allow the polygon to reach the horizontal axis at both ends of the distribution.

Add two classes

How can we turn a frequency polygon into a histogram? A frequency polygon is simply a line graph that connects the midpoints of all the bars in a histogram. Therefore, we can reproduce the histogram by drawing vertical lines from the bounds of the classes

Converting a frequency polygon to a histogram

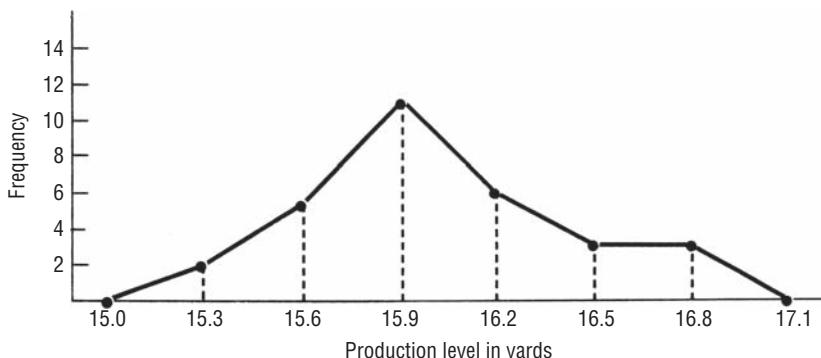


FIGURE 2-8 FREQUENCY POLYGON OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS USING 0.3-YARD CLASS INTERVALS

(as marked on the horizontal axis) and connecting them with horizontal lines at the heights of the polygon at each midpoint. We have done this with dotted lines in Figure 2-9.

A frequency polygon that uses the relative frequency of data points in each of the classes rather than the actual number of points is called a *relative frequency polygon*. The relative frequency polygon has the same shape as the frequency polygon made from the same data set but a different scale of values on the vertical axis. Rather than the absolute number of observations, the scale is the number of observations in each class as a fraction of the total number of observations.

Histograms and frequency polygons are similar. Why do we need both? The advantages of histograms are

Constructing a relative frequency polygon

Advantages of histograms

1. The rectangle clearly shows each separate class in the distribution.
2. The area of each rectangle, relative to all the other rectangles, shows the proportion of the total number of observations that occur in that class.

Frequency polygons, however, have certain advantages, too.

Advantages of polygons

1. The frequency polygon is simpler than its histogram counterpart.
2. It sketches an outline of the data pattern more clearly.
3. The polygon becomes increasingly smooth and curvelike as we increase the number of classes and the number of observations.

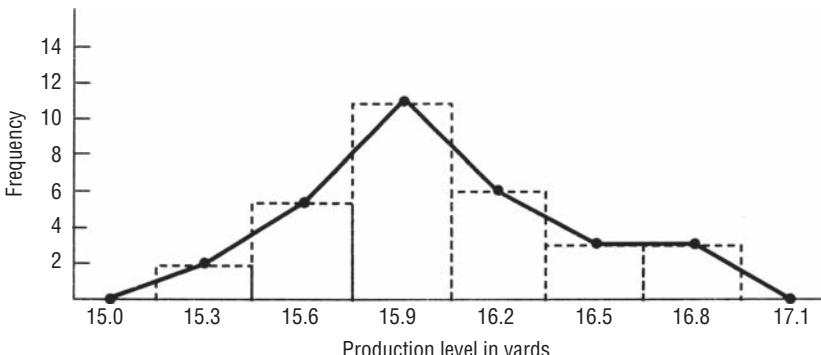


FIGURE 2-9 HISTOGRAM DRAWN FROM THE POINTS OF THE FREQUENCY POLYGON IN FIGURE 2-8

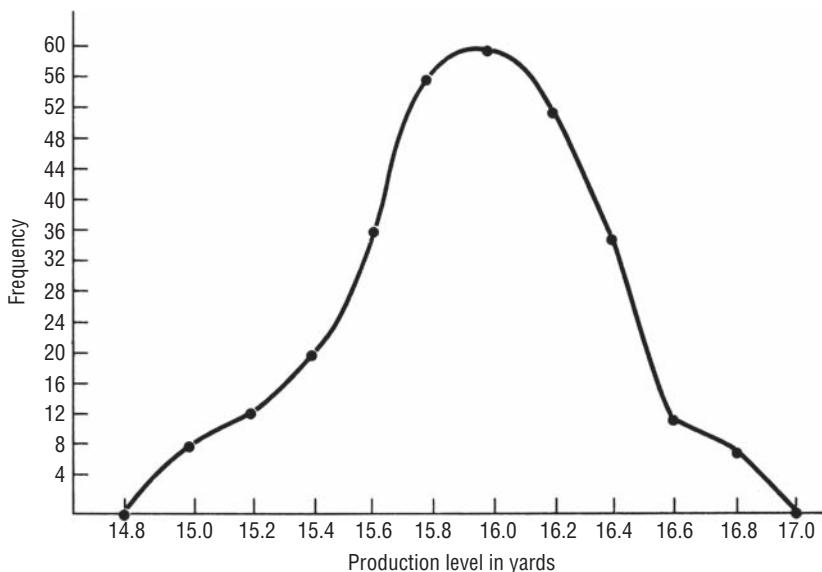


FIGURE 2-10 FREQUENCY CURVE OF PRODUCTION LEVELS IN A SAMPLE OF 300 CARPET LOOMS USING 0.2-YARD INTERVALS

A polygon such as the one we have just described, smoothed by added classes and data points, is called a *frequency curve*. In Figure 2-10, we have used our carpet-loom example, but we have increased the number of observations to 300 and the number of classes to 10. Notice that we have connected the points with curved lines to approximate the way the polygon would look if we had a very large number of data points and very small class intervals.

Creating a frequency curve

TABLE 2-15 CUMULATIVE “LESS-THAN” FREQUENCY DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS

Class	Cumulative Frequency
Less than 15.2	0
Less than 15.5	2
Less than 15.8	7
Less than 16.1	18
Less than 16.4	24
Less than 16.7	27
Less than 17.0	30

Ogives

A *cumulative frequency distribution* enables us to see how many observations lie above or below certain values, rather than merely recording the number of items within intervals. For example, if we wish to know how many looms made less than 17.0 yards we can use a table recording the cumulative “less-than” frequencies in our sample, such as Table 2-15.

A graph of a cumulative frequency distribution is called an *ogive* (pronounced “oh-jive”). The ogive for the cumulative distribution in Table 2-15 is shown in Figure 2-11. The plotted points represent the number of looms having less production

Cumulative frequency distribution defined

A “less-than” ogive

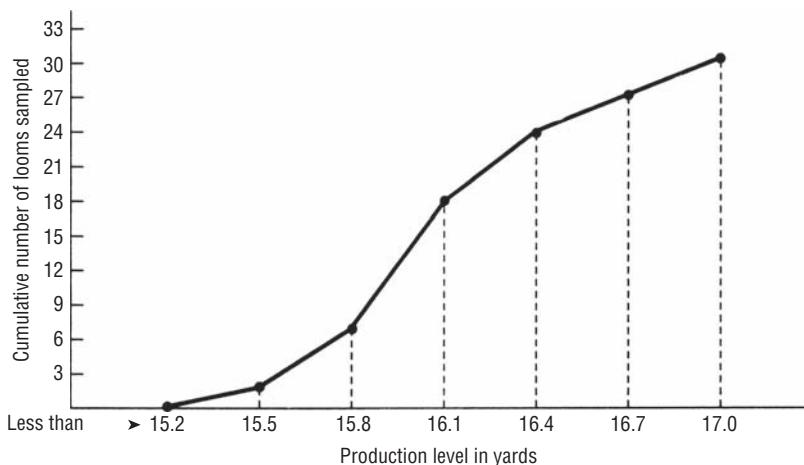


FIGURE 2-11 “LESS-THAN” OGIVE OF THE DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS

than the number of yards shown on the horizontal axis. Notice that the lower bound of the classes in the table becomes the upper bound of the cumulative distribution of the ogive.

Occasionally, the information we are using is presented in terms of “more-than” frequencies. The appropriate ogive for such information would slope down and to the right, instead of up and to the right as it did in Figure 2-11.

We can construct an ogive of a relative frequency distribution in the same manner in which we drew the ogive of an absolute frequency distribution in Figure 2-11. There will be one change—the vertical scale. As in Figure 2-7, on page 37, this scale must mark the *fraction* of the total number of observations that falls into each class.

To construct a cumulative “less-than” ogive in terms of relative frequencies, we can refer to a relative frequency distribution (such as Figure 2-7) and set up a table using the data (such as Table 2-16). Then

Ogives of relative frequencies

TABLE 2-16 CUMULATIVE RELATIVE FREQUENCY DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS

Class	Cumulative Frequency	Cumulative Relative Frequency
Less than 15.2	0	0.00
Less than 15.5	2	0.07
Less than 15.8	7	0.23
Less than 16.1	18	0.60
Less than 16.4	24	0.80
Less than 16.7	27	0.90
Less than 17.0	30	1.00

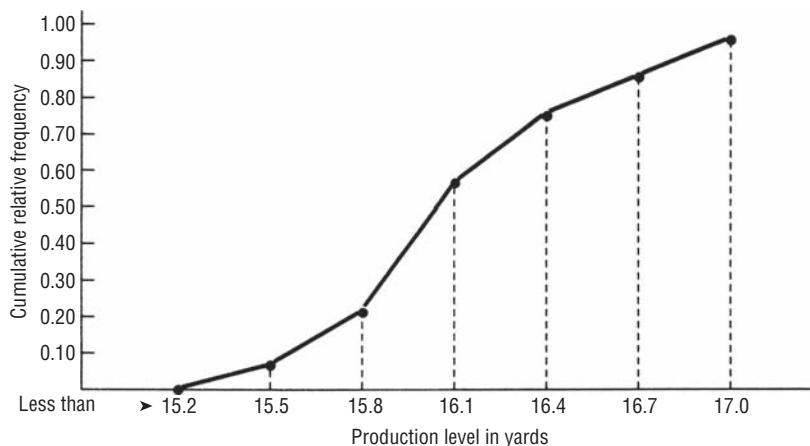


FIGURE 2-12 “LESS-THAN” OGIVE OF THE DISTRIBUTION OF PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS USING RELATIVE FREQUENCIES

we can convert the figures there to an ogive (as in Figure 2-12). Notice that Figures 2-11 and 2-12 are equivalent except for the left-hand vertical axis.

Suppose we now draw a line perpendicular to the vertical axis at the 0.50 mark to intersect our ogive. (We have done this in Figure 2-13.) In this way, we can read an approximate value of 16.0 for the production level in the fifteenth loom of an array of the 30.

Thus, we are back to the first data arrangement discussed in this chapter. From the data array, we can construct frequency distributions. From frequency distributions, we can construct cumulative frequency distributions. From these, we can graph an ogive. And from this ogive, we can approximate the values we had in the data array. However, we cannot normally recover the *exact* original data from any of the graphic representations we have discussed.

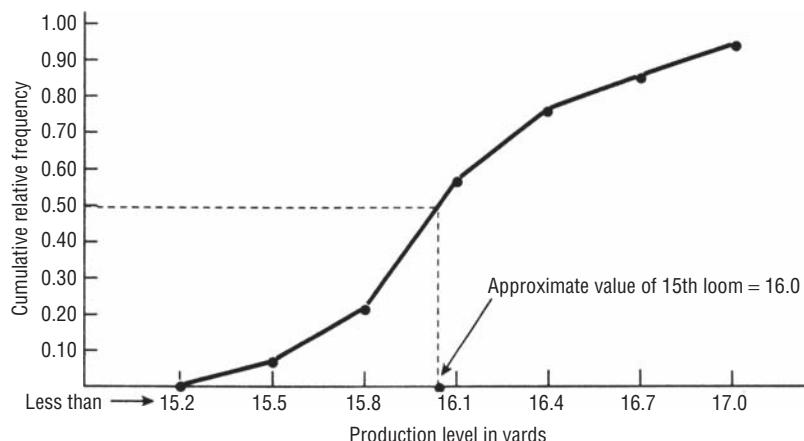
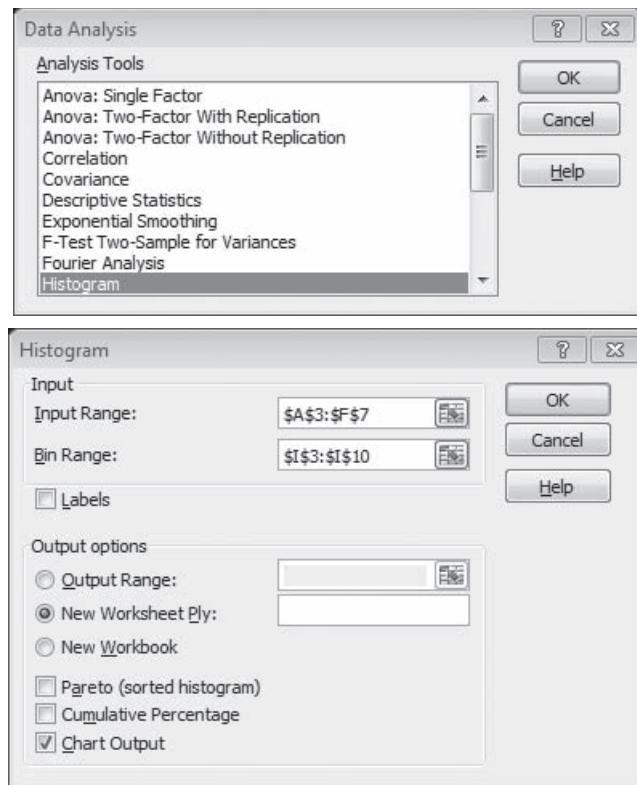


FIGURE 2-13 “LESS-THAN” OGIVE OF THE DISTRIBUTION OF THE PRODUCTION LEVELS IN A SAMPLE OF 30 CARPET LOOMS, INDICATING THE APPROXIMATE MIDDLE VALUE IN THE ORIGINAL DATA ARRAY

Using Statistical Packages to Graph Frequency Distribution: Histogram

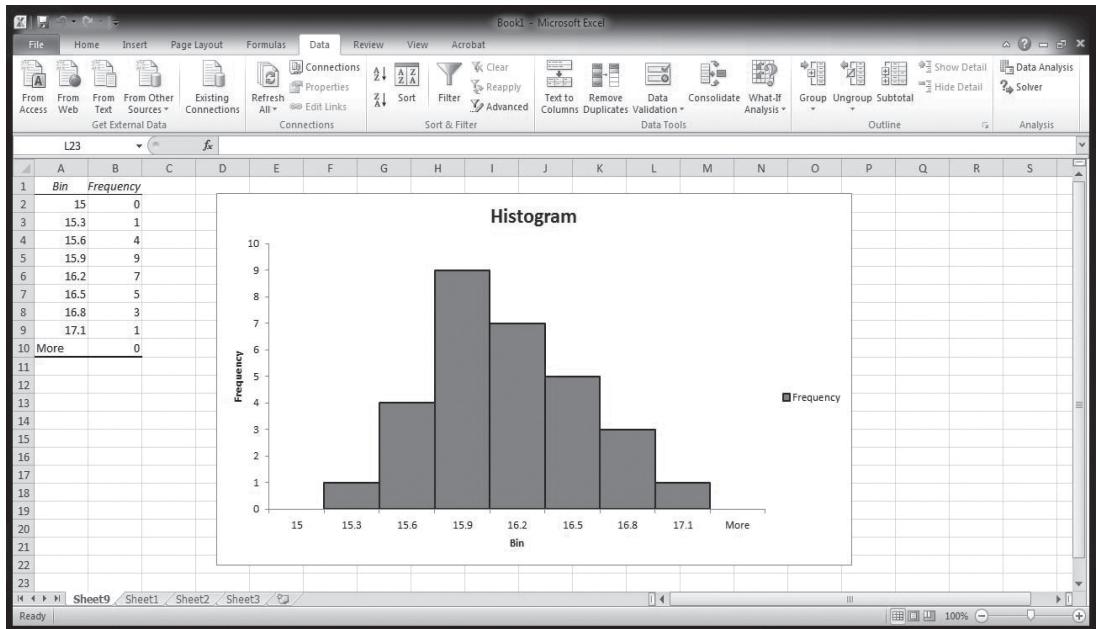
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																			
2																			
3	16.2	16.8	15.9	15.6	15.9	16.6													
4	15.8	16	16	15.7	15.9	15.6													
5	15.8	16.4	16.3	16	16.8	15.6													
6	15.8	15.2	16	16.2	15.4	16.9													
7	16.3	15.9	16.4	16.1	15.7	16.3													
8																			
9																			
10																			
11																			

Above data is sample of daily production in meters of 30 carpet looms and the desired mid values for creating histogram.

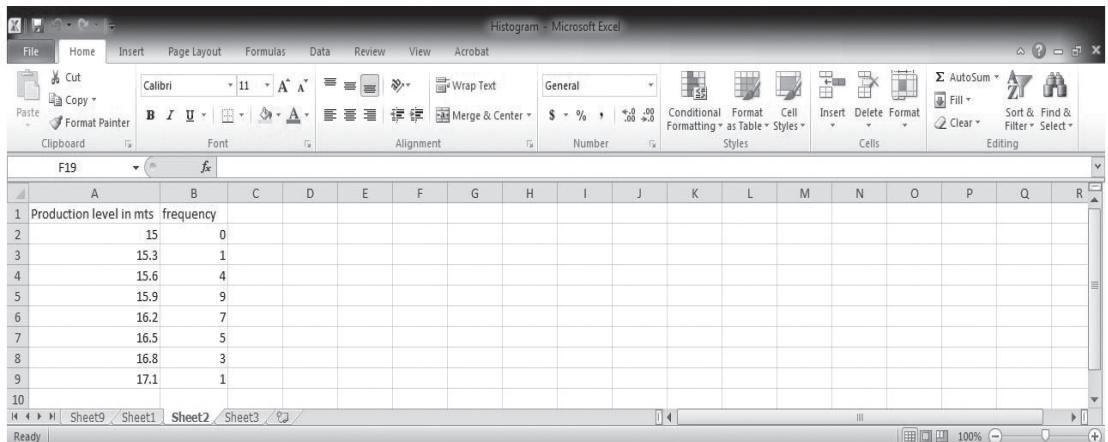


For histogram go to **DATA>DATA ANALYSIS >HISTOGARAM>DEFINE INPUT RANGE, BIN RANGE (mid values)> SELECT CHART OUTPUT>OK.**

Now for correcting generated histogram click on any data series and go to Format Data Series and set gap width as zero, go to border style and set width as 2.

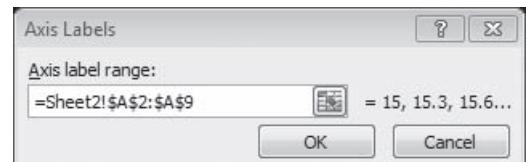
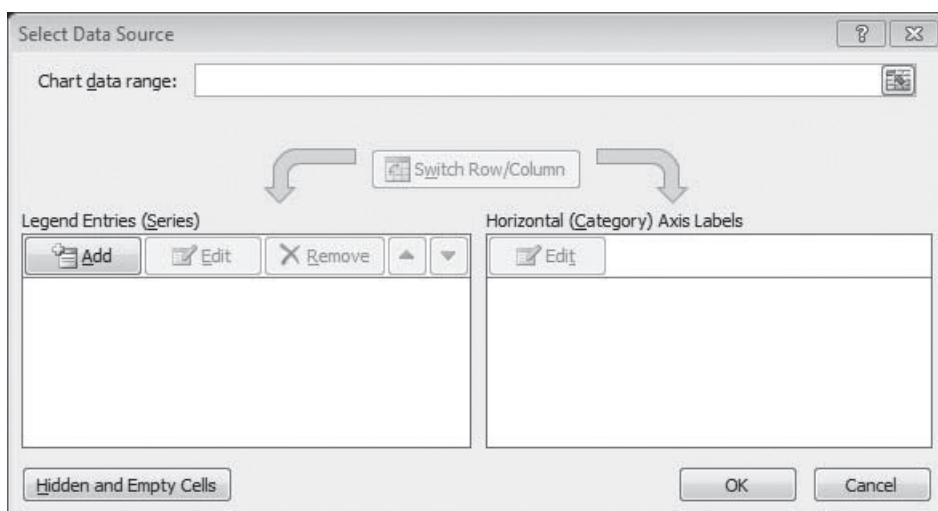
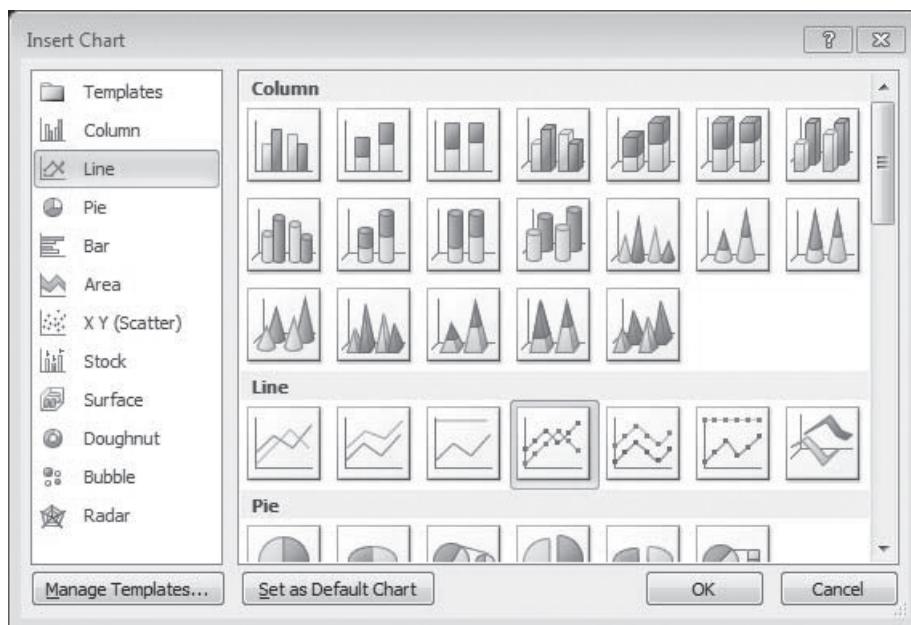


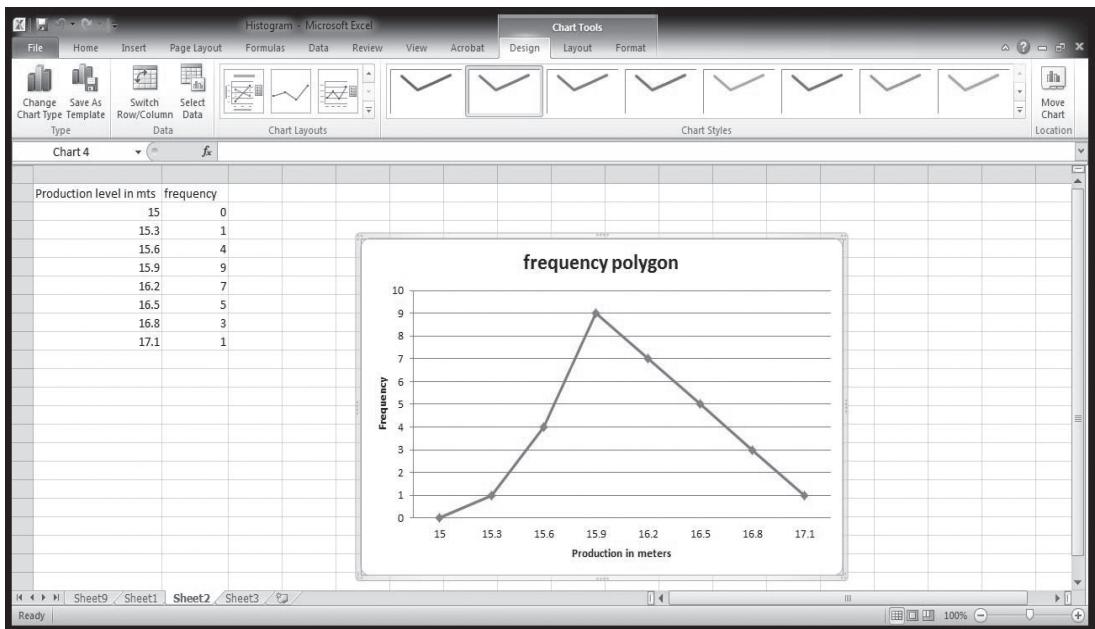
Frequency Polygon



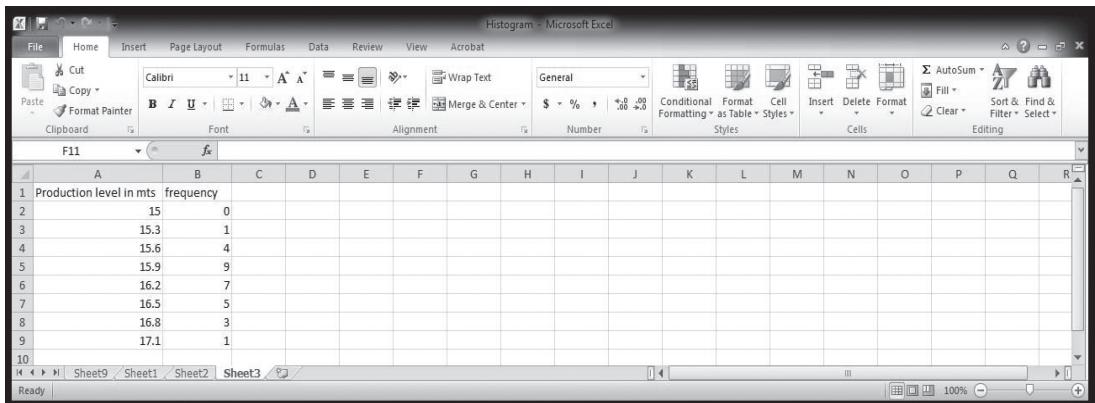
Above data is sample of daily production in meters of 30 carpet looms and the desired mid values for creating frequency polygon.

For Frequency Polygon go to **Insert>Chart>Line>Line with Markers>Select Data Source>Add Legend Entries>Select Series Name>Select Series Value> Add Horizontal Axis Label**



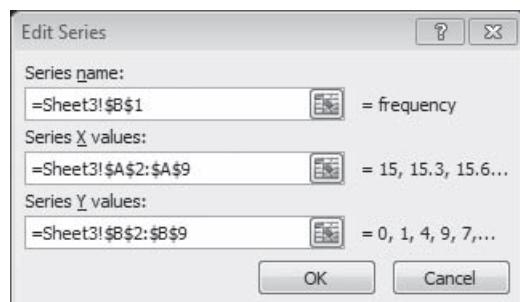
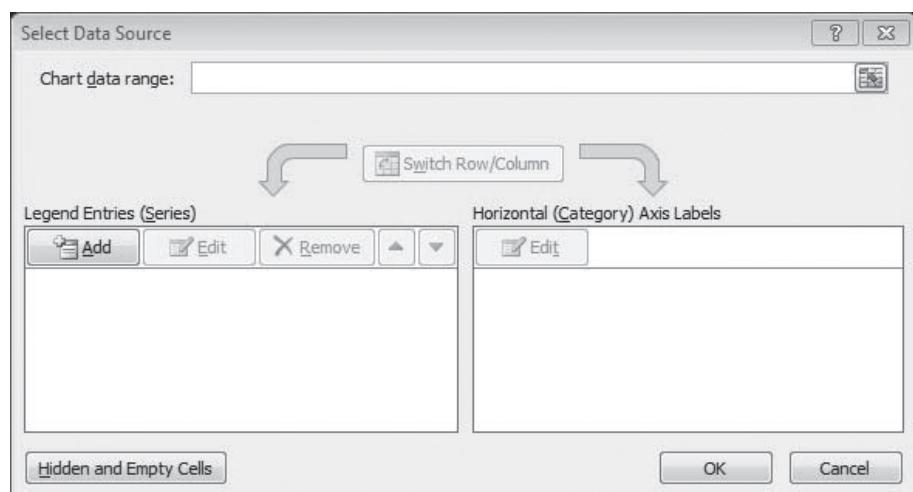
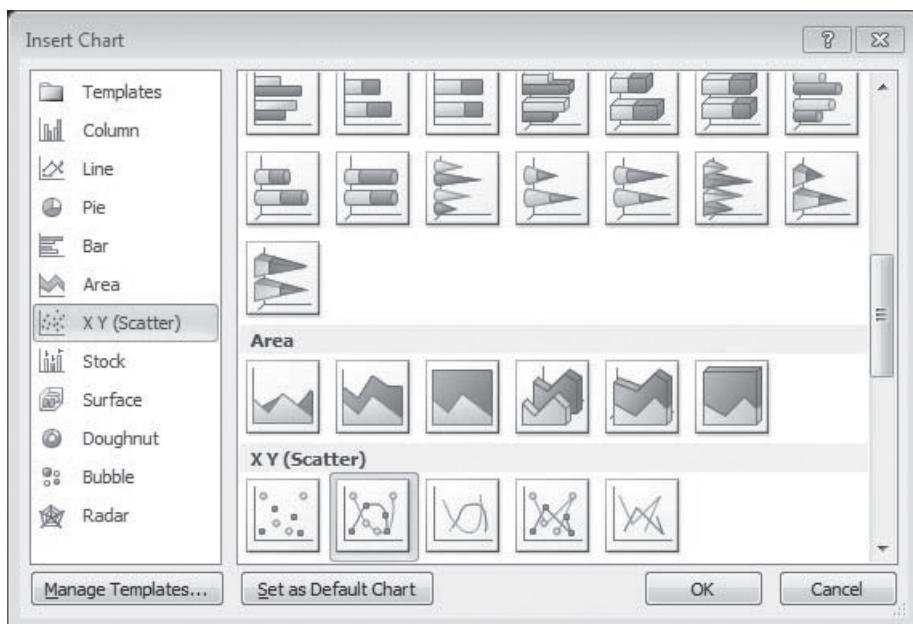


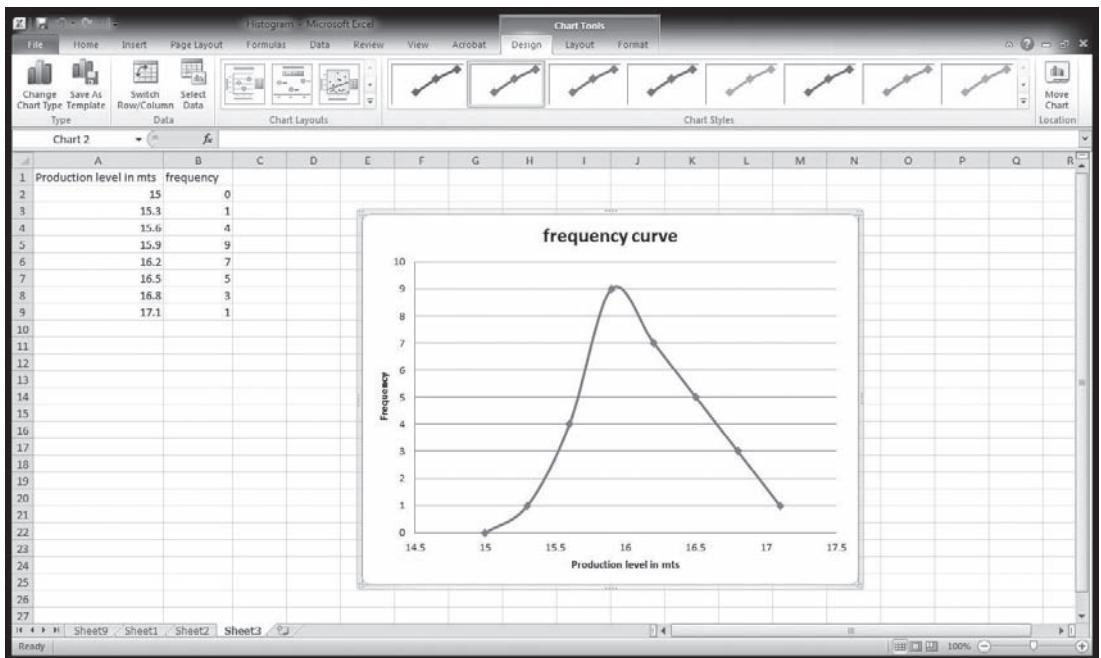
Frequency Curve



Above data is sample of daily production in meters of 30 carpet looms and the desired mid values for creating frequency curve.

For Frequency Polygon go to Insert>Chart>XY (Scatter)>Scatter with smooth line and markers> Select Data>Select Data Source>Legend Entries>Give Series X Values>Give Series Y Values

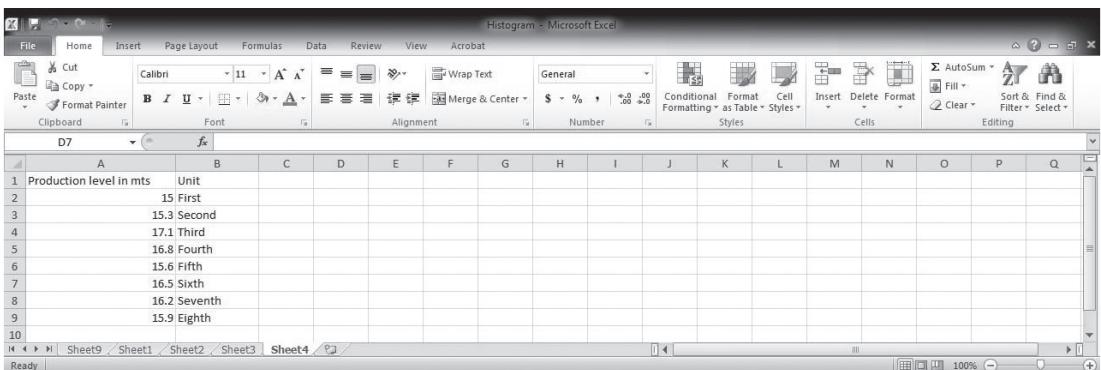


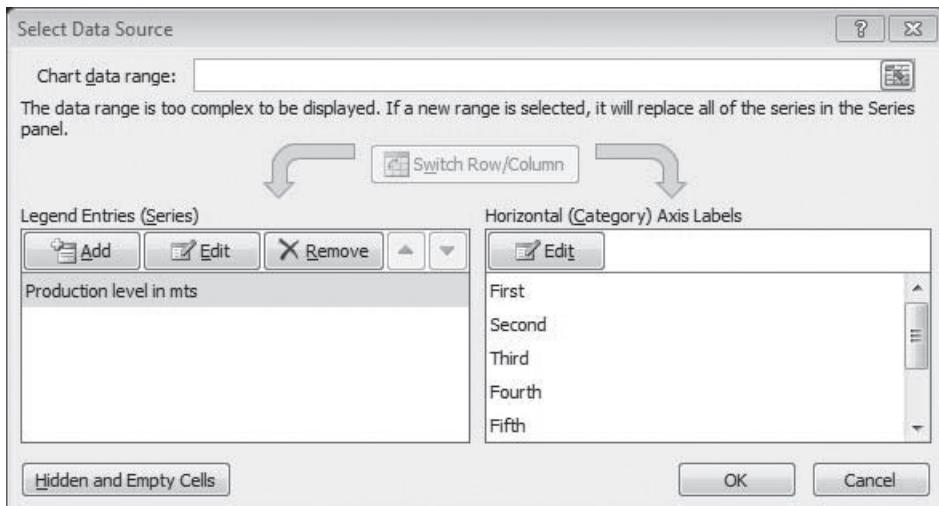
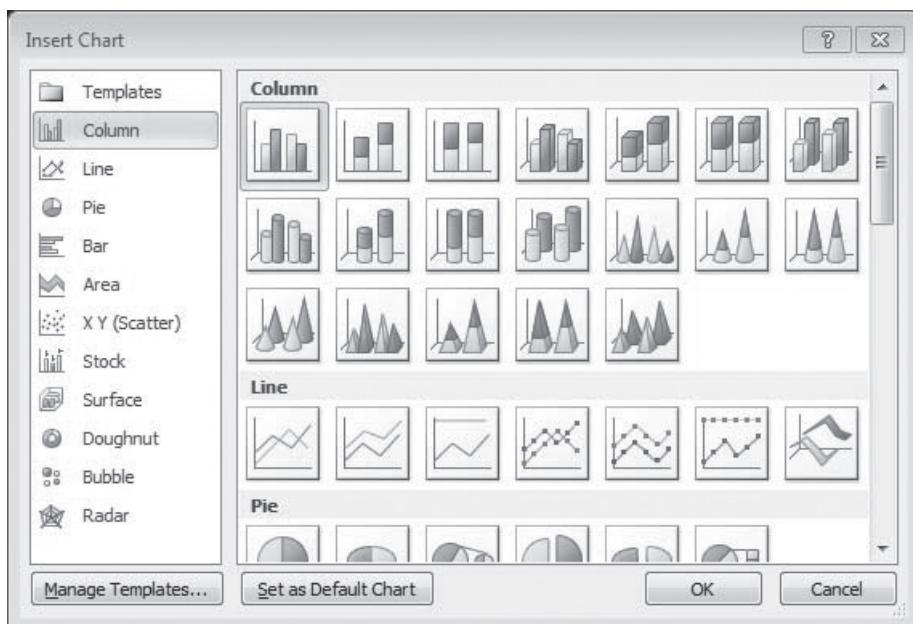


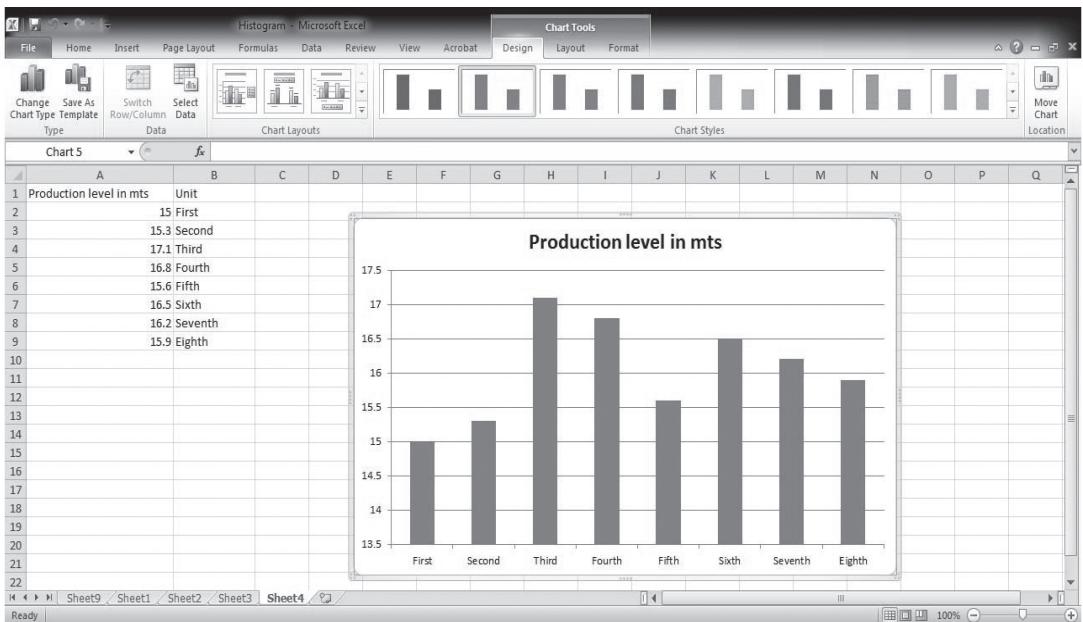
Bar Chart

Below data is sample unit wise production in meters of 30 carpet looms for creating bar chart.

For Bar chart go to **Insert>Chart>Column>Clustered Column>Select Data>Add Legend Entries>Add Horizontal Axis Label**

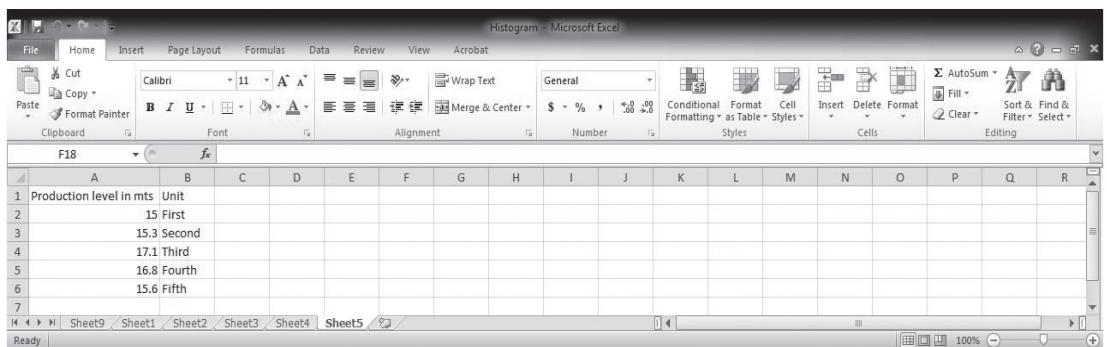




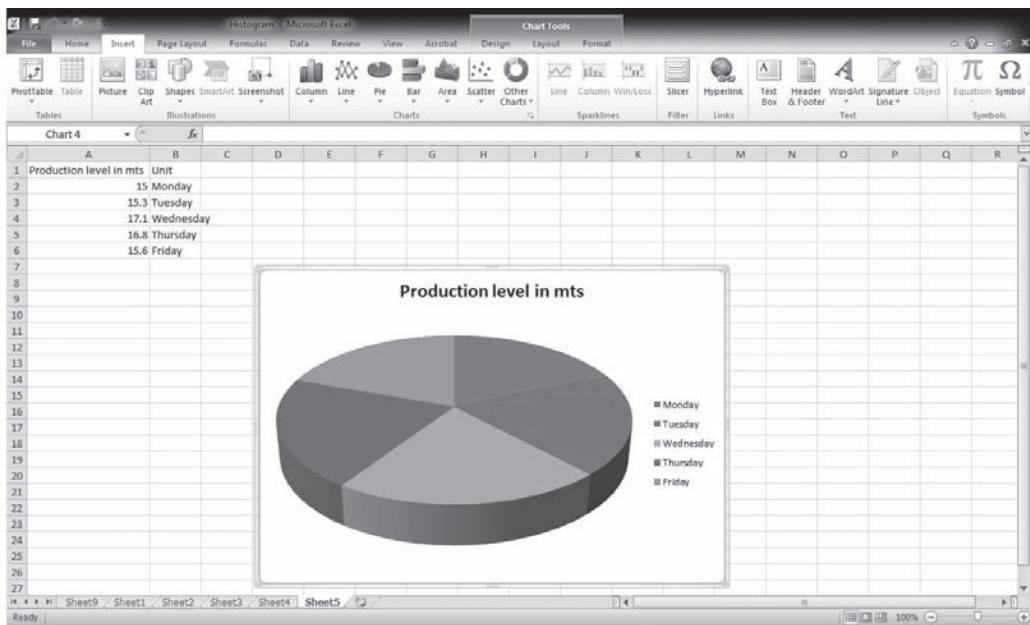
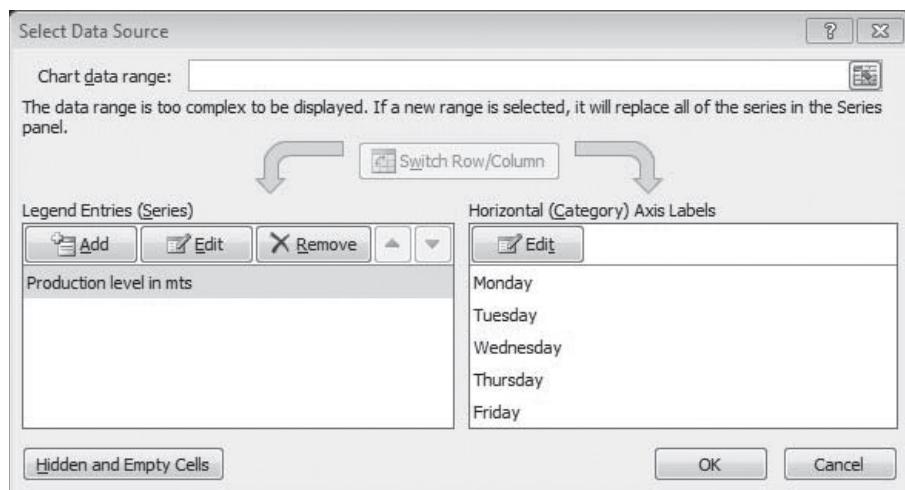


Pie Chart

Below data is sample weekday production in meters of 30 carpet looms for creating pie chart.



For Pie Chart go to **Insert>Chart>Pie>Pie in 3D>Select Data>Add Legend Entries>Add Horizontal Axis Label**



HINTS & ASSUMPTIONS

Whoever said “a picture is worth a thousand words” understood intuitively what we have been covering in this section. Using graphic methods to display data gives us a quick sense of patterns and trends and what portion of our data is above or below a certain value. Warning: Some publications print graphic displays of data (histograms) in a way that is confusing by using a vertical axis that doesn’t go all the way to zero. Be aware when you see one of these that small differences have been made to look too large, and that the pattern you are seeing is misleading.

EXERCISES 2.5

Self-Check Exercises

- SC 2-5** Here is a frequency distribution of the weight of 150 people who used a ski lift a certain day. Construct a histogram for these data.

Class	Frequency	Class	Frequency
75–89	10	150–164	23
90–104	11	165–179	9
105–119	23	180–194	9
120–134	26	195–209	6
135–149	31	210–224	2

- (a) What can you see from the histogram about the data that was not immediately apparent from the frequency distribution?
- (b) If each ski lift chair holds two people but is limited in total safe weight capacity to 400 pounds, what can the operator do to maximize the people capacity of the ski lift without exceeding the safe weight capacity of a chair? Do the data support your proposal?
- SC 2-6** Central Carolina Hospital has the following data representing weight in pounds at birth of 200 premature babies.

Class	Frequency	Class	Frequency
0.5–0.9	10	2.5–2.9	29
1.0–1.4	19	3.0–3.4	34
1.5–1.9	24	3.5–3.9	40
2.0–2.4	27	4.0–4.4	17

Construct an ogive that will help you answer these questions:

- (a) What was the approximate middle value in the original data set?
- (b) If premature babies under 3.0 pounds are normally kept in an incubator for several days as a precaution, about what percentage of Central's premature babies will need an incubator?

Applications

- 2-34** Here is a frequency distribution of the length of phone calls made by 175 people during a Labor Day weekend. Construct a histogram for these data.

Length in Minutes	Frequency
1–7	45
8–14	32
15–21	34
22–28	22
29–35	16
36–42	12
43–49	9
50–56	5

- (a) Describe the general shape of the histogram. Does there appear to be a pattern?
 (b) Suppose all the people were making their calls from a room that had 10 different phones, and each person knew which time class the call would belong to. Suggest an ordering so that all calls can be completed as fast as possible.
 (c) Does the order affect the length of time to complete all calls?

2-35 Golden Acres is a homeowners' association that operates a trailer park outside Orlando, Florida, where retirees keep their winter homes. In addition to lot rents, a monthly facility fee of \$12 is charged for social activities at the clubhouse. One board member has noted that many of the older residents never attend the clubhouse functions, and has proposed waiving the fee for association members over age 60. A survey of 25 residents reported the following ages:

66	65	96	80	71
93	66	96	75	61
69	61	51	84	58
73	77	89	69	92
57	56	55	78	96

Construct an ogive that will help you answer these questions:

- (a) Roughly what proportion of residents would be eligible for no fee?
 (b) Approximately what fee would the board have to charge to the remaining (fee-paying) residents to cover the same total cost of running the clubhouse?

2-36 Homer Willis, a fishing boat captain from Salter Path, North Carolina, believes that the break-even catch on his boats is 5,000 pounds per trip. Here are data on a sample of catches on 20 fishing trips Homer's boats have made recently:

6,500	6,700	3,400	3,600	2,000
7,000	5,600	4,500	8,000	5,000
4,600	8,100	6,500	9,000	4,200
4,800	7,000	7,500	6,000	5,400

Construct an ogive that will help you answer these questions:

- (a) Roughly what proportion of the trips breaks even for Homer?
 (b) What is the approximate middle value in the data array for Homer's boats?
 (c) What catch do Homer's boats exceed 80 percent of the time?

2-37 The Massachusetts Friends of Fish has the following data representing pollutants (in parts per million) at 150 sites in the state:

Pollutants (in ppm)	Frequency	Pollutants (in ppm)	Frequency
5.0– 8.9	14	25.0–28.9	16
9.0–12.9	16	29.0–32.9	9
13.0–16.9	28	33.0–36.9	7
17.0–20.9	36	37.0–40.9	4
21.0–24.9	20		

Construct an ogive that will help you answer the following questions:

- (a) Below what value (approximately) do the lowest one-fourth of these observations fall?
 (b) If the Friends of Fish heavily monitor all sites with more than 30 ppm of pollutants, what percentage of sites will be heavily monitored?

- 2-38** Before constructing a dam on the Colorado River, the U.S. Army Corps of Engineers performed a series of tests to measure the water flow past the proposed location of the dam. The results of the testing were used to construct the following frequency distribution:

River Flow (Thousands of Gallons per Minute)	Frequency
1,001–1,050	7
1,051–1,100	21
1,101–1,150	32
1,151–1,200	49
1,201–1,250	58
1,251–1,300	41
1,301–1,350	27
1,351–1,400	11
Total	246

- (a) Use the data given in the table to construct a “more-than” cumulative frequency distribution and ogive.
 (b) Use the data given in the table to construct a “less-than” cumulative frequency distribution and ogive.
 (c) Use your ogive to estimate what proportion of the flow occurs at less than 1,300 thousands of gallons per minute.
- 2-39** Pamela Mason, a consultant for a small local brokerage firm, was attempting to design investment programs attractive to senior citizens. She knew that if potential customers could obtain a certain level of return, they would be willing to risk an investment, but below a certain level, they would be reluctant. From a group of 50 subjects, she obtained the following data regarding the various levels of return required for each subject to invest \$1,000:

Indifference Point	Frequency	Indifference Point	Frequency
\$70–74	2	\$ 90– 94	11
75–79	5	95– 99	3
80–84	10	100–104	3
85–89	14	105–109	2

- (a) Construct both “more-than” and “less-than” cumulative relative frequency distributions.
 (b) Graph the 2 distributions in part (a) into relative frequency ogives.
- 2-40** At a newspaper office, the time required to set the entire front page in type was recorded for 50 days. The data, to the nearest tenth of a minute, are given below.

20.8	22.8	21.9	22.0	20.7	20.9	25.0	22.2	22.8	20.1
25.3	20.7	22.5	21.2	23.8	23.3	20.9	22.9	23.5	19.5
23.7	20.3	23.6	19.0	25.1	25.0	19.5	24.1	24.2	21.8
21.3	21.5	23.1	19.9	24.2	24.1	19.8	23.9	22.8	23.9
19.7	24.2	23.8	20.7	23.8	24.3	21.1	20.9	21.6	22.7

- (a) Arrange the data in an array from lowest to highest.
 (b) Construct a frequency distribution and a “less-than” cumulative frequency distribution from the data, using intervals of 0.8 minute.

- (c) Construct a frequency polygon from the data.
 (d) Construct a “less-than” ogive from the data.
 (e) From your ogive, estimate what percentage of the time the front page can be set in less than 24 minutes.

2-41 Chien-Ling Lee owns a CD store specializing in spoken-word recordings. Lee has 35 months of gross sales data, arranged as a frequency distribution.

Monthly Sales	Frequency	Monthly Sales	Frequency
\$10,000–12,499	2	\$20,000–22,499	6
12,500–14,999	4	22,500–24,999	8
15,000–17,499	7	25,000–27,499	2
17,500–19,999	5	27,500–29,999	1

- (a) Construct a relative frequency distribution.
 (b) Construct, on the same graph, a relative frequency histogram and a relative frequency polygon.

2-42 The National Association of Real Estate Sellers has collected these data on a sample of 130 salespeople representing their total commission earnings annually:

Earnings	Frequency
\$ 5,000 or less	5
\$ 5,001–\$10,000	9
\$10,001–\$15,000	11
\$15,001–\$20,000	33
\$20,001–\$30,000	37
\$30,001–\$40,000	19
\$40,001–\$50,000	9
Over \$50,000	7

Construct an ogive that will help you answer these questions.

- (a) About what proportion of the salespeople earns more than \$25,000?
 (b) About what does the “middle” salesperson in the sample earn?
 (c) Approximately how much could a real estate salesperson whose performance was about 25 percent from the top expect to earn annually?

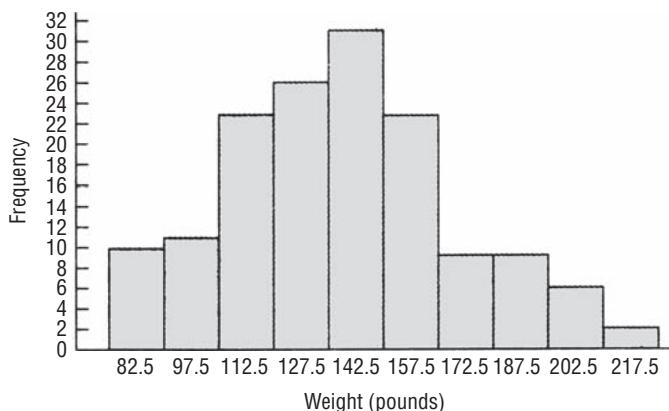
2-43 Springfield is a college town with the usual parking problems. The city allows people who have received tickets for illegally parked cars to come in and make their case to an administrative officer and have the ticket voided. The town’s administrative officer collected the following frequency distribution for the time spent on each appeal:

Minutes Spent on Appeal	Frequency	Minutes Spent on Appeal	Frequency
Less than 2	30	8–9	70
2–3	40	10–11	50
4–5	40	12–13	50
6–7	90	14–15	30
			400

- (a) Construct a “less-than” cumulative frequency distribution.
- (b) Construct an ogive based on part (a).
- (c) The town administrator will consider streamlining the paperwork for the appeal process if more than 50 percent of appeals take longer than 4 minutes. What is the percentage taking more than 4 minutes? What is the approximate time for the 200th (midpoint) appeal?

Worked-Out Answers to Self-Check Exercises

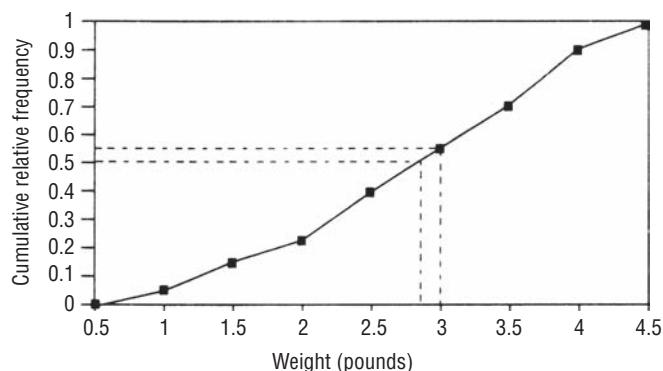
SC 2-5



- (a) The lower tail of the distribution is fatter (has more observations in it) than the upper tail.
- (b) Because there are so few people who weigh 180 pounds or more, the operator can afford to pair each person who appears to be heavy with a lighter person. This can be done without greatly delaying any individual's turn at the lift.

C 2-6

Class	Cumulative Relative Frequency	Class	Cumulative Relative Frequency
0.5–0.9	0.050	2.5–2.9	0.545
1.0–1.4	0.145	3.0–3.4	0.715
1.5–1.9	0.265	3.5–3.9	0.915
2.0–2.4	0.400	4.0–4.4	1.000



- (a) The middle value was about 2.8 pounds.
- (b) About 55 percent will need incubators.

STATISTICS AT WORK

Loveland Computers

Case 2: Arranging Data New Year's Day 1995, found Lee Azko staring out the window, watching a light dusting of snow fall on the Denver suburbs. Lee had graduated early from the University of Colorado, one semester short of the usual 4 years, thanks to a handful of advanced placement credits from high school. Lee was both excited and apprehensive that the next day would be the start of a serious job search for a well-trained business major, with little experience in the real world.

Contemplation of the future was interrupted by a phone call from Lee's uncle. "I was going to call you anyway to congratulate you on finishing school early. But I have another reason for calling—some things have come up in the business, and it looks as if I need someone to crunch some numbers in a hurry. Why don't you drive up tomorrow and I'll tell you what I have in mind."

Lee knew that Uncle Walter's company, Loveland Computers, had been growing by leaps and bounds. Walter Azko had developed the computer company from a strange background. Unlike Lee, Walter never finished college. "I was making too much money to stay in school," he used to explain. Walter had traveled extensively in the Far East with his parents, so it was only natural that he would begin an importing business while still a student at Boulder. He imported just about anything that could be sold cheaply and that would appeal to students: furniture, gifts, household utensils, and some clothing.

On one buying trip to Taiwan in the early 1980s, Walter was offered some personal computers. Looking back, they were awful. Not much memory and no hard drive, but they were dirt cheap and Walter soon sold them to "tekkies" at the university. The computer business grew, and within 2 years, Walter sold his retail importing business and concentrated solely on importing and selling computers.

Walter's first move was to lease a commercial building in Loveland, Colorado, where rents were much cheaper than in Boulder. From this location, he could market directly to students at the Universities at Boulder, Fort Collins, and Greeley. About an hour north of Denver's Stapleton International Airport, Loveland was a convenient site for imports coming by airfreight and a good place to recruit part-time workers. The name Loveland Computers seemed a natural.

At first, Walter Azko acted as his own sales staff, personally delivering computers from the back of his car. Walter made every sale on price alone and word-of-mouth referrals supplemented a few ads placed in the college newspapers. Because he sold directly to students and enthusiasts, it seemed that he was the only game in town. Walter's niche seemed to be an altogether different market from the one being reached by the industry giants. At the top end of the market for PCs, IBM was using expensive retail distribution, targeting the business market. And Apple was defending its high-price strategy with easy "point-and-click" graphical computing that couldn't be matched by IBM-compatible machines.

Azko began reading computer magazines and found he wasn't the only box shop (the industry name for a company that shipped boxes of computers to users with little or no additional service). One or two other companies had found cheap overseas suppliers and they were pursuing a mail-order strategy. Walter thought customers would be reluctant to buy such an expensive—and novel—piece of equipment sight unseen, but the arrival of a new shipment of computers with preinstalled hard disk drives gave him the motivation to run a few ads of his own.

So Loveland Computers joined the ranks of the national mail-order box shops, and by 1988, the company was one of the two dozen companies in this market. The mail-order companies together shared about the same percentage of the market as "Big Blue" (IBM) was maintaining: about 20 percent. But the market for PCs was huge and growing rapidly. By 1993, Loveland Computers regularly booked sales of \$10 million a quarter, and even at discount prices, profits regularly amounted to 6 percent of sales. Uncle Walter had become a rich man.

Along the way, Walter Azko realized that to give customers exactly what they wanted, there were advantages in assembling computers at his ever-expanding Loveland facility. He never saw himself as a manufacturer—just an assembler of premade parts such as drive controllers and power supplies. With his contacts with overseas manufacturers, Walter was able to hunt around for the best prices, so Loveland Computers' costs remained low.

To configure new machines and to help with specifications, Walter hired a bright young engineer, Gratia Delaguardia. Gratia knew hardware: She had completed several development projects for Storage Technology. In only a few years at Loveland Computers, she built a development staff of more than two dozen and was rewarded with a partnership in the business.

Loveland Computers had a few setbacks due to misjudging demand. Walter Azko was always optimistic about sales so inventory of components was often much greater than needed. Once or twice there were embarrassing “write-downs,” such as when a shipment of power supplies turned out to be useless because they produced too little current for Loveland’s latest model. Gratia Delaguardia had concluded that Loveland ought to be able to manage the supplies better, but it seemed difficult to predict what the market would be like from one month to the next.

After a sleepless night, Lee Azko met with Loveland Computers’ founder and president. “Come and sit over here by the window—you can see my new Mercedes 500 SL sports car,” Walter Azko said, welcoming his young visitor. “Let me tell you my problem. You know that things move pretty fast around here. Seems like each model lasts about 6 months and then we replace it with something fancier. Up to this point, I’ve pretty much relied on the local bank for financing. But this is a hot business and we’re getting some attention from folks on Wall Street. We may be doing a ‘private placement’—that’s where we’d raise money for expansion from one or two well-heeled investors or banks—and then, later on, we might want to take the company public. Thing is, they want to know a whole lot about our sales growth: how much is coming from which products and so on. They want to know how long each model lasts, what we should project for next year. Now, of course, I have monthly sales reports going back almost to the beginning. The good news is, it’s all on disk. The bad news is, we kept changing our formats so it’s very difficult to compare numbers. And, of course, no one wants to flip through, say, 48 months of reports. Your job is to organize it all so it makes sense when these city slickers come to town in their corporate jet.”

“When would I start, Uncle?” asked Lee Azko, quite taken aback by the task ahead.

“You’ve already started,” snapped Walter. “It’s when you finish that’s important. These folks are due in next Monday.”

Lee made a mental note to cancel a ski trip planned for the weekend and pulled out a notepad and started to sketch out a plan.

Study Questions: What information should Lee gather, other than financial information relating to sales and income? What format will present the company’s rapid growth most clearly in a 45-minute business presentation?

CHAPTER REVIEW

Terms Introduced in Chapter 2

Continuous Data Data that may progress from one class to the next without a break and may be expressed by either whole numbers or fractions.

Cumulative Frequency Distribution A tabular display of data showing how many observations lie above, or below, certain values.

Data A collection of any number of related observations on one or more variables.

Data Array The arrangement of raw data by observations in either ascending or descending order.

Data Point A single observation from a data set.

Data Set A collection of data.

Discrete Classes Data that do not progress from one class to the next without a break; that is, where classes represent distinct categories or counts and may be represented by whole numbers.

Frequency Curve A frequency polygon smoothed by adding classes and data points to a data set.

Frequency Distribution An organized display of data that shows the number of observations from the data set that falls into each of a set of mutually exclusive and collectively exhaustive classes.

Frequency Polygon A line graph connecting the midpoints of each class in a data set, plotted at a height corresponding to the frequency of the class.

Histogram A graph of a data set, composed of a series of rectangles, each proportional in width to the range of values in a class and proportional in height to the number of items falling in the class, or the fraction of items in the class.

Ogive A graph of a cumulative frequency distribution.

Open-Ended Class A class that allows either the upper or lower end of a quantitative classification scheme to be limitless.

Population A collection of all the elements we are studying and about which we are trying to draw conclusions.

Raw Data Information before it is arranged or analyzed by statistical methods.

Relative Frequency Distribution The display of a data set that shows the fraction or percentage of the total data set that falls into each of a set of mutually exclusive and collectively exhaustive classes.

Representative Sample A sample that contains the relevant characteristics of the population in the same proportions as they are included in that population.

Sample A collection of some, but not all, of the elements of the population under study, used to describe the population.

Equations Introduced in Chapter 2

$$2-1 \quad \text{Width of class intervals} = \frac{\text{Next unit value after largest value in data} - \text{Smallest value in data}}{\text{Total number of class intervals}} \quad \text{p. 28}$$

To arrange raw data, decide the number of classes into which you will divide the data (normally between 6 and 15), and then use Equation 2-1 to determine the *width of class intervals of equal size*. This formula uses the next value of the same units because it measures the interval between the first value of one class and the first value of the next class.

Review and Application Exercises

- 2-44 The following set of raw data gives income and education level for a sample of individuals. Would rearranging the data help us to draw some conclusions? Rearrange the data in a way that makes them more meaningful.

Income	Education	Income	Education	Income	Education
\$17,000	High school	\$ 21,200	B.S.	\$17,200	2 years college
20,800	B.S.	28,000	B.S.	19,600	B.A.
27,000	M.A.	30,200	High school	36,200	M.S.
70,000	M.D.	22,400	2 years college	14,400	1 year college
29,000	Ph.D.	100,000	M.D.	18,400	2 years college
14,400	10th grade	76,000	Law degree	34,400	B.A.
19,000	High school	44,000	Ph.D.	26,000	High school
23,200	M.A.	17,600	11th grade	52,000	Law degree
30,400	High school	25,800	High school	64,000	Ph.D.
25,600	B.A.	20,200	1 year college	32,800	B.S.

2-45 All 50 states send the following information to the Department of Labor: the average number of workers absent daily during the 13 weeks of a financial quarter, and the percentage of absentees for each state. Is this an example of raw data? Explain.

2.46 The Nebraska Department of Agriculture has these data representing weekly growth (in inches) on samples of newly planted spring corn:

0.4	1.9	1.5	0.9	0.3	1.6	0.4	1.5	1.2	0.8
0.9	0.7	0.9	0.7	0.9	1.5	0.5	1.5	1.7	1.8

- (a) Arrange the data in an array from highest to lowest.
- (b) Construct a relative frequency distribution using intervals of 0.25.
- (c) From what you have done so far, what conclusions can you come to about growth in this sample?
- (d) Construct an ogive that will help you determine what proportion of the corn grew at more than 1.0 inch a week.
- (e) What was the approximate weekly growth rate of the middle item in the data array?

2-47 The National Safety Council randomly sampled the tread depth of 60 right front tires on passenger vehicles stopped at a rest area on an interstate highway. From its data, it constructed the following frequency distribution:

Tread Depth (Inches)	Frequency	Tread Depth (Inches)	Frequency
$\frac{16}{32}$ (new tire)	5	$\frac{4}{32} - \frac{6}{32}$	7
$\frac{13}{32} - \frac{15}{32}$	10	$\frac{1}{32} - \frac{3}{32}$	4
$\frac{10}{32} - \frac{12}{32}$	20	$\frac{0}{32}$ bald	2
$\frac{7}{32} - \frac{9}{32}$	12		

- (a) Approximately what was the tread depth of the thirtieth tire in the data array?
- (b) If a tread depth less than $\frac{7}{32}$ inch is considered dangerous, approximately what proportion of the tires on the road are unsafe?

2-48 The High Point Fastener Company produces 15 basic items. The company keeps records on the number of each item produced per month in order to examine the relative production

levels. Records show the following numbers of each item were produced by the company for the last month of 20 operating days:

9,897	10,052	10,028	9,722	9,908
10,098	10,587	9,872	9,956	9,928
10,123	10,507	9,910	9,992	10,237

Construct an ogive that will help you answer these questions.

- (a) On how many of its items did production exceed the break-even point of 10,000 units?
- (b) What production level did 75 percent of its items exceed that month?
- (c) What production level did 90 percent of its items exceed that month?

2-49

The administrator of a hospital has ordered a study of the amount of time a patient must wait before being treated by emergency room personnel. The following data were collected during a typical day:

Waiting Time (Minutes)									
12	16	21	20	24	3	11	17	29	18
26	4	7	14	25	1	27	15	16	5

- (a) Arrange the data in an array from lowest to highest. What comment can you make about patient waiting time from your data array?
- (b) Now construct a frequency distribution using 6 classes. What additional interpretation can you give to the data from the frequency distribution?
- (c) From an ogive, state how long 75 percent of the patients should expect to wait based on these data.

2-50

Of what additional value is a relative frequency distribution once you have already constructed a frequency distribution?

2-51

Below are the weights of an entire population of 100 NFL football players.

226	198	210	233	222	175	215	191	201	175
264	204	193	244	180	185	190	216	178	190
174	183	201	238	232	257	236	222	213	207
233	205	180	267	236	186	192	245	218	193
189	180	175	184	234	234	180	252	201	187
155	175	196	172	248	198	226	185	180	175
217	190	212	198	212	228	184	219	196	212
220	213	191	170	258	192	194	180	243	230
180	135	243	180	209	202	242	259	238	227
207	218	230	224	228	188	210	205	197	169

- (a) Select two samples: one sample of the first 10 elements, and another sample of the largest 10 elements.
- (b) Are the two samples equally representative of the population? If not, which sample is more representative, and why?
- (c) Under what conditions would the sample of the largest 10 elements be as representative as the sample of the first 10 elements?

2-52

In the population under study, there are 2,000 women and 8,000 men. If we are to select a sample of 250 individuals from this population, how many should be women to make our sample considered strictly representative?

- 2-53** The U.S. Department of Labor publishes several classifications of the unemployment rate, as well as the rate itself. Recently, the unemployment rate was 6.8 percent. The department reported the following educational categories:

Level of Education	Relative Frequency (% of Those Unemployed)
Did not complete high school	35%
Received high school diploma	31
Attended college but did not receive a degree	16
Received a college degree	9
Attended graduate school but did not receive a degree	6
Received a graduate degree	3
Total	100%

Using these data, construct a relative frequency histogram.

- 2-54** Using the relative frequency distribution given in Exercise 2-63, construct a relative frequency histogram and polygon. For the purposes of the present exercise, assume that the upper limit of the last class is \$51.00.

- 2-55** Consider the following information about March 1992 nonfarm employment (in thousands of workers) in the United States, including Puerto Rico and the Virgin Islands:

Alabama	1,639.0	Nebraska	730.6
Alaska	235.5	Nevada	638.4
Arizona	1,510.0	New Hampshire	466.5
Arkansas	951.1	New Jersey	3,390.7
California	12,324.3	New Mexico	583.3
Colorado	1,552.7	New York	7,666.4
Connecticut	1,510.6	North Carolina	3,068.3
Delaware	335.2	North Dakota	271.0
District of Columbia	667.0	Ohio	4,709.9
Florida	5,322.8	Oklahoma	1,196.9
Georgia	2,927.1	Oregon	1,245.6
Hawaii	546.3	Pennsylvania	4,992.1
Idaho	400.4	Rhode Island	413.2
Illinois	5,146.2	South Carolina	1,494.6
Indiana	2,496.3	South Dakota	295.6
Iowa	1,229.2	Tennessee	2,178.6
Kansas	1,108.3	Texas	7,209.7
Kentucky	1,474.8	Utah	752.2
Louisiana	1,617.5	Vermont	244.8
Maine	500.0	Virginia	2,792.4
Maryland	2,037.3	Washington	2,165.8
Massachusetts	2,751.6	West Virginia	622.1
Michigan	3,828.9	Wisconsin	2,272.1
Minnesota	2,117.1	Wyoming	198.0
Mississippi	940.9	Puerto Rico	842.4
Missouri	2,275.9	Virgin Islands	42.4
Montana	299.3		

Source: Sharon R. Cohany, "Employment Data," *Monthly Labor Review* 115(6), (June 1992): 80–82.

- (a) Arrange the data into 10 equal-width, mutually exclusive classes.
- (b) Determine the frequency and relative frequency within each class.
- (c) Are these data discrete or continuous?
- (d) Construct a “less-than” cumulative frequency distribution and ogive for the relative frequency distribution in part (b).
- (e) Based on the ogive constructed in part (d), what proportion of states have nonfarm employment greater than 3 million?

2-56 Using the frequency distribution given in Exercise 2-57 for miles per day of jogging, construct an ogive that will help you estimate what proportion of the joggers are averaging 4.0 miles or fewer daily.

2-57 A sports psychologist studying the effect of jogging on college students’ grades collected data from a group of college joggers. Along with some other variables, he recorded the average number of miles run per day. He compiled his results into the following distribution:

Miles per Day	Frequency
1.00–1.39	32
1.40–1.79	43
1.80–2.19	81
2.20–2.59	122
2.60–2.99	131
3.00–3.39	130
3.40–3.79	111
3.80–4.19	95
4.20–4.59	82
4.60–4.99	47
5.00 and up	53
	927

- (a) Construct an ogive that will tell you approximately how many miles a day the middle jogger runs.
- (b) From the ogive you constructed in part (a), approximately what proportion of college joggers run at least 3.0 miles a day?

2-58 A behavioral researcher studying the success of college students in their careers conducts interviews with 100 Ivy League undergraduates, half men and half women, as the basis for the study. Comment on the adequacy of this survey.

2-59 If the following age groups are included in the proportions indicated, how many of each age group should be included in a sample of 3,000 people to make the sample representative?

Age Group	Relative Proportion in Population
12–17	0.17
18–23	0.31
24–29	0.27
30–35	0.21
36+	0.04

- 2-60** State University has three campuses, each with its own business school. Last year, State's business professors published numerous articles in prestigious professional journals, and the board of regents counted these articles as a measure of the productivity of each department.

Journal Number	Number of Publications	Campus	Journal Number	Number of Publications	Campus
9	3	North	14	20	South
12	6	North	10	18	South
3	12	South	3	12	West
15	8	West	5	6	North
2	9	West	7	5	North
5	15	South	7	15	West
1	2	North	6	2	North
15	5	West	2	3	West
12	3	North	9	1	North
11	4	North	11	8	North
7	9	North	14	10	West
6	10	West	8	17	South

- (a) Construct a frequency distribution and a relative frequency distribution by journal.
- (b) Construct a frequency distribution and a relative frequency distribution by university branch.
- (c) Construct a frequency distribution and a relative frequency distribution by number of publications (using intervals of 3).
- (d) Briefly interpret your results.

- 2-61** A reporter wants to know how the cost of compliance with the Americans with Disabilities Act (ADA) has affected national hiring practices and sends out a form letter to 2,000 businesses in the same ZIP code as the magazine's editorial offices. A total of 880 responses are received. Comment on the data available in these responses in terms of the five tests for data.

- 2-62** With each appliance that Central Electric produces, the company includes a warranty card for the purchaser. In addition to validating the warranty and furnishing the company with the purchaser's name and address, the card also asks for certain other information that is used for marketing studies. For each of the numbered blanks on the card, determine the most likely characteristics of the categories that would be used by the company to record the information. In particular, would they be (1) quantitative or qualitative, (2) continuous or discrete, (3) open-ended or closed? Briefly state the reasoning behind your answers.

Name _____	Marital Status _____ (3)
Address _____	Where was appliance purchased?
City _____ State _____	(4) _____
Zip Code _____	Why was appliance purchased?
Age (1) _____ Yearly Income (2) _____	(5) _____

- 2-63** The following relative frequency distribution resulted from a study of the dollar amounts spent per visit by customers at a supermarket:

Amount Spent	Relative Frequency
\$ 0-\$ 5.99	1%
6.00-\$10.99	3
11.00-\$15.99	4
16.00-\$20.99	6
21.00-\$25.99	7
26.00-\$30.99	9
31.00-\$35.99	11
36.00-\$40.99	19
41.00-\$45.99	32
46.00 and above	8
Total	100%

Determine the class marks (midpoints) for each of the intervals.

- 2-64** The following responses were given by two groups of hospital patients, one receiving a new treatment, the other receiving a standard treatment for an illness. The question asked was, "What degree of discomfort are you experiencing?"

Group 1			Group 2		
Mild	Moderate	Severe	Moderate	Mild	Severe
None	Severe	Mild	Severe	None	Moderate
Moderate	Mild	Mild	Mild	Moderate	Moderate
Mild	Moderate	None	Moderate	Mild	Severe
Moderate	Mild	Mild	Severe	Moderate	Moderate
None	Moderate	Severe	Severe	Mild	Moderate

Suggest a better way to display these data. Explain why it is better.

- 2-65** The production manager of the Browner Bearing Company posted final worker performance ratings based on total units produced, percentages of rejects, and total hours worked. Is this an example of raw data? Why or why not? If not, what would the raw data be in this situation?

- 2-66** The head of a large business department wanted to classify the specialties of its 67 members. He asked Peter Wilson, a Ph.D. candidate, to get the information from the faculty members' publications. Peter compiled the following:

Specialty	Faculty Members Publishing
Accounting only	1
Marketing only	5
Statistics only	4
Finance only	2
Accounting and marketing	7

(continued)

Specialty	Faculty Members Publishing
Accounting and statistics	6
Accounting and finance	3
Marketing and finance	8
Statistics and finance	9
Statistics and marketing	21
No publications	$\frac{1}{67}$
	67

Construct a relative frequency distribution for the *types* of specialties. (*Hint:* The categories of your distribution will be mutually exclusive, but any individual may fall into several categories.)

- 2-67** Lesley Niles, a summer intern at the Internet Financial Services Corporation, has been asked to investigate the low participation rates in the company's 401(k) investment program. Niles read an article in *The Wall Street Journal* commenting on families' second wage-earner income as a determinant of plan participation. Niles went from office to office and interviewed executives eligible to participate. None of the executives reported a spouse with second income over \$35,000 and many families had no second income. To examine the situation, Niles decides to construct both frequency and relative frequency distributions.
- Develop a continuous, closed distribution with \$5,000 intervals.
 - Develop a continuous distribution open at both ends, with 6 categories. You may relax the requirement for \$5,000 intervals for the open-ended categories.

- 2-68** The Kawahondi Computer Company compiled data regarding the number of interviews required for each of its 40 salespeople to make a sale. Following are a frequency distribution and a relative frequency distribution of the number of interviews required per salesperson per sale. Fill in the missing data.

Number of Interviews (Classes)	Frequency	Relative Frequency
0–10	?	0.075
11–20	1	?
21–30	4	?
31–40	?	?
41–50	2	?
51–60	?	0.175
61–70	?	0.225
71–80	5	?
81–90	?	0.000
91–100	<u>?</u>	<u>0.025</u>
	<u>?</u>	<u>?</u>

- 2-69** A. T. Cline, the mine superintendent of the Grover Coal Co., has recorded the amount of time per workshift that Section Crew #3 shuts down its machinery for on-the-spot adjustments, repairs, and moving. Here are the records for the crew's last 35 shifts:

60	72	126	110	91	115	112
80	66	101	75	93	129	105
113	121	93	87	119	111	97
102	116	114	107	113	119	100
110	99	139	108	128	84	99

- (a) Arrange the data in an array from highest to lowest.
 (b) If Cline believes that a typical amount of downtime per shift is 108 minutes, how many of Crew #3's last 35 shifts exceeded this limit? How many were under the limit?
 (c) Construct a relative frequency distribution with 10-minute intervals.
 (d) Does your frequency distribution indicate that Cline should be concerned?
- 2-70** Cline has obtained information on Section Crew #3's coal production per shift for the same 35-shift period discussed in Exercise 2-69. The values are in tons of coal mined per shift:

356	331	299	391	364	317	386
360	281	360	402	411	390	362
311	357	300	375	427	370	383
322	380	353	371	400	379	380
369	393	377	389	430	340	368

- (a) Construct a relative frequency distribution with six equal intervals.
 (b) If Cline considers 330 to 380 tons per shift to be an expected range of output, how many of the crew's shifts produced less than expected? How many did better than expected?
 (c) Does this information affect the conclusions you reached from the preceding problem on equipment downtime?

- 2-71** Virginia Suboleski is an aircraft maintenance supervisor. A recent delivery of bolts from a new supplier caught the eye of a clerk. Suboleski sent 25 of the bolts to a testing lab to determine the force necessary to break each of the bolts. In thousands of pounds of force, the results are as follows:

147.8	137.4	125.2	141.1	145.7
119.9	133.3	142.3	138.7	125.7
142.0	130.8	129.8	141.2	134.9
125.0	128.9	142.0	118.6	133.0
151.1	125.7	126.3	140.9	138.2

- (a) Arrange the data into an array from highest to lowest.
 (b) What proportion of the bolts withstood at least 120,000 pounds of force? What proportion withstood at least 150,000 pounds?
 (c) If Suboleski knows that these bolts when installed on aircraft are subjected to up to 140,000 pounds of force, what proportion of the sample bolts would have failed in use? What should Suboleski recommend the company do about continuing to order from the new supplier?

- 2-72** The telephone system used by PHM, a mail-order company, keeps track of how many customers tried to call the toll-free ordering line but could not get through because all the firm's lines were busy. This number, called the phone overflow rate, is expressed as a percentage of the total number of calls taken in a given week. Mrs. Loy has used the overflow data for the last year to prepare the following frequency distribution:

Overflow Rate	Frequency	Overflow Rate	Frequency
0.00–2.50%	3	12.51–15.00%	4
2.51–5.00%	7	17.51–20.00%	3
5.00–7.50%	13	20.01–22.51%	2
7.51–10.00%	10	22.51–25.50%	2
10.00–12.50%	6	25.51 or greater	2
		52 Total number of weeks	

List and explain errors you can find in Mrs. Loy's distribution.

- 2-73** Hanna Equipment Co. sells process equipment to agricultural companies in developing countries. A recent office fire burned two staff members and destroyed most of Hanna's business records. Karl Slayden has just been hired to help rebuild the company. He has found sales records for the last 2 months:

Country	# of Sales	Country	# of Sales	Country	# of Sales
1	3	7	4	13	1
2	1	8	9	14	1
3	1	9	5	15	5
4	8	10	1	16	6
5	3	11	3	17	6
6	5	12	7	18	2
19	2	23	1	27	1
20	1	24	7	28	5
21	1	25	3		
22	2	26	1		

- (a) Arrange the sales data in an array from highest to lowest.
- (b) Construct two relative frequency distributions of number of sales, one with 3 classes and one with 9 classes. Compare the two. If Slayden knows nothing about Hanna's sales patterns, think about the conclusions he might draw from each about country-to-country sales variability.

- 2-74** Jeanne Moreno is analyzing the waiting times for cars passing through a large expressway toll plaza that is severely clogged and accident-prone in the morning. Information was collected on the number of minutes that 3,000 consecutive drivers waited in line at the toll gates:

Minutes of Waiting	Frequency	Minutes of Waiting	Frequency
less than 1	75	9–10.99	709
1–2.99	183	11–12.99	539
3–4.99	294	13–14.99	164
5–6.99	350	15–16.99	106
7–8.99	580		

- (a) Construct a “less-than” cumulative frequency and cumulative relative frequency distribution.
 (b) Construct an ogive based on part (a). What percentage of the drivers had to wait more than 4 minutes in line? 8 minutes?

2-75 Maribor Cement Company of Montevideo, Uruguay, hired Delbert Olsen, an American manufacturing consultant, to help design and install various production reporting systems for its concrete roof tile factory. For example, today Maribor made 7,000 tiles and had a breakage rate during production of 2 percent. To measure daily tile output and breakage rate, Olsen has set up equally spaced classes for each. The class marks (midpoints of the class intervals) for daily tile output are 4,900, 5,500, 6,100, 6,700, 7,300, and 7,900. The class marks for breakage rates are 0.70, 2.10, 3.50, 4.90, 6.30, and 7.70.

- (a) What are the upper and lower boundaries of the classes for the daily tile output?
 (b) What are the upper and lower boundaries of the classes for the breakage rate?

2-76 BMT, Inc., manufactures performance equipment for cars used in various types of racing. It has gathered the following information on the number of models of engines in different size categories used in the racing market it serves:

Class (Engine Size in Cubic Inches)	Frequency (# of Models)
101–150	1
151–200	7
201–250	7
251–300	8
301–350	17
351–400	16
401–450	15
451–500	7

Construct a cumulative relative frequency distribution that will help you answer these questions:

- (a) Seventy percent of the engine models available are larger than about what size?
 (b) What was the approximate middle value in the original data set?
 (c) If BMT has designed a fuel-injection system that can be used on racing engines up to 400 cubic inches, about what percentage of the engine models available will not be able to use BMT’s system?

2-77 A business group is supporting the addition of a light-rail shuttle in the central business district and has two competing bids with different numbers of seats in each car. They arrange a fact-finding trip to Denver, and in a meeting they are given the following frequency distribution of number of passengers per car:

Number of Passengers	Frequency
1–10	20
11–20	18
21–30	11
31–40	8
41–50	3
51–60	1

- (a) One bid proposes light-rail cars with 30 seats and 10 standees. What percentage of the total observations are more than 30 and less than 41 passengers?
- (b) The business group members have been told that street cars with fewer than 11 passengers are uneconomical to operate and more than 30 passengers lead to poor customer satisfaction. What proportion of trips would be economical and satisfying?

2-78 Refer to the toll plaza problem in Exercise 2-74. Jeanne Moreno's employer, the state Department of Transportation, recently worked with a nearby complex of steel mills, with 5,000 employees, to modify the complex's shift changeover schedule so that shift changes do not coincide with the morning rush hour. Moreno wants an initial comparison to see whether waiting times at the toll plaza appear to have dropped. Here are the waiting times observed for 3,000 consecutive drivers after the mill schedule change:

Minutes of Waiting	Frequency
less than 1	177
1– 2.99	238
3– 4.99	578
5– 6.99	800
7– 8.99	713
9–10.99	326
11–12.99	159
13–14.99	9
15–16.99	0
	3,000

- (a) Construct a “less-than” cumulative frequency and cumulative relative frequency distribution.
- (b) Construct an ogive based on part (a). What percentage of the drivers had to wait more than 4 minutes in line? 8 minutes?
- (c) Compare your results with your answers to Exercise 2-74. Is there an obvious difference in waiting times?

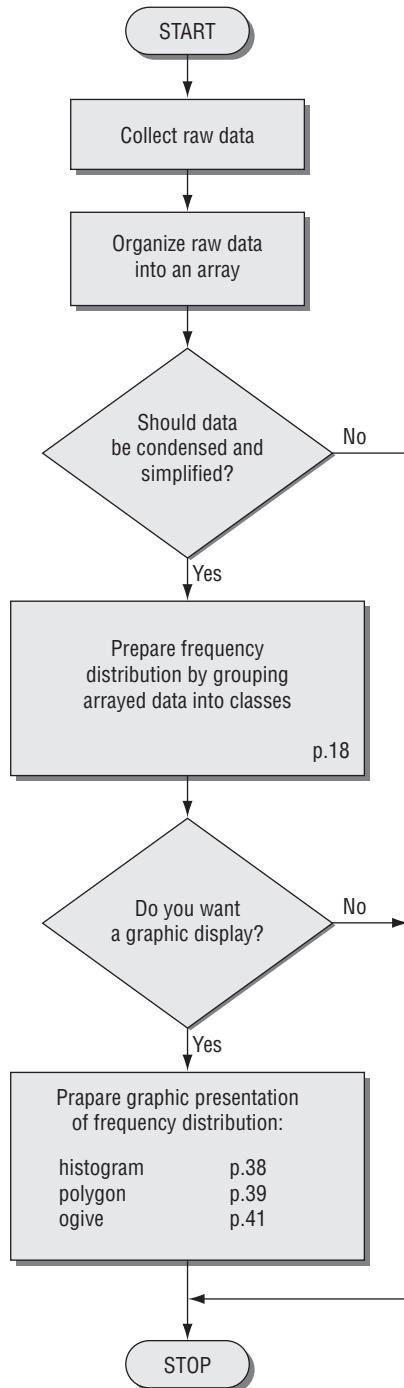


Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Construct a pie chart showing the distribution of type of bank account held by the people in the banks. (Question 2)
2. Construct a bar chart showing the frequency of usage of e-banking by the customers. (Question 5)
3. Construct a bar chart comparing the level of satisfaction with e-services across the different age group of customers. (Question 9 vs Question 14)
4. Draw an appropriate chart depicting the problems faced in e-banking and the promptness with which they are solved. (Question 10 vs Question 12)
5. Draw an appropriate diagram to study the gap in the expected and observed e-banking services provided by the banks to their customers. (Question 7 & 8)



Flow Chart: Arranging Data to Convey Meaning



Measures of Central Tendency and Dispersion in Frequency Distributions

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To use summary statistics to describe collections of data
 - To use the mean, median, and mode to describe how data “bunch up”
 - To use the range, variance, and standard deviation to describe how data “spread out”
 - To examine computer-based exploratory data analysis to see other useful ways to summarize data
-

CHAPTER CONTENTS

- | | |
|---|--|
| 3.1 Summary Statistics 74 | 3.10 Relative Dispersion: The Coefficient of Variation 132 |
| 3.2 A Measure of Central Tendency:
The Arithmetic Mean 77 | 3.11 Descriptive Statistics Using Msexcel & SPSS 136 |
| 3.3 A Second Measure of Central Tendency:
The Weighted Mean 87 | ■ Statistics at Work 140 |
| 3.4 A Third Measure of Central Tendency:
The Geometric Mean 92 | ■ Terms Introduced in Chapter 3 141 |
| 3.5 A Fourth Measure of Central Tendency:
The Median 96 | ■ Equations Introduced in Chapter 3 142 |
| 3.6 A Final Measure of Central Tendency:
The Mode 104 | ■ Review and Application Exercises 145 |
| 3.7 Dispersion: Why It Is Important 111 | ■ Flow Charts: Measures of Central Tendency and Dispersion 151 |
| 3.8 Ranges: Useful Measures of
Dispersion 113 | |
| 3.9 Dispersion: Average Deviation
Measures 119 | |

The vice president of marketing of a fast-food chain is studying the sales performance of the 100 stores in his eastern district and has compiled this frequency distribution of annual sales:

Sales (000s)	Frequency	Sales (000s)	Frequency
700–799	4	1,300–1,399	13
800–899	7	1,400–1,499	10
900–999	8	1,500–1,599	9
1,000–1,099	10	1,600–1,699	7
1,100–1,199	12	1,700–1,799	2
1,200–1,299	17	1,800–1,899	1

The vice president would like to compare the eastern district with the other three districts in the country. To do so, he will summarize the distribution, with an eye toward getting information about the central tendency of the data. This chapter also discusses how he can measure the variability in a distribution and thus get a much better feel for the data. ■

3.1 SUMMARY STATISTICS

In Chapter 2, we constructed tables and graphs from raw data. The resulting “pictures” of frequency distributions illustrated trends and patterns in the data. In most cases, however, we need more exact measures. In these cases, we can use single numbers called *summary statistics* to describe characteristics of a data set.

Summary statistics, central tendency, and dispersion

Two of these characteristics are particularly important to decision makers: *central tendency* and *dispersion*.

Central Tendency Central tendency is the middle point of a distribution. *Measures of central tendency* are also called *measures of location*. In Figure 3-1, the central location of curve B lies to the right of those of curve A and curve C. Notice that the central location of curve A is equal to that of curve C.

Middle of a data set

Dispersion Dispersion is the spread of the data in a distribution, that is, the extent to which the observations are scattered. Notice that curve A in Figure 3-2 has a wider spread, or dispersion, than curve B.

Spread of a data set

There are two other characteristics of data sets that provide useful information: *skewness* and *kurtosis*. Although the derivation of specific statistics to measure these characteristics is beyond the scope of this book, a general understanding of what each means will be helpful.

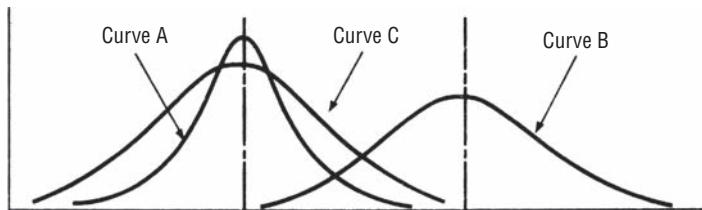
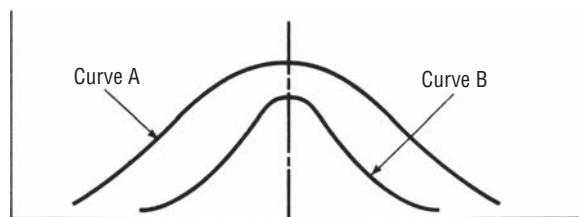


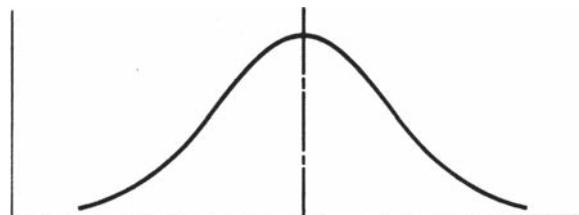
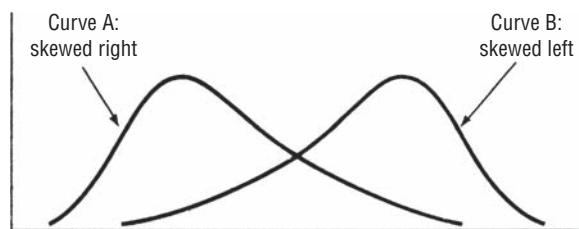
FIGURE 3-1 COMPARISON OF CENTRAL LOCATION OF THREE CURVES

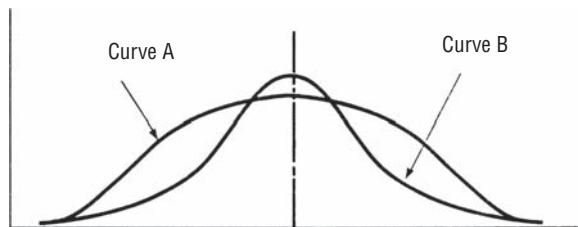
**FIGURE 3-2** COMPARISON OF DISPERSION OF TWO CURVES

Skewness Curves representing the data points in the data set may be either symmetrical or skewed. **Symmetrical curves**, like the one in Figure 3-3, are such that a vertical line drawn from the center of the curve to the horizontal axis divides the area of the curve into two equal parts. Each part is the mirror image of the other.

Curves A and B in Figure 3-4 are *skewed* curves. They are **Skewness of a data set** skewed because values in their frequency distributions are concentrated at either the low end or the high end of the measuring scale on the horizontal axis. The values are not equally distributed. Curve A is skewed to the right (or *positively skewed*) because it tails off toward the high end of the scale. Curve B is just the opposite. It is skewed to the left (*negatively skewed*) because it tails off toward the low end of the scale.

Curve A might represent the frequency distribution of the number of days' supply on hand in the wholesale fruit business. The curve would be skewed to the right, with many values at the low end and few at the high, because the inventory must turn over rapidly. Similarly, curve B could represent the frequency of the number of days a real-estate broker requires to sell a house. It would be skewed to the left, with many values at the high end and few at the low, because the inventory of houses turns over very slowly.

**FIGURE 3-3** SYMMETRICAL CURVE**FIGURE 3-4** COMPARISON OF TWO SKEWED CURVES

**FIGURE 3-5 TWO CURVES WITH THE SAME CENTRAL LOCATION BUT DIFFERENT KURTOSIS**

Kurtosis When we measure the *kurtosis* of a distribution, we are measuring its peakedness. In Figure 3-5, for example, curves A and B differ only in that one is more peaked than the other.

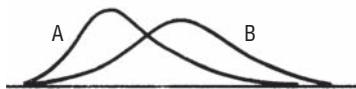
Peakedness of a data set

They have the same central location and dispersion, and both are symmetrical. Statisticians say that the two curves have different degrees of kurtosis.

EXERCISES 3.1

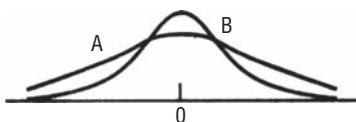
Basic Concepts

- 3-1 Draw three curves, all symmetrical but with different dispersions.
- 3-2 Draw three curves, all symmetrical and with the same dispersion, but with the following central locations:
 - (a) 0.0
 - (b) 1.0
 - (c) -1.0
- 3-3 Draw a curve that would be a good representation of the grades on a statistics test in a poorly prepared class and another or a well-prepared class.
- 3-4 For the following distributions, indicate which distribution
 - (a) Has the larger average value.
 - (b) Is more likely to produce a small value than a large value.
 - (c) Is the better representation of the distribution of ages at a rock concert.
 - (d) Is the better representation of the distribution of the times patients have to wait at a doctor's office.

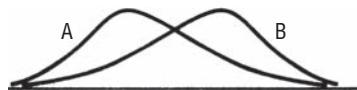


For the next two distributions, indicate which distribution, if any

- (e) Has values more evenly distributed across the range of possible values.
- (f) Is more likely to produce a value near 0.
- (g) Has a greater likelihood of producing positive values than negative values.



- 3-5** If the following two curves represent the distribution of scores for a group of students on two tests, which test appears to be more difficult for the students, A or B? Explain.



3.2 A MEASURE OF CENTRAL TENDENCY: THE ARITHMETIC MEAN

Most of the time when we refer to the “average” of something, we are talking about its arithmetic mean. This is true in cases such as the average winter temperature in New York City, the average life of a flashlight battery, and the average corn yield from an acre of land.

Table 3-1 presents data describing the number of days the generators at a power station on Lake Ico are out of service owing to regular maintenance or some malfunction. To find the arithmetic mean, we sum the values and divide by the number of observations:

$$\begin{aligned}\text{Arithmetic mean} &= \frac{7 + 23 + 4 + 8 + 2 + 12 + 6 + 13 + 9 + 4}{10} \\ &= \frac{88}{10} \\ &= 8.8 \text{ days}\end{aligned}$$

In this 1-year period, the generators were out of service for an average of 8.8 days. With this figure, the power plant manager has a reasonable single measure of the behavior of *all* her generators.

Conventional Symbols

To write equations for these measures of frequency distributions, we need to learn the mathematical notations used by statisticians. A *sample* of a population consists of n observations (a lowercase n) with a mean of \bar{x} (read x-bar). Remember that the measures we compute for a sample are called *statistics*.

The notation is different when we are computing measures for the entire *population*, that is, for the group containing every element we are describing. The mean of a population is symbolized by μ , which is the Greek letter *mu*. The number of

Characteristics of a sample are called statistics

Characteristics of a population are called parameters

TABLE 3-1 DOWNTIME OF GENERATORS AT LAKE ICO STATION

GENERATOR	1	2	3	4	5	6	7	8	9	10
DAYS OUT OF SERVICE	7	23	4	8	2	12	6	13	9	4

elements in a population is denoted by the capital italic letter N . Generally in statistics, we use italicized Roman letters to symbolize sample information and Greek letters to symbolize population information.

Calculating the Mean from Ungrouped Data

In the example, the average of 8.8 days would be μ (the population mean) if the 10 generators are the entire population. It would be \bar{x} (the sample mean) if the 10 generators are a sample drawn from a larger population of generators. To write the formulas for these two means, we combine our mathematical symbols and the steps we used to determine the arithmetic mean. If we add the values of the observations and divide this sum by the number of observations, we will get

Finding the population and sample means

Population Arithmetic Mean

Sum of values of all observations

$$\mu = \frac{\sum x}{N} \quad [3-1]$$

Number of elements in the population

and

Sample Arithmetic Mean

Sum of values of all observations

$$\bar{x} = \frac{\sum x}{n} \quad [3-2]$$

Number of elements in the sample

Because μ is the *population arithmetic mean*, we use N to indicate that we divide by the number of observations or elements in the population. Similarly, \bar{x} is the *sample arithmetic mean* and n is the number of observations in the sample. The Greek letter sigma, Σ , indicates that all the values of x are summed together.

Another example: Table 3-2 lists the percentile increase in SAT verbal scores shown by seven different students taking an SAT preparatory course.

TABLE 3-2 SAT VERBAL SCORES

STUDENT	1	2	3	4	5	6	7
INCREASE	9	7	7	6	4	4	2

We compute the mean of this sample of seven students as follows:

$$\bar{x} = \frac{\sum x}{n} \quad [3-2]$$

$$= \frac{9 + 7 + 7 + 6 + 4 + 4 + 2}{7}$$

$$= \frac{39}{7}$$

$$= 5.6 \text{ points per student} \leftarrow \text{Sample mean}$$

Notice that to calculate this mean, we added every observation. Statisticians call this kind of data *ungrouped* data. The computations were not difficult because our sample size was small. But suppose we are dealing with the weights of 5,000 head of cattle and prefer not to add each of our data points separately. Or suppose we have access to only the frequency distribution of the data, not to every individual observation. In these cases, we will need a different way to calculate the arithmetic mean.

Dealing with ungrouped data

Calculating the Mean from Grouped Data

A frequency distribution consists of data that are grouped by classes. Each value of an observation falls somewhere in one of the classes. Unlike the SAT example, we do not know the separate values of every observation. Suppose we have a frequency distribution (illustrated in Table 3-3) of average monthly checking-account balances of 600 customers at a branch bank. From the information in this table, we can easily compute an *estimate* of the value of the mean of these grouped data. It is an estimate because we do not use all 600 data points in the sample. Had we used the original, ungrouped data, we could have calculated the actual value of the mean, but only by averaging the 600 separate values. For ease of calculation, we must give up accuracy.

Dealing with grouped data

To find the arithmetic mean of grouped data, we first calculate the midpoint of each class. To make midpoints come out in whole cents, we round up. Thus, for example, the midpoint for the first class becomes 25.00, rather than 24.995. Then we multiply each midpoint by the frequency of observations in that class, sum all these results, and divide the sum by the total number of observations in the sample. The formula looks like this:

Estimating the mean

Calculating the mean

Sample Arithmetic Mean of Grouped Data

$$\bar{x} = \frac{\Sigma(f \times x)}{n} \quad [3-3]$$

where

- \bar{x} = sample mean
- Σ = symbol meaning “the sum of”

TABLE 3-3 AVERAGE MONTHLY BALANCE OF 600 CUSTOMERS

Class (Dollars)	Frequency
0–49.99	78
50.00–99.99	123
100.00–149.99	187
150.00–199.99	82
200.00–249.99	51
250.00–299.99	47
300.00–349.99	13
350.00–399.99	9
400.00–449.99	6
450.00–499.99	4
	600

- f = frequency (number of observations) in each class
- x = midpoint for each class in the sample
- n = number of observations in the sample

Table 3-4 illustrates how to calculate the arithmetic mean from our grouped data, using Equation 3-3.

TABLE 3-4 CALCULATION OF ARITHMETIC SAMPLE MEAN FROM GROUPED DATA IN TABLE 3-3

Class (Dollars) (1)	Midpoint (x) (2)	Frequency (f) (3)	$f \times x$ (3) \times (2)
0–49.99	25.00	×	78 = 1,950
50.00–99.99	75.00	×	123 = 9,225
100.00–149.99	125.00	×	187 = 23,375
150.00–199.99	175.00	×	82 = 14,350
200.00–249.99	225.00	×	51 = 11,475
250.00–299.99	275.00	×	47 = 12,925
300.00–349.00	325.00	×	13 = 4,225
350.00–399.99	375.00	×	9 = 3,375
400.00–449.99	425.00	×	6 = 2,550
450.00–499.99	475.00	×	4 = 1,900
		$\Sigma f = n = 600$	$85,350 \leftarrow \Sigma(f \times x)$

$$\bar{x} = \frac{\Sigma(f \times x)}{n} \quad [3-3]$$

$$= \frac{85,350}{600}$$

$$= 142.25 \longleftarrow \text{Sample mean (dollars)}$$

In our sample of 600 customers, the average monthly checking-account balance is \$142.25. This is our approximation from the frequency distribution. Notice that because we did not know every data point in the sample, we assumed that every value in a class was equal to its midpoint. Our results, then, can only approximate the actual average monthly balance.

We make an assumption

Coding

In situations where a computer is not available and we have to do the arithmetic by hand, we can further simplify our calculation of the mean from grouped data. Using a technique called *coding*, we eliminate the problem of large or inconvenient midpoints. Instead of using the actual midpoints to perform our calculations, we can assign small-value consecutive integers (whole numbers) called *codes* to each of the midpoints. The integer zero can be assigned anywhere, but to keep the integers small, we will assign zero to the midpoint in the *middle* (or the one nearest to the middle) of the frequency distribution. Then we can assign negative integers to values smaller than that midpoint and positive integers to those larger, as follows:

Assigning codes to the midpoints

Class	1–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45
Code (u)	-4	-3	-2	-1	0	1	2	3	4
					↑ x_0				

Symbolically, statisticians use x_0 to represent the midpoint that is assigned the code 0, and u for the coded midpoint. The following formula is used to determine the sample mean using codes:

Calculating the mean from grouped data using codes

Sample Arithmetic Mean of Grouped Data Using Codes

$$\bar{x} = x_0 + w \frac{\sum(u \times f)}{n} \quad [3-4]$$

where

- \bar{x} = mean of sample
- x_0 = value of the midpoint assigned the code 0
- w = numerical width of the class interval
- u = code assigned to each class
- f = frequency or number of observations in each class
- n = total number of observations in the sample

Keep in mind that $\sum(u \times f)$ simply means that we (1) multiply u by f for every class in the frequency distribution and (2) sum all of these products. Table 3-5 illustrates how to code the midpoints and find the sample mean of the annual snowfall (in inches) over 20 years in Harlan, Kentucky.

TABLE 3-5 ANNUAL SNOWFALL IN HARLAN, KENTUCKY

Class (1)	Midpoint (x) (2)	Code (u) (3)	Frequency (f) (4)	$u \times f$ (3) \times (4)
0–7	3.5	-2	\times	2 = -4
8–15	11.5	-1	\times	6 = -6
16–23	19.5 $\leftarrow x_0$	0	\times	3 = 0
24–31	27.5	1	\times	5 = 5
32–39	35.5	2	\times	2 = 4
40–47	43.5	3	\times	$\frac{2}{= 6}$
$\Sigma f = n = 20$				$5 \leftarrow \Sigma(u \times f)$
$\bar{x} = x_0 + w \frac{\sum(u \times f)}{n}$ $= 19.5 + 8 \left(\frac{5}{20} \right)$ $= 19.5 + 2 = 21.5 \leftarrow \text{Average annual snowfall}$				

Advantages and Disadvantages of the Arithmetic Mean

The arithmetic mean, as a single number representing a whole data set, has important advantages. First, its concept is familiar to most people and intuitively clear. Second, every data set has a mean. It is a measure that can be calculated, and it is unique because every data set has one and only one mean. Finally, the mean is useful for performing statistical procedures such as comparing the means from several data sets (a procedure we will carry out in Chapter 9).

Advantages of the mean

Yet, like any statistical measure, the arithmetic mean has disadvantages of which we must be aware. **First**, although the mean is reliable in that it reflects all the values in the data set, it may also be affected by extreme values that are not representative of the rest of the data. Notice that if the seven members of a track team have times in a mile race shown in Table 3-6, the mean time is

Three disadvantages of the mean

$$\begin{aligned}\mu &= \frac{\sum x}{N} & [3-1] \\ &= \frac{4.2 + 4.3 + 4.7 + 4.8 + 5.0 + 5.1 + 9.0}{7} \\ &= \frac{37.1}{7} \\ &= 5.3 \text{ minutes} \leftarrow \text{Population mean}\end{aligned}$$

If we compute a mean time for the first six members, however, and exclude the 9.0 value, the answer is about 4.7 minutes. The one *extreme value* of 9.0 distorts the value we get for the mean. It would be more representative to calculate the mean *without* including such an extreme value.

A **second** problem with the mean is the same one we encountered with our 600 checking-account balances: It is tedious to compute the mean because we *do* use every data point in our calculation (unless, of course, we take the short-cut method of using grouped data to approximate the mean).

The **third** disadvantage is that we are unable to compute the mean for a data set that has open-ended classes at either the high or low end of the scale. Suppose the data in Table 3-6 had been arranged in the frequency distribution shown in Table 3-7. We could not compute a mean value for these data because of the open-ended class of “5.4 and above.” We have no way of knowing whether the value is 5.4, near 5.4, or far above 5.4.

TABLE 3-6 TIMES FOR TRACK-TEAM MEMBERS IN A 1-MILE RACE

MEMBER	1	2	3	4	5	6	7
TIME IN MINUTES	4.2	4.3	4.7	4.8	5.0	5.1	9.0

TABLE 3-7 TIMES FOR TRACK-TEAM MEMBERS IN A 1-MILE RACE

CLASS IN MINUTES	4.2–4.5	4.6–4.9	5.0–5.3	5.4 and above
FREQUENCY	2	2	2	1

HINTS & ASSUMPTIONS

The mean (or average) *can* be an excellent measure of central tendency (how data group around the middle point of a distribution). But unless the mean is truly representative of the data from which it was computed, we are violating an important assumption. Warning: If there are very high or very low values in the data that don’t look like most of the data, the mean is *not* representative. Fortunately there are measures that can be calculated that don’t suffer from this shortcoming. A helpful hint in choosing which one of these to compute is to look at the data points.

EXERCISES 3.2

Self-Check Exercises

- SC 3-1** The frequency distribution below represents the weights in pounds of a sample of packages carried last month by a small airfreight company.

Class	Frequency	Class	Frequency
10.0–10.9	1	15.0–15.9	11
11.0–11.9	4	16.0–16.9	8
12.0–12.9	6	17.0–17.9	7
13.0–13.9	8	18.0–18.9	6
14.0–14.9	12	19.0–19.9	2

- (a) Compute the sample mean using Equation 3-3.
- (b) Compute the sample mean using the coding method (Equation 3-4) with 0 assigned to the fourth class.
- (c) Repeat part (b) with 0 assigned to the sixth class.
- (d) Explain why your answers in parts (b) and (c) are the same.

SC 3-2 Davis Furniture Company has a revolving credit agreement with the First National Bank. The loan showed the following ending monthly balances last year:

Jan.	\$121,300	Apr.	\$72,800	July	\$58,700	Oct.	\$52,800
Feb.	\$112,300	May	\$72,800	Aug.	\$61,100	Nov.	\$49,200
Mar.	\$72,800	June	\$57,300	Sept.	\$50,400	Dec.	\$46,100

The company is eligible for a reduced rate of interest if its average monthly balance is over \$65,000. Does it qualify?

Applications

3-6 Child-Care Community Nursery is eligible for a county social services grant as long as the average age of its children stays below 9. If these data represent the ages of all the children currently attending Child-Care, do they qualify for the grant?

8 5 9 10 9 12 7 12 13 7 8

3-7 Child-Care Community Nursery can continue to be supported by the county social services office as long as the average annual income of the families whose children attend the nursery is below \$12,500. The family incomes of the attending children are

\$14,500	\$15,600	\$12,500	\$8,600	\$7,800	
\$6,500	\$5,900	\$10,200	\$8,800	\$14,300	\$13,900

- (a) Does Child-Care qualify now for county support?
- (b) If the answer to part (a) is no, by how much must the average family income fall for it to qualify?
- (c) If the answer to part (a) is yes, by how much can average family income rise and Child-Care still stay eligible?

3-8 These data represent the ages of patients admitted to a small hospital on February 28, 1996:

85	75	66	43	40
88	80	56	56	67
89	83	65	53	75
87	83	52	44	48

- (a) Construct a frequency distribution with classes 40–49, 50–59, etc.
- (b) Compute the sample mean from the frequency distribution.
- (c) Compute the sample mean from the raw data.
- (d) Compare parts (b) and (c) and comment on your answer.

- 3-9** The frequency distribution below represents the time in seconds needed to serve a sample of customers by cashiers at BullsEye Discount Store in December 1996.

Time (in seconds)	Frequency
20–29	6
30–39	16
40–49	21
50–59	29
60–69	25
70–79	22
80–89	11
90–99	7
100–109	4
110–119	0
120–129	2

- (a) Compute the sample mean using Equation 3-3.
 (b) Compute the sample mean using the coding method (Equation 3-4) with 0 assigned to the 70–79 class.
- 3-10** The owner of Pets ‘R Us is interested in building a new store. The owner will build if the average number of animals sold during the first 6 months of 1995 is at least 300 and the overall monthly average for the year is at least 285. The data for 1995 are as follows:

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
234	216	195	400	315	274	302	291	275	300	375	450

What is the owner’s decision and why?

- 3-11** A cosmetics manufacturer recently purchased a machine to fill 3-ounce cologne bottles. To test the accuracy of the machine’s volume setting, 18 trial bottles were run. The resulting volumes (in ounces) for the trials were as follows:

3.02	2.89	2.92	2.84	2.90	2.97	2.95	2.94	2.93
3.01	2.97	2.95	2.90	2.94	2.96	2.99	2.99	2.97

The company does not normally recalibrate the filling machine for this cologne if the average volume is within 0.04 of 3.00 ounces. Should it recalibrate?

- 3-12** The production manager of Hinton Press is determining the average time needed to photograph one printing plate. Using a stopwatch and observing the platemakers, he collects the following times (in seconds)

20.4	20.0	22.2	23.8	21.3	25.1	21.2	22.9	28.2	24.3
22.0	24.7	25.7	24.9	22.7	24.4	24.3	23.6	23.2	21.0

An average per-plate time of less than 23.0 seconds indicates satisfactory productivity. Should the production manager be concerned?

- 3-13** National Tire Company holds reserve funds in short-term marketable securities. The ending daily balance (in millions) of the marketable securities account for 2 weeks is shown below:

Week 1	\$1.973	\$1.970	\$1.972	\$1.975	\$1.976
Week 2	1.969	1.892	1.893	1.887	1.895

What was the average (mean) amount invested in marketable securities during

- (a) The first week?
- (b) The second week?
- (c) The 2-week period?
- (d) An average balance over the 2 weeks of more than \$1.970 million would qualify National for higher interest rates. Does it qualify?
- (e) If the answer to part (c) is less than \$1.970 million, by how much would the last day's invested amount have to rise to qualify the company for the higher interest rates?
- (f) If the answer to part (c) is more than \$1,970 million, how much could the company treasurer withdraw from reserve funds on the last day and still qualify for the higher interest rates?

- 3-14** M. T. Smith travels the eastern United States as a sales representative for a textbook publisher. She is paid on a commission basis related to volume. Her quarterly earnings over the last 3 years are given below.

	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
Year 1	\$10,000	\$ 5,000	\$25,000	\$15,000
Year 2	20,000	10,000	20,000	10,000
Year 3	30,000	15,000	45,000	50,000

- (a) Calculate separately M. T.'s average earnings in each of the four quarters.
- (b) Calculate separately M. T.'s average quarterly earnings in each of the 3 years.
- (c) Show that the mean of the four numbers you found in part (a) is equal to the mean of the three numbers you found in part (b). Furthermore, show that both these numbers equal the mean of all 12 numbers in the data table. (This is M. T.'s average quarterly income over 3 years.)

- 3-15** Lillian Tyson has been the chairperson of the county library committee for 10 years. She contends that during her tenure she has managed the book-mobile repair budget better than her predecessor did. Here are data for bookmobile repair for 15 years:

Year	Town Budget	Year	Town Budget	Year	Town Budget
1992	\$30,000	1987	\$24,000	1982	\$30,000
1991	28,000	1986	19,000	1981	20,000
1990	25,000	1985	21,000	1980	15,000
1989	27,000	1984	22,000	1979	10,000
1988	26,000	1983	24,000	1978	9,000

- (a) Calculate the average annual budget for the last 5 years (1988–1992).
- (b) Calculate the average annual budget for her first 5 years in office (1983–1987).
- (c) Calculate the average annual budget for the 5 years before she was elected (1978–1982).
- (d) Based on the answers you found for parts (a), (b), and (c), do you think that there has been a decreasing or increasing trend in the annual budget? Has she been saving the county money?

Worked-Out Answers to Self-Check Exercises

SC 3-1

Class	Frequency (f)	(a)		(b)		(c)	
		Midpoint (x)	f × x	u	u × f	u	u × f
10.0–10.9	1	10.5	10.5	-3	-3	-5	-5
11.0–11.9	4	11.5	46.0	-2	-8	-4	-16
12.0–12.9	6	12.5	75.0	-1	-6	-3	-18
13.0–13.9	8	13.5	108.0	0	0	-2	-16
14.0–14.9	12	14.5	174.0	1	12	-1	-12
15.0–15.9	11	15.5	170.5	2	22	0	0
16.0–16.9	8	16.5	132.0	3	24	1	8
17.0–17.9	7	17.5	122.5	4	28	2	14
18.0–18.9	6	18.5	111.0	5	30	3	18
19.0–19.9	2	19.5	39.0	6	12	4	8
	65		988.5		111		-19

(a) $\bar{x} = \frac{\sum(f \times x)}{n} = \frac{988.5}{65} = 15.2077$ pounds

(b) $\bar{x} = x_0 + w \frac{\sum(u \times f)}{n} = 13.5 + \frac{1.0(111)}{65} = 15.2077$ pounds

(c) $\bar{x} = x_0 + w \frac{\sum(u \times f)}{n} = 15.5 + \frac{1.0(-19)}{65} = 15.2077$ pounds

- (d) Shifting the class assigned the code of 0 up by k classes replaces x_0 by $x_0 + kw$ and changes each code from u to $u - k$. But because

$$\begin{aligned}\bar{x}_b &= x_0 + w \frac{\sum(u \times f)}{n} = (x_0 + kw) - kw + w \frac{\sum(u \times f)}{n} \\ &= (x_0 + kw) + w \frac{\sum(u - k)f}{n} = \bar{x}_c\end{aligned}$$

we see that it does not matter which class is assigned the code 0.

SC 3-2 $\bar{x} = \frac{\sum x}{n} = \frac{827,600}{12} = \$68,967$

Because this exceeds \$65,000, they do qualify for the reduced interest rate.

3.3 A SECOND MEASURE OF CENTRAL TENDENCY: THE WEIGHTED MEAN

The weighted mean enables us to calculate an average that takes into account the *importance* of each value to the overall total. Consider, for example, the company in Table 3-8, which uses

A **weighted mean**

TABLE 3-8 LABOR INPUT IN MANUFACTURING PROCESS

Grade of Labor	Hourly Wage (x)	Labor Hours per Unit of Output	
		Product 1	Product 2
Unskilled	\$5.00	1	4
Semiskilled	7.00	2	3
Skilled	9.00	5	3

three grades of labor—unskilled, semiskilled, and skilled—to produce two end products. The company wants to know the average cost of labor per hour for each of the products.

A simple arithmetic average of the labor wage rates would be

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} & [3-2] \\ &= \frac{\$5 + \$7 + \$9}{3} \\ &= \frac{\$21}{3} \\ &= \$7.00/\text{hour}\end{aligned}$$

Using this average rate, we would compute the labor cost of one unit of product 1 to be $\$7(1 + 2 + 5) = \56 and of one unit of product 2 to be $\$7(4 + 3 + 3) = \70 . But these answers are incorrect.

In this case, the arithmetic mean is incorrect

To be correct, the answers must take into account that different amounts of each grade of labor are used. We can determine the correct answers in the following manner. For product 1, the total labor cost per unit is $(\$5 \times 1) + (\$7 \times 2) + (\$9 \times 5) = \64 , and, since there are 8 hours of labor input, the average labor cost per hour is $\$64/8 = \8.00 per hour. For product 2, the total labor cost per unit is $(\$5 \times 4) + (\$7 \times 3) + (\$9 \times 3) = \68 , for an average labor cost per hour of $\$68/10$, or $\$6.80$ per hour.

Another way to calculate the correct average cost per hour for the two products is to take a *weighted average* of the cost of the three grades of labor. To do this, we weight the hourly wage for each grade by its proportion of the total labor required to produce the product. One unit of product 1, for example, requires 8 hours of labor. Unskilled labor uses $\frac{1}{8}$ of this time, semiskilled labor uses $\frac{2}{8}$ of this time, and skilled labor requires $\frac{5}{8}$ of this time. If we use these fractions as our weights, then one hour of labor for product 1 costs an average of

$$\left(\frac{1}{8} \times \$5\right) + \left(\frac{2}{8} \times \$7\right) + \left(\frac{5}{8} \times \$9\right) = \$8.00 / \text{hour}$$

The correct answer is the weighted mean

Similarly, a unit of product 2 requires 10 labor hours, of which $\frac{4}{10}$ is used for unskilled labor, $\frac{3}{10}$ for semiskilled labor, and $\frac{3}{10}$ for skilled labor. By using these fractions as weights, one hour of labor for product 2 costs

$$\left(\frac{4}{10} \times \$5\right) + \left(\frac{3}{10} \times \$7\right) + \left(\frac{3}{10} \times \$9\right) = \$6.80 / \text{hour}$$

Calculating the weighted mean

Thus, we see that the weighted averages give the correct values for the average hourly labor costs of the two products because **they take into account that different amounts of each grade of labor are used in the products.**

Symbolically, the formula for calculating the weighted average is

Weight Mean

$$\bar{x}_w = \frac{\sum(w \times x)}{\sum w} \quad [3-5]$$

where

- \bar{x}_w = symbol for the weighted mean*
- w = weight assigned to each observation ($\frac{1}{8}$, $\frac{2}{8}$, and $\frac{5}{8}$, for product 1 and $\frac{1}{10}$, $\frac{3}{10}$, and $\frac{3}{10}$, for product 2 in our example)
- $\sum(w \times x)$ = sum of the weight of each element times that element
- $\sum w$ = sum of all of the weights

If we apply Equation 3-5 to product 1 in our labor-cost example, we find

$$\begin{aligned} \bar{x}_w &= \frac{\sum(w \times x)}{\sum w} && [3-5] \\ &= \frac{\left(\frac{1}{8} \times \$5\right) + \left(\frac{2}{8} \times \$7\right) + \left(\frac{5}{8} \times \$9\right)}{\frac{1}{8} + \frac{2}{8} + \frac{5}{8}} \\ &= \frac{\$8}{1} \\ &= \$8.00/\text{hour} \end{aligned}$$

Notice that Equation 3-5 states more formally something we have done previously. When we calculated the arithmetic mean from grouped data (page 79), we actually found a weighted mean, using the midpoints for the x values and the frequencies of each class as the weights. We divided this answer by the sum of all the frequencies, which is the same as dividing by the sum of all the weights.

The arithmetic mean of grouped data: the weighted mean

In like manner, *any* mean computed from all the values in a data set according to Equation 3-1 or 3-2 is really a weighted average of the components of the data set. What those components are, of course, determines what the mean measures. In a factory, for example, we could determine the weighted mean

*The symbol \bar{x}_w is read *x-bar sub w*. The lowercase w is called a subscript and is a reminder that this is not an ordinary mean but one that is weighted according to the relative importance of the values of x .

of all the wages (skilled, semiskilled, and unskilled) or of the wages of men workers, women workers, or union and nonunion members.

HINTS & ASSUMPTIONS

Distinguish between *distinct values* and *individual observations* in a data set, since several observations can have the same value. If values occur with different frequencies, the arithmetic mean of the *values* (as opposed to the arithmetic mean of the *observations*) may not be an accurate measure of central tendency. In such cases, we need to use the weighted mean of the values. If you are using an average value to make a decision, ask how it was calculated. If the values in the sample do not appear with the same frequency, insist on a weighted mean as the correct basis for your decision.

EXERCISES 3.3

Self-Check Exercises

- SC 3-3** Dave's Giveaway Store advertises, "If our average prices are not equal or lower than everyone else's, you get it free." One of Dave's customers came into the store one day and threw on the counter bills of sale for six items she bought from a competitor for an average price less than Dave's. The items cost

\$1.29 \$2.97 \$3.49 \$5.00 \$7.50 \$10.95

Dave's prices for the same six items are \$1.35, \$2.89, \$3.19, \$4.98, \$7.59 and \$11.50. Dave told the customer, "My ad refers to a weighted average price of these items. Our average is lower because our sales of these items have been:"

7 9 12 8 6 3

Is Dave getting himself into or out of trouble by talking about weighted averages?

- SC 3-4** Bennett Distribution Company, a subsidiary of a major appliance manufacturer, is forecasting regional sales for next year. The Atlantic branch, with current yearly sales of \$193.8 million, is expected to achieve a sales growth of 7.25 percent; the Midwest branch, with current sales of \$79.3 million, is expected to grow by 8.20 percent; and the Pacific branch, with sales of \$57.5 million, is expected to increase sales by 7.15 percent. What is the average rate of sales growth forecasted for next year?

Applications

- 3-16** A professor has decided to use a weighted average in figuring final grades for his seminar students. The homework average will count for 20 percent of a student's grade; the midterm, 25 percent; the final, 35 percent; the term paper, 10 percent; and quizzes, 10 percent. From the following data, compute the final average for the five students in the seminar.

Student	Homework	Quizzes	Paper	Midterm	Final
1	85	89	94	87	90
2	78	84	88	91	92
3	94	88	93	86	89
4	82	79	88	84	93
5	95	90	92	82	88

3-17 Jim's Videotaping Service recently placed an order for VHS videotape. Jim ordered 6 cases of High-Grade, 4 cases of Performance High-Grade, 8 cases of Standard, 3 cases of High Standard, and 1 case of Low Grade. Each case contains 24 tapes. Suppose a case of High-Grade costs \$28, Performance High-Grade costs \$36, Standard costs \$16, High Standard costs \$18, and Low costs \$6.

- (a) What is the average cost per case to Jim?
- (b) What is the average cost per tape to Jim?
- (c) Suppose Jim will sell any tape for \$1.25. Is this a good business practice for Jim?
- (d) How would your answer to parts (a)–(c) change if there were 48 tapes per case?

3-18 Keyes Home Furnishings ran six local newspaper advertisements during December. The following frequency distribution resulted:

NUMBER OF TIMES SUBSCRIBER SAW AD DURING DECEMBER	0	1	2	3	4	5	6
FREQUENCY	897	1,082	1,325	814	307	253	198

3-19 What is the average number of times a subscriber saw a Keyes advertisement during December? The Nelson Window Company has manufacturing plants in five U.S. cities: Orlando, Minneapolis, Dallas, Pittsburgh, and Seattle. The production forecast for the next year has been completed. The Orlando division, with yearly production of 72 million windows, is predicting an 11.5 percent increase. The Pittsburgh division, with yearly production of 62 million, should grow by 6.4 percent. The Seattle division, with yearly production of 48 million, should also grow by 6.4 percent. The Minneapolis and Dallas divisions, with yearly productions of 89 and 94 million windows, respectively, are expecting to decrease production in the coming year by 9.7 and 18.2 percent, respectively. What is the average rate of change in production for the Nelson Window Company for the next year?

3-20 The U.S. Postal Service handles seven basic types of letters and cards: third class, second class, first class, air mail, special delivery, registered, and certified. The mail volume during 1977 is given in the following table:

Type of Mailing	Ounces Delivered (in millions)	Price per Ounce
Third class	16,400	\$0.05
Second class	24,100	0.08
First class	77,600	0.13
Air mail	1,900	0.17
Special delivery	1,300	0.35
Registered	750	0.40
Certified	800	0.45

What was the average revenue per ounce for these services during the year?

- 3-21** Matthews, Young and Associates, a management consulting firm, has four types of professionals on its staff: managing consultants, senior associates, field staff, and office staff. Average rates charged to consulting clients for the work of each of these professional categories are \$75/hour, \$40/hour, \$30/hour, and \$15/hour. Office records indicate the following number of hours billed last year in each category: 8,000, 14,000, 24,000, and 35,000. If Matthews, Young is trying to come up with an average billing rate for estimating client charges for next year, what would you suggest they do and what do you think is an appropriate rate?

Worked-Out Answers to Self-Check Exercises

SC 3-3 With unweighted averages, we get

$$\bar{x}_c = \frac{\sum x}{n} = \frac{31.20}{6} = \$5.20 \text{ at the competition}$$

$$\bar{x}_D = \frac{31.50}{6} = \$5.25 \text{ at Dave's}$$

With weighted averages, we get

$$\begin{aligned}\bar{x}_c &= \frac{\sum(w \times x)}{\sum w} \\ &= \frac{7(1.29) + 9(2.97) + 12(3.49) + 8(5.00) + 6(7.50) + 3(10.95)}{7 + 9 + 12 + 8 + 6 + 3} \\ &= \frac{195.49}{45} = \$4.344 \text{ at the competition}\end{aligned}$$

$$\begin{aligned}\bar{x}_D &= \frac{7(1.35) + 9(2.89) + 12(3.19) + 8(4.98) + 6(7.59) + 3(11.50)}{7 + 9 + 12 + 8 + 6 + 3} \\ &= \frac{193.62}{45} = \$4.303 \text{ at Dave's}\end{aligned}$$

Although Dave is technically correct, the word *average* in popular usage is equivalent to *unweighted average* in technical usage, and the typical customer will surely be angry with Dave's assertion (whether he or she understands the technical point or not).

$$\begin{aligned}\text{SC 3-4 } \bar{x}_w &= \frac{\sum(w \times x)}{\sum w} = \frac{193.8(7.25) + 79.3(8.20) + 57.5(7.15)}{193.8 + 79.3 + 57.5} \\ &= \frac{2466.435}{330.6} = 7.46\%\end{aligned}$$

3.4 A THIRD MEASURE OF CENTRAL TENDENCY: THE GEOMETRIC MEAN

Sometimes when we are dealing with quantities that change over a period of time, we need to know an average rate of change, such as an average growth rate over a period of several years. In

Finding the growth rate: The geometric mean

TABLE 3-9 GROWTH OF \$100 DEPOSIT IN A SAVINGS ACCOUNT

Year	Interest Rate	Growth Factor	Savings at End of Year
1	7%	1.07	\$107.00
2	8	1.08	115.56
3	10	1.10	127.12
4	12	1.12	142.37
5	18	1.18	168.00

such cases, the simple arithmetic mean is inappropriate, because it gives the wrong answers. What we need to find is the *geometric mean*, simply called the G.M.

Consider, for example, the growth of a savings account. Suppose we deposit \$100 initially and let it accrue interest at varying rates for 5 years. The growth is summarized in Table 3-9.

The entry labeled “growth factor” is equal to

$$1 + \frac{\text{interest rate}}{100}$$

The growth factor is the amount by which we multiply the savings at the beginning of the year to get the savings at the end of the year. The simple arithmetic mean growth factor would be $(1.07 + 1.08 + 1.10 + 1.12 + 1.18)/5 = 1.11$, which corresponds to an average interest rate of 11 percent per year. If the bank gives interest at a constant rate of 11 percent per year, however, a \$100 deposit would grow in five years to

$$\$100 \times 1.11 \times 1.11 \times 1.11 \times 1.11 \times 1.11 = \$168.51$$

In this case, the arithmetic mean growth rate is incorrect

Table 3-9 shows that the actual figure is only \$168.00. Thus, the correct average growth factor must be slightly less than 1.11.

To find the correct average growth factor, we can multiply together the 5 years’ growth factors and then take the fifth root of the product—the number that, when multiplied by itself four times, is equal to the product we started with. The result is the *geometric mean growth rate*, which is the appropriate average to use here. The formula for finding the geometric mean of a series of numbers is

Calculating the geometric mean

Geometric Mean

Number of x values

$$\text{G.M.} = \sqrt[x]{\text{product of all } x \text{ values}}$$

[3-6]

If we apply this equation to our savings-account problem, we can determine that 1.1093 is the correct average growth factor.

$$\begin{aligned}
 \text{G.M.} &= \sqrt[n]{\text{product of all } x \text{ values}} & [3-6] \\
 &= \sqrt[5]{1.07 \times 1.08 \times 1.10 \times 1.12 \times 1.18} \\
 &= \sqrt[5]{1.679965} \\
 &= 1.1093 \longleftarrow \text{Average growth factor (the geometric mean of the 5 growth factors)}
 \end{aligned}$$

Notice that the correct average interest rate of 10.93 percent per year obtained with the geometric mean is very close to the incorrect average rate of 11 percent obtained with the arithmetic mean. This happens because the interest rates are relatively small. Be careful however, not to be tempted to use the arithmetic mean instead of the more complicated geometric mean. The following example demonstrates why.

Warning: use the appropriate mean

In highly inflationary economies, banks must pay high interest rates to attract savings. Suppose that over 5 years in an unbelievably inflationary economy, banks pay interest at annual rates of 100, 200, 250, 300, and 400 percent, which correspond to growth factors of 2, 3, 3.5, 4, and 5. (We've calculated these growth factors just as we did in Table 3-9.)

In 5 years, an initial deposit of \$100 would grow to $\$100 \times 2 \times 3 \times 3.5 \times 4 \times 5 = \$42,000$. The arithmetic mean growth factor is $(2 + 3 + 3.5 + 4 + 5)/5$, or 3.5. This corresponds to an average interest rate of 250 percent. Yet if the banks actually gave interest at a constant rate of 250 percent per year, then \$100 would grow to \$52,521.88 in 5 years:

$$\$100 \times 3.5 \times 3.5 \times 3.5 \times 3.5 \times 3.5 = \$52,521.88.$$

This answer exceeds the actual \$42,000 by more than \$10,500, a sizable error.

Let's use the formula for finding the geometric mean of a series of numbers to determine the correct growth factor:

$$\begin{aligned}
 \text{G.M.} &= \sqrt[5]{\text{product of all } x \text{ values}} & [3-6] \\
 &= \sqrt[5]{2 \times 3 \times 3.5 \times 4 \times 5} \\
 &= \sqrt[5]{420} \\
 &= 3.347 \longleftarrow \text{Average growth factor}
 \end{aligned}$$

This growth factor corresponds to an average interest rate of 235 percent per year. In this case, the use of the appropriate mean *does* make a significant difference.

HINTS & ASSUMPTIONS

We use the geometric mean to show multiplicative effects over time in compound interest and inflation calculations. In certain situations, answers using the arithmetic mean and the geometric mean will not be too far apart, but even a small difference can generate a poor decision. A good working hint is to use the geometric mean whenever you are calculating the average percentage change in some variable over time. When you see a value for the average increase in inflation, for example, ask whether it's a geometric mean and be warned that if it's not, you are dealing with an incorrect value.

EXERCISES 3.4

Self-Check Exercises

- SC 3-5** The growth in bad-debt expense for Johnston Office Supply Company over the last few years follows. Calculate the average percentage increase in bad-debt expense over this time period. If this rate continues, estimate the percentage increase in bad debts for 1997, relative to 1995.

1989	1990	1991	1992	1993	1994	1995
0.11	0.09	0.075	0.08	0.095	0.108	0.120

- SC 3-6** Realistic Stereo Shops marks up its merchandise 35 percent above the cost of its latest additions to stock. Until 4 months ago, the Dynamic 400-S VHS recorder had been \$300. During the last 4 months Realistic has received 4 monthly shipments of this recorder at these unit costs: \$275, \$250, \$240, and \$225. At what average rate per month has Realistic's retail price for this unit been decreasing during these 4 months?

Applications

- 3-22** Hayes Textiles has shown the following percentage increase in net worth over the last 5 years:

1992	1993	1994	1995	1996
5%	10.5%	9.0%	6.0%	7.5%

What is the average percentage increase in net worth over the 5-year period?

- 3-23** MacroSwift, the U.S.-based computer software giant, has posted an increase in net worth during 7 of the last 9 years. Calculate the average percentage change in net worth over this time period. Assuming similar conditions in the years to come, estimate the percentage change for 1998, relative to 1996.

1988	1989	1990	1991	1992	1993	1994	1995	1996
0.11	0.09	0.07	0.08	-0.04	0.14	0.11	-0.03	0.06

- 3-24** The Birch Company, a manufacturer of electrical circuit boards, has manufactured the following number of units over the past 5 years:

1992	1993	1994	1995	1996
12,500	13,250	14,310	15,741	17,630

Calculate the average percentage increase in units produced over this time period, and use this to estimate production for 1999.

- 3-25** Bob Headen is calculating the average growth factor for his stereo store over the last 6 years. Using a geometric mean, he comes up with an answer of 1.24. Individual growth factors for the first 5 years were 1.19, 1.35, 1.23, 1.19, and 1.30, but Bob lost the records for the sixth year, after he calculated the mean. What was it?

- 3-26** Over a 3-week period, a store owner purchased \$120 worth of acrylic sheeting for new display cases in three equal purchases of \$40 each. The first purchase was at \$1.00 per square foot; the second, \$1.10; and the third, \$1.15. What was the average weekly rate of increase in the price per square foot paid for the sheeting?

- 3-27** Lisa's Quick Stop has been attracting customers by selling milk at a price 2 percent below that of the main grocery store in town. Given below are Lisa's prices for a gallon of milk for a 2-month period. What was the average rate of change in price at Lisa's Quick Stop?

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
\$2.30	\$2.42	\$2.36	\$2.49	\$2.24	\$2.36	\$2.42	\$2.49

- 3-28** Industrial Suppliers, Inc., keeps records on the cost of processing a purchase order. Over the last 5 years, this cost has been \$55.00, \$58.00, \$61.00, \$65.00, and \$66.00. What has Industrial's average percentage increase been over this period? If this average rate stays the same for 3 more years, what will it cost Industrial to process a purchase order at that time?
- 3-29** A sociologist has been studying the yearly changes in the number of convicts assigned to the largest correctional facility in the state. His data are expressed in terms of the percentage increase in the number of prisoners (a negative number indicates a percentage decrease). The sociologist's most recent data are as follows:

1991	1992	1993	1994	1995	1996
-4%	5%	10%	3%	6%	-5%

- (a) Calculate the average percentage increase using only the 1992–1995 data.
- (b) Rework part (a) using the data from all 6 years.
- (c) A new penal code was passed in 1990. Previously, the prison population grew at a rate of about 2 percent per year. What seems to be the effect of the new code?

Worked-Out Answers to Self-Check Exercises

SC 3-5 $G.M. = \sqrt[7]{1.11(1.09)(1.075)(1.08)(1.095)(1.108)(1.12)} = \sqrt[7]{1.908769992} = 1.09675$

The average increase is 9.675 percent per year. The estimate for bad debt expenses in 1997 is $(1.09675)^2 - 1 = .2029$, i.e., 20.29 percent higher than in 1995.

SC 3-6 The monthly growth factors are $275/300 = 0.9167$, $250/275 = 0.9091$, $240/250 = 0.9600$, and $225/240 = 0.9375$, so

$$G.M. = \sqrt[4]{0.9167(0.9091)(0.9600)(0.9375)} = \sqrt[4]{0.7500} = 0.9306 = 1 - 0.0694$$

The price has been decreasing at an average rate of 6.94 percent per month.

3.5 A FOURTH MEASURE OF CENTRAL TENDENCY: THE MEDIAN

The *median* is a measure of central tendency different from any of the means we have discussed so far. The median is a single value from the data set that measures the central item in the data.

Median defined

This single item is the *middlemost* or *most central* item in the set of numbers. Half of the items lie above this point, and the other half lie below it.

Calculating the Median from Ungrouped Data

To find the median of a data set, first array the data in ascending or descending order. If the data set contains an *odd* number of items, the middle item of the array is the median. If there is an *even* number of items, the median is the average of the two middle items. In formal language, the median is

Finding the median of ungrouped data

Median
Number of x values  $\text{Median} = \left(\frac{n + 1}{2} \right) \text{th item in a data array}$ [3-7]

Suppose we wish to find the median of seven items in a data array. According to Equation 3-7, the median is the $(7 + 1)/2 = 4$ th item in the array. If we apply this to our previous example of the times for seven members of a track team, we discover that the fourth element in the array is 4.8 minutes. This is the median time for the track team. Notice that unlike the arithmetic mean we calculated earlier, the median we calculated in Table 3-10 was *not* distorted by the presence of the last value (9.0). This value could have been 15.0 or even 45.0 minutes, and the median would have been the same!

An odd number of items

Now let's calculate the median for an array with an even number of items. Consider the data shown in Table 3-11 concerning the number of patients treated daily in the emergency room of a hospital. The data are arrayed in descending order. The median of this data set would be

The median is not distorted by extreme values

An even number of items

$$\begin{aligned}
 \text{Median} &= \left(\frac{n + 1}{2} \right) \text{th item in a data array} \\
 &= \frac{8 + 1}{2} \\
 &= 4.5\text{th item}
 \end{aligned}
 \quad [3-7]$$

Because the median is the 4.5th element in the array, we need to average the fourth and fifth elements. The fourth element in Table 3-11 is 43 and the fifth is 35. The average of these two elements is equal to

TABLE 3-10 TIMES FOR TRACK-TEAM MEMBERS

ITEM IN DATA ARRAY	1	2	3	4	5	6	7
TIME IN MINUTES	4.2	4.3	4.7	4.8	5.0	5.1	9.0
				↑	Median		

TABLE 3-11 PATIENTS TREATED IN EMERGENCY ROOM ON 8 CONSECUTIVE DAYS

ITEM IN DATA ARRAY	1	2	3	4	5	6	7	8
NUMBER OF PATIENTS	86	52	49	43	35	31	30	11
↑ Median of 39								

$(43 + 35)/2$, or 39. Therefore, 39 is the median number of patients treated in the emergency room per day during the 8-day period.

Calculating the Median from Grouped Data

Often, we have access to data only after they have been grouped in a frequency distribution. For example, we do not know every observation that led to the construction of Table 3-12, the data on 600 bank customers originally introduced earlier. Instead, we have 10 class intervals and a record of the frequencies with which the observations appear in each of the intervals.

Finding the median of grouped data

Nevertheless, we can compute the median checking-account balance of these 600 customers by determining which of the 10 class intervals *contains* the median. To do this, we must add the frequencies in the frequency column in Table 3-12 until we reach the $(n + 1)/2$ th item. Because there are 600 accounts, the value for $(n + 1)/2$ is 300.5 (the average of the 300th and 301st items). The problem is to find the class intervals containing the 300th and 301st elements. The cumulative frequency for the first two classes is only $78 + 123 = 201$. But when we moved to the third class interval, 187 elements are added to 201, for a total of 388. Therefore, the 300th and 301st observations must be located in this third class (the interval from \$100.00 to \$149.99).

Locate the median class

The *median class* for this data set contains 187 items. If we assume that these 187 items begin at \$100.00 and are *evenly spaced over the entire class interval* from \$100.00 to \$149.99, then we can interpolate and find values for the 300th and 301st items. First, we determine that the 300th item is the 99th element in the median class:

TABLE 3-12 AVERAGE MONTHLY BALANCES FOR 600 CUSTOMERS

Class in Dollars	Frequency
0–49.99	78
50.00–99.99	123
100.00–149.99	187 Median class
150.00–199.99	82
200.00–249.99	51
250.00–299.99	47
300.00–349.99	13
350.00–399.99	9
400.00–449.99	6
450.00–499.99	2
	600

$$300 - 201 \text{ [items in the first two classes]} = 99$$

and that the 301st item is the 100th element in the median class:

$$301 - 201 = 100$$

Then we can calculate the *width* of the 187 equal steps from \$100.00 to \$149.99, as follows:

$$\begin{array}{ccc} \text{First item of next class} & & \text{First item of median class} \\ \searrow & & \swarrow \\ \frac{\$150.00 - \$100.00}{187} & = & \$0.267 \text{ in width} \end{array}$$

Now, if there are 187 steps of \$0.267 each and if 98 steps will take us to the 99th item, then the 99th item is

$$(\$0.267 \times 98) + \$100 = \$126.17$$

and the 100th item is one additional step:

$$\$126.17 + \$0.267 = \$126.44$$

Therefore, we can use \$126.17 and \$126.44 as the values of the 300th and 301st items, respectively.

The actual median for this data set is the value of the 300.5th item, that is, the average of the 300th and 301st items. This average is

$$\frac{\$126.17 + \$126.44}{2} = \$126.30$$

This figure (\$126.30) is the median monthly checking account balance, as estimated from the grouped data in Table 3-12.

In summary, we can calculate the median of grouped data as follows:

1. Use Equation 3-7 to determine which element in the distribution is center-most (in this case, the average of the 300th and 301st items). *Steps for finding the median of grouped data*
2. Add the frequencies in each class to find the class that contains that center-most element (the third class, or \$100.00–\$149.99).
3. Determine the number of elements in the class (187) and the location in the class of the median element (item 300 was the 99th element; item 301, the 100th element).
4. Learn the width of each step in the median class by dividing the class interval by the number of elements in the class (width = \$0.267).
5. Determine the number of steps from the lower bound of the median class to the appropriate item for the median (98 steps for the 99th element; 99 steps for the 100th element).
6. Calculate the estimated value of the median element by multiplying the number of steps to the median element times the width of each step and by adding the result to the lower bound of the median class ($\$100 + 98 \times \$0.267 = \$126.17$; $\$126.17 + \$0.267 = \$126.44$).
7. If, as in our example, there is an even number of elements in the distribution, average the values of the median element calculated in step 6 (\$126.30).

To shorten this procedure, statisticians use an equation to determine the median of grouped data. For a sample, this equation would be

An easier method

Sample Median of Grouped Data

$$\tilde{m} = \left(\frac{(n+1)/2 - (F+1)}{f_m} \right) w + L_m \quad [3-8]$$

where

- \tilde{m} = sample median
- n = total number of items in the distribution
- F = sum of all the class frequencies *up to*, but *not including*, the median class
- f_m = frequency of the median class
- w = class-interval width
- L_m = lower limit of the median-class interval

If we use Equation 3-8 to compute the median of our sample of checking-account balances, then $n = 600$, $F = 201$, $f_m = 187$, $w = \$50$, and $L_m = \$100$.

$$\begin{aligned}\tilde{m} &= \left(\frac{(n+1)/2 - (F+1)}{f_m} \right) w + L_m \\ &= \left(\frac{601/2 - 202}{187} \right) \$50 + \$100 \\ &= \left(\frac{98.5}{187} \right) \$50 + \$100 \\ &= (0.527)(\$50) + \$100 \\ &= \$126.35 \leftarrow \text{Estimated sample median}\end{aligned}\quad [3-8]$$

The slight difference between this answer and our answer calculated the long way is due to rounding.

Advantages and Disadvantages of the Median

The median has several advantages over the mean. The most important, demonstrated in our track-team example in Table 3-10, is that extreme values do not affect the median as strongly as they do the mean. The median is easy to understand and can be calculated from any kind of data—even for grouped data with open-ended classes such as the frequency distribution in Table 3-7—unless the median falls in an open-ended class.

Advantages of the median

We can find the median even when our data are qualitative descriptions such as color or sharpness, rather than numbers. Suppose, for example, we have five runs of a printing press, the results from which must be rated according to sharpness of the image. We can array the results from best to worst: extremely

sharp, very sharp, sharp, slightly blurred, and very blurred. The median of the five ratings is the $(5+1)/2$, or the third rating (sharp).

The median has some disadvantages as well. Certain statistical procedures that use the median are more complex than those that use the mean. Also, because the median is the value at the average position, we must array the data before we can perform any calculations. This is time consuming for any data set with a large number of elements. Therefore, if we want to use a sample statistic as an estimate of a population parameter, the mean is easier to use than the median. Chapter 7 will discuss estimation in detail.

Disadvantages of the median

HINTS & ASSUMPTIONS

In using the median, there is good news and bad news. The good news is that it is fairly quick to calculate and it avoids the effect of very large and very small values. The bad news is that you do give up some accuracy by choosing a single value to represent a distribution. With the values 2, 4, 5, 40, 100, 213, and 347, the median is 40, which has no apparent relationship to any of the other values in the distribution. Warning: Before you do any calculating, take a commonsense look at the data themselves. If the distribution looks unusual, just about anything you calculate from it will have shortcomings.

EXERCISES 3.5

Self-Check

SC 3-7 Swifty Markets compares prices charged for identical items in all of its food stores. Here are the prices charged by each store for a pound of bacon last week:

\$1.08 0.98 1.09 1.24 1.33 1.14 1.55 1.08 1.22 1.05

- Calculate the median price per pound.
- Calculate the mean price per pound.
- Which value is the better measure of the central tendency of these data?

SC 3-8 For the following frequency distribution, determine

- The median class.
- The number of the item that represents the median.
- The width of the equal steps in the median class.
- The estimated value of the median for these data.

Class	Frequency
100–149.5	12
150–199.5	14
200–249.5	27
250–299.5	58

Class	Frequency
300–349.5	72
350–399.5	63
400–449.5	36
450–499.5	18

Applications

- 3-30** Meridian Trucking maintains mileage records on all of its rolling equipment. Here are weekly mileage records for its trucks.

810	450	756	789	210	657	589	488	876	689
1,450	560	469	890	987	559	788	943	447	775

- (a) Calculate the median miles a truck traveled.
- (b) Calculate the mean for the 20 trucks.
- (c) Compare parts (a) and (b) and explain which one is a better measure of the central tendency of the data.

- 3-31** The North Carolina Consumers' Bureau has conducted a survey of cable television providers in the state. Here are the number of channels they offer in basic service:

32	28	31	15	25	14	12	29	22	28	29	32	33	24	26	8	35
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	---	----

- (a) Calculate the median number of channels provided.
- (b) Calculate the mean number of channels provided.
- (c) Which value is the better measure of the central tendency of these data?

- 3-32** For the following frequency distribution,

- (a) Which number item represents the median?
- (b) Which class contains the median?
- (c) What is the width of the equal steps in the median class?
- (d) What is the estimated value of the median for these data?
- (e) Use Equation 3-8 to estimate the median for the data. Are your two estimates close to one another?

Class	Frequency	Class	Frequency
10–19.5	8	60–69.5	52
20–29.5	15	70–79.5	84
30–39.5	23	80–89.5	97
40–49.5	37	90–99.5	16
50–59.5	46	100 or over	5

- 3-33** The following data represent weights of gamefish caught on the charter boat *Slickdrifter*:

Class	Frequency
0–24.9	5
25–49.9	13
50–74.9	16
75–99.9	8
100–124.9	6

- (a) Use Equation 3-8 to estimate the median weight of the fish caught.
- (b) Use Equation 3-3 to compute the mean for these data.

- (c) Compare parts (a) and (b) and comment on which is the better measure of the central tendency of these data.

- 3-34** The Chicago Transit Authority thinks that excessive speed on its buses increases maintenance cost. It believes that a reasonable median time from O'Hare Airport to John Hancock Center is about 30 minutes. From the following sample data (in minutes) can you help them determine whether the buses have been driven at excessive speeds? If you conclude from these data that they have, what explanation might you get from the bus drivers?

17	32	21	22
29	19	29	34
33	22	28	33
52	29	43	39
44	34	30	41

- 3-35** Mark Merritt, manager of Quality Upholstery Company, is researching the amount of material used in the firm's upholstery jobs. The amount varies between jobs, owing to different furniture styles and sizes. Merritt gathers the following data (in yards) from the jobs completed last week. Calculate the median yardage used on a job last week.

5¼	6¼	6	7¾	9¼	9½	10½
5¾	6	6¼	8	9½	9¾	10¼
5½	5¾	6½	8¼	9¾	10¼	10¾
5¾	5¾	7	8½	9¾	10½	10¾
6	5¾	7½	9	9¼	9¾	10

If there are 150 jobs scheduled in the next 3 weeks, use the median to predict how many yards of material will be required.

- 3-36** If insurance claims for automobile accidents follow the distribution given, determine the median using the method outlined on page 94. Verify that you get the same answer using Equation 3-8.

Amount of Claim (\$)	Frequency	Amount of Claim (\$)	Frequency
less than 250	52	750–999.99	1,776
250–499.99	337	1,000 and above	1,492
500–749.99	1,066		

- 3-37** A researcher obtained the following answers to a statement on an evaluation survey: strongly disagree, disagree, mildly disagree, agree somewhat, agree, strongly agree. Of the six answers, which is the median?

Worked-Out Answers to Self-Check Exercises

- SC 3-7** We first arrange the prices in ascending order:

0.98 1.05 1.08 1.08 1.09 1.14 1.22 1.24 1.33 1.55

(a) Median = $\frac{1.09 + 1.14}{2} = \1.115 , the average of items 5 and 6

$$(b) \bar{x} = \frac{\sum x}{n} = \frac{11.76}{10} = \$1.176$$

- (c) Because the data are skewed slightly, the median might be a bit better than the mean, but there really isn't very much difference.

SC 3-8

Class	Frequency	Cumulative Frequency
100–149.5	12	12
150–199.5	14	26
200–249.5	27	53
250–299.5	58	111
300–349.5	72	183
350–399.5	63	246
400–449.5	36	282
450–499.5	18	300

- (a) Median class = 300–349.5
 (b) Average of 150th and 151st
 (c) Step width = $50/72 = .6944$
 (d) $300 + 38(.6944) = 326.3872$ (150th)

$$300 + 39(.6944) = \frac{327.0816}{653.4688} \text{ (151st)}$$

$$\text{Median} = \frac{653.4688}{2} = 326.7344$$

3.6 A FINAL MEASURE OF CENTRAL TENDENCY: THE MODE

The mode is a measure of central tendency that is different from the mean but somewhat like the median because it is not actually calculated by the ordinary processes of arithmetic. The mode is *the value that is repeated most often in the data set*.

Mode defined

As in every other aspect of life, chance can play a role in the arrangement of data. Sometimes chance causes a single unrepresentative item to be repeated often enough to be the most frequent value in the data set. For this reason, we rarely use the mode of ungrouped data as a measure of central tendency. Table 3-13, for example, shows the number of delivery trips per day made by a Redi-mix concrete plant. The modal value is 15 because it occurs more often than any other value (three times). A mode of 15 implies that the plant activity is higher than 6.7 (6.7 is the answer we'd get if we calculated the mean). The mode tells us that 15 is the most frequent number of trips, but it fails to let us know that most of the values are under 10.

Risks in using the mode of ungrouped data

TABLE 3-13 DELIVERY TRIPS PER DAY IN ONE 20-DAY PERIOD

Trips Arrayed in Ascending Order					
0	2	5	7	15	← Mode
0	2	5	7	15	
1	4	6	8	15	
1	4	6	12	19	

Now let's group these data into a frequency distribution, as we have done in Table 3-14. If we select the class with the most observations, which we can call the *modal class*, we would choose 4–7 trips. This class is more representative of the activity of the plant than is the mode of 15 trips per day. For this reason, whenever we use the mode as a measure of the central tendency of a data set, we should calculate the mode from grouped data.

TABLE 3-14 FREQUENCY DISTRIBUTION OF DELIVERY TRIPS

CLASS IN NUMBER OF TRIPS	0–3	4–7	8–11	12 and more
FREQUENCY	6	8	1	5
		↑		
			Modal class	

Finding the modal class of grouped data

Calculating the Mode from Grouped Data

When data are already grouped in a frequency distribution, we must assume that the mode is located in the class with the most items, that is, the class with the highest frequency. To determine a single value for the mode from this modal class, we use Equation 3-9:

Mode
$Mo = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w \quad [3-9]$

where

- L_{Mo} = lower limit of the modal class
- d_1 = frequency of the modal class minus the frequency of the class *directly below it*
- d_2 = frequency of the modal class minus the frequency of the class *directly above it*
- w = width of the modal class interval

If we use Equation 3-9 to compute the mode of our checking-account balances (see Table 3-12), then $L_{Mo} = \$100$, $d_1 = 187 - 123 = 64$, $d_2 = 187 - 82 = 105$, and $w = \$50$.

$$\begin{aligned}
 Mo &= L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w \\
 &= \$100 + \frac{64}{64 + 105} \$50 \\
 &= \$100 + (0.38)(\$50) \\
 &= \$100 + \$19 \\
 &= \$119.00 \leftarrow \text{Mode}
 \end{aligned} \quad [3-9]$$

Our answer of \$119 is the estimate of the mode.

Multimodal Distributions

What happens when we have two different values that *each* appear the greatest number of times of any values in the data set? Table 3-15 shows the billing errors for one 20-day period in a hospital office. Notice that both 1 and 4 appear the greatest number of times in the data set. They each appear three times. This distribution, then, has two modes and is called a *bimodal distribution*.

In Figure 3-6, we have graphed the data in Table 3-15. Notice that there are *two* highest points on the graph. They occur at the values of 1 and 4 billing errors. The distribution in Figure 3-7 is also called bimodal, even though the two highest points are not equal. Clearly, these points are higher than the neighboring values in the frequency with which they are observed.

Bimodal distributions

TABLE 3-15 BILLING ERRORS PER DAY IN 20-DAY PERIOD

Errors Arrayed in Ascending Order			
0	2	6	9
0	4	6	9
1 }	4 }	← Mode	7 10
1 }	4 }		8 12
1	5	8	12

Advantages and Disadvantages of the Mode

The mode, like the median, can be used as a central location for qualitative as well as quantitative data. If a printing press turns out five impressions, which we rate “very sharp,” “sharp,” “sharp,” “sharp,” and “blurred,” then the modal value is “sharp.” Similarly, we can talk about modal styles when, for example, furniture customers prefer Early American furniture to other styles.

Advantages of the mode

Also like the median, **the mode is not unduly affected by extreme values**. Even if the high values are very high and the low values very low, we choose the most frequent value of the data set to be the modal value. We can use the mode no matter how large, how small, or how spread out the values in the data set happen to be.

A third advantage of the mode is that we can use it even when one or more of the classes are open ended. Notice, for example, that Table 3-14 contains the open-ended class “12 trips and more.”

Despite these advantages, the mode is not used as often to measure central tendency as are the mean and median. Too often, there is no modal value because the data set contains no values

Disadvantages of the mode

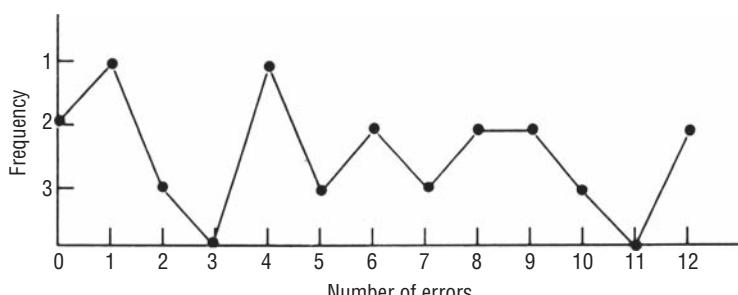


FIGURE 3-6 DATA IN TABLE 3-15 SHOWING THE BIMODAL DISTRIBUTION

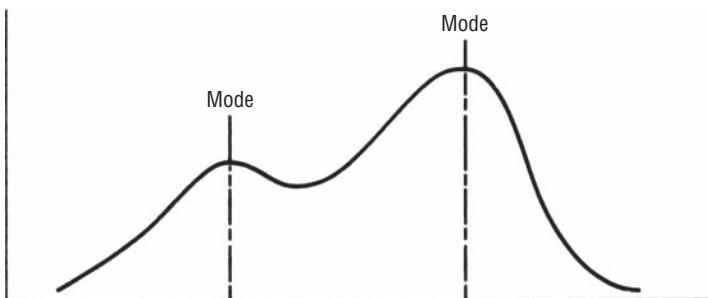


FIGURE 3-7 BIMODAL DISTRIBUTION WITH TWO UNEQUAL MODES

that occur more than once. Other times, every value is the mode, because every value occurs the same number of times. Clearly, the mode is a useless measure in these cases. Another disadvantage is that when data sets contain two, three, or many modes, they are difficult to interpret and compare.

Comparing the Mean, Median, and Mode

When we work statistical problems, we must decide whether to use the mean, the median, or the mode as the measure of central tendency. Symmetrical distributions that contain only one mode always have the same value for the mean, the median, and the mode. In these cases, we need not choose the measure of central tendency because the choice has been made for us.

*Mean, median, and mode
are identical in a symmetrical
distribution*

In a positively skewed distribution (one skewed to the right), as illustrated in Figure 3-8(a), the mode is at the highest point of the distribution, the median is to the right of that, and the mean is to the right of both the median and mode.

In a negatively skewed distribution (one skewed to the left), as illustrated in Figure 3-8(b), the mode is still at the highest point of the distribution, the median is to the left of that, and the mean is to the left of both the median and mode.

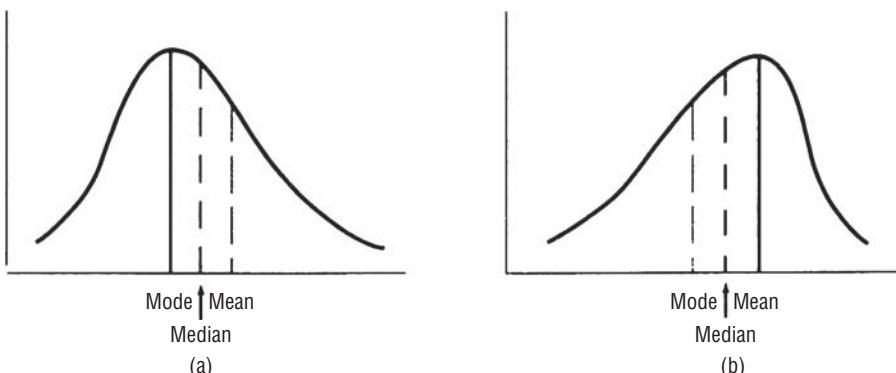


FIGURE 3-8 POSITIVELY (A) AND NEGATIVELY (B) SKEWED DISTRIBUTIONS, ILLUSTRATING RELATIVE POSITIONS OF MEAN, MEDIAN, AND MODE.

When the population is skewed negatively or positively, the median is often the best measure of location because it is always between the mean and the mode. The median is not as highly influenced by the frequency of occurrence of a single value as is the mode, nor is it pulled by extreme values as is the mean.

The median may be the best location measure in skewed distributions

Otherwise, there are no universal guidelines for applying the mean, median, or mode as the measure of central tendency for different populations. Each case must be judged independently, according to the guidelines we have discussed.

HINTS & ASSUMPTIONS

Hint: In trying to decide on the uses of the various means, the median, and the mode, think about practical situations in which each of them would make more sense. If you are averaging a small group of factory wages fairly near each other, the arithmetic mean is very accurate and fast. If there are 500 new houses in a development all within \$10,000 of each other in value, then the median is much quicker and quite accurate too. Dealing with the cumulative effects of inflation or interest requires the geometric mean if you want accuracy. A common-sense example: Although it's true that the average family has 1.65 children, automobile designers will make better decisions by using the modal value of 2.0 kids.

EXERCISES 3.6

Self-Check Exercises

SC 3-9 Here are the ages in years of the cars worked on by the Village Autohaus last week:

5 6 3 6 11 7 9 10 2 4 10 6 2 1 5

- (a) Compute the mode for this data set.
- (b) Compute the mean of the data set.
- (c) Compare parts (a) and (b) and comment on which is the better measure of the central tendency of the data.

SC 3-10 The ages of a sample of the students attending Sandhills Community College this semester are:

19	17	15	20	23	41	33	21	18	20
18	33	32	29	24	19	18	20	17	22
55	19	22	25	28	30	44	19	20	39

- (a) Construct a frequency distribution with intervals 15–19, 20–24, 25–29, 30–34, and 35 and older.
- (b) Estimate the modal value using Equation 3-9.
- (c) Now compute the mean of the raw data.
- (d) Compare your answers in parts (b) and (c) and comment on which of the two is the better measure of the central tendency of these data and why.

Applications

- 3-38** A librarian polled 20 different people as they left the library and asked them how many books they checked out. Here are the responses:

1 0 2 2 3 4 2 1 2 0 2 2 3 1 0 7 3 5 4 2

- (a) Compute the mode for this data set.
- (b) Compute the mean for this data set.
- (c) Graph the data by plotting frequency versus number checked out. Is the mean or the mode a better measure of the central tendency of the data?

- 3-39** The ages of residents of Twin Lakes Retirement Village have this frequency distribution:

Class	Frequency
47–51.9	4
52–56.9	9
57–61.9	13
62–66.9	42
67–71.9	39
72–76.9	20
77–81.9	9

Estimate the modal value of the distribution using Equation 3-9.

- 3-40** What are the modal values for the following distributions?

(a) Hair Color	Black	Brunette	Redhead	Blonde			
Frequency	11	24	6	18			
(b) Blood Type	AB	O	A	B			
Frequency	4	12	35	16			
(c) Day of Birth	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
Frequency	22	10	32	17	13	32	14

- 3-41** The numbers of apartments in 27 apartment complexes in Cary, North Carolina, are given below.

91	79	66	98	127	139	154	147	192
88	97	92	87	142	127	184	145	162
95	89	86	98	145	129	149	158	241

- (a) Construct a frequency distribution using intervals 66–87, 88–109,..., 220–241.
- (b) Estimate the modal value using Equation 3-9.
- (c) Compute the mean of the raw data.
- (d) Compare your answers in parts (b) and (c) and comment on which of the two is the better measure of central tendency of these data and why.

- 3-42** Estimate the mode for the distribution given in Exercise 3-36.

- 3-43** The number of solar heating systems available to the public is quite large, and their heat-storage capacities are quite varied. Here is a distribution of heat-storage capacity (in days) of 28 systems that were tested recently by University Laboratories, Inc.:

Days	Frequency
0–0.99	2
1–1.99	4
2–2.99	6
3–3.99	7
4–4.99	5
5–5.99	3
<u>6–6.99</u>	<u>1</u>

University Laboratories, Inc., knows that its report on the tests will be widely circulated and used as the basis for tax legislation on solar-heat allowances. It therefore wants the measures it uses to be as reflective of the data as possible.

- (a) Compute the mean for these data.
- (b) Compute the mode for these data.
- (c) Compute the median for these data.
- (d) Select the answer among parts (a), (b), and (c) that best reflects the central tendency of the test data and justify your choice.

- 3-44** Ed Grant is the director of the Student Financial Aid Office at Wilderness College. He has used available data on the summer earnings of all students who have applied to his office for financial aid to develop the following frequency distribution:

Summer Earnings	Number of Students
\$ 0–499	231
500–999	304
1,000–1,499	400
1,500–1,999	296
2,000–2,499	123
2,500–2,999	68
3,000 or more	23

- (a) Find the modal class for Ed's data.
- (b) Use Equation 3-9 to find the mode for Ed's data.
- (c) If student aid is restricted to those whose summer earnings were at least 10 percent lower than the modal summer earnings, how many of the applicants qualify?

Worked-Out Answers to Self-Check Exercises

- SC 3-9** (a) Mode = 6

(b)
$$\bar{x} = \frac{\Sigma x}{n} = \frac{87}{15} = 5.8$$

- (c) Because the modal frequency is only 3 and because the data are reasonably symmetric, the mean is the better measure of central tendency.

Class	15–19	20–24	25–29	30–34	≥ 35
Frequency	10	9	3	4	4

(b) $Mo = L_{Mo} + \frac{d_1}{d_1 + d_2} w = 15 + \left(\frac{10}{10+1} \right) 5 = 19.55$

(c) $\bar{x} = \frac{\Sigma x}{n} = \frac{760}{30} = 25.33$

(d) Because this distribution is very skewed, the mode is a better measure of central tendency.

3.7 DISPERSION: WHY IT IS IMPORTANT

Early in this chapter, in Figure 3-2, we illustrated two sets of data with the same central location but with one more spread out than the other. This is true of the three distributions in Figure 3-9. The mean of all three curves is the same, but curve A has less spread (or *variability*) than curve B, and curve B has less variability than curve C. If we measure only the mean of these three distributions, we will miss an important difference among the three curves. Likewise for any data, the mean, the median, and the mode tell us only part of what we need to know about the characteristics of the data. To increase our understanding of the pattern of the data, we must also measure its *dispersion*—its spread, or variability.

Need to measure dispersion or variability

Why is the dispersion of the distribution such an important characteristic to understand and measure? **First**, it gives us additional information that enables us to judge the reliability of our measure of the central tendency. If data are widely dispersed, such as those in curve C in Figure 3-9, the central location is less representative of the data as a whole than it would be for data more closely centered around the mean, as in curve A. **Second**, because there are problems peculiar to widely dispersed data, we must be able to recognize that data are widely dispersed before we can tackle those problems. **Third**, we may wish to compare dispersions of various samples. If a wide spread of values away from the center is undesirable or presents an unacceptable risk, we need to be able to recognize and avoid choosing the distributions with the greatest dispersion.

Uses of dispersion measures

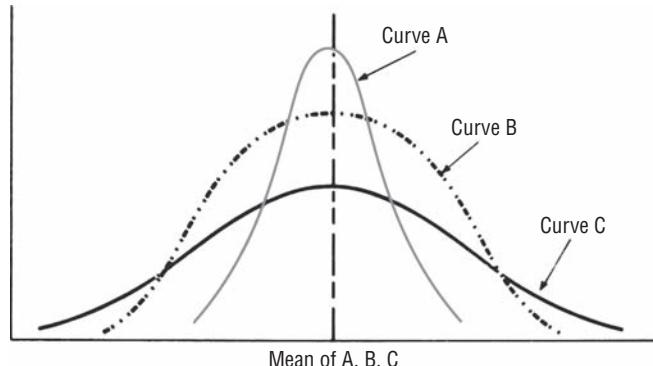


FIGURE 3-9 THREE CURVES WITH THE SAME MEAN BUT DIFFERENT VARIABILITIES

Financial analysts are concerned about the dispersion of a firm's earnings. Widely dispersed earnings—those varying from extremely high to low or even negative levels—indicate a higher risk to stockholders and creditors than do earnings remaining relatively stable. Similarly, quality control experts analyze the dispersion of a product's quality levels. A drug that is average in purity but ranges from very pure to highly impure may endanger lives.

Financial use and quality-control use

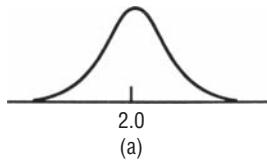
HINTS & ASSUMPTIONS

Airline seat manufacturers make an assumption about the shape of the average flyer. In some coach sections, it's common to find seat widths of only 19". If you weigh 250 pounds and wear a size 22 dress, sitting in a 19" seat is like putting on a tight shoe. It's O.K. to make this assumption for an airliner, but ignoring the dispersion (or spread) of the data gets you in trouble in football. A team that averages 3.6 yards per play should theoretically win every game because 3.6×4 plays is more than the 10 yards necessary to retain possession. Alas, bad luck comes to us all, and the theoretically unbeatable average of 3.6 yards is affected by the occasional 20-yard loss. Warning: Don't put too much stock in averages unless you know that the dispersion is small. A recruiter for the U.S. Air Force looking for pilot trainees who average 6' tall would get fired if he showed up with one who was 4' and another who was 8'. Under "reason for termination" on his personnel file, it should say "disregarded dispersion."

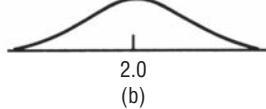
EXERCISES 3.7

Basic Concepts

- 3-45** For which of the following distributions is the mean more representative of the data as a whole? Why?



(a)



(b)

- 3-46** Which of the following is not a valid reason for measuring the dispersion of a distribution?
- It provides an indication of the reliability of the statistic used to measure central tendency.
 - It enables us to compare several samples with similar averages.
 - It uses more data in describing a distribution.
 - It draws attention to problems associated with very small or very large variability in distributions.

Applications

- 3-47** To measure scholastic achievement, educators need to test students' levels of knowledge and ability. Taking students' individual differences into account, teachers can plan their curricula better. The curves that follow represent distributions based on previous scores of two different tests. Which would you select as the better for the teachers' purpose?



- 3-48** A firm using two different methods to ship orders to its customers found the following distributions of delivery time for the two methods, based on past records. From available evidence, which shipment method would you recommend?



- 3-49** Of the 3 curves shown in Figure 3-9, choose one that would best describe the distribution of values for the ages of the following groups: members of Congress, newly elected members of the House of Representatives, and the chairpersons of major congressional committees. In making your choices, disregard the common mean of the curves in Figure 3-9 and consider only the variability of the distributions. Briefly state your reasons for your choices.

- 3-50** How do you think the concept of variability might apply to an investigation that the Federal Trade Commission (FTC) is conducting into possible price fixing by a group of manufacturers?

- 3-51** Choose which of the three curves shown in Figure 3-9 best describes the distribution of the following characteristics of various groups. Make your choices only on the basis of the variability of the distributions. Briefly state a reason for each choice.

- (a) The number of points scored by each player in a professional basketball league during an 80-game season.
- (b) The salary of each of 100 people working at roughly equivalent jobs in the federal government.
- (c) The grade-point average of each of the 15,000 students at a major state university.
- (d) The salary of each of 100 people working at roughly equivalent jobs in a private corporation.
- (e) The grade-point average of each student at a major state university who has been accepted for graduate school.
- (f) The percentage of shots made by each player in a professional basketball league during an 80-game season.

3.8 RANGES: USEFUL MEASURES OF DISPERSION

Dispersion may be measured in terms of the difference between two values selected from the data set. In this section, we shall study three of these so-called *distance measures*: the range, the interfractile range, and the interquartile range.

Three distance measures

Range

The *range* is the difference between the highest and lowest observed values. In equation form, we can say

Defining and computing the range

TABLE 3-16 ANNUAL PAYMENTS FROM BLUE CROSS-BLUE SHIELD (000S OMITTED)

CUMBERLAND	863	903	957	1,041	1,138	1,204
	1,354	1,624	1,698	1,745	1,802	1,883
VALLEY FALLS	490	540	560	570	590	600
	610	620	630	660	670	690

Range
Range = $\frac{\text{value of highest observation} - \text{value of lowest observation}}{[3-10]}$

Using this equation, we compare the ranges of annual payments from Blue Cross–Blue Shield received by the two hospitals illustrated in Table 3-16.

The range of annual payments to Cumberland is \$1,883,000 – \$863,000 = \$1,020,000. For Valley Falls, the range is \$690,000 – \$490,000 = \$200,000.

The range is easy to understand and to find, but its usefulness as a measure of dispersion is limited. The range considers only the highest and lowest values of a distribution and fails to take account of any other observation in the data set. As a result, it ignores the nature of the variation among all the other observations, and it is heavily influenced by extreme values. Because it measures only two values, the range is likely to change drastically from one sample to the next in a given population, even though the values that fall between the highest and lowest values may be quite similar. Keep in mind, too, that open-ended distributions have no range because no “highest” or “lowest” value exists in the open-ended class.

Characteristics of the range

Interfractile Range

In a frequency distribution, a given fraction or proportion of the data lie at or below a *fractile*. The median, for example, is the 0.5 fractile, because half the data set is less than or equal to this value. You will notice that fractiles are similar to percentages. In any distribution, 25 percent of the data lie at or below the 0.25 fractile; likewise, 25 percent of the data lie at or below the 25th percentile. The *interfractile range* is a measure of the spread between two fractiles in a frequency distribution, that is, the difference between the values of the two fractiles.

Fractiles

Suppose we wish to find the interfractile range between the first and second *thirds* of Cumberland’s receipts from Blue Cross–Blue Shield. We begin by dividing the observations into thirds, as we have done in Table 3-17. Each third contains four items ($\frac{1}{3}$ of the total of 12 items). Therefore, 33 $\frac{1}{3}$ percent of the items lie at \$1,041,000 or below it, and 66 $\frac{2}{3}$ percent are less than or equal to \$1,624,000. Now we can calculate the interfractile range between the $\frac{1}{3}$ and $\frac{2}{3}$ fractiles by subtracting the value \$1,041,000 from the value \$1,624,000. This

Meaning of the interfractile range

Calculating the interfractile range

TABLE 3-17 BLUE CROSS-BLUE SHIELD ANNUAL PAYMENTS TO CUMBERLAND HOSPITAL (000S OMITTED)

First Third	Second Third	Last Third
863	1,138	1,698
903	1,204	1,745
957	1,354	1,802
1,041 ← $\frac{1}{3}$ fractile	1,624 ← $\frac{2}{3}$ fractile	1,883

difference of \$583,000 is the spread between the top of the first third of the payments and the top of the second third.

Fractiles have special names, depending on the number of equal parts into which they divide the data. Fractiles that divide the data into 10 equal parts are called *deciles*. *Quartiles* divide the data into four equal parts. *Percentiles* divide the data into 100 equal parts.

Special fractiles: deciles, quartiles, and percentiles

Interquartile Range

The interquartile range measures approximately how far from the median we must go on either side before we can include one-half the values of the data set. To compute this range, we divide our data into four parts, each of which contains 25 percent of the items in the distribution. The *quartiles* are then the highest values in each of these four parts, and the *interquartile range* is the difference between the values of the first and third quartiles:

Computing the interquartile range

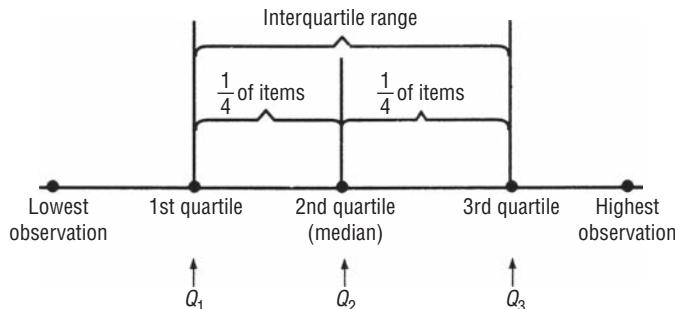
Interquartile Range

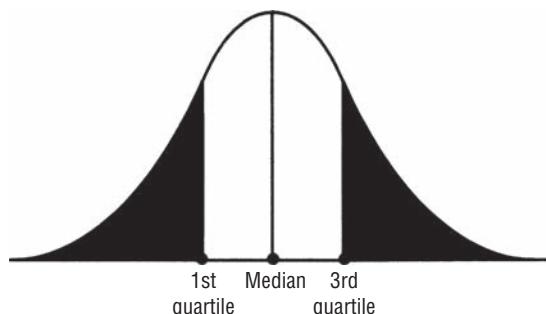
$$\text{Interquartile range} = Q_3 - Q_1$$

[3-11]

Figure 3-10 shows the concept of the interquartile range graphically. Notice in that figure that the widths of the four quartiles need not be the same.

In Figure 3-11, another illustration of quartiles, the quartiles divide the area under the distribution into four equal parts, each containing 25 percent of the area.

**FIGURE 3-10** INTERQUARTILE RANGE

**FIGURE 3-11 QUARTILES****HINTS & ASSUMPTIONS**

Fractile is a term used more by statisticians than by the rest of us, who are more familiar with 100 fractiles, or percentiles, especially when our percentile score on the SAT, the GMAT, or the LSAT is involved. When we get that letter indicating that our percentile score was 35, we know that 35 percent of those taking the test did worse than we did. The meaning of the range is easier to understand especially when the professor publishes the highest and lowest scores on the next statistics test. Hint: All of these terms help us deal with dispersion in data. If all the values look pretty much alike, then spending time computing dispersion values may not add much. If the data really spread out, betting your job on the average without considering dispersion is risky!

EXERCISES 3.8**Self-Check Exercises**

SC3-11 Here are student scores on a history quiz. Find the 80th percentile.

95	81	59	68	100	92	75	67	85	79
71	88	100	94	87	65	93	72	83	91

SC 3-12 The Casual Life Insurance Company is considering purchasing a new fleet of company cars. The financial department's director, Tom Dawkins, sampled 40 employees to determine the number of miles each drove over a 1-year period. The results of the study follow. Calculate the range and interquartile range.

3,600	4,200	4,700	4,900	5,300	5,700	6,700	7,300
7,700	8,100	8,300	8,400	8,700	8,700	8,900	9,300
9,500	9,500	9,700	10,000	10,300	10,500	10,700	10,800
11,000	11,300	11,300	11,800	12,100	12,700	12,900	13,100
13,500	13,800	14,600	14,900	16,300	17,200	18,500	20,300

Basic Concepts

- 3-52** For the following data, compute the interquartile range.

99	75	84	61	33	45	66	97	69	55
72	91	74	93	54	76	52	91	77	68

- 3-53** For the sample that follows, compute the

- (a) Range.
- (b) Interfractile range between the 20th and 80th percentiles.
- (c) Interquartile range.

2,549	3,897	3,661	2,697	2,200	3,812	2,228	3,891	2,668	2,268
3,692	2,145	2,653	3,249	2,841	3,469	3,268	2,598	3,842	3,362

Applications

- 3-54** Here are the high temperature readings during June 1995 in Phoenix, Arizona. Find the 70th percentile.

84	86	78	69	94	95	94	98	89	87	88	89	92	99	102
94	92	96	89	88	87	88	84	82	88	94	97	99	102	105

- 3-55** These are the total fares (in dollars) collected Tuesday by the 20 taxis belonging to City Transit, Ltd.

147	95	93	127	143	101	123	83	135	129
185	92	115	126	157	93	133	51	125	132

Compute the range of these data and comment on whether you think it is a useful measure of dispersion.

- 3-56** Redi-Mix Incorporated kept the following record of time (to the nearest 100th of a minute) its truck waited at the job to unload. Calculate the range and the interquartile range.

0.10	0.45	0.50	0.32	0.89	1.20	0.53	0.67	0.58	0.48
0.23	0.77	0.12	0.66	0.59	0.95	1.10	0.83	0.69	0.51

- 3-57** Warlington Appliances has developed a new combination blender-crock-pot. In a marketing demonstration, a price survey determined that most of those sampled would be willing to pay around \$60, with a surprisingly small interquartile range of \$14.00. In an attempt to replicate the results, the demonstration and accompanying survey were repeated. The marketing department hoped to find an even smaller interquartile range. The data follow. Was its hope realized?

52	35	48	46	43	40	61	49	57	58	65	46
72	69	38	37	55	52	50	31	41	60	45	41
55	38	51	49	46	43	64	52	60	61	68	49
69	66	35	34	52	49	47	28	38	57	42	38

- 3-58** MacroSwift has decided to develop a new software program designed for CEOs and other high-level executives. MacroSwift did not want to develop a program that required too much hard-drive space, so they polled 36 executives to determine the amount of available space on their PCs. The results are given below in megabytes.

6.3	6.7	7.9	8.4	9.7	10.6	12.4	19.4	29.1	42.6
59.8	97.6	100.4	120.6	135.5	148.6	178.6	200.1	229.6	284.6
305.6	315.6	325.9	347.5	358.6	397.8	405.6	415.9	427.8	428.6
439.5	440.9	472.3	475.9	477.2	502.6				

Calculate the range and interquartile range.

- 3-59** The New Mexico State Highway Department is charged with maintaining all state roads in good condition. One measure of condition is the number of cracks present in each 100 feet of roadway. From the department's yearly sample, the following data were obtained:

4	7	8	9	9	10	11	12	12	13
13	13	13	14	14	14	15	15	16	16
16	16	16	17	17	17	18	18	19	19

Calculate the interfractile ranges between the 20th, 40th, 60th, and 80th percentiles.

- 3-60** Ted Nichol is a statistical analyst who reports directly to the highest levels of management at Research Incorporated. He helped design the company slogan: "If you can't find the answer, then RESEARCH!" Ted has just received some disturbing data: the monthly dollar volume of research contracts that the company has won for the past year. Ideally, these monthly numbers should be fairly stable because too much fluctuation in the amount of work to be done can result in an inordinate amount of hiring and firing of employees. Ted's data (in thousands of dollars) follow:

253	104	633	57	500	201
43	380	467	162	220	302

Calculate the following:

- The interfractile range between the second and eighth deciles.
- The median, Q_1 , and Q_3 .
- The interquartile range.

Worked-Out Answers to Self-Check Exercises

- SC 3-11** First we arrange the data in increasing order:

59	65	67	68	71	72	75	79	81	83
85	87	88	91	92	93	94	95	100	100

The 16th of these (or 93) is the 80th percentile.

- SC 3-12** Range = $20,300 - 3,600 = 16,700$ miles

Interquartile range = $Q_3 - Q_1 = 12,700 - 8,100 = 4,600$ miles

3.9 DISPERSION: AVERAGE DEVIATION MEASURES

The most comprehensive descriptions of dispersion are those that deal with the average deviation from some measure of central tendency. Two of these measures are important to our study of statistics: the *variance* and the *standard deviation*. Both of these tell us an average distance of any observation in the data set from the mean of the distribution.

Two measures of average deviation

Population Variance

Every population has a variance, which is symbolized by σ^2 (sigma squared). To calculate the population variance, we divide the sum of the squared distances between the mean and each item in the population by the total number of items in the population. By squaring each distance, we make each number positive and, at the same time, assign more weight to the larger deviations (deviation is the distance between the mean and a value).

The formula for calculating the variance is

Variance

Formula for the variance of a population

Population Variance

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} = \frac{\Sigma x^2}{N} - \mu^2 \quad [3-12]$$

where

- σ^2 = population variance
- x = item or observation
- μ = population mean
- N = total number of items in the population
- Σ = sum of all the values $(x - \mu)^2$ or all the values x^2

In Equation 3-12, the middle expression $\frac{\Sigma(x - \mu)^2}{N}$, is the definition of σ^2 . The last expression, $\frac{\Sigma x^2}{N} - \mu^2$,

is *mathematically* equivalent to the definition but is often much more convenient to use if we must actually compute the value of σ^2 , since it frees us from calculating the deviations from the mean. However, when the x values are large and the $x - \mu$ values are small, it may be more convenient to use the middle expression, $\frac{\Sigma(x - \mu)^2}{N}$, to compute σ^2 . Before we can use this formula in an example, we need to discuss an important problem concerning the variance. In solving that problem, we will learn what the standard deviation is and how to calculate it. Then we can return to the variance itself.

Earlier, when we calculated the range, the answers were expressed in the same units as the data. (In our examples, the units were “thousands of dollars of payments.”) For the variance, however, the units are the *squares of the units* of the data—for example, “squared dollars” or “dollars squared.” Squared dollars or dollars squared are not intuitively

Units in which the variance is expressed cause a problem

clear or easily interpreted. For this reason, we have to make a significant change in the variance to compute a useful measure of deviation, one that does not give us a problem with units of measure and thus is less confusing. **This measure is called the standard deviation, and it is the square root of the variance.** The square root of 100 dollars squared is 10 dollars because we take the square root of both the value and the units in which it is measured. The standard deviation, then, is in units that are the same as the original data.

Population Standard Deviation

The population standard deviation, or σ , is simply the square root of the population variance. Because the variance is the average of the squared distances of the observations from the mean, **the standard deviation is the square root of the average of the squared distances of the observations from the mean.** While the variance is expressed in the square of the units used in the data, the standard deviation is in the same units as those used in the data. The formula for the standard deviation is

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{\frac{\sum x^2}{N} - \mu^2} \quad [3-13]$$

- x = observation
- μ = population mean
- N = total number of elements in the population
- \sum = sum of all the values $(x - \mu)^2$, or all the values x^2
- σ = population standard deviation
- σ^2 = population variance

The square root of a positive number may be either positive or negative because $a^2 = (-a)^2$. When taking the square root of the variance to calculate the standard deviation, however, statisticians consider only the positive square root.

To calculate either the variance or the standard deviation, we construct a table, using every element of the population. If we have a population of fifteen vials of compound produced in one day and we test each vial to determine its purity, our data might look like Table 3-18. In Table 3-19, we show how to use these data to compute the mean ($0.166 = 2.49/15$), the column (1) sum divided by N), the deviation of each value from the mean (column 3), the square of the deviation of each value from the mean (column 4), and the sum of the squared deviations. From this, we can compute the variance, which is 0.0034 percent squared. (Table 3-19 also computes σ^2 using the second half of Equation 3-12, $= \frac{\sum x^2}{N} - \mu^2$). Note

Relationship of standard deviation to the variance

standard deviation is the square root of the average of the squared distances of the observations from the mean.

Use the positive square root

Computing the standard deviation

TABLE 3-18 RESULTS OF PURITY TEST ON COMPOUNDS

Observed Percentage Impurity				
0.04	0.14	0.17	0.19	0.22
0.06	0.14	0.17	0.21	0.24
0.12	0.15	0.18	0.21	0.25

TABLE 3-19 DETERMINATION OF THE VARIANCE AND STANDARD DEVIATION OF PERCENT IMPURITY OF COMPOUNDS

Observation (x) (1)	Mean $\mu = 2.49/15$ (2)	Deviation $(x - \mu)$ (3) = (1) - (2)	Deviation Squared $(x - \mu)^2$ (4) = $ (1) - (2) ^2$	Observation Squared (x^2) (5) = (1) ²
0.04	—	0.166	= -0.126	0.016
0.06	—	0.166	= -0.106	0.0036
0.12	—	0.166	= -0.046	0.0144
0.14	—	0.166	= -0.026	0.0196
0.14	—	0.166	= -0.026	0.0196
0.15	—	0.166	= -0.016	0.0025
0.17	—	0.166	= 0.004	0.000
0.17	—	0.166	= 0.004	0.000
0.18	—	0.166	= 0.014	0.000
0.19	—	0.166	= 0.024	0.001
0.21	—	0.166	= 0.044	0.002
0.21	—	0.166	= 0.044	0.002
0.22	—	0.166	= 0.054	0.003
0.24	—	0.166	= 0.074	0.005
0.25	—	0.166	= 0.084	0.007
2.49 ← $\sum x$			0.051 ← $\sum (x - \mu)^2$	0.4643 ← $\sum x^2$
$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$	[3-12]	← OR →	$\sigma^2 = \frac{\Sigma x^2}{N} - \mu^2$	[3-12]
$= \frac{0.051}{15}$			$= \frac{0.4643}{15} - (0.166)^2$	
= 0.0034 percent squared			= 0.0034 percent squared	
$\sigma = \sqrt{\sigma^2}$	[3-13]			
$= \sqrt{.0034}$				
= 0.058 percent				

that we get the same result but do a bit less work, since we do not have to compute the deviations from the mean.) Taking the square root of σ^2 , we can compute the standard deviation, 0.058 percent.

Uses of the Standard Deviation

The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are

Chebyshev's theorem

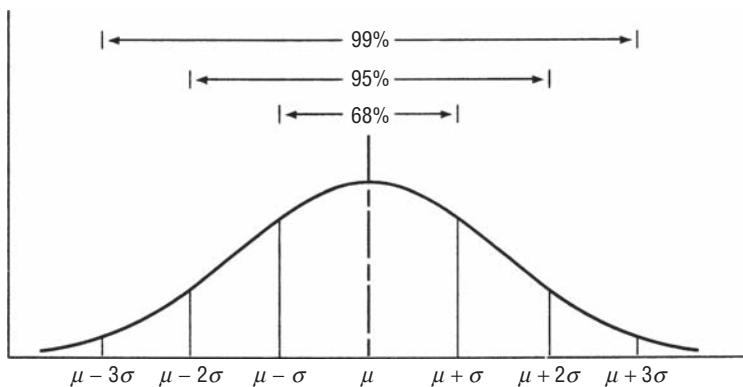


FIGURE 3-12 LOCATION OF OBSERVATIONS AROUND THE MEAN OF A BELL-SHAPED FREQUENCY DISTRIBUTION

located in relation to the mean. We can do this according to a theorem devised by the Russian mathematician P. L. Chebyshev (1821–1894). Chebyshev's theorem says that no matter what the shape of the distribution, at least 75 percent of the values will fall within ± 2 standard deviations from the mean of the distribution, and at least 89 percent of the values will lie within ± 3 standard deviations from the mean.

We can measure with even more precision the percentage of items that fall within specific ranges under a symmetrical, bell-shaped curve such as the one in Figure 3-12. In these cases, we can say that:

1. About 68 percent of the values in the population will fall within ± 1 standard deviation from the mean.
2. About 95 percent of the values will lie within ± 2 standard deviations from the mean.
3. About 99 percent of the values will be in an interval ranging from 3 standard deviations below the mean to 3 standard deviations above the mean.

In the light of Chebyshev's theorem, let's analyze the data in Table 3-19. There, the mean impurity of the 15 vials of compound is 0.166 percent, and the standard deviation is 0.058 percent. Chebyshev's theorem tells us that at least 75 percent of the values (at least 11 of our 15 items) are between $0.166 - 2(0.058) = 0.050$ and $0.166 + 2(0.058) = 0.282$. In fact, 93 percent of the values (14 of the 15 values) are actually in that interval. Notice that the distribution is reasonably symmetrical and that 93 percent is close to the theoretical 95 percent for an interval of plus and minus 2 standard deviations from the mean of a bell-shaped curve.

Using Chebyshev's theorem

The standard deviation is also useful in describing how far individual items in a distribution depart from the mean of the distribution. A measure called the *standard score* gives us the number of standard deviations a particular observation lies below or above the mean. If we let x symbolize the observation, the standard score computed from population data is

Concept of the standard score

Standard Score

$$\text{Population standard score} = \frac{x - \mu}{\sigma} \quad [3-14]$$

where

- x = observation from the population
- μ = population mean
- σ = population standard deviation

Suppose we observe a vial of compound that is 0.108 percent impure. Because our population has a mean of 0.166 and a standard deviation of 0.058, an observation of 0.108 would have a standard score of –1:

$$\begin{aligned}\text{Standard score} &= \frac{x - \mu}{\sigma} & [3-14] \\ &= \frac{0.108 - 0.166}{0.058} \\ &= -\frac{0.058}{0.058} \\ &= -1\end{aligned}$$

An observed impurity of 0.282 percent would have a standard score of +2:

$$\begin{aligned}\text{Standard score} &= \frac{x - \mu}{\sigma} & [3-14] \\ &= \frac{0.282 - 0.166}{0.058} \\ &= \frac{0.116}{0.058} \\ &= 2\end{aligned}$$

The standard score indicates that an impurity of 0.282 percent deviates from the mean by $2(0.058) = 0.116$ unit, which is equal to +2 in terms of the number of standard deviations away from the mean.

Calculating the standard score

Interpreting the standard score

Calculation of Variance and Standard Deviation Using Grouped Data

In our chapter-opening example, data on sales of 100 fast-food restaurants were already grouped in a frequency distribution. With such data, we can use the following formulas to calculate the variance and the standard deviation:

Calculating the variance and standard deviation for grouped data

Variance of Grouped Data

$$\sigma^2 = \frac{\sum f(x - \mu)^2}{N} = \frac{\sum fx^2}{N} - \mu^2 & [3-15]$$

Standard Deviation Grouped Data

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum f(x - \mu)^2}{N}} = \sqrt{\frac{\sum fx^2}{N} - \mu^2} \quad [3-16]$$

where

- σ^2 = population variance
- σ = population standard deviation
- f = frequency of each of the classes
- x = midpoint for each class
- μ = population mean
- N = size of the population

Table 3-20 shows how to apply these equations to find the variance and standard deviation of the sales of 100 fast-food restaurants.

We leave it as an exercise for the curious reader to verify that the second half of Equation 3-15, $\frac{\sum fx^2}{N} - \mu^2$, will yield the same value of σ^2 .

Now we are ready to compute the sample statistics that are analogous to the population variance σ^2 and the population standard deviation σ . These are the sample variance s^2 and the sample standard deviation s . In the next section, you'll notice we are changing from Greek letters (which denote population parameters) to the Roman letters of sample statistics.

*Switching to sample variance
and sample standard deviation*

Sample Standard Deviation

To compute the sample variance and the sample standard deviation, we use the same formulas Equations 3-12 and 3-13, replacing μ with \bar{x} and N with $n - 1$. The formulas look like this:

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{\sum x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1} \quad [3-17]$$

Computing the sample standard deviation

Sample Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1}} \quad [3-18]$$

TABLE 3-20 DETERMINATION OF THE VARIANCE AND STANDARD DEVIATION OF SALES OF 100 FAST-FOOD RESTAURANTS IN THE EASTERN DISTRICT (000S OMITTED)

Class	Midpoint <i>x</i> (1)	Frequency <i>f</i> (2)	<i>f</i> × <i>x</i> (3) = (2) × (1)	Mean μ (4)	<i>x</i> − μ (1) − (4)	$(x - \mu)^2$ [(1) − (4)] ²	$f(x - \mu)^2$ (2) × [(1) − (4)] ²
700–799	750	4	3,000	1,250	−500	250,000	1,000,000
800–899	850	7	5,950	1,250	−400	160,000	1,120,000
900–999	950	8	7,600	1,250	−300	90,000	720,000
1,000–1,099	1,050	10	10,500	1,250	−200	40,000	400,000
1,100–1,199	1,150	12	13,800	1,250	−100	10,000	120,000
1,200–1,299	1,250	17	21,250	1,250	0	0	0
1,300–1,399	1,350	13	17,550	1,250	100	10,000	130,000
1,400–1,499	1,450	10	14,500	1,250	200	40,000	400,000
1,500–1,599	1,550	9	13,950	1,250	300	90,000	810,000
1,600–1,699	1,650	7	11,550	1,250	400	160,000	1,120,000
1,700–1,799	1,750	2	3,500	1,250	500	250,000	500,000
1,800–1,899	1,850	1	1,850	1,250	600	360,000	360,000
		100	125,000				6,680,000

$$\bar{x} = \frac{\sum(f \times x)}{n} \quad [3-3]$$

$$= \frac{125,000}{100}$$

$$= 1,250 \text{ (thousands of dollars)} \leftarrow \text{Mean}$$

$$\sigma^2 = \frac{\sum f(x - \mu)^2}{N} \quad [3-15]$$

$$= \frac{6,680,000}{100}$$

$$= 66,800 \text{ (or } 66,800 \text{ [thousands of dollars]}^2\text{)} \leftarrow \text{Variance}$$

$$\sigma = \sqrt{\sigma^2} \quad [3-16]$$

$$= \sqrt{66,800}$$

$$= 258.5 \leftarrow \text{Standard deviation} = \$258,500$$

where

- s^2 = sample variance
- s = sample standard deviation
- x = value of each of the n observations
- \bar{x} = mean of the sample
- $n - 1$ = number of observations in the sample minus 1

Why do we use $n - 1$ as the denominator instead of n ? Statisticians can prove that if we take many samples from a given population, find the sample variance (s^2) for each sample, and average each of these together, then this average tends not to equal the population variance, σ^2 , unless we use $n - 1$ as the denominator. In Chapter 7, we shall learn the statistical explanation of why this is true.

Use of $n - 1$ as the denominator

Equations 3-17 and 3-18 enable us to find the sample variance and the sample standard deviation of the annual Blue Cross–Blue Shield payments to Cumberland Hospital in Table 3-21; note that both halves of Equation 3-17 yield the same result.

Calculating sample variance and standard deviation for hospital data

TABLE 3-21 DETERMINATION OF THE SAMPLE VARIANCE AND STANDARD DEVIATION OF ANNUAL BLUE CROSS-BLUE SHIELD PAYMENTS TO CUMBERLAND HOSPITAL (000S OMITTED)

Observation (x) (1)	Mean (\bar{x}) (2)	$x - \bar{x}$ (1) - (2)	$(x - \bar{x})^2$ [(1) - (2)] ²	x^2 (1) ²
863	1,351	-488	238,144	744,769
903	1,351	-448	200,704	815,409
957	1,351	-394	155,236	915,849
1,041	1,351	-310	96,100	1,083,681
1,138	1,351	-213	45,369	1,295,044
1,204	1,351	-147	21,609	1,449,616
1,354	1,351	3	9	1,833,316
1,624	1,351	273	74,529	2,637,376
1,698	1,351	347	120,409	2,883,204
1,745	1,351	394	155,236	3,045,025
1,802	1,351	451	203,401	3,247,204
1,883	1,351	532	283,024	3,545,689
$\sum (x - \bar{x})^2 \rightarrow 1,593,770$			$23,496,182 \leftarrow \sum x^2$	

$$\left\{ s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \right. \quad [3-17]$$

$$= \frac{1,593,770}{11} \quad [3-17]$$

= 144,888 (or 144,888 [thousands of dollars]²) ← Sample variance

$$s = \sqrt{s^2} \quad [3-18]$$

$$= \sqrt{144,888}$$

= 380.64 (that is, \$380,640) ← Sample standard deviation

$$\left\{ s^2 = \frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1} \right. \quad [3-17]$$

$$= \frac{23,496,182}{11} - \frac{12(1,351)^2}{11} \quad [3-17]$$

$$= \frac{1,593,770}{11}$$

$$= 144,888$$

OR

Just as we used the population standard deviation to derive population standard scores, we may also use the sample deviation to compute sample standard scores. These sample standard scores tell us how many standard deviations a particular sample observation lies below or above the sample mean. The appropriate formula is

Computing sample standard Scores

Standard Score of an Item in a Sample

$$\text{Sample standard score} = \frac{x - \bar{x}}{s} \quad [3-19]$$

where

- x = observation from the sample
- \bar{x} = sample mean
- s = sample standard deviation

In the example we just did, we see that the observation 863 corresponds to a standard score of -1.28:

$$\begin{aligned} \text{Sample standard score} &= \frac{x - \bar{x}}{s} \\ &= \frac{863 - 1,351}{380.64} \\ &= \frac{-488}{380.64} \\ &= -1.28 \end{aligned} \quad [3-19]$$

This section has demonstrated why the standard deviation is the measure of dispersion used most often. We can use it to compare distributions and to compute standard scores, an important element of statistical inference to be discussed later. Like the variance, the standard deviation takes into account every observation in the data set. But the standard deviation has some disadvantages, too. It is not as easy to calculate as the range, and it cannot be computed from open-ended distributions. In addition, extreme values in the data set distort the value of the standard deviation, although to a lesser extent than they do the range.

HINTS & ASSUMPTIONS

We assume when we calculate and use the standard deviation that there are not too many very large or very small values in the data set because we know that the standard deviation uses every value, and such extreme values will distort the answer. Hint: Forgetting whether to use N or $n - 1$ as the denominator for samples and populations can be avoided by associating the *smaller* value ($n - 1$) with the *smaller* set (the sample).

EXERCISES 3.9

Self-Check Exercises

SC 3-13 Talent, Ltd., a Hollywood casting company, is selecting a group of extras for a movie. The ages of the first 20 men to be interviewed are

50	56	55	49	52	57	56	57	56	59
54	55	61	60	51	59	62	52	54	49

The director of the movie wants men whose ages are fairly tightly grouped around 55 years. Being a statistics buff of sorts, the director suggests that a standard deviation of 3 years would be acceptable. Does this group of extras qualify?

SC 3-14 In an attempt to estimate potential future demand, the National Motor Company did a study asking married couples how many cars the average energy-minded family should own in 1998. For each couple, National averaged the husband's and wife's responses to get the overall couple response. The answers were then tabulated:

Number of cars	0	0.5	1.0	1.5	2.0	2.5
Frequency	2	14	23	7	4	2

- (a) Calculate the variance and the standard deviation.
- (b) Since the distribution is roughly bell-shaped, how many of the observations should theoretically fall between 0.5 and 1.5? Between 0 and 2? How many actually do fall in those intervals?

Applications

3-61 The head chef of The Flying Taco has just received two dozen tomatoes from her supplier, but she isn't ready to accept them. She knows from the invoice that the average weight of a tomato is 7.5 ounces, but she insists that all be of uniform weight. She will accept them only if the average weight is 7.5 ounces and the standard deviation is less than 0.5 ounce. Here are the weights of the tomatoes

6.3	7.2	7.3	8.1	7.8	6.8	7.5	7.8	7.2	7.5	8.1	8.2
8.0	7.4	7.6	7.7	7.6	7.4	7.5	8.4	7.4	7.6	6.2	7.4

What is the chef's decision and why?

3-62 These data are a sample of the daily production rate of fiberglass boats from Hydrosport, Ltd., a Miami manufacturer:

17	21	18	27	17	21	20	22	18	23
----	----	----	----	----	----	----	----	----	----

The company production manager feels that a standard deviation of more than three boats a day indicates unacceptable production-rate variations. Should she be concerned about plant-production rates?

3-63 A set of 60 observations has a mean of 66.8, a variance of 12.60, and an unknown distribution shape.

- (a) Between what values should at least 75 percent of the observations fall, according to Chebyshev's theorem?

- (b) If the distribution is symmetrical and bell-shaped, approximately how many observations should be found in the interval 59.7 to 73.9?
- (c) Find the standard scores for the following observations from the distribution: 61.45, 75.37, 84.65, and 51.50.

3-64 The number of checks cashed each day at the five branches of The Bank of Orange County during the past month had the following frequency distribution:

Class	Frequency
0–199	10
200–399	13
400–599	17
600–799	42
800–999	18

Hank Spivey, director of operations for the bank, knows that a standard deviation in check cashing of more than 200 checks per day creates staffing and organizational problems at the branches because of the uneven workload. Should Hank worry about staffing next month?

3-65 The Federal Reserve Board has given permission to all member banks to raise interest rates $\frac{1}{2}$ percent for all depositors. Old rates for passbook savings were $5\frac{1}{4}$ percent; for certificates of deposit (CDs): 1-year CD, $7\frac{1}{2}$ percent; 18-month CD, $8\frac{3}{4}$ percent; 2-year CD, $9\frac{1}{2}$ percent; 3-year CD, $10\frac{1}{2}$ percent; and 5-year CD, 11 percent. The president of the First State Bank wants to know what the characteristics of the new distribution of rates will be if a full $\frac{1}{2}$ percent is added to all rates. How are the new characteristics related to the old ones?

3-66 The administrator of a Georgia hospital surveyed the number of days 200 randomly chosen patients stayed in the hospital following an operation. The data are:

Hospital stay in days	1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24
Frequency	18	90	44	21	9	9	4	5

- (a) Calculate the standard deviation and mean.
- (b) According to Chebyshev's theorem, how many stays should be between 0 and 17 days? How many are actually in that interval?
- (c) Because the distribution is roughly bell-shaped, how many stays can we expect between 0 and 17 days?

3-67 FundInfo provides information to its subscribers to enable them to evaluate the performance of mutual funds they are considering as potential investment vehicles. A recent survey of funds whose stated investment goal was growth and income produced the following data on total annual rate of return over the past five years:

Annual return (%)	11.0–11.9	12.0–12.9	13.0–13.9	14.0–14.9	15.0–15.9	16.0–16.9	17.0–17.9	18.0–18.9
Frequency	2	2	8	10	11	8	3	1

- (a) Calculate the mean, variance, and standard deviation of the annual rate of return for this sample of 45 funds.
- (b) According to Chebyshev's theorem, between what values should at least 75 percent of the sample observations fall? What percentage of the observations actually do fall in that interval?

- (c) Because the distribution is roughly bell-shaped, between what values would you expect to find 68 percent of the observations? What percentage of the observations actually do fall in that interval?

3-68 Nell Berman, owner of the Earthbred Bakery, said that the average weekly production level of her company was 11,398 loaves, and the variance was 49,729. If the data used to compute the results were collected for 32 weeks, during how many weeks was the production level below 11,175? Above 11,844?

3-69 The Creative Illusion Advertising Company has three offices in three cities. Wage rates differ from state to state. In the Washington, D.C. office, the average wage increase for the past year was \$1,500, and the standard deviation was \$400. In the New York office, the average raise was \$3,760, and the standard deviation was \$622. In Durham, N.C., the average increase was \$850, and the standard deviation was \$95. Three employees were interviewed. The Washington employee received a raise of \$1,100; the New York employee, a raise of \$3,200; and the Durham employee, a raise of \$500. Which of the three had the smallest raise in relation to the mean and standard deviation of his office?

3-70 American Foods heavily markets three different products nationally. One of the underlying objectives of each of the product's advertisements is to make consumers recognize that American Foods makes the product. To measure how well each ad implants recognition, a group of consumers was asked to identify as quickly as possible the company responsible for a long list of products. The first American Foods product had an average latency of 2.5 seconds, and a standard deviation of 0.004 second. The second had an average latency of 2.8 seconds, and a standard deviation of 0.006 second. The third had an average latency of 3.7 seconds, and a standard deviation of 0.09 second. One particular subject had the following latencies: 2.495 for the first, 2.79 for the second, and 3.90 for the third. For which product was this subject farthest from average performance, in standard deviation units?

3-71 Sid Levinson is a doctor who specializes in the knowledge and effective use of pain-killing drugs for the seriously ill. In order to know approximately how many nurses and office personnel to employ, he has begun to keep track of the number of patients he sees each week. Each week his office manager records the number of seriously ill patients and the number of routine patients. Sid has reason to believe that the number of routine patients per week would look like a bell-shaped curve if he had enough data. (This is not true of seriously ill patients.) However, he has been collecting data for only the past five weeks.

Seriously ill patients	33	50	22	27	48
Routine patients	34	31	37	36	27

- (a) Calculate the mean and variance for the number of seriously ill patients per week. Use Chebyshev's theorem to find boundaries within which the "middle 75 percent" of numbers of seriously ill patients per week should fall.
- (b) Calculate the mean, variance, and standard deviation for the number of routine patients per week. Within what boundaries should the "middle 68 percent" of these weekly numbers fall?

3-72 The superintendent of any local school district has two major problems: A tough job dealing with the elected school board is the first, and the second is the need to be always prepared to look for a new job because of the first problem. Tom Langley, superintendent of School District 18, is no exception. He has learned the value of understanding all numbers in any budget

and being able to use them to his advantage. This year, the school board has proposed a media research budget of \$350,000. From past experience, Tom knows that actual spending always exceeds the budget proposal, and the amount by which it exceeds the proposal has a mean of \$40,000 and variance of 100,000,000 dollars squared. Tom learned about Chebyshev's theorem in college, and he thinks that this might be useful in finding a range of values within which the actual expenditure would fall 75 percent of the time in years when the budget proposal is the same as this year. Do Tom a favor and find this range.

3-73

Bea Reele, a well-known clinical psychologist, keeps very accurate data on all her patients. From these data, she has developed four categories within which to place all her patients: child, young adult, adult, and elderly. For each category, she has computed the mean IQ and the variance of IQs within that category. These numbers are given in the following table. If on a certain day Bea saw four patients (one from each category), and the IQs of those patients were as follows: child, 90; young adult, 92; adult, 100; elderly, 98; then which of the patients had the IQ farthest above the mean, in standard deviation units, for that particular category?

Category	Mean IQ	IQ Variance
Child	110	81
Young adult	90	64
Adult	95	49
Elderly	90	121

Worked-Out Answers to Self-Check Exercises

SC 3-13

x	$x - \bar{x}$	$(x - \bar{x})^2$	x	$x - \bar{x}$	$(x - \bar{x})^2$
50	-5.2	27.04	54	-1.2	1.44
56	0.8	0.64	55	-0.2	0.04
55	-0.2	0.04	61	5.8	33.64
49	-6.2	38.44	60	4.8	23.04
52	-3.2	10.24	51	-4.2	17.64
57	1.8	3.24	59	3.8	14.44
56	0.8	0.64	62	6.8	46.24
57	1.8	3.24	52	-3.2	10.24
56	0.8	0.64	54	-1.2	1.44
59	3.8	14.44	49	-6.2	38.44
1,104			285.20		

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1,104}{20} = 55.2 \text{ years, which is close to the desired 55 years}$$

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}} = \sqrt{\frac{285.20}{19}} = 3.874 \text{ years, which shows more variability than desired}$$

SC 3-14 (a)	# of cars <i>x</i>	Frequency <i>f</i>	<i>f</i> × <i>x</i>	<i>x</i> − \bar{x}	$(x - \bar{x})^2$	<i>f</i> (<i>x</i> − \bar{x}) ²
	0	2	0	-1.0288	1.0585	2.1170
	0.5	14	7	-0.5288	0.2797	3.9155
	1	23	23	-0.0288	0.0008	0.0191
	1.5	7	10.5	0.4712	0.2220	1.5539
	2	4	8	0.9712	0.9431	3.7726
	2.5	2	5	1.4712	2.1643	4.3286
		52	53.5			15.7067

$$\bar{x} = \frac{\Sigma x}{n} = \frac{53.5}{52} \quad 1.0288 \text{ cars}$$

$$s^2 = \frac{\Sigma f(x - \bar{x})^2}{n - 1} = \frac{15.707}{51} = 0.3080 \quad \text{so} \quad s = \sqrt{0.3080} = 0.55 \text{ car}$$

- (b) (0.5, 1.5) is approximately $\bar{x} \pm s$, so about 68 percent of the data, or $0.68(52) = 35.36$ observations should fall in this range. In fact, 44 observations fall into this interval.
 (0, 2) is approximately $\bar{x} \pm 2s$, so about 95 percent of the data, or $0.95(52) = 49.4$ observations should fall in this range. In fact, 50 observations fall into this interval.

3.10 RELATIVE DISPERSION: THE COEFFICIENT OF VARIATION

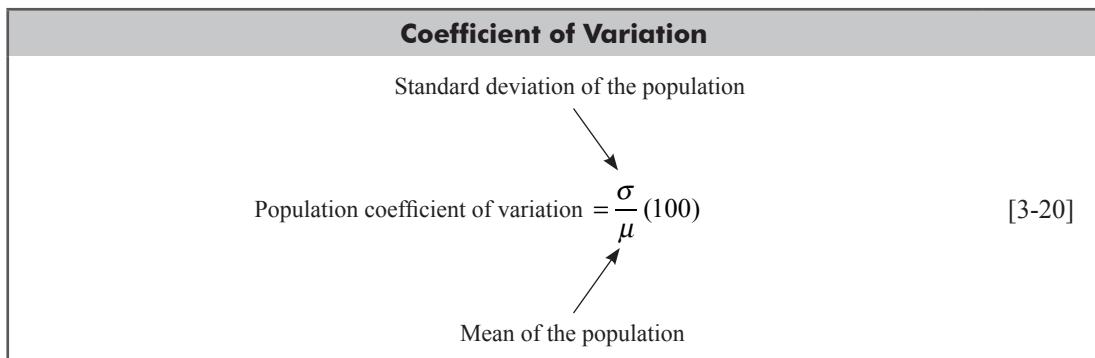
The standard deviation is an *absolute* measure of dispersion that expresses variation in the same units as the original data. The annual Blue Cross–Blue Shield payments to Cumberland Hospital (Table 3-21) have a standard deviation of \$380,640. The annual Blue Cross–Blue Shield payments to Valley Falls Hospital (Table 3-16) have a standard deviation (which you can compute) of \$57,390. Can we compare the values of these two standard deviations? Unfortunately, no.

The standard deviation cannot be the sole basis for comparing two distributions. If we have a standard deviation of 10 and a mean of 5, the values vary by an amount twice as large as the mean itself. On the other hand, if we have a standard deviation of 10 and a mean of 5,000, the variation relative to the mean is insignificant. Therefore, we cannot know the dispersion of a set of data until we know the standard deviation, the mean, *and* how the standard deviation compares with the mean.

What we need is a *relative* measure that will give us a feel for the magnitude of the deviation relative to the magnitude of the mean. The *coefficient of variation* is one such relative measure of dispersion. It relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean. The unit of measure, then, is “percent” rather than the same units as the original data. For a population, the formula for the coefficient of variation is

Shortcomings of the standard deviation

The coefficient of variation, a relative measure



Using this formula in an example, we may suppose that each day, laboratory technician A completes on average 40 analyses with a standard deviation of 5. Technician B completes on average 160 analyses per day with a standard deviation of 15. Which employee shows less variability?

At first glance, it appears that technician B has three times more variation in the output rate than technician A. But B completes analyses at a rate four times faster than A. Taking all this information into account, we can compute the coefficient of variation for both technicians:

$$\begin{aligned} \text{Coefficient of variation} &= \frac{\sigma}{\mu} (100) & [3-20] \\ &= \frac{5}{40} (100) & \text{Computing the coefficient of} \\ & & \text{variation} \\ &= 12.5\% \leftarrow \text{For technician A} \end{aligned}$$

and

$$\begin{aligned} \text{Coefficient of variation} &= \frac{15}{160} (100) \\ &= 9.4\% \leftarrow \text{For technician B} \end{aligned}$$

So we find that technician B, who has more *absolute* variation in output than technician A, has less *relative* variation because the mean output for B is much greater than for A.

For large data sets, we use the computer to calculate our measures of central tendency and variability. In Figure 3-13, we have used Minitab to compute some of these summary statistics for the grade data in Appendix 10. The statistics are shown for each section as well as for the course as a whole.

Using the computer to compute measures of central tendency and variability

In Figure 3-14, we have used Minitab to calculate several measures of central tendency and variability for the earnings data in Appendix 11. The statistics are given for all 224 companies together, and they are also broken down by stock exchange (1 = OTC, 2 = ASE, 3 = NYSE). The statistic TRMEAN is a “trimmed mean,” a mean calculated with the top 5 percent and bottom 5 percent of the data omitted. This helps to alleviate the distortion caused by the extreme values from which the ordinary arithmetic mean suffers.

HINTS & ASSUMPTIONS

The concept and usefulness of the coefficient of variation are quickly evident if you try to compare overweight men with overweight women. Suppose a group of men and women are all 20 pounds overweight. The 20 pounds is not a good measure of the excessive weight. Average weight for men is about 160 pounds, and average weight for women is about 120 pounds. Using a simple ratio, we can see that the women are 20/120, or about 16.7 percent overweight but the men are 20/160, or about 12.5 percent overweight. Although the coefficient of variation is a bit more complex than our simple ratio example, the concept is the same: We use it to compare the amount of variation in data groups that have different means. *Warning:* Don't compare the dispersion in data sets by using their standard deviations unless their means are close to each other.

EXERCISES 3.10**Self-Check Exercises**

SC 3-15 Bassart Electronics is considering employing one of two training programs. Two groups were trained for the same task. Group 1 was trained by program A; group 2, by program B. For the first group, the times required to train the employees had an average of 32.11 hours and a variance of 68.09. In the second group, the average was 19.75 hours and the variance was 71.14. Which training program has less relative variability in its performance?

SC 3-16 Southeastern Stereos, a wholesaler, was contemplating becoming the supplier to three retailers, but inventory shortages have forced Southeastern to select only one. Southeastern's credit manager is evaluating the credit record of these three retailers. Over the past 5 years, these retailers' accounts receivable have been outstanding for the following average number of days. The credit manager feels that consistency, in addition to lowest average, is important. Based on relative dispersion, which retailer would make the best customer?

Lee	62.2	61.8	63.4	63.0	61.7
Forrest	62.5	61.9	62.8	63.0	60.7
Davis	62.0	61.9	63.0	63.9	61.5

Applications

3-74 The weights of the Baltimore Bullets professional football team have a mean of 224 pounds with a standard deviation of 18 pounds, while the mean weight and standard deviation of their Sunday opponent, the Chicago Trailblazers, are 195 and 12, respectively. Which team exhibits the greater relative dispersion in weights?

3-75 The university has decided to test three new kinds of lightbulbs. They have three identical rooms to use in the experiment. Bulb 1 has an average life-time of 1,470 hours and a variance of 156. Bulb 2 has an average lifetime of 1,400 hours and a variance of 81. Bulb 3 has an average lifetime of 1,350 hours and a standard deviation of 6 hours. Rank the bulbs in terms of relative variability. Which was the best bulb?

3-76 Students' ages in the regular daytime M.B.A. program and the evening program of Central University are described by these two samples:

Regular M.B.A.	23	29	27	22	24	21	25	26	27	24
Evening M.B.A.	27	34	30	29	28	30	34	35	28	29

If homogeneity of the class is a positive factor in learning, use a measure of relative variability to suggest which of the two groups will be easier to teach.

- 3-77 There are a number of possible measures of sales performance, including how consistent a salesperson is in meeting established sales goals. The data that follow represent the percentage of goal met by each of three salespeople over the last 5 years.

Patricia	88	68	89	92	103
John	76	88	90	86	79
Frank	104	88	118	88	123

- (a) Which salesperson is the most consistent?
- (b) Comment on the adequacy of using a measure of consistency along with percentage of sales goal met to evaluate sales performance.
- (c) Can you suggest a more appropriate alternative measure of consistency?

- 3-78 The board of directors of Gothic Products is considering acquiring one of two companies and is closely examining the management of each company in regard to their inclinations toward risk. During the past five years, the first company's returns on investments had an average of 28.0 percent and a standard deviation of 5.3 percent. The second company's returns on investments had an average of 37.8 percent and a standard deviation of 4.8 percent. If we consider risk to be associated with greater relative dispersion, which of these two companies has pursued a riskier strategy?

- 3-79 A drug company that supplies hospitals with premeasured doses of certain medications uses different machines for medications requiring different dosage amounts. One machine, designed to produce doses of 100 cc, has as its mean dose 100 cc, and a standard deviation of 5.2 cc. Another machine produces premeasured amounts of 180 cc of medication and has a standard deviation of 8.6 cc. Which machine has the lower accuracy from the standpoint of relative dispersion?

- 3-80 HumanPower, the temporary employment agency, has tested many people's data entry skills. Infotech needs a data entry person, and the person needs to be not only quick but also consistent. HumanPower pulls the speed records for 4 employees with the data given below in terms of number of correct entries per minute. Which employee is best for Infotech based on relative dispersion?

John	63	66	68	62	69	72
Jeff	68	67	66	67	69	
Mary	62	79	75	59	72	84
Tammy	64	68	58	57	59	

- 3-81 Wyatt Seed Company sells three grades of Early White Sugar corn seed, distinguished according to the consistency of germination of the seeds. The state seed testing laboratory has a sample of each grade of seed and its test results on the number of seeds that germinated out of packages of 100 are as follows:

Grade I (Regular)	88	91	92	89	79
Grade II (Extra)	87	92	88	90	92
Grade III (Super)	90	89	79	93	88

Does Wyatt's grading of its seeds make sense?

- 3-82** Sunray Appliance Company has just completed a study of three possible assembly-line configurations for producing its best-selling two-slice toaster. Configuration I has yielded a mean time to construct a toaster of 34.8 minutes, and a standard deviation of 4.8 minutes. Configuration II has yielded a mean of 25.5 minutes, and a standard deviation of 7.5 minutes. Configuration III has yielded a mean of 37.5 minutes, and a standard deviation of 3.8 minutes. Which assembly-line configuration has the least relative variation in the time it takes to construct a toaster?

Worked-Out Answers to Self-Check Exercises

SC 3-15 Program A: $CV = \frac{\sigma}{\mu}(100) = \frac{\sqrt{68.09}(100)}{32.11} = 25.7$ percent

Program B: $CV = \frac{\sigma}{\mu}(100) = \frac{\sqrt{71.14}(100)}{19.75} = 42.7$ percent

Program A has less relative variability.

SC 3-16 Lee: $\bar{x} = 62.42$ $s = 0.7497$ $CV = (s/\bar{x})(100) = \frac{0.7497(100)}{62.42} = 1.20$ percent

Forrest: $\bar{x} = 62.18$ $s = 0.9257$ $CV = (s/\bar{x})(100) = \frac{0.9257(100)}{62.18} = 1.49$ percent

Davis: $\bar{x} = 62.46$ $s = 0.9762$ $CV = (s/\bar{x})(100) = \frac{0.9762(100)}{62.46} = 1.56$ percent

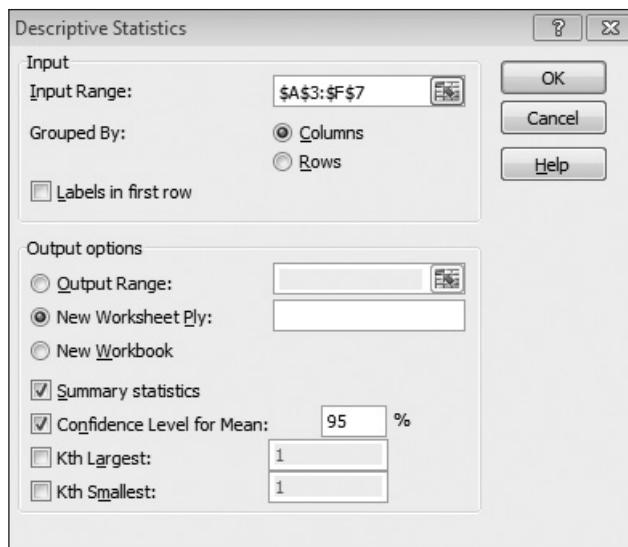
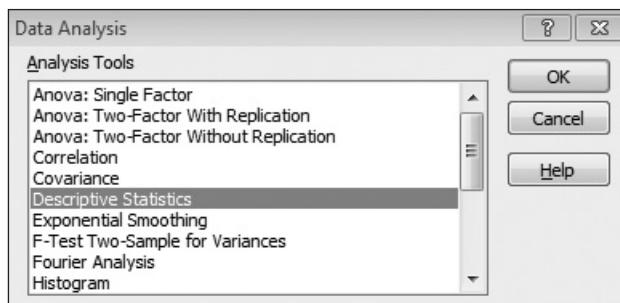
Based on relative dispersion, Lee would be the best customer, but there really isn't much difference among the three of them.

3.11 DESCRIPTIVE STATISTICS USING MSEXCEL & SPSS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																			
2	Sample of daily production in meters of 30 carpet looms																		
3	16.2	16.8	15.9	15.6	15.9	16.6													
4	15.8	16	16	15.7	15.9	15.6													
5	15.8	16.4	16.3	16	16.8	15.6													
6	15.8	15.2	16	16.2	15.4	16.9													
7	16.3	15.9	16.4	16.1	15.7	16.3													
8																			

Above data is sample of daily production in meters of 30 carpet looms for calculating measure of central tendency and dispersion.

For Measure of central tendency and dispersion go to **Data>Data Analysis>Descriptive Statistics>Give Data Range>Select summary statistics and CI for mean**



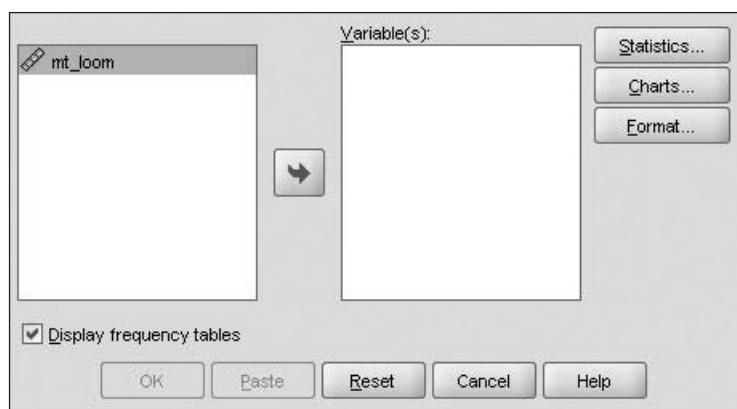
Chapter 3 - Microsoft Excel

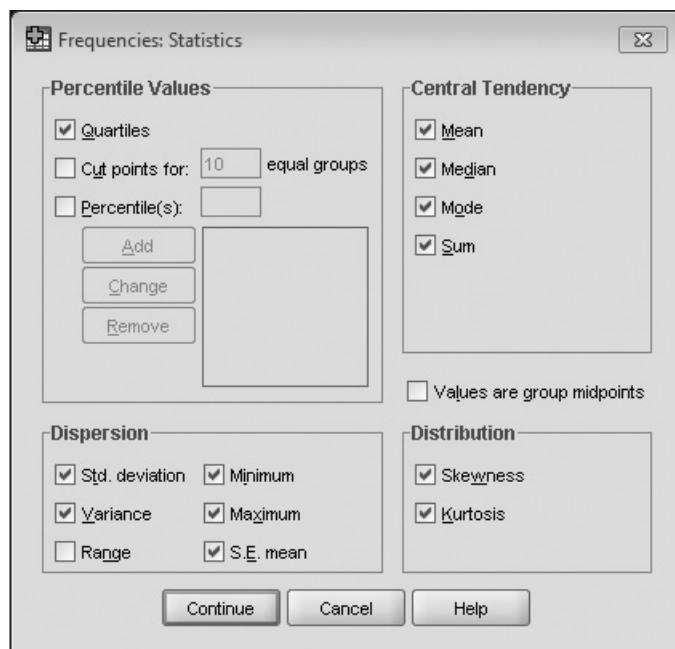
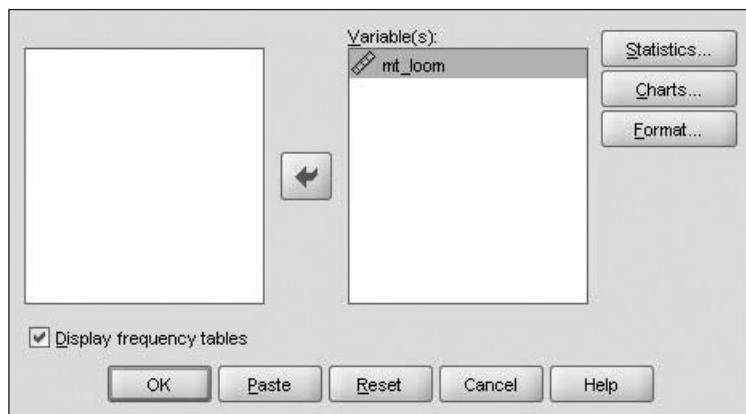
The screenshot shows a Microsoft Excel spreadsheet titled 'Chapter 3 - Microsoft Excel'. The data is organized into columns labeled 'Column1' through 'Column6'. Row 3 contains the mean values for each column: 15.98, 16.06, 16.12, 15.92, 15.94, and 16.2. Row 4 contains the standard deviation values: 0.111355, 0.267582, 0.096954, 0.115758, 0.233666, and 0.262679. Row 5 contains the median values: 15.8, 15.8, 16, 16, 15.9, and 16.3. Row 6 contains the mode values: 15.8, 15.8, IN/A, IN/A, 15.9, and 15.6. Row 7 contains the standard error values: 0.248998, 0.248998, 0.598331, 0.216795, 0.258844, and 0.522494. Row 8 contains the sample variance values: 0.062, 0.062, 0.358, 0.047, 0.067, and 0.273. Row 9 contains the kurtosis values: -2.80437, -2.80437, 0.396523, -2.36797, 2.656281, and -2.48645. Row 10 contains the skewness values: 0.6959578, -0.38095, -0.38095, 0.559407, -0.36327, and 1.357255. Row 11 contains the range values: 0.5, 1.6, 0.5, 0.6, 1.4, and 1.3. Row 12 contains the minimum values: 15.8, 15.8, 15.2, 15.9, 15.4, and 15.6. Row 13 contains the maximum values: 16.3, 16.3, 16.8, 16.4, 16.8, and 16.9. Row 14 contains the sum values: 79.9, 80.3, 80.6, 79.6, 79.7, and 81. Row 15 contains the count values: 5, 5, 5, 5, 5, and 5. Row 16 contains the confidence interval values: 0.309172, 0.742926, 0.269186, 0.321397, 0.648762, and 0.729312. The Excel ribbon at the top includes tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Acrobat. The Data tab is active, showing various tools like Refresh, Connections, Sort & Filter, Text to Columns, Data Tools, and What-If Analysis.

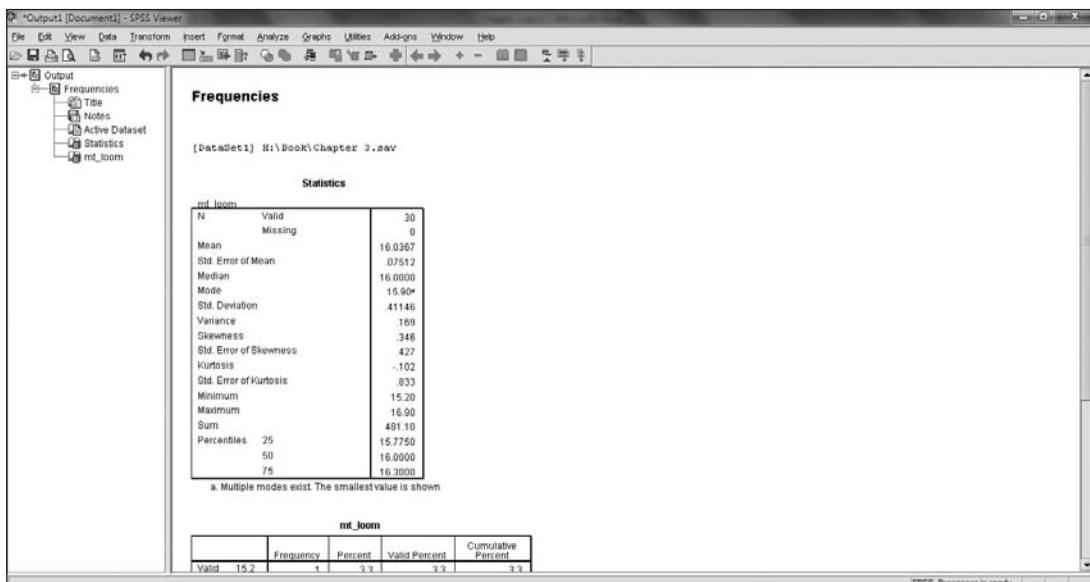
	Column1	Column2	Column3	Column4	Column5	Column6
1	Mean	15.98	Mean	16.06	Mean	16.12
2	Standard	0.111355	Standard	0.267582	Standard	0.096954
3	Median	15.8	Median	15.8	Median	16
4	Mode	15.8	Mode	IN/A	Mode	16
5	Standard	0.248998	Standard	0.598331	Standard	0.216795
6	Sample	0.062	Sample	0.358	Sample	0.047
7	Variance	0.062	Variance	0.062	Variance	0.047
8	Kurtosis	-2.80437	Kurtosis	0.396523	Kurtosis	-2.36797
9	Skewness	0.6959578	Skewness	-0.38095	Skewness	0.559407
10	Range	0.5	Range	1.6	Range	0.5
11	Minimum	15.8	Minimum	15.2	Minimum	15.9
12	Maximum	16.3	Maximum	16.8	Maximum	16.4
13	Sum	79.9	Sum	80.3	Sum	80.6
14	Count	5	Count	5	Count	5
15	Confident	0.309172	Confident	0.742926	Confident	0.269186
16	Confident	0.321397	Confident	0.648762	Confident	0.729312

For calculating measure of central tendency and dispersion in SPSS 16.0 Go to **Analyze>Descriptive Statistics>Frequencies>In statistics select desired measure of central tendency and measure of dispersion.**

	mt_loom	var													
1	16.20														
2	15.80														
3	15.00														
4	15.80														
5	16.30														
6	16.80														
7	16.00														
8	16.40														
9	15.20														
10	15.90														
11	15.90														
12	16.00														
13	16.30														
14	16.00														
15	16.40														
16	15.60														
17	15.70														
18	16.00														
19	16.20														
20	16.10														
21	15.90														
22	15.90														
23	16.00														
24	15.40														
25	15.70														
26	16.60														







STATISTICS AT WORK

Loveland Computers

Case 3: Central Tendency and Dispersion “Not bad for a few days’ work, Lee,” Uncle Walter congratulated his new assistant as he flipped through 12 pages of tables, charts, and graphs. Monday morning had come all too soon for Lee.

“Well, Nunc,” replied Lee, with a familiarity possible only in a family firm, “it took a few all-nighters. But I’ve set things up so that we won’t have to go through this kind of agony in the future. I’ve archived all the old data on diskettes in a common format, and I’ve kept the last 3 years on the hard drive. More important, I’ve set up some common reporting formats for each product line so the data will be collected in a consistent manner from here on out. And with the 3D spreadsheet, I can easily sum them together and give you data by month or by quarter.” Warming to his audience, Lee flipped to the last page and showed a simple pie chart. “Here’s the beauty of this business: You can show those New Yorkers that your average gross margin (you know, revenue minus your cost of goods sold) is 28 percent. That should impress them.”

“Well maybe yes and maybe no,” commented Gratia Delaguardia, Walter Azko’s partner, who had just walked in. If Walter was known for his charm and his “street smarts,” Gratia certainly earned the title of “the brains” of this outfit. “You’re probably mixing up apples and oranges there. Some of the low-speed PCs don’t have that large a gross margin any more. The profit is a little thin, but at least it’s predictable. With the new technologies, we make a huge margin on our ‘hit’ products, but there are others where we had to cut prices to get rid of them. You’ll remember our first ‘portable’ that weighed more than 50 pounds, Walt.”

“I try to forget that one,” responded the CEO tersely. “But, Lee, Gratia has a point. Don’t you think you ought to break out new products—say, products within their first 6 months on sale—versus the

established lines. See if the gross margins look different and whether they're all over the place like Gratia says. I'm off to the airport to pick up the investment folks. See what you can whip up by the time I get back."

Study Questions: The spreadsheet program Lee is using has many built-in statistical functions. Which ones should Lee use to answer the questions about gross margins? How might the data be presented, and how will this help the new investors in their decision making? What limitations are there on assuming a bell-shaped distribution for "percentage" data?

CHAPTER REVIEW

Terms Introduced in Chapter 3

Bimodal Distribution A distribution of data points in which two values occur more frequently than the rest of the values in the data set.

Boxplot A graphical EDA technique used to highlight the center and extremes of a data set.

Chebyshev's Theorem No matter what the shape of a distribution, at least 75 percent of the values in the population will fall within 2 standard deviations of the mean and at least 89 percent will fall within 3 standard deviations.

Coding A method of calculating the mean for grouped data by recoding values of class midpoints to more simple values.

Coefficient of Variation A relative measure of dispersion, comparable across distributions, that expresses the standard deviation as a percentage of the mean.

Deciles Fractiles that divide the data into 10 equal parts.

Dispersion The spread or variability in a set of data.

Distance Measure A measure of dispersion in terms of the difference between two values in the data set.

Exploratory Data Analysis (EDA) Methods for analyzing data that require very few prior assumptions.

Fractile In a frequency distribution, the location of a value at or above a given fraction of the data.

Geometric Mean A measure of central tendency used to measure the average rate of change or growth for some quantity, computed by taking the n th root of the product of n values representing change.

Interfractile Range A measure of the spread between two fractiles in a distribution, that is, the difference between the values of two fractiles.

Interquartile Range The difference between the values of the first and the third quartiles; this difference indicates the range of the middle half of the data set.

Kurtosis The degree of peakedness of a distribution of points.

Mean A central tendency measure representing the arithmetic average of a set of observations.

Measure of Central Tendency A measure indicating the value to be expected of a typical or middle data point.

Measure of Dispersion A measure describing how the observations in a data set are scattered or spread out.

Median The middle point of a data set, a measure of location that divides the data set into halves.

Median Class The class in a frequency distribution that contains the median value for a data set.

Mode The value most often repeated in the data set. It is represented by the highest point in the distribution curve of a data set.

Parameters Numerical values that describe the characteristics of a whole population, commonly represented by Greek letters.

Percentiles Fractiles that divide the data into 100 equal parts.

Quartiles Fractiles that divide the data into four equal parts.

Range The distance between the highest and lowest values in a data set.

Skewness The extent to which a distribution of data points is concentrated at one end or the other; the lack of symmetry.

Standard Deviation The positive square root of the variance; a measure of dispersion in the same units as the original data, rather than in the squared units of the variance.

Standard Score Expressing an observation in terms of standard deviation units above or below the mean; that is, the transformation of an observation by subtracting the mean and dividing by the standard deviation.

Statistics Numerical measures describing the characteristics of a sample. Represented by Roman letters.

Stem and Leaf Display A histogram-like display used in EDA to group data, while still displaying all the original values.

Summary Statistics Single numbers that describe certain characteristics of a data set.

Symmetrical A characteristic of a distribution in which each half is the mirror image of the other half.

Variance A measure of the average squared distance between the mean and each item in the population.

Weighted Mean An average calculated to take into account the importance of each value to the overall total, that is, an average in which each observation value is weighted by some index of its importance.

Equations Introduced in Chapter 3

$$3-1 \quad \mu = \frac{\sum x}{N} \quad \text{p. 79}$$

The *population arithmetic mean* is equal to the sum of the values of all the elements in the population ($\sum x$) divided by the number of elements in the population (N).

$$3-2 \quad \bar{x} = \frac{\sum x}{n} \quad \text{p. 79}$$

To calculate the *sample arithmetic mean*, sum the values of all the elements in the sample ($\sum x$) and divide by the number of elements in the sample (n).

$$3-3 \quad \bar{x} = \frac{\sum(f \times x)}{n} \quad \text{p. 79}$$

To find the *sample arithmetic mean of grouped data*, calculate the midpoints (x) for each class in the sample. Then multiply each midpoint by the frequency (f) of observations in the class, sum (\sum) all these results, and divide by the total number of observations in the sample (n).

$$3-4 \quad \bar{x} = x_0 + w \frac{\sum(u \times f)}{n} \quad \text{p. 81}$$

This formula enables us to calculate the *sample arithmetic mean of grouped data* using codes to eliminate dealing with large or inconvenient midpoints. Assign these codes (u) as follows: Give the value of zero to the middle midpoint (called x_0), positive consecutive integers to midpoints larger than x_0 , and negative consecutive integers to smaller midpoints. Then, multiply the code assigned to each class (u) by the frequency (f) of observations in the class and sum (\sum) all these products. Divide this result by the total number of observations in the sample (n),

multiply by the numerical width of the class interval (w), and add the value of the midpoint assigned the code zero (x_0).

$$3-5 \quad \bar{x}_w = \frac{\sum(w \times x)}{\sum w} \quad \text{p. 89}$$

The *weighted mean*, \bar{x}_w , is an average that takes into account how important each value is to the overall total. We can calculate this average by multiplying the weight, or proportion, of each element (w) by that element (x), summing the results (\sum), and dividing this amount by the sum of all the weights ($\sum w$).

$$3-6 \quad \text{G.M.} = \sqrt[n]{\text{product of all } x \text{ values}} \quad \text{p. 94}$$

The *geometric mean*, or G.M., is appropriate to use whenever we need to measure the average rate of change (the growth rate) over a period of time. In this equation, n is equal to the number of x values dealt with in the problem.

$$3-7 \quad \text{Median} = \left(\frac{n+1}{2} \right) \text{th item in a data array} \quad \text{p. 97}$$

where n = number of items in the data array

The *median* is a single value that measures the central item in the data set. Half the items lie above the median, half below it. If the data set contains an odd number of items, the middle item of the array is the median. For an even number of items, the median is the average of the two middle items. Use this formula when the data are ungrouped.

$$3-8 \quad \tilde{m} = \left(\frac{(n+1)/2 - (F+1)}{f_m} \right) w + L_m \quad \text{p. 100}$$

This formula enables us to find the *sample median of grouped data*. In it, n equals the total number of items in the distribution; F equals the sum of all the class frequencies up to, but not including, the median class; f_m is the frequency of observations in the median class; w is the class-interval width; and L_m is the lower limit of the median class interval.

$$3-9 \quad Mo = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w \quad \text{p. 105}$$

The *mode* is that value most often repeated in the data set. To find the *mode of grouped data* (symbolized Mo), use this formula and let L_{Mo} = lower limit of the modal class; d_1 = frequency of the modal class minus the frequency of the class directly below it; d_2 = frequency of the modal class minus the frequency of the class directly above it; and w = width of the modal class interval.

$$3-10 \quad \text{Range} = \frac{\text{Value of highest observation} - \text{Value of lowest observation}}{} \quad \text{p. 114}$$

The *range* is the difference between the highest and lowest values in a frequency distribution.

$$3-11 \quad \text{Interquartile range} = Q_3 - Q_1 \quad \text{p. 115}$$

The *interquartile range* measures approximately how far from the median we must go on either side before we can include one-half the values of the data set. To compute this range, divide the data into four equal parts. The *quartiles* (Q) are the highest values in each of these four parts. The *interquartile range* is the difference between the values of the first and third quartiles (Q_1 and Q_3).

3-12

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} = \frac{\Sigma x^2}{N} - \mu^2$$

p. 119

This formula enables us to calculate the *population variance*, a measure of the average *squared* distance between the mean and each item in the population. The middle expression, $\frac{\Sigma(x - \mu)^2}{N}$, is the definition of σ^2 . The last expression, $\frac{\Sigma x^2}{N} - \mu^2$, is mathematically equivalent to the definition but is often much more convenient to use because it frees us from calculating the deviations from the mean.

3-13

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x - \mu)^2}{N}} = \sqrt{\frac{\Sigma x^2}{N} - \mu^2}$$

p. 120

The population standard deviation, σ , is the square root of the population variance. It is a more useful parameter than the variance because it is expressed in the same units as the data (whereas the units of the variance are the squares of the units of the data). The standard deviation is always the *positive* square root of the variance.

3-14

$$\text{Population standard score} = \frac{x - \mu}{\sigma}$$

p. 122

The *standard score* of an observation is the number of standard deviations the observation lies below or above the mean of the distribution. The standard score enables us to make comparisons between distribution items that differ in order of magnitude or in the units used. Use Equation 3-14 to find the standard score of an item in a *population*.

3-15

$$\sigma^2 = \frac{\Sigma f(x - \mu)^2}{N} = \frac{\Sigma fx^2}{N} - \mu^2$$

p. 123

This formula in either form enables us to calculate the *variance of data already grouped* in a frequency distribution. Here, f represents the frequency of the class and x represents the midpoint.

3-16

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}} = \sqrt{\frac{\Sigma fx^2}{N} - \mu^2}$$

p. 124

Take the square root of the variance and you have the *standard deviation using grouped data*.

3-17

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{\Sigma x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1}$$

p. 124

To compute the *sample variance*, use the same formula as Equation 3-12, replacing μ with \bar{x} and N with $n - 1$. Chapter 7 contains an explanation of why we use $n - 1$ rather than n to calculate the sample variance.

3-18

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{\Sigma x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1}}$$

p. 124

The *sample standard deviation* is the square root of the sample variance. It is similar to Equation 3-13, except that μ is replaced by the sample mean \bar{x} and N is changed to $n - 1$.

3-19

$$\text{Sample standard score} = \frac{x - \bar{x}}{s}$$

p. 127

Use this equation to find the standard score of an item in a *sample*.

3-20

$$\text{Population coefficient of variation} = \frac{\sigma}{\mu} (100)$$

p. 133

The *coefficient of variation* is a relative measure of dispersion that enables us to compare two distributions. It relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean.

Review and Application Exercises

- 3-83 The weights and measures department of a state agriculture department measured the amount of granola sold in 4-ounce packets and recorded the following data:

4.01	4.00	4.02	4.02	4.03	4.00	3.98	3.99	3.99	4.01
3.99	3.98	3.97	4.00	4.02	4.01	4.02	4.00	4.01	3.99

If the sample is typical of all granola snacks marketed by this manufacturer, what is the range of weights in 95 percent of the packages?

- 3-84 How would you react to this statement from a football fan: "The Rockland Raiders average 3.6 yards a carry in their ground game. Since they need only 10 yards for a first down, and they have four plays to get it, they can't miss if they just stick to their ground game."

- 3-85 How would you reply to the following statement: "Variability is not an important factor because even though the outcome is more uncertain, you still have an equal chance of falling either above or below the median. Therefore, on average, the outcome will be the same."

- 3-86 Following are three general sections of one year's defense budget, each of which was allocated the same amount of funding by Congress:

- (a) Officer salaries (total).
- (b) Aircraft maintenance.
- (c) Food purchases (total).

Considering the distribution of possible outcomes for the funds actually spent in each of these areas, match each section to one of the curves in Figure 3-9. Support your answers.

- 3-87 Ed's Sports Equipment Company stocks two grades of fishing line. Data on each line are

Mean Test Strength (lb)		Standard Deviation
Master	40	Exact value unknown, but estimated to be quite large
Super	30	Exact value unknown, but estimated to be quite small

If you are going fishing for bluefish, which have been averaging 25 pounds this season, with which line would you probably land more fish?

- 3-88 The VP of sales for Vanguard Products has been studying records regarding the performances of his sales reps. He has noticed that in the last 2 years, the average level of sales per sales rep has remained the same, while the distribution of the sales levels has widened. Salespeople's sales levels from this period have significantly larger variations from the mean than in any of the previous 2-year periods for which he has records. What conclusions might be drawn from these observations?

- 3-89 New cars sold in December at eight Ford dealers within 50 miles of Canton, Ohio, can be described by this data set:

200	156	231	222	96	289	126	308
-----	-----	-----	-----	----	-----	-----	-----

- (a) Compute the range, interquartile range, and standard deviation of these data.
 (b) Which of the three measures you have computed in part (a) best describes the variability of these data?

3-90 Two economists are studying fluctuations in the price of gold. One is examining the period of 1968–1972. The other is examining the period of 1975–1979. What differences would you expect to find in the variability of their data?

3-91 The Downhill Ski Boot Company runs two assembly lines in its plant. The production manager is interested in improving the consistency of the line with the greater variation. Line number 1 has a monthly average of 11,350 units, and a standard deviation of 1,050. Line number 2 has a monthly average of 9,935, and a standard deviation of 1,010. Which line has the greater relative dispersion?

3-92 The Fish and Game station on Lake Wylie keeps records of fish caught on the lake and reports its finding to the National Fish and Game Service. The catch in pounds for the last 20 days was:

101	132	145	144	130	88	156	188	169	130
90	140	130	139	99	100	208	192	165	216

Calculate the range, variance, and standard deviation for these data. In this instance, is the range a good measure of the variability? Why?

3-93 The owner of Records Anonymous, a large record retailer, uses two different formulas for predicting monthly sales. The first formula has an average miss of 700 records, and a standard deviation of 35 records. The second formula has an average miss of 300 records, and a standard deviation of 16. Which formula is relatively less accurate?

3-94 Using the following population data, calculate the interquartile range, variance, and standard deviation. What do your answers tell you about the cost behavior of heating fuel?

Average Heating Fuel Cost per Gallon for Eight States							
1.89	1.66	1.77	1.83	1.71	1.68	1.69	1.73

3-95 The following are the average numbers of New York City police officers on duty each day between 8 P.M. and midnight in the borough of Manhattan:

Mon.	2,950	Wed.	2,900	Fri.	3,285	Sun.	2,975
Tues.	2,900	Thurs.	2,980	Sat.	3,430		

- (a) Would either the variance or the standard deviation be a good measure of the variability of these data?
 (b) What in the staffing pattern caused you to answer part (a) the way you did?

3-96 A psychologist wrote a computer program to simulate the way a person responds to a standard IQ test. To test the program, he gave the computer 15 different forms of a popular IQ test and computed its IQ from each form.

IQ Values				
134	136	137	138	138
143	144	144	145	146
146	146	147	148	153

- (a) Calculate the mean and standard deviation of the IQ scores.
 (b) According to Chebyshev's theorem, how many of the values should be between 132.44 and 153.56? How many are actually in that interval?

3-97 Liquid Concrete delivers ready-mixed concrete from 40 trucks. The number of cubic yards delivered by each truck on one day was as follows:

Cubic Yards									
11.9	12.8	14.6	15.8	13.7	9.9	18.8	16.9	10.4	9.1
17.1	13.0	18.6	16.0	13.9	14.7	17.7	12.1	18.0	17.8
19.0	13.3	12.4	9.3	14.2	15.0	19.3	10.6	11.2	9.6
13.6	14.5	19.6	16.6	12.7	15.3	10.9	18.3	17.4	16.3

List the values in each decile. Eighty percent of trucks delivered fewer than _____ cubic yards.

3-98 Baseball attendance at the Baltimore Eagles' last 10 home games looked like this:

20,100	24,500	31,600	28,400	49,500
19,350	25,600	30,600	11,300	28,560

- (a) Compute the range, variance, and standard deviation for these data.
 (b) Are any of your answers in part (a) an accurate portrayal of the variability in the attendance data?
 (c) What other measure of variability might be a better measure?
 (d) Compute the value of the measure you suggest in part (c).

3-99 Matthews, Young and Associates, a Chapel Hill consulting firm, has these records indicating the number of days each of its ten staff consultants billed last year:

212	220	230	210	228	229	231	219	221	222
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

- (a) Without computing the value of any of these measures, which of them would you guess would give you more information about this distribution: range or standard deviation?
 (b) Considering the difficulty and time of computing each of the measures you reviewed in part (a), which one would you suggest is better?
 (c) What will cause you to change your mind about your choice?

3-100 Larsen Equipment Rental provides contractors with tools they need for just a few days, such as concrete saws. When equipment is broken during a rental, it must be taken out of service until a repair is made. Often this can be done quickly, but there are sometimes delays while parts are ordered. Analysis of time lost for servicing is useful in planning for inventory. The records of downtime for last year were:

Equipment Group	Days Out of Service	Equipment Group	Days Out of Service
1	2		8
2	19		9
3	14		10
4	21		11
5	5		12
6	7		13
7	11		14

- (a) What was last year's mean downtime for the equipment groups?
 (b) What was the median?

3-101 Larsen (see Exercise 3-102) has just gotten the following additional information:

Equipment Group	Pieces of Machinery	Equipment Group	Pieces of Machinery
1	1	8	5
2	3	9	8
3	1	10	2
4	4	11	2
5	2	12	6
6	1	13	1
7	1	14	1

- (a) What is the average downtime per piece of machinery?
 (b) What is the average downtime per piece of machinery for each group when classified by group?
 (c) How many groups had a higher-than-average downtime per piece of machinery?

3-102 Compare and contrast the central position and skewness of the distributions of the readership volume in numbers of readers per issue for all nationally distributed
 (a) Monthly magazines.
 (b) Weekly news magazines.
 (c) Monthly medical journals.

3-103 Compare and contrast the central tendency and skewness of the distributions of the amount of taxes paid (in dollars) for all
 (a) Individuals filing federal returns in the United States, where the top tax bracket is 28 percent.
 (b) Individuals paying state income taxes in North Carolina, where the top tax bracket is 7 percent.
 (c) Individuals paying airport taxes (contained in the price of the airplane ticket) at JFK International Airport in New York City.

3-104 Allison Barrett does statistical analyses for an automobile racing team. Here are the fuel consumption figures in miles per gallon for the team's cars in recent races:

4.77	6.11	6.11	5.05	5.99	4.91	5.27	6.01
5.75	4.89	6.05	5.22	6.02	5.24	6.11	5.02

- (a) Calculate the median fuel consumption.
 (b) Calculate the mean fuel consumption.
 (c) Group the data into five equally sized classes. What is the fuel consumption value of the modal class?
 (d) Which of the three measures of central tendency is best for Allison to use when she orders fuel? Explain.

- 3-105** Claire Chavez, an Internal Revenue Service analyst, has been asked to describe the “average” American taxpayer in terms of gross annual income. She has summary data grouping taxpayers into different income classes. Which measure of central tendency should she use?
- 3-106** Emmot Bulb Co. sells a grab bag of flower bulbs. The bags are sold by weight; thus, the number of bulbs in each can vary depending on the varieties included. The number of bulbs in each of 20 bags sampled were:

21	33	37	56	47
36	23	26	33	37
25	33	32	47	34
26	37	37	43	45

- (a) What are the mean and median number of bulbs per bag?
 (b) Based on your answer, what can you conclude about the shape of the distribution of number of bulbs per bag?
- 3-107** An engineer tested nine samples of each of three designs of a certain bearing for a new electrical winch. The following data are the number of hours it took for each bearing to fail when the winch motor was run continuously at maximum output, with a load on the winch equivalent to 1.9 times the intended capacity.

Design		
A	B	C
16	18	31
16	27	16
53	23	42
15	21	20
31	22	18
17	26	17
14	39	16
30	17	15
20	28	19

- (a) Calculate the mean and median for each group.
 (b) Based on your answer, which design is best and why?
- 3-108** Table Spice Co. is installing a screener in one stage of its new processing plant to separate leaves, dirt, and insect parts from a certain expensive spice seed that it receives in bulk from growers. The firm can use a coarse 3.5-millimeter mesh screen or a finer 3-millimeter mesh. The smaller mesh will remove more debris but also will remove more seeds. The larger mesh will pass debris and remove fewer seeds. Table Spice has the following information from a sample of pieces of debris.

Debris Size (in millimeters)	Frequency
1.0 or less	12
1.01–1.5	129
1.51–2.0	186
2.01–2.5	275
2.51–3.0	341
3.01–3.5	422
3.51–4.0	6,287
4.01–4.5	8,163
4.51–5.0	6,212
5.01–5.5	2,416
more than 5.5	1,019

- (a) What are the median debris size and the modal class size?
 (b) Which screen would you use based on part (a) if you wanted to remove at least half of the debris?

3-109 The following is the average amount of money each major airline operator spend per passenger on baggage handling:

Airlines	Amount (in Rs '00)
Katar Airlines	3.17
Lusiana Airlines	6.00
Go-Deigo Airlines	2.41
Splice Jet	7.93
Indiana Airlines	5.90
East-West Airlines	1.76
India Konnect Airlines	0.98
Ethos Airlines	6.77
Air Malaya	7.15

What is the mean baggage handling cost per passenger? What is the median baggage handling cost per passenger? A new airline is planning to start operations. Which of the above two average it should consider for planning purpose and why?

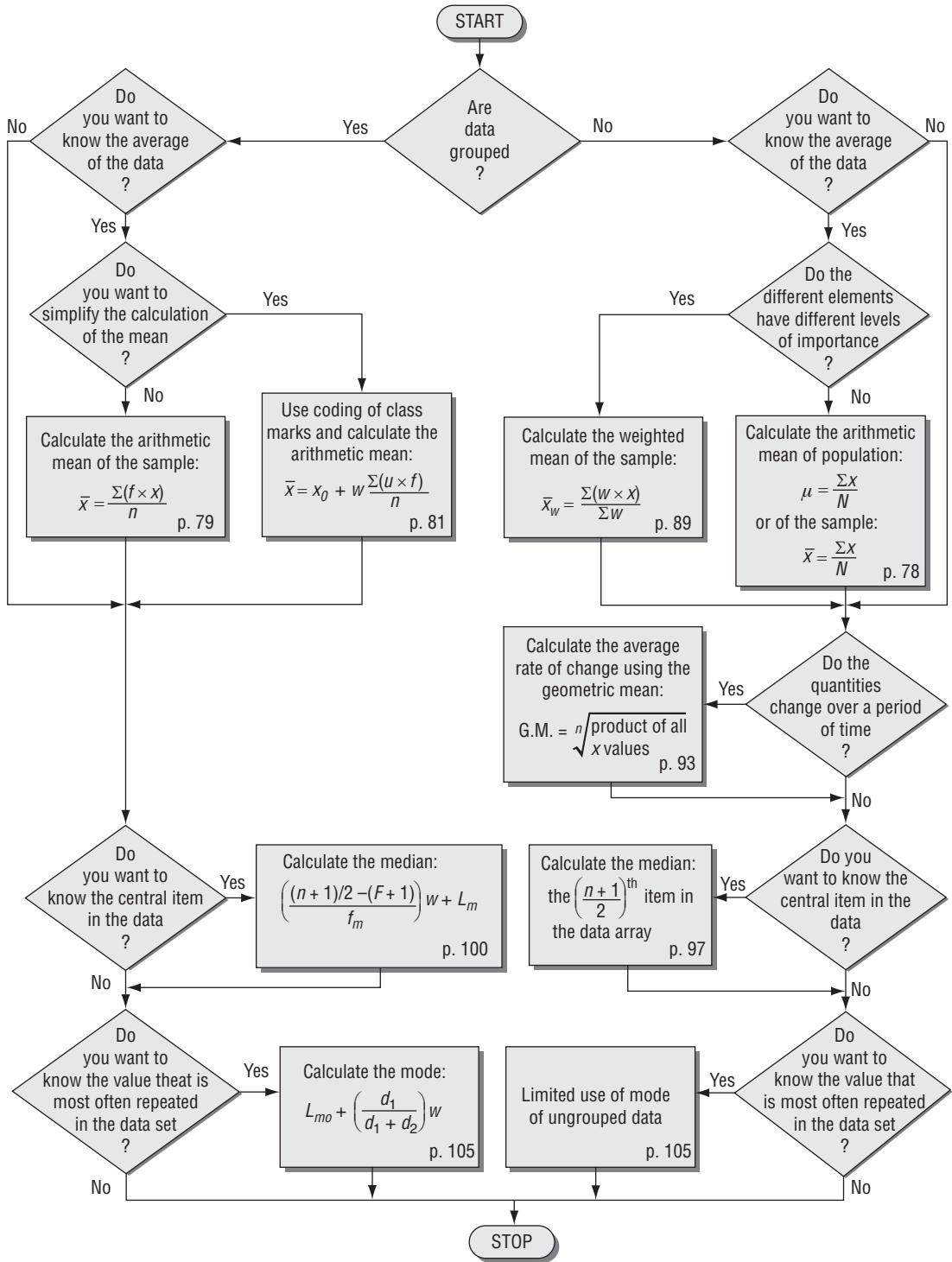


Questions on Running Case: SURYA Bank Pvt. Ltd.

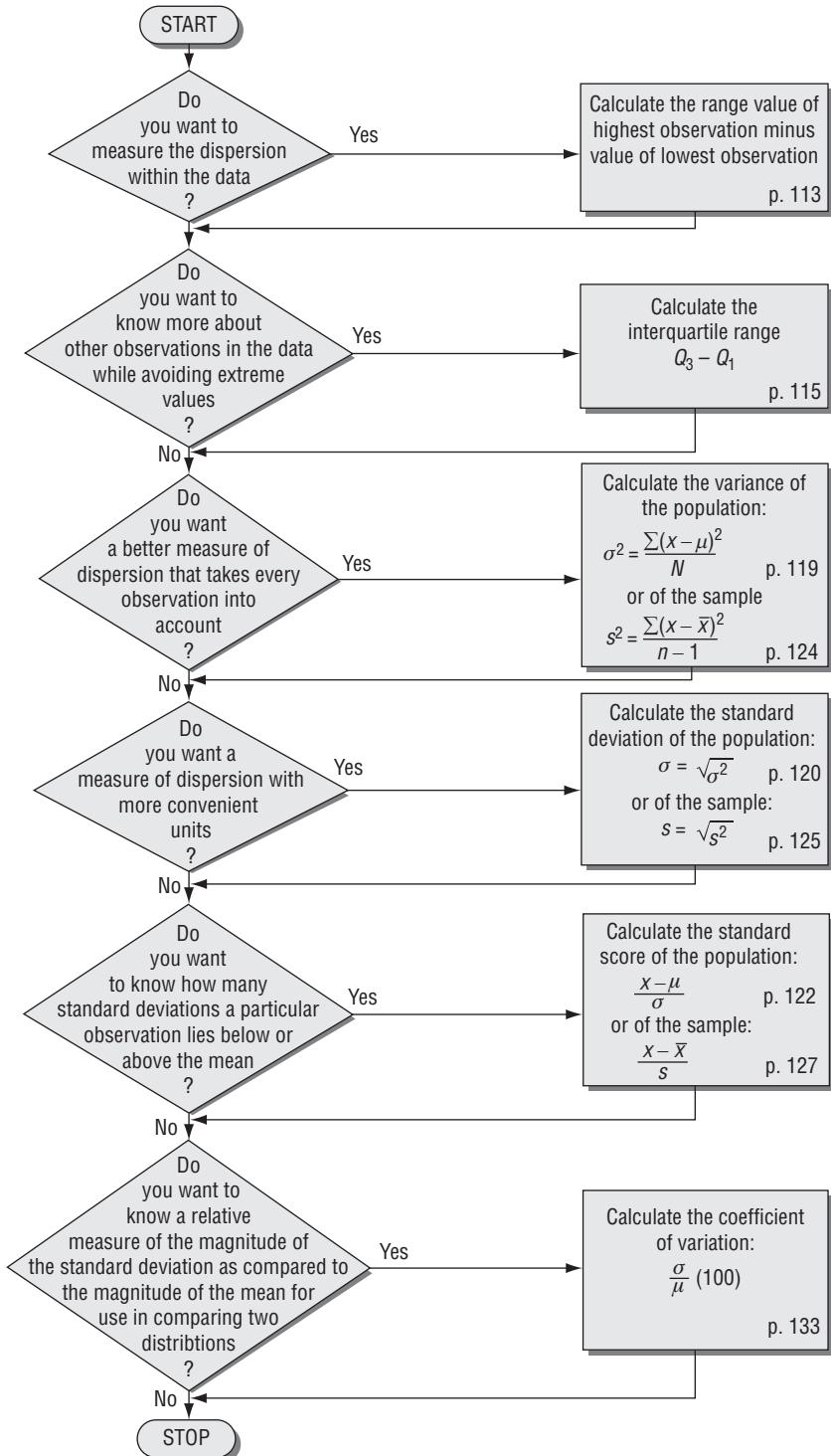
- Calculate the mean, variance, coefficient of skewness and kurtosis for the different modes which help in creating the customer awareness about e-banking. Compare the results of the different modes of creating awareness (Q4).
- Which of the e-banking facilities, on an average, influences the customer most while selecting the bank?(Q7)
- Which facility has the highest variability (Q7).
- Comment on the average satisfaction level of the customers with the e-services provided by their banks. Also calculate the variance and coefficient of skewness of the satisfaction level of the customers (Q12).



Flow Charts: Measures of Central Tendency and Dispersion



Flow Charts: Measures of Central Tendency and Dispersion



4 Probability I: Introductory Ideas

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To examine the use of probability theory in decision making
 - To explain the different ways probabilities arise
 - To develop rules for calculating different kinds of probabilities
 - To use probabilities to take new information into account: the definition and use of Bayes' theorem
-

CHAPTER CONTENTS

- | | |
|--|---|
| 4.1 Probability: The Study of Odds and Ends 154 | ■ Statistics at Work 197 |
| 4.2 Basic Terminology in Probability 155 | ■ Terms Introduced in Chapter 4 199 |
| 4.3 Three Types of Probability 158 | ■ Equations Introduced in Chapter 4 199 |
| 4.4 Probability Rules 165 | ■ Review and Application Exercises 201 |
| 4.5 Probabilities under Conditions of Statistical Independence 171 | ■ Flow Chart: Probability I: Introductory Ideas 208 |
| 4.6 Probabilities under Conditions of Statistical Dependence 179 | |
| 4.7 Revising Prior Estimates of Probabilities: Bayes' Theorem 189 | |

Gamblers have used odds to make bets during most of recorded history. But it wasn't until the seventeenth century that French nobleman Antoine Gombauld (1607–1684) sought a mathematical basis for success at the dice tables. He asked French mathematician Blaise Pascal (1623–1662), "What are the odds of rolling two sixes at least once in twenty-four rolls of a pair of dice?" Pascal solved the problem, having become interested in the idea of probabilities as was Gombauld. They shared their ideas with the famous mathematician Pierre de Fermat (1601–1665), and the letters written by these three constitute the first academic journal in probability theory. We have no record of the degree of success enjoyed by these gentlemen at the dice tables, but we do know that their curiosity and research introduced many of the concepts we shall study in this chapter and the next. ■

4.1 PROBABILITY: THE STUDY OF ODDS AND ENDS

Jacob Bernoulli (1654–1705), Abraham de Moivre (1667–1754), the Reverend Thomas Bayes (1702–1761), and Joseph Lagrange (1736–1813) developed probability formulas and techniques. In the nineteenth century, Pierre Simon, Marquis de Laplace (1749–1827), unified all these early ideas and compiled the first general theory of probability.

Early probability theorists

Probability theory was successfully applied at the gambling tables and, more relevant to our study, eventually to social and economic problems. The insurance industry, which emerged in the nineteenth century, required precise knowledge about the risk of loss in order to calculate premiums. Within 50 years, many learning centers were studying probability as a tool for understanding social phenomena. Today, the mathematical theory of probability is the basis for statistical applications in both social and decision-making research.

Need for probability theory

Probability is a part of our everyday lives. In personal and managerial decisions, we face uncertainty and use probability theory whether or not we admit the use of something so sophisticated. When we hear a weather forecast of a 70 percent chance of rain, we change our plans from a picnic to a pool game. Playing bridge, we make some probability estimate before attempting a finesse. Managers who deal with inventories of highly styled women's clothing must wonder about the chances that sales will reach or exceed a certain level, and the buyer who stocks up on skateboards considers the probability of the life of this particular fad. Before Muhammad Ali's highly publicized fight with Leon Spinks, Ali was reputed to have said, "I'll give you **odds** I'm still the greatest when it's over." And when you begin to study for the inevitable quiz attached to the use of this book, you may ask yourself, "What are the chances the professor will ask us to recall something about the history of probability theory?"

Examples of the use of probability theory

We live in a world in which we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of probability theory. In many instances, we, as concerned citizens, will have some knowledge about the possible outcomes of a decision. By organizing this information and considering it systematically, we will be able to recognize our assumptions, communicate our reasoning to others, and make a sounder decision than we could by using a shot-in-the-dark approach.

EXERCISES 4.1

Applications

- 4-1 The insurance industry uses probability theory to calculate premium rates, but life insurers know for certain that every policyholder is going to die. Does this mean that probability theory does not apply to the life insurance business? Explain.
- 4-2 “Use of this product may be hazardous to your health. This product contains saccharin, which has been determined to cause cancer in laboratory animals.” How might probability theory have played a part in this statement?
- 4-3 Is there really any such thing as an “uncalculated risk”? Explain.
- 4-4 A well-known soft drink company decides to alter the formula of its oldest and most popular product. How might probability theory be involved in such a decision?

4.2 BASIC TERMINOLOGY IN PROBABILITY

In our day-to-day life involving decision-making problems, we encounter two broad types of problems. These problems can be categorized into two types of models: **Deterministic Models** and **Random or Probabilistic Models**. **Deterministic Models** cover those situations, where everything related to the situation is known with certainty to the decision-maker, when decision is to be made. Whereas in **Probabilistic Models**, the totality of the outcomes is known but it can not be certain, which particular outcome will appear. So, there is always some uncertainty involved in decision-making.

In Deterministic Models, frequency distribution or descriptive statistics measures are used to arrive at a decision. Similarly, in random situations, probability and probability distributions are used to make decisions. So, probability can also be defined as a measure of uncertainty.

In general, probability is the chance something will happen. Probabilities are expressed as fractions ($\frac{1}{6}$, $\frac{1}{2}$, $\frac{2}{3}$) or as decimals (0.167, 0.500, 0.889) between zero and 1. Assigning a probability of zero means that something can never happen; a probability of 1 indicates that something will always happen.

In probability theory, an *event* is one or more of the possible outcomes of doing something. If we toss a coin, getting a tail would be an *event*, and getting a head would be another event. Similarly, if we are drawing from a deck of cards, selecting the ace of spades would be an event. An example of an event closer to your life, perhaps, is being picked from a class of 100 students to answer a question. When we hear the frightening predictions of highway traffic deaths, we hope not to be one of those events.

The activity that produces such an event is referred to in probability theory as an *experiment*. Using this formal language, we could ask the question, “In a coin-toss *experiment*, what is the probability of the event *head*?” And, of course, if it is a fair coin with an equal chance of coming down on either side (and no chance of landing on its edge), we would answer “ $\frac{1}{2}$ ” or “0.5.” The set of all possible outcomes of an experiment is called the *sample space* for the experiment. In the coin-toss experiment, the sample space is

$$S = \{\text{head, tail}\}$$

In the card-drawing experiment, the sample space has 52 members: ace of hearts, deuce of hearts, and so on.

Most of us are less excited about coins or cards than we are interested in questions such as “What are the chances of making that plane connection?” or “What are my chances of getting a second job interview?” In short, we are concerned with the chances that an event will happen.

Events are said to be *mutually exclusive* if one and only one of them can take place at a time. Consider again our example of the coin. We have two possible outcomes, heads and tails. On any toss, either heads or tails may turn up, but not both. As a result, the events heads and tails on a single toss are said to be mutually exclusive. Similarly, you will either pass or fail this course or, before the course is over, you may drop it without a grade. Only one of those three outcomes can happen; they are said to be mutually exclusive events. The crucial question to ask in deciding whether events are really mutually exclusive is, “Can two or more of these events occur at one time?” If the answer is yes, the events are *not* mutually exclusive.

Mutually exclusive events

When a list of the possible events that can result from an experiment includes every possible outcome, the list is said to be *collectively exhaustive*. In our coin example, the list “head and tail” is collectively exhaustive (unless, of course, the coin stands on its edge when we toss it). In a presidential campaign, the list of outcomes “Democratic candidate and Republican candidate” is *not* a collectively exhaustive list of outcomes, because an independent candidate or the candidate of another party could conceivably win.

A collectively exhaustive list

Let us consider a situation, total number of possible outcomes related to the situation is “N”, out of them “m” are the number of outcomes where the desired event “E” has occurred. So, “N-m” is the number of outcomes where the desired event has not occurred.

Odd in favor and against

Hence, we may define:

$$\begin{aligned}\text{Odds in favor of happening of } E &= m : N-m \\ \text{Odds against the happening of } E &= N-m : m\end{aligned}$$

This concept is related to the concept of Probability as:

$$\text{Probability of happening of the event } E = m/N$$

Ex: A cricket match is to be played between two teams CX Club and TE Club. A cricket analyst has predicated that the odds in favor of CX Club winning the match are 4:3. This prediction is based upon the historical records and upon the current strengths and weaknesses of the two teams. So, if a cricket fan is interested in knowing the chances that CX will win the match, then the desired chances would be $\frac{4}{7}$.

EXERCISES 4.2**Self-Check Exercises**

SC 4-1 Give a collectively exhaustive list of the possible outcomes of tossing two dice.

SC 4-2 Give the probability for each of the following totals in the rolling of two dice: 1, 2, 5, 6, 7, 10, and 11.

Basic Concepts

- 4-5** Which of the following are pairs of mutually exclusive events in the drawing of one card from a standard deck of 52?
- A heart and a queen.
 - A club and a red card.

- (c) An even number and a spade.
 (d) An ace and an even number.

Which of the following are mutually exclusive outcomes in the rolling of two dice?

- (a) A total of 5 points and a 5 on one die.
 (b) A total of 7 points and an even number of points on both dice.
 (c) A total of 8 points and an odd number of points on both dice.
 (d) A total of 9 points and a 2 on one die.
 (e) A total of 10 points and a 4 on one die.

- 4-6** A batter “takes” (does not swing at) each of the pitches he sees. Give the sample space of outcomes for the following experiments in terms of balls and strikes:
- (a) Two pitches.
 (b) Three pitches.

Applications

- 4-7** Consider a stack of nine cards, all spades, numbered 2 through 10, and a die. Give a collectively exhaustive list of the possible outcomes of rolling the die and picking one card. How many elements are there in the sample space?
- 4-8** Consider the stack of cards and the die discussed in Exercise 4-7. Give the probability for each of the following totals in the sum of the roll of the die and the value of the card drawn:

2 3 8 9 12 14 16

- 4-9** In a recent meeting of union members supporting Joe Royal for union president, Royal’s leading supporter said “chances are good” that Royal will defeat the single opponent facing him in the election.
- (a) What are the “events” that could take place with regard to the election?
 (b) Is your list collectively exhaustive? Are the events in your list mutually exclusive?
 (c) Disregarding the supporter’s comments and knowing no additional information, what probabilities would you assign to each of your events?
- 4-10** Southern Bell is considering the distribution of funds for a campaign to increase long-distance calls within North Carolina. The following table lists the markets that the company considers worthy of focused promotions:

Market Segment	Cost of Special Campaign Aimed at Group
Minorities	\$350,000
Businesspeople	\$550,000
Women	\$250,000
Professionals and white-collar workers	\$200,000
Blue-collar workers	\$250,000

There is up to \$800,000 available for these special campaigns.

- (a) Are the market segments listed in the table collectively exhaustive? Are they mutually exclusive?
 (b) Make a collectively exhaustive and mutually exclusive list of the possible events of the spending decision.
 (c) Suppose the company has decided to spend the entire \$800,000 on special campaigns. Does this change your answer to part (b)? If so, what is your new answer?

Worked-Out Answers to Self-Check Exercises

SC 4-1 (Die 1, Die 2)

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

SC 4-2 $P(1) = 0/36$, $P(2) = 1/36$, $P(5) = 4/36$, $P(6) = 5/36$, $P(7) = 6/36$, $P(10) = 3/36$, $P(11) = 2/36$.

4.3 THREE TYPES OF PROBABILITY

There are three basic ways of classifying probability. These three represent rather different conceptual approaches to the study of probability theory; in fact, experts disagree about which approach is the proper one to use. Let us begin by defining the

1. Classical approach
2. Relative frequency approach
3. Subjective approach

Classical Probability

Classical probability defines the probability that an event will occur as

Classical probability defined

Probability of an Event

$$\text{Probability of an event} = \frac{\text{number of outcomes where the event occurs}}{\text{total number of possible outcomes}} \quad [4-1]$$

It must be emphasized that in order for Eq. 4-1 to be valid, each of the possible outcomes must be equally likely. This is a rather complex way of defining something that may seem intuitively obvious to us, but we can use it to write our coin-toss and dice-rolling examples in symbolic form. First, we would state the question, “What is the probability of getting a head on one toss?” as

$$P(\text{Head})$$

Then, using formal terms, we get

$$P(\text{Head}) = \frac{1}{1+1} = \frac{1}{2}$$

Number of outcomes of one toss where the event occurs
(in this case, the number hat will produce a head)

Total number of possible outcomes
of one toss (a head or a tail)

And for the dice-rolling example:

$$P(5) = \frac{1}{1+1+1+1+1+1} = \frac{1}{6}$$

Number of outcomes of one roll
of the die that will produce a 5

Total number of possible outcomes of one
roll of the die (getting a 1, a 2, a 3, a 4, a 5,
or a 6)

Classical probability is often called *a priori* probability because if we keep using orderly examples such as fair coins, unbiased dice, and standard decks of cards, we can state the answer in advance (*a priori*) without tossing a coin, rolling a die, or drawing a card. We do not have to perform experiments to make our probability statements about fair coins, standard card decks, and unbiased dice. Instead, we can make statements based on logical reasoning before any experiments take place.

A priori probability

This approach assumes a number of assumptions, in defining the probability. So, if those assumptions are included then the complete definition should be: Probability of an event may be defined as the ratio of number of outcomes where the event occurs (favorable outcomes) to the total number of possible outcomes, provided these outcomes are equally likely (the chances of happening of all outcomes are equal), exhaustive (the totality of all outcomes are known and defined) and mutually exclusive (happening of one outcome results in non-happening of others). If these assumptions related to the outcomes are not followed, then this approach can not be applied in determining the probability.

Shortcomings of the classical approach

This approach to probability is useful when we deal with card games, dice games, coin tosses, and the like, but has serious problems when we try to apply it to the less orderly decision problems we encounter in management. The classical approach to probability assumes a world that does not exist. It assumes away situations that are very unlikely but that could conceivably happen. Such occurrences as a coin landing on its edge, your classroom burning down during a discussion of probabilities, and your eating pizza while on a business trip at the North Pole are all extremely unlikely but not impossible. Nevertheless, the classical approach assumes them all away. Classical probability also assumes a kind of symmetry about the world, and that assumption can get us into trouble. Real-life situations, disorderly and unlikely as they often are, make it useful to define probabilities in other ways.

Relative Frequency of Occurrence

Suppose we begin asking ourselves complex questions such as, “What is the probability that I will live to be 85?” or “What are the chances that I will blow one of my stereo speakers if I turn my 200-watt amplifier up to wide open?” or “What is the probability that the location of a new paper plant on the river near our town will cause a substantial fish kill?” We quickly see that we may not be able to state in advance, without experimentation, what these probabilities are. Other approaches may be more useful.

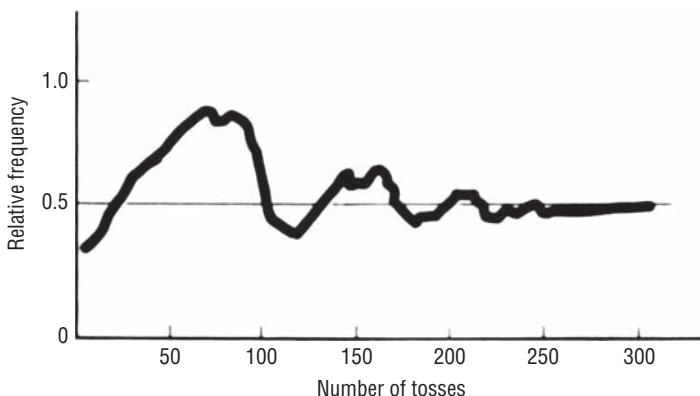
In the 1800s, British statisticians, interested in a theoretical foundation for calculating risk of losses in life insurance and commercial insurance, began defining probabilities from statistical data collected on births and deaths. Today, this approach is called the *relative frequency of occurrence*. It defines probability as either:

Probability redefined

1. The observed relative frequency of an event in a very large number of trials, or
2. The proportion of times that an event occurs in the long run when conditions are stable.

This method uses the relative frequencies of past occurrences as probabilities. We determine how often something has happened in the past and use that figure to predict the probability that it will happen again in the future. Let us look at an example. Suppose an insurance company knows from past actuarial data that of all males 40 years old, about 60 out of every 100,000 will die within a 1-year period. Using this method, the company estimates the probability of

Using the relative frequency of occurrence approach

**FIGURE 4-1 RELATIVE FREQUENCY OF OCCURRENCE OF HEADS IN 300 TOSSES OF A FAIR COIN**

death for that age group as

$$\frac{60}{100,000}, \text{ or } 0.0006$$

A second characteristic of probabilities established by the relative frequency of occurrence method can be shown by tossing one of our fair coins 300 times. Figure 4-1 illustrates the outcomes of these 300 tosses. Here we can see that although the proportion of heads was far from 0.5 in the first 100 tosses, it seemed to stabilize and approach 0.5 as the number of tosses increased. In statistical language, we would say that the relative frequency becomes stable as the number of tosses becomes large (if we are tossing the coin under uniform conditions). Thus, when we use the relative frequency approach to establish probabilities, our probability figure will gain accuracy as we increase the number of observations. Of course, this improved accuracy is not free; although more tosses of our coin will produce a more accurate probability of heads occurring, we must bear the time and the cost of additional observations.

More trials, greater accuracy

Suppose an event is capable of being repeated sufficiently large number of times "N", and the frequency of the desired outcome is "f". Then relative frequency of the outcome is " $\frac{f}{N}$ ". The limiting value of the relative frequency can be used to define probability of the outcome.

One difficulty with the relative frequency approach is that people often use it without evaluating a sufficient number of outcomes. If you heard someone say, "My aunt and uncle got the flu this year, and they are both over 65, so everyone in that age bracket will probably get the flu," you would know that your friend did not base his assumptions on enough evidence. His observations were insufficient data for establishing a relative frequency of occurrence probability.

A limitation of relative frequency

This approach of defining probability is better than the Classical Approach, as it is not based on assumptions of mutually exclusive, equally likely and exhaustive. The drawback of using this approach is that it requires the event to be capable of being repeated large number of times. Moreover, one can not be certain that after how many occurrences, the relative frequency may stabilize. In the real and business world, we have to take decisions on those events which occur only once or not so frequent and the environmental conditions related to the situation might change. These factors restrict the use of this approach in real life decision making.

But what about a different kind of estimate, one that seems not to be based on statistics at all? Suppose your school's basketball team lost the first 10 games of the year. You were a loyal fan, however, and bet \$100 that your team would beat Indiana's in the eleventh game. To everyone's surprise, you won your bet. We would have difficulty convincing you that you were statistically incorrect. And you would be right to be skeptical about our argument. Perhaps, without knowing that you did so, you may have based your bet on the statistical foundation described in the next approach to establishing probabilities.

Subjective Probabilities

The relative frequency approach can't deal with specific or unique situations, which are typical of the business or management world. **Subjective probability defined**

So, the probability approach dealing with such unique situations should be based upon some belief or educated guess of the decision maker.

Subjective probabilities are based on the beliefs of the person making the probability assessment. In fact, subjective probability can be defined as the probability assigned to an event by an individual, based on whatever evidence is available. This evidence may be in the form of relative frequency of past occurrences, or it may be just an educated guess. Probably the earliest subjective probability estimate of the likelihood of rain occurred when someone's Aunt Bess said, "My corns hurt; I think we're in for a downpour." Subjective assessments of probability permit the widest flexibility of the three concepts we have discussed. The decision maker can use whatever evidence is available and temper this with personal feelings about the situation.

Subjective probability assignments are often found when events occur only once or at most a very few times. Say that it is your job to interview and select a new social services caseworker. You have narrowed your choice to three people. Each has an attractive appearance, a high level of energy, abounding self-confidence, a record of past accomplishments, and a state of mind that seems to welcome challenges. What are the chances each will relate to clients successfully? Answering this question and choosing among the three will require you to assign a subjective probability to each person's potential.

Here is one more illustration of this kind of probability assignment. A judge is deciding whether to allow the construction **Using the subjective approach** of a nuclear power plant on a site where there is some evidence of a geological fault. He must ask himself, "What is the probability of a major nuclear accident at this location?" The fact that there is no relative frequency of occurrence evidence of previous accidents at this location does not excuse him from making a decision. He must use his best judgment in trying to determine the subjective probabilities of a nuclear accident.

Because most higher-level social and managerial decisions are concerned with specific, unique situations, rather than with a long series of identical situations, decision makers at this level make considerable use of subjective probabilities.

The subjective approach to assigning probabilities was introduced in 1926 by Frank Ramsey in his book *The Foundation of Mathematics and Other Logical Essays*. The concept was further developed by Bernard Koopman, Richard Good, and Leonard Savage, names that appeared regularly in advanced work in this field. Professor Savage pointed out that two reasonable people faced with the same evidence could easily come up with quite different subjective probabilities for the same event. The two people who made opposing bets on the outcome of the Indiana basketball game would understand quite well what he meant.

HINTS & ASSUMPTIONS

Warning: In classical probability problems, be sure to check whether the situation is “with replacement” after each draw or “without replacement.” The chance of drawing an ace from a 52-card deck on the first draw is $\frac{4}{52}$, or about .077. If you draw one and it is replaced, the odds of drawing an ace on the second draw are the same, $\frac{4}{52}$. However, without replacement, the odds change to $\frac{3}{51}$ if the first card was *not* an ace, or to $\frac{3}{51}$ if the first card *was* an ace. In assigning subjective probabilities, it’s normal for two different people to come up with different probabilities for the same event; that’s the result of experience and time (we often call this combination “wisdom”). In assigning probabilities using the relative frequency of occurrence method, be sure you have observed an adequate number of outcomes. Just because red hasn’t come up in 9 spins of the roulette wheel, you shouldn’t bet next semester’s tuition on black this spin!

EXERCISES 4.3**Self-Check Exercises**

- SC 4-3** Union shop steward B. Lou Khollar has drafted a set of wage and benefit demands to be presented to management. To get an idea of worker support for the package, he randomly polls the two largest groups of workers at his plant, the machinists (M) and the inspectors (I). He polls 30 of each group with the following results:

Opinion of Package	M	I
Strongly support	9	10
Mildly support	11	3
Undecided	2	2
Mildly oppose	4	8
Strongly oppose	<u>4</u>	<u>7</u>
	<u>30</u>	<u>30</u>

- (a) What is the probability that a machinist randomly selected from the polled group mildly supports the package?
- (b) What is the probability that an inspector randomly selected from the polled group is undecided about the package?
- (c) What is the probability that a worker (machinist or inspector) randomly selected from the polled group strongly or mildly supports the package?
- (d) What types of probability estimates are these?

- SC 4-4** Classify the following probability estimates as to their type (classical, relative frequency, or subjective):

- (a) The probability of scoring on a penalty shot in ice hockey is 0.47.
- (b) The probability that the current mayor will resign is 0.85.
- (c) The probability of rolling two sixes with two dice is $\frac{1}{36}$.
- (d) The probability that a president elected in a year ending in zero will die in office is $\frac{1}{10}$.
- (e) The probability that you will go to Europe this year is 0.14.

Basic Concepts

- 4-11** Determine the probabilities of the following events in drawing a card from a standard deck of 52 cards:
- A seven.
 - A black card.
 - An ace or a king.
 - A black two or a black three.
 - A red face card (king, queen, or jack).
- What type of probability estimates are these?
- 4-12** During a recent bridge game, once the lead card had been played and the dummy's hand revealed, the declarer took a moment to count up the number of cards in each suit with the results given below:

Suit	We	They
Spades	6	7
Hearts	8	5
Diamonds	4	9
Clubs	8	5
	<u>26</u>	<u>26</u>

- What is the probability that a card randomly selected from the We team's hand is a spade?
- What is the probability that a card randomly selected from the They team's hand is a club?
- What is the probability that a card randomly selected from all the cards is either a spade or heart?
- If this type of analysis were repeated for every hand many times, what would be the long-run probability that a card drawn from the We team's hand is a spade?

Applications

- 4-13** Below is a frequency distribution of annual sales commissions from a survey of 300 media salespeople.

Annual Commission	Frequency
\$ 0–4,999	15
5,000–9,999	25
10,000–14,999	35
15,000–19,999	125
20,000–24,999	70
25,000+	30

Based on this information, what is the probability that a media salesperson makes a commission: (a) between \$5,000 and \$10,000, (b) less than \$15,000, (c) more than \$20,000, and (d) between \$15,000 and \$20,000.

- 4-14** General Buck Turgidson is preparing to make his annual budget presentation to the U.S. Senate and is speculating about his chances of getting all or part of his requested budget

approved. From his 20 years of experience in making these requests, he has deduced that his chances of getting between 50 and 74 percent of his budget approved are twice as good as those of getting between 75 and 99 percent approved, and two and one-half times as good as those of getting between 25 and 49 percent approved. Further, the general believes that there is no chance of less than 25 percent of his budget being approved. Finally, the entire budget has been approved only once during the general's tenure, and the general does not expect this pattern to change. What are the probabilities of 0–24 percent, 25–49 percent, 50–74 percent, 75–99 percent, and 100 percent approval, according to the general?

- 4-15** The office manager of an insurance company has the following data on the functioning of the copiers in the office:

Copier	Days Functioning	Days Out of Service
1	209	51
2	217	43
3	258	2
4	229	31
5	247	13

What is the probability of a copier being out of service based on these data?

- 4-16** Classify the following probability estimates as classical, relative frequency, or subjective:
- The probability the Cubs will win the World Series this year is 0.175.
 - The probability tuition will increase next year is 0.95.
 - The probability that you will win the lottery is 0.00062.
 - The probability a randomly selected flight will arrive on time is 0.875.
 - The probability of tossing a coin twice and observing two heads is 0.25.
 - The probability that your car will start on a very cold day is 0.97.

Worked-Out Answers to Self-Check Exercises

SC 4-3 (a) $P(\text{Machinist mildly supports}) = \frac{\text{number of machinists in "mildly support" class}}{\text{total number of machinists polled}} = 11/30$

(b) $P(\text{Inspector undecided}) = \frac{\text{number of inspectors in "undecided" class}}{\text{total number of inspectors polled}} = 2/30 = 1/15$

Opinion	Frequency (combined)
SS	19
MS	14
U	4
MO	12
SO	11
	60

$$P(\text{Strongly or mildly support}) = (19 + 14)/60 = 33/60 = 11/20$$

- (d) Relative frequency.
SC 4-4 (a) Relative frequency. (b) Subjective. (c) Classical.
(d) Relative frequency. (e) Subjective.

4.4 PROBABILITY RULES

Most managers who use probabilities are concerned with two conditions:

1. The case where one event *or* another will occur
2. The situation where two or more events will *both* occur

We are interested in the first case when we ask, “What is the probability that today’s demand will exceed our inventory?” To illustrate the second situation, we could ask, “What is the probability that today’s demand will exceed our inventory *and* that more than 10 percent of our sales force will not report for work?” In the sections to follow, we shall illustrate methods of determining answers to questions such as these under a variety of conditions.

Some Commonly Used Symbols, Definitions, and Rules

Symbol for a Marginal Probability In probability theory, we use symbols to simplify the presentation of ideas. As we discussed earlier in this chapter, the probability of the event A is expressed as

Probability of Event A Happening				
$P(A)$	= the	probability	of	event A happening

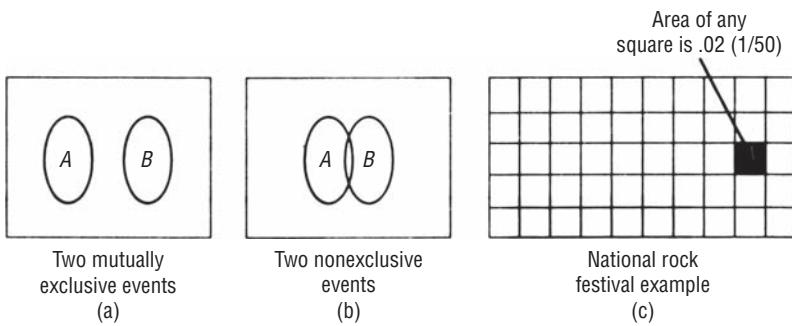
A *single* probability means that only one event can take place. It is called a *marginal* or *unconditional probability*. To illustrate, let us suppose that 50 members of a school class drew tickets to see which student would get a free trip to the National Rock Festival. Any one of the students could calculate his or her chances of winning as:

$$\begin{aligned} P(\text{Winning}) &= \frac{1}{50} \\ &= 0.02 \end{aligned}$$

In this case, a student’s chance is 1 in 50 because we are certain that the possible events are mutually exclusive, that is, only one student can win at a time.

There is a nice diagrammatic way to illustrate this example and other probability concepts. We use a pictorial representation called a *Venn diagram*, after the nineteenth-century English mathematician John Venn. In these diagrams, the entire sample space is represented by a rectangle, and events are represented by parts of the rectangle. If two events are *mutually exclusive*, their parts of the rectangle will not overlap each other, as shown in Figure 4-2(a). If two events are *not* mutually exclusive, their parts of the rectangle *will* overlap, as in Figure 4-2(b).

Because probabilities behave a lot like areas, we shall let the rectangle have an area of 1 (because the probability of *something* happening is 1). Then the probability of an event is the area of *its* part of the rectangle. Figure 4-2(c) illustrates this for the National Rock Festival example. There the rectangle is divided into 50 equal, nonoverlapping parts.

**FIGURE 4-2 SOME VENN DIAGRAMS**

Addition Rule of Probabilistic Events If two events are not mutually exclusive, it is possible for both events to occur. In these cases, our addition rule must be modified. For example, what is the probability of drawing either an ace or a heart from a deck of cards? Obviously, the events ace and heart can occur together because we could draw the ace of hearts. Thus, ace and heart are not mutually exclusive events. We must adjust our Equation 4-3 to avoid double counting, that is, we have to *reduce* the probability of drawing either an ace or a heart by the chance that we could draw both of them together. As a result, the correct equation for the probability of one or more of two events that are not mutually exclusive is

Probability of one or more events not mutually exclusive

Addition Rule of Probabilistic Events

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

Probability of A happening Probability of A and B happening together
 Probability of A or B happening when A and B are not mutually exclusive Probability of B happening

[4-2]

A Venn diagram illustrating Equation 4-2 is given in Figure 4-3. There, the event A or B is outlined with a heavy line. The event A and B is the cross-hatched wedge in the middle. If we add the areas of circles A and B , we *double count* the area of the wedge, and so we must subtract it to make sure it is counted only once.

Using Equation 4-2 to determine the probability of drawing either an ace or a heart, we can calculate:

$$\begin{aligned} P(\text{Ace or Heart}) &= P(\text{Ace}) + P(\text{Heart}) - P(\text{Ace and Heart}) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} \\ &= \frac{16}{52} \text{ or } \frac{4}{13} \end{aligned}$$

Let's do a second example. The employees of a certain company have elected five of their number to represent them on the employee-management productivity council. Profiles of the five are as follows:

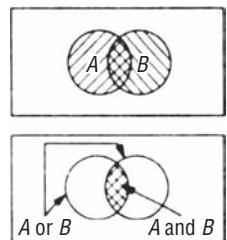


FIGURE 4-3 VENN DIAGRAM FOR THE ADDITION RULE FOR TWO EVENTS NOT MUTUALLY EXCLUSIVE

1. male	age 30
2. male	32
3. female	45
4. female	20
5. male	40

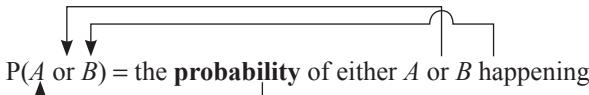
This group decides to elect a spokesperson by drawing a name from a hat. Our question is, “What is the probability the spokesperson will be *either* female *or* over 35?” Using Equation 4-2, we can set up the solution to our question like this:

$$\begin{aligned} P(\text{Female or Over 35}) &= P(\text{Female}) + P(\text{Over 35}) - P(\text{Female and Over 35}) \\ &= \frac{2}{5} + \frac{2}{5} - \frac{1}{5} \\ &= \frac{3}{5} \end{aligned}$$

We can check our work by inspection and see that of the five people in the group, three would fit the requirements of being either female or over 35.

Addition Rule for Mutually Exclusive Events Often, however, we are interested in the probability that one thing *or* another will occur. If these two events are mutually exclusive, we can express this probability using the addition rule for mutually exclusive events. This rule is expressed symbolically as

Probability of one or more mutually exclusive events



and is calculated as follows:

Probability of Either A or B Happening

$$P(A \text{ or } B) = P(A) + P(B)$$

[4-3]

This addition rule is illustrated by the Venn diagram in Figure 4-4, where we note that the area in the two circles together (denoting the event *A* or *B*) is the sum of the areas of the circle denoting the event *A* and the circle denoting the event *B*.

Now to use this formula in an example. Five equally capable students are waiting for a summer job interview with a company that has announced that it will hire only one of the five by random drawing. The group consists of Bill, Helen, John, Sally, and Walter. If our question is, “What is the probability that John will be the candidate?” we can use Equation 4-1 and give the answer.

$$\begin{aligned} P(\text{John}) &= \frac{1}{5} \\ &= 0.02 \end{aligned}$$

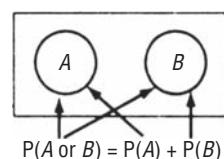


FIGURE 4-4 VENN DIAGRAM FOR THE ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

However, if we ask, “What is the probability that either John or Sally will be the candidate?” we would use Equation 4-3:

$$\begin{aligned} P(\text{John or Sally}) &= P(\text{John}) + P(\text{Sally}) \\ &= \frac{1}{5} + \frac{1}{5} \\ &= \frac{2}{5} \\ &= 0.4 \end{aligned}$$

Let’s calculate the probability of two or more events happening once more. Table 4-1 contains data on the sizes of families in a certain town. We are interested in the question, “What is the probability that a family chosen at random from this town will have four or more children (that is, four, five, six or more children)?” Using Equation 4-3, we can calculate the answer as

$$\begin{aligned} P(4, 5, 6 \text{ or more}) &= P(4) + P(5) + P(6 \text{ or more}) \\ &= 0.15 + 0.10 + 0.05 \\ &= 0.30 \end{aligned}$$

There is an important special case of Equation 4-3. For any event A , either A happens or it doesn’t. So the events A and $\text{not } A$ are exclusive and exhaustive. Applying Equation 4-3 yields the result

$$P(A) + P(\text{not } A) = 1$$

or, equivalently,

$$P(A) = 1 - P(\text{not } A)$$

A special case of Equation 4-3

For example, referring back to Table 4-1, the probability of a family’s having five or fewer children is most easily obtained by subtracting from 1 the probability of the family’s having six or more children, and thus is seen to be 0.95.

TABLE 4.1 FAMILY-SIZE DATA

NUMBER OF CHILDREN	0	1	2	3	4	5	6 or more
PROPORTION OF FAMILIES HAVING THIS MANY CHILDREN	0.05	0.10	0.30	0.25	0.15	0.10	0.05

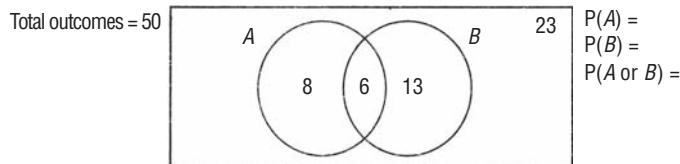
HINTS & ASSUMPTIONS

John Venn’s diagrams are a useful way to avoid errors when you apply the addition rule for events that are and are not mutually exclusive. The most common error here is double counting. Hint: In applying the addition rule for mutually exclusive events, we’re looking for a probability of one event or another and overlap is *not* a problem. However, with non-mutually exclusive events, both can occur together and we need to reduce our probability by the chance that they *could*. Thus, we subtract the overlap or cross-hatched area in the Venn diagram to get the correct value.

EXERCISES 4.4

Self-Check Exercises

- SC 4-5** From the following Venn diagram, which indicates the number of outcomes of an experiment corresponding to each event and the number of outcomes that do not correspond to either event, give the probabilities indicated.



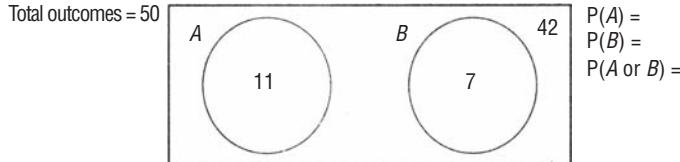
- SC 4-6** An inspector of the Alaska Pipeline has the task of comparing the reliability of two pumping stations. Each station is susceptible to two kinds of failure: pump failure and leakage. When either (or both) occur, the station must be shut down. The data at hand indicate that the following probabilities prevail:

Station	$P(\text{Pump Failure})$	$P(\text{Leakage})$	$P(\text{Both})$
1	0.07	0.10	0
2	0.09	0.12	0.06

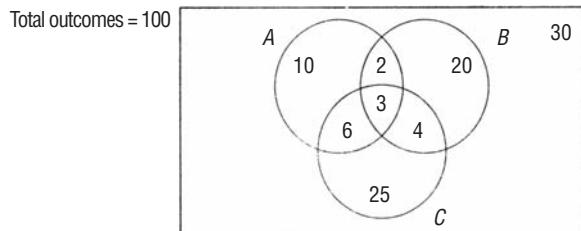
Which station has the higher probability of being shut down?

Basic Concepts

- 4-17** From the following Venn diagram, which indicates the number of outcomes of an experiment corresponding to each event and the number of outcomes that do not correspond to either event, give the probabilities indicated:



- 4-18** Using this Venn diagram, give the probabilities indicated:



$$\begin{array}{lll} P(A) = & P(B) = & P(C) = \\ P(A \text{ or } B) = & P(A \text{ or } C) = & P(B \text{ but not } (A \text{ or } C)) \end{array}$$

- 4-19** In this section, two expressions were developed for the probability of either of two events, A or B , occurring. Referring to Equations 4-2 and 4-3:
- What can you say about the probability of A and B occurring simultaneously when A and B are *mutually exclusive*?
 - Develop an expression for the probability that at least one of three events, A , B , or C , could occur, that is, $P(A \text{ or } B \text{ or } C)$. Do not assume that A , B , and C are mutually exclusive of each other.
 - Rewrite your expression for the case in which A and B are mutually exclusive, but A and C and B and C are not mutually exclusive.
 - Rewrite your expression for the case in which A and B and A and C are mutually exclusive, but not B and C .
 - Rewrite your expression for the case in which A , B , and C are mutually exclusive of the others.

Applications

- 4.20** An employee at Infotech must enter product information into the computer. The employee may use a light pen that transmits the information to the PC along with the keyboard to issue commands, or fill out a bubble sheet and feed it directly into the old mainframe. Historically, we know the following probabilities:

$$P(\text{Light pen will fail}) = 0.025$$

$$P(\text{PC keyboard will fail}) = 0.15$$

$$P(\text{Light pen and PC keyboard will fail}) = 0.005$$

$$P(\text{Mainframe will fail}) = 0.25$$

Data can be entered into the PC only if both the light pen and keyboard are functioning.

- What is the probability that the employee can use the PC to enter data?
- What is the probability that either the PC fails or the mainframe fails? Assume they cannot both fail at the same time.

- 4-21** The HAL Corporation wishes to improve the resistance of its personal computer to disk-drive and keyboard failures. At present, the design of the computer is such that disk-drive failures occur only one-third as often as keyboard failures. The probability of simultaneous disk-drive and keyboard failures is 0.05.

- If the computer is 80 percent resistant to disk-drive and/or keyboard failure, how low must the disk-drive failure probability be?
- If the keyboard is improved so that it fails only twice as often as the disk-drive (and the simultaneous failure probability is still 0.05), will the disk-drive failure probability from part (a) yield a resistance to disk-drive and/or keyboard failure higher or lower than 90 percent?

- 4-22** The Herr-McFee Company, which produces nuclear fuel rods, must X-ray and inspect each rod before shipping. Karen Wood, an inspector, has noted that for every 1,000 fuel rods she inspects, 10 have interior flaws, 8 have casing flaws, and 5 have both flaws. In her quarterly report, Karen must include the probability of flaws in fuel rods. What is this probability?

Worked-Out Answers to Self-Check Exercises

SC 4-5 $P(A) = 14/50 = 0.28$ $P(B) = 19/50 = 0.38$

$$P(A \text{ or } B) = \frac{14}{50} + \frac{19}{50} - \frac{6}{50} = 0.54$$

SC 4-6 $P(\text{Failure}) = P(\text{Pump failure or leakage})$

$$\text{Station 1: } 0.07 + 0.1 - 0 = 0.17 \quad \text{Station 2: } 0.09 + 0.12 - 0.06 = 0.15$$

Thus, Station 1 has the higher probability of being shut down.

4.5 PROBABILITIES UNDER CONDITIONS OF STATISTICAL INDEPENDENCE

When two events happen, the outcome of the first event may or may not have an effect on the outcome of the second event. That is, the events may be either dependent or independent. In this section, we examine events that are *statistically independent*: The occurrence of one event *has no effect* on the probability of the occurrence of any other event. There are three types of probabilities under statistical independence:

Independence defined

1. Marginal
2. Joint
3. Conditional

Marginal Probabilities under Statistical Independence

As we explained previously, a marginal or unconditional probability is the simple probability of the occurrence of an event. In a fair coin toss, $P(H) = 0.5$, and $P(T) = 0.5$; that is, the probability of heads equals 0.5 and the probability of tails equals 0.5. This is true for every toss, no matter how many tosses have been made or what their outcomes have been. Every toss stands alone and is in no way connected with any other toss. Thus, the outcome of *each* toss of a fair coin is an event that is statistically independent of the outcomes of *every other* toss of the coin.

Marginal probability of independent events

Imagine that we have a biased or unfair coin that has been altered in such a way that heads occurs 0.90 of the time and tails 0.10 of the time. On each individual toss, $P(H) = 0.90$, and $P(T) = 0.10$. The outcome of any particular toss is completely unrelated to the outcomes of the tosses that may precede or follow it. The outcomes of several tosses of *this* coin are statistically independent events too, even though the coin is biased.

Joint Probabilities under Statistical Independence

The probability of two or more independent events occurring together or in succession is the product of their marginal probabilities. Mathematically, this is stated (for two events):

Multiplication rule for joint, independent events

Joint Probability of Two Independent Events

$$P(AB) = P(A) \times P(B)$$

[4-4]

where

- $P(AB)$ = probability of events A and B occurring together or in succession; this is known as a *joint probability*
- $P(A)$ = marginal probability of event A occurring
- $P(B)$ = marginal probability of event B occurring

In terms of the fair coin example, the probability of heads on two successive tosses is the probability of heads on the first toss (which we shall call H_1) times the probability of heads on the second toss (H_2). That is, $P(H_1H_2) = P(H_1) \times P(H_2)$. We have shown that the events are statistically independent, because the probability of any outcome is not affected by any preceding outcome. Therefore, the probability of heads on any toss is 0.5, and $P(H_1H_2) = 0.5 \times 0.5 = 0.25$. Thus, the probability of heads on two successive tosses is 0.25.

The fair coin example

Likewise, the probability of getting three heads on three successive tosses is $P(H_1H_2H_3) = 0.5 \times 0.5 \times 0.5 = 0.125$.

Assume next that we are going to toss an unfair coin that has $P(H) = 0.8$ and $P(T) = 0.2$. The events (outcomes) are independent, because the probabilities of all tosses are exactly the same—the individual tosses are completely separate and in no way affected by any other toss or outcome. Suppose our question is, “What is the probability of getting three heads on three successive tosses?” We use Equation 4-4 and discover that:

$$P(H_1H_2H_3) = P(H_1) \times P(H_2) \times P(H_3) = 0.8 \times 0.8 \times 0.8 = 0.512$$

Now let us ask the probability of getting three tails on three successive tosses:

$$P(T_1T_2T_3) = P(T_1) \times P(T_2) \times P(T_3) = 0.2 \times 0.2 \times 0.2 = 0.008$$

Note that these two probabilities do not add up to 1 because the events $H_1H_2H_3$ and $T_1T_2T_3$ do not constitute a collectively exhaustive list. They are mutually exclusive, because if one occurs, the other cannot.

We can make the probabilities of events even more explicit using a *probability tree*. Figure 4-5 is a probability tree showing

Constructing a probability tree

the possible outcomes and their respective probabilities for one toss of a fair coin.

For toss 1, we have two possible outcomes, heads and tails, each with a probability of 0.5. Assume that the outcome of toss 1 is heads. We toss again. The second toss has two possible outcomes, heads and tails, each with a probability of 0.5. In Figure 4-6, we add these two branches of the tree.

Next we consider the possibility that the outcome of toss 1 is tails. Then the second toss must stem from

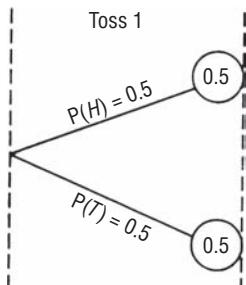
One toss, two possible outcomes

FIGURE 4-5 PROBABILITY TREE OF ONE TOSS

Two tosses, four possible outcomes

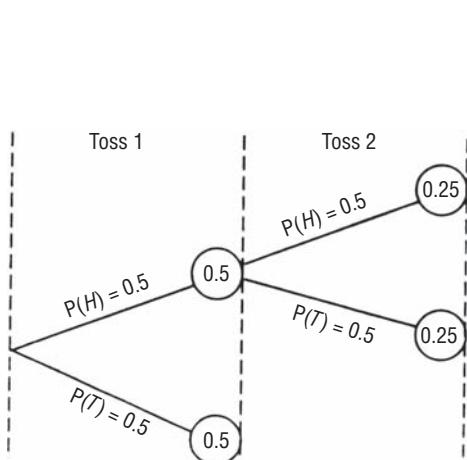


FIGURE 4-6 PROBABILITY TREE OF A PARTIAL SECOND TOSS

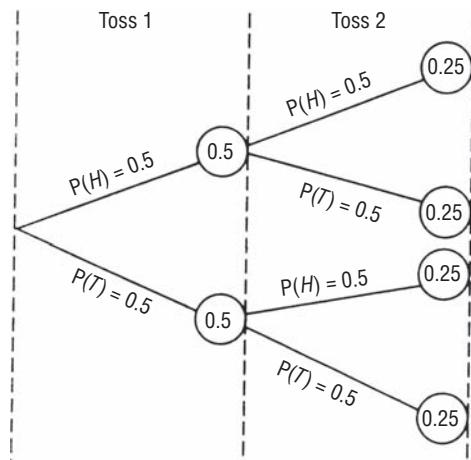


FIGURE 4-7 PROBABILITY TREE OF TWO TOSSES

the lower branch representing toss 1. Thus, in Figure 4-7, we add two more branches to the tree. Notice that on two tosses, we have four possible outcomes: H_1H_2 , H_1T_2 , T_1H_2 , and T_1T_2 (remember the subscripts indicate the toss number, so that T_2 , for example, means tails on toss 2). Thus, after two tosses, we may arrive at any one of four possible points. Because we are going to toss three times, we must add more branches to the tree.

Assuming that we have had heads on the first two tosses, we are now ready to begin adding branches for the third toss. As before, the two possible outcomes are heads and tails, each with a probability of 0.5. The first step is shown in Figure 4-8. The additional branches are added in exactly the same

Three tosses, eight possible outcomes

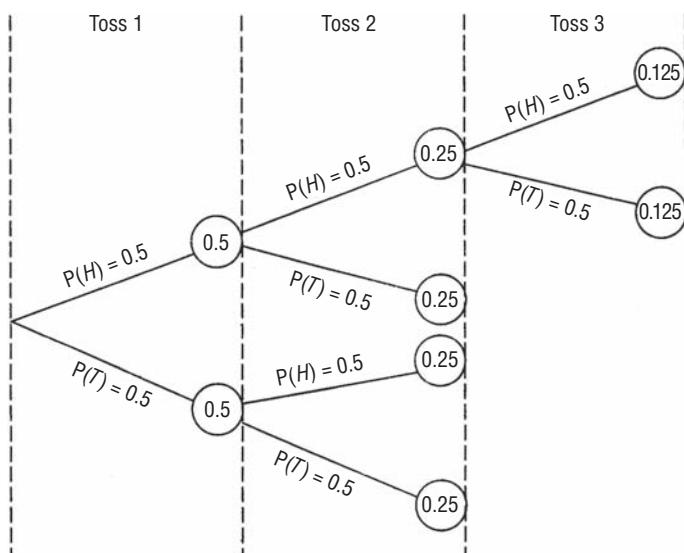
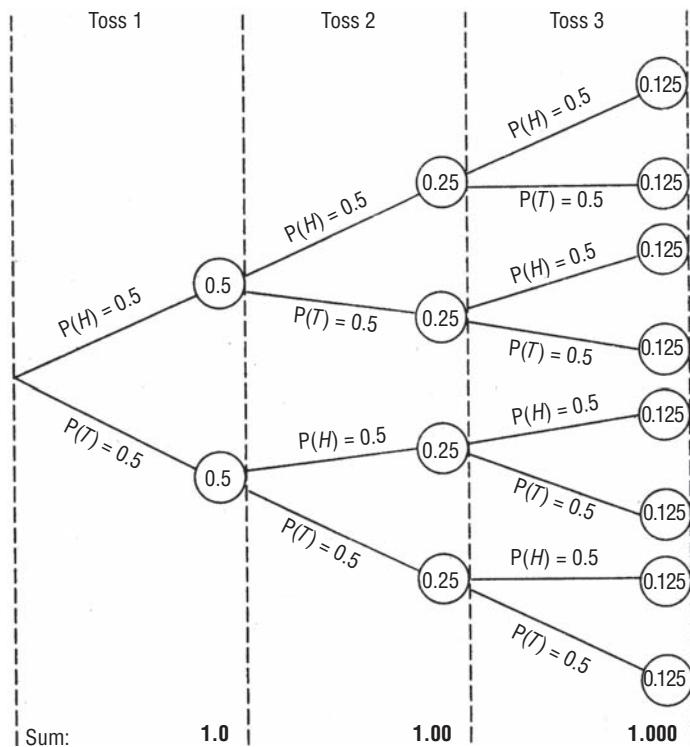


FIGURE 4-8 PROBABILITY TREE OF PARTIAL THIRD TOSS

**FIGURE 4-9 COMPLETED PROBABILITY TREE**

manner. The completed probability tree is shown in Figure 4-9. Notice that both heads and tails have a probability of 0.5 of occurring no matter how far from the origin (first toss) any particular toss may be. **This follows from our definition of independence: No event is affected by the events preceding or following it.**

Suppose we are going to toss a fair coin and want to know the probability that all three tosses will result in heads. Expressing the problem symbolically, we want to know $P(H_1H_2H_3)$. From the mathematical definition of the joint probability of independent events, we know that

$$P(H_1H_2H_3) = P(H_1) \times P(H_2) \times P(H_3) = 0.5 \times 0.5 \times 0.5 = 0.125$$

We could have read this answer from the probability tree in Figure 4-9 by following the branches giving $H_1H_2H_3$.

Try solving these problems using the probability tree in Figure 4-9.

Example 1 What is the probability of getting tails, heads, tails *in that order* on three successive tosses of a fair coin?

Outcomes in a particular order

Solution $P(T_1H_2T_3) = P(T_1) \times P(H_2) \times P(T_3) = 0.125$. Following the prescribed path on the probability tree will give us the same answer.

TABLE 4-2 LISTS OF OUTCOMES

1 Toss		2 Tosses		3 Tosses	
Possible Outcomes	Probability	Possible Outcomes	Probability	Possible Outcomes	Probability
H_1	0.5	H_1H_2	0.25	$H_1H_2H_3$	0.125
T_1	0.5	H_1T_2	0.25	$H_1H_2T_3$	0.125
	$\frac{1.0}{}$	T_1H_2	0.25	$H_1T_2H_3$	0.125
	$\frac{1.0}{}$	T_1T_2	$\frac{0.25}{}$	$T_1H_2H_3$	0.125
The sum of the probabilities of all the possible outcomes must always equal 1		$\frac{1.00}{}$		$T_1H_2T_3$	0.125
				$T_1T_2H_3$	0.125
				$T_1T_2T_3$	0.125
					$\frac{0.125}{}$
				$\frac{1.00}{}$	

Example 2 What is the probability of getting tails, tails, heads *in that order* on three successive tosses of a fair coin?

Solution If we follow the branches giving tails on the first toss, tails on the second toss, and heads on the third toss, we arrive at the probability of 0.125. Thus, $P(T_1T_2H_3) = 0.125$.

It is important to notice that the probability of arriving at a given point by a given route is *not* the same as the probability of, say, heads on the third toss. $P(H_1T_2H_3) = 0.125$, but $P(H_3) = 0.5$. The first is a case of *joint probability* that is, the probability of getting heads on the first toss, tails on the second, and heads on the third. The latter, by contrast, is simply the *marginal probability* of getting heads on a particular toss, in this instance toss 3.

Notice that the sum of the probabilities of all the possible outcomes for each toss is 1. This results from the fact that we have mutually exclusive and collectively exhaustive lists of outcomes. These are given in Table 4-2.

Example 3 What is the probability of *at least* two heads on three tosses? **Outcomes in terms of "at least"**

Solution Recalling that the probabilities of mutually exclusive events are additive, we can note the possible ways that at least two heads on three tosses can occur, and we can sum their individual probabilities. The outcomes satisfying the requirement are $H_1H_2H_3$, $H_1H_2T_3$, $H_1T_2H_3$, and $T_1H_2H_3$. Because each of these has an individual probability of 0.125, the sum is 0.5. Thus, the probability of at least two heads on three tosses is 0.5.

Example 4 What is the probability of *at least* one tail on three tosses?

Solution There is only one case in which no tails occur, namely $H_1H_2H_3$. Therefore, we can simply subtract for the answer:

$$1 - P(H_1H_2H_3) = 1 - 0.125 = 0.875$$

The probability of at least one tail occurring in three successive tosses is 0.875.

Example 5 What is the probability of *at least* one head on two tosses?

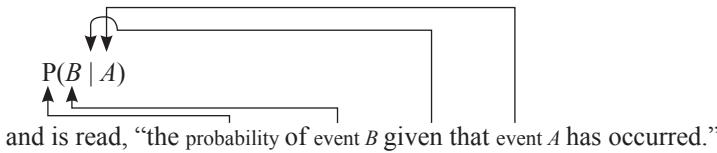
Solution The possible ways at least one head may occur are H_1H_2 , H_1T_2 , T_1H_2 . Each of these has a probability of 0.25. Therefore, the probability of at least one head on two tosses is 0.75. Alternatively, we could consider the case in which no head occurs—namely, $T_1 T_2$ —and subtract its probability from 1; that is,

$$1 - P(T_1 T_2) = 1 - 0.25 = 0.75$$

Conditional Probabilities under Statistical Independence

Thus far, we have considered two types of probabilities, **Conditional probability** marginal (or unconditional) probability and joint probability.

Symbolically, marginal probability is $P(A)$ and joint probability is $P(AB)$. Besides these two, there is one other type of probability, known as *conditional probability*. Symbolically, conditional probability is written



Conditional probability is the probability that a second event (B) will occur *if* a first event (A) has already happened.

For statistically independent events, the conditional probability of event B given that event A has occurred is simply the probability of event B :

Conditional probability of independent events

Conditional Probability for Statistically independent Events

$$P(B | A) = P(B)$$

[4-5]

At first glance, this may seem to be contradictory. Remember, however, that by definition, independent events are those whose probabilities are in no way affected by the occurrence of each other. In fact, statistical independence is defined symbolically as the condition in which $P(B | A) = P(B)$.

We can understand conditional probability better by solving an illustrative problem. Our question is, “What is the probability that the second toss of a fair coin will result in heads, given that heads resulted on the first toss?” Symbolically, this is written as $P(H_2 | H_1)$. Remember that for two independent events, the results of the first toss have absolutely no effect on the results of the second toss. Because the probabilities of heads and tails are identical for every toss, the probability of heads on the second toss is 0.5. Thus, we must say that $P(H_2 | H_1) = 0.5$.

Table 4-3 summarizes the three types of probabilities and their mathematical formulas under conditions of statistical independence.

TABLE 4-3 PROBABILITIES UNDER STATISTICAL INDEPENDENCE

Type of Probability	Symbol	Formula
Marginal	$P(A)$	$P(A)$
Joint	$P(AB)$	$P(A) \times P(B)$
Conditional	$P(B A)$	$P(B)$

HINTS & ASSUMPTIONS

Warning: In statistical independence, our assumption is that events are not related. In a series of coin toss examples, this is true, but in a series of business decisions, there may be a relationship among them. At the very least, you learn from the outcome of each decision and that knowledge affects your next decision. Before calculating conditional or joint probabilities in business situations while assuming independence, be careful you have considered some of the ways that experience affects future judgment.

EXERCISES 4.5

Self-Check Exercise

- SC 4-7** What is the probability that in selecting two cards one at a time from a deck with replacement, the second card is
- A face card, given that the first card was red?
 - An ace, given that the first card was a face card?
 - A black jack, given that the first card was a red ace?
- SC 4-8** Sol O'Tarry, a prison administrator, has been reviewing the prison records on attempted escapes by inmates. He has data covering the last 45 years that the prison has been open, arranged by seasons. The data are summarized in the table:

Attempted Escapes	Winter	Spring	Summer	Fall
0	3	2	1	0
1–5	15	10	11	12
6–10	15	12	11	16
11–15	5	8	7	7
16–20	3	4	6	5
21–25	2	4	5	3
More than 25	<u>2</u>	<u>5</u>	<u>4</u>	<u>2</u>
	45	45	45	45

- What is the probability that in a year selected at random, the number of escapes was between 16 and 20 during the winter?
- What is the probability that more than 10 escapes were attempted during a randomly chosen summer season?
- What is the probability that between 11 and 20 escapes were attempted during a randomly chosen season? (*Hint:* Group the data together.)

Basic Concepts

- 4-23** What is the probability that a couple's second child will be
- A boy, given that their first child was a girl?
 - A girl, given that their first child was a girl?

- 4-24** In rolling two dice, what is the probability of rolling
 (a) A total of 7 on the first roll, followed by a total of 11 on the second roll?
 (b) A total of 21 on the first two rolls combined?
 (c) A total of 6 on the first three rolls combined?
- 4-25** A bag contains 32 marbles: 4 are red, 9 are black, 12 are blue, 6 are yellow, and 1 is purple. Marbles are drawn one at a time with replacement. What is the probability that
 (a) The second marble is yellow given the first one was yellow?
 (b) The second marble is yellow given the first one was black?
 (c) The third marble is purple given both the first and second were purple?
- 4-26** George, Richard, Paul, and John play the following game. Each man takes one of four balls numbered 1 through 4 from an urn. The man who draws ball 4 loses. The other three return their balls to the urn and draw again. Now the one who draws ball 3 loses. The other two return their balls to the urn and draw again. The man who draws ball 1 wins the game.
 (a) What is the probability that John does not lose in the first two draws?
 (b) What is the probability that Paul wins the game?

Applications

- 4-27** The health department routinely conducts two independent inspections of each restaurant, with the restaurant passing only if both inspectors pass it. Inspector A is very experienced, and, hence, passes only 2 percent of restaurants that actually do have health code violations. Inspector B is less experienced and passes 7 percent of restaurants with violations. What is the probability that
 (a) Inspector A passes a restaurant, given that inspector B has found a violation?
 (b) Inspector B passes a restaurant with a violation, given that inspector A passes it?
 (c) A restaurant with a violation is passed by the health department?
- 4-28** The four floodgates of a small hydroelectric dam fail and are repaired independently of each other. From experience, it's known that each floodgate is out of order 4 percent of the time.
 (a) If floodgate 1 is out of order, what is the probability that floodgates 2 and 3 are out of order?
 (b) During a tour of the dam, you are told that the chances of all four floodgates being out of order are less than 1 in 5,000,000. Is this statement true?
- 4-29** Rob Rales is preparing a report that his employer, the Titre Corporation, will eventually deliver to the Federal Aviation Administration. First, the report must be approved by Rob's group leader, department head, and division chief (in that order). Rob knows from experience that the three managers act independently. Further, he knows that his group leader approves 85 percent of his reports, his department head approves 80 percent of the reports written by Rob that reach him, and his division chief approves 82 percent of Rob's work.
 (a) What is the probability that the first version of Rob's report is submitted to the FAA?
 (b) What is the probability that the first version of Rob's report is approved by his group leader and department head, but is not approved by his division chief?

- 4-30** A grocery store is reviewing its restocking policies and has analyzed the number of half-gallon containers of orange juice sold each day for the past month. The data are given below:

Number Sold	Morning	Afternoon	Evening
0–19	3	8	2
20–39	3	4	3
40–59	12	6	4
60–79	4	9	9
80–99	5	3	6
100 or more	3	0	6
	30	30	30

- (a) What is the probability that on a randomly selected day the number of cartons of orange juice sold in the evening is between 80 and 99?
 (b) What is the probability that 39 or fewer cartons were sold during a randomly selected afternoon?
 (c) What is the probability that either 0–19 or 100 or more cartons were sold in a randomly selected morning?
- 4-31** Bill Borde, top advertising executive for Grapevine Concepts, has just launched a publicity campaign for a new restaurant in town, The Black Angus. Bill has just installed four billboards on a highway outside of town, and he knows from experience the probabilities that each will be noticed by a randomly chosen motorist. The probability of the first billboard's being noticed by a motorist is 0.75. The probability of the second's being noticed is 0.82, the third has a probability of 0.87 of being noticed, and the probability of the fourth sign's being noticed is 0.9. Assuming that the event that a motorist notices any particular billboard is independent of whether or not he notices the others, what is the probability that
 (a) All four billboards will be noticed by a randomly chosen motorist?
 (b) The first and fourth, but not the second and third billboards will be noticed?
 (c) Exactly one of the billboards will be noticed?
 (d) None of the billboards will be noticed?
 (e) The third and fourth billboards won't be noticed?

Worked-Out Answers to Self-Check Exercises

- SC 4-7** (a) $P(\text{Face}_2 \mid \text{Red}_1) = 12/52 = 3/13$
 (b) $P(\text{Ace}_2 \mid \text{Face}_1) = 4/52 = 1/13$
 (c) $P(\text{Black jack}_2 \mid \text{Red ace}_1) = 2/52 = 1/26$

- SC 4-8** (a) $3/45 = 1/15$
 (b) $(7 + 6 + 5 + 4)/45 = 22/45$
 (c) $(8 + 12 + 13 + 12)/180 = 45/180 = 1/4$

4.6 PROBABILITIES UNDER CONDITIONS OF STATISTICAL DEPENDENCE

Statistical dependence exists when the probability of some event is dependent on or affected by the occurrence of some

Dependence defined

other event. Just as with independent events, the types of probabilities under statistical dependence are

1. Conditional
2. Joint
3. Marginal

Conditional Probabilities under Statistical Dependence

Conditional and joint probabilities under statistical dependence are more involved than marginal probabilities are. We shall discuss conditional probabilities first, because the concept of joint probabilities is best illustrated by using conditional probabilities as a basis.

Assume that we have one box containing 10 balls distributed as follows:

Examples of conditional probability of dependent events

- Three are colored and dotted.
- One is colored and striped.
- Two are gray and dotted.
- Four are gray and striped.

The probability of drawing any one ball from this box is 0.1, since there are 10 balls, each with equal probability of being drawn. The discussion of the following examples will be facilitated by reference to Table 4-4 and to Figure 4-10, which shows the contents of the box in diagram form.

Example 1 Suppose someone draws a colored ball from the box. What is the probability that it is dotted? What is the probability it is striped?

Solution This question can be expressed symbolically as $P(D|C)$, or “What is the conditional probability that this ball is dotted, given that it is colored?”

We have been told that the ball that was drawn is colored. Therefore, to calculate the probability that the ball is dotted, we will ignore *all* the gray balls and concern ourselves with the colored

TABLE 4-4 COLOR AND CONFIGURATION OF 10 BALLS

Event	Probability of Event
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
10	0.1

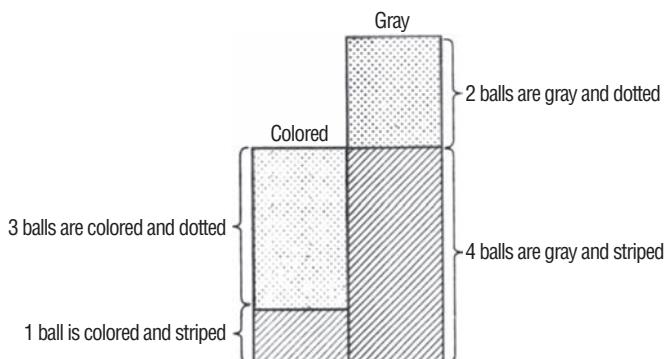


FIGURE 4-10 CONTENTS OF THE BOX

balls only. In diagram form, we consider only what is shown in Figure 4-11.

From the statement of the problem, we know that there are four colored balls, three of which are dotted and one of which is striped. Our problem is now to find the simple probabilities of dotted and striped. To do so, we divide the number of balls in each category by the total number of colored balls:

$$P(D|C) = \frac{3}{5} = 0.75$$

$$P(S|C) = \frac{1}{4} = 0.25$$

1.00

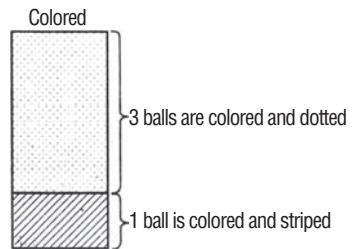


FIGURE 4-11 PROBABILITY OF DOTTED AND STRIPED, GIVEN COLORED

In other words, three-fourths of the colored balls are dotted and one-fourth of the colored balls are striped. Thus, the probability of dotted, given that the ball is colored, is 0.75. Likewise, the probability of striped, given that the ball is colored, is 0.25.

Now we can see how our reasoning will enable us to develop the formula for conditional probability under statistical dependence. We can first assure ourselves that these events *are* statistically dependent by observing that the color of the balls determines the probabilities that they are either striped or dotted. For example, a gray ball is more likely to be striped than a colored ball is. Since color affects the probability of striped or dotted, these two events are dependent.

To calculate the probability of dotted given colored, $P(D|C)$, we divided the probability of colored and dotted balls (3 out of 10, or 0.3) by the probability of colored balls (4 out of 10, or 0.4):

$$P(D|C) = \frac{P(DC)}{P(C)}$$

Expressed as a general formula using the letters A and B to represent the two events, the equation is

Conditional Probability for Statistically Dependent Events

$$P(B|A) = \frac{P(BA)}{P(A)} \quad [4-6]$$

This is the formula for *conditional probability under statistical dependence*.

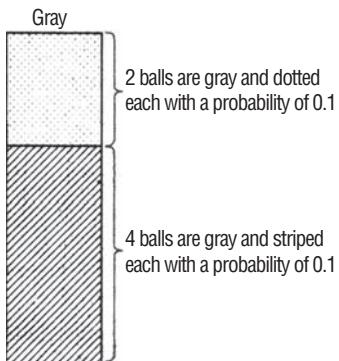
Example 2 Continuing with our example of the colored and gray balls, let's answer the questions, "What is $P(D|G)$?" and "What is $P(S|G)$?"

Solution

$$P(D|G) = \frac{P(DG)}{P(G)} = \frac{0.2}{0.6} = \frac{1}{3}$$

$$P(S|G) = \frac{P(SG)}{P(G)} = \frac{0.4}{0.6} = \frac{2}{3}$$

1.0

**FIGURE 4-12 PROBABILITY OF DOTTED AND STRIPED, GIVEN GRAY**

The problem is shown diagrammatically in Figure 4-12.

The total probability of gray is 0.6 (6 out of 10 balls). To determine the probability that the ball (which we know is gray) will be dotted, we divide the probability of gray and dotted (0.2) by the probability of gray (0.6), or $0.2/0.6 = 1/3$. Similarly, to determine the probability that the ball will be striped, we divide the probability of gray and striped (0.4) by the probability of gray (0.6), or $0.4/0.6 = 2/3$.

Example 3 Calculate $P(G|D)$ and $P(C|D)$.

Solution Figure 4-13 shows the contents of the box arranged according to the striped or dotted markings on the balls. Because we have been told that the ball that was drawn is dotted, we can disregard striped and consider only dotted.

Now see Figure 4-14, showing the probabilities of colored and gray, given dotted. Notice that the relative proportions of the two are 0.4 to 0.6. The calculations used to arrive at these proportions were

$$P(G|D) = \frac{P(GD)}{P(D)} = \frac{0.2}{0.5} = 0.4$$

$$P(C|D) = \frac{P(CD)}{P(D)} = \frac{0.3}{0.5} = 0.6$$

1.0

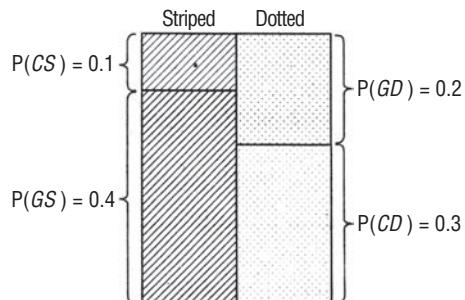
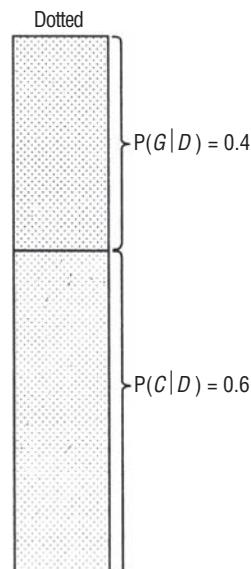
Example 4 Calculate $P(C|S)$ and $P(G|S)$.

Solution

$$P(C|S) = \frac{P(CS)}{P(S)} = \frac{0.1}{0.5} = 0.2$$

$$P(G|S) = \frac{P(GS)}{P(S)} = \frac{0.4}{0.5} = 0.8$$

1.0

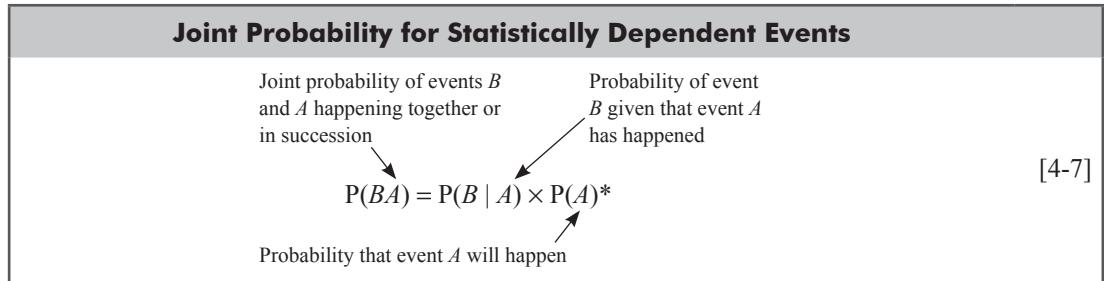
**FIGURE 4-13 CONTENTS OF THE BOX ARRANGED BY CONFIGURATION, STRIPED AND DOTTED****FIGURE 4-14 PROBABILITY OF COLORED AND GRAY, GIVEN DOTTED**

Joint Probabilities under Statistical Dependence

We have shown that the formula for conditional probability under conditions of statistical dependence is

$$P(B|A) = \frac{P(BA)}{P(A)} \quad [4-6]$$

If we solve this for $P(BA)$ by cross multiplication, we have the formula for *joint probability under conditions of statistical dependence*:



[4-7]

Notice that this formula is *not* $P(BA) = P(B) \times P(A)$, as it would be under conditions of statistical independence.

Converting the general formula $P(BA) = P(B|A) \times P(A)$ to our example and to the terms of colored, gray, dotted, and striped, we have $P(CD) = P(C|D) \times P(D)$, or $P(CD) = 0.6 \times 0.5 = 0.3$. Here, 0.6 is the probability of colored, given dotted (computed in Example 3 above) and 0.5 is the probability of dotted (also computed in Example 3).

$P(CD) = 0.3$ can be verified in Table 4-4, where we originally arrived at the probability by inspection: Three balls out of 10 are colored and dotted.

The following joint probabilities are computed in the same manner and can also be substantiated by reference to Table 4-4.

Several examples

$$P(CS) = P(C|S) \times P(S) = 0.2 \times 0.5 = 0.1$$

$$P(GD) = P(G|D) \times P(D) = 0.4 \times 0.5 = 0.2$$

$$P(GS) = P(G|S) \times P(S) = 0.8 \times 0.5 = 0.4$$

Marginal Probabilities under Statistical Dependence

Marginal probabilities under statistical dependence are computed by summing up the probabilities of all the joint events in which the simple event occurs. In the example above, we can compute the marginal probability of the event colored by summing the probabilities of the two joint events in which colored occurred:

$$P(C) = P(CD) + P(CS) = 0.3 + 0.1 = 0.4$$

*To find the joint probability of events A and B , you could also use the formula $P(BA) = P(AB) = P(A|B) \times P(B)$. This is because $BA = AB$.

TABLE 4-5 PROBABILITIES UNDER STATISTICAL INDEPENDENCE AND DEPENDENCE

Type of Probability	Symbol	Formula under Statistical Independence	Formula under Statistical Dependence
Marginal	$P(A)$	$P(A)$	Sum of the probabilities of the joint events in which A occurs
Joint	$P(AB)$	$P(A) \times P(B)$	$P(A B) \times P(B)$
	or $P(BA)$	$P(B) \times P(A)$	$P(B A) \times P(A)$
Conditional	$P(B A)$	$P(B)$	$\frac{P(BA)}{P(A)}$
	or $P(A B)$	$P(A)$	$\frac{P(AB)}{P(B)}$

Similarly, the marginal probability of the event gray can be computed by summing the probabilities of the two joint events in which gray occurred:

$$P(G) = P(GD) + P(GS) = 0.2 + 0.4 = 0.6$$

In like manner, we can compute the marginal probability of the event dotted by summing the probabilities of the two joint events in which dotted occurred:

$$P(D) = P(CD) + P(GD) = 0.3 + 0.2 = 0.5$$

And, finally, the marginal probability of the event striped can be computed by summing the probabilities of the two joint events in which gray occurred:

$$P(S) = P(CS) + P(GS) = 0.1 + 0.4 = 0.5$$

These four marginal probabilities, $P(C) = 0.4$, $P(G) = 0.6$, $P(D) = 0.5$, and $P(S) = 0.5$, can be verified by inspection of Table 4-4 on page 180.

We have now considered the three types of probability (conditional, joint, and marginal) under conditions of statistical dependence. Table 4-5 provides a résumé of our development of probabilities under both statistical independence and statistical dependence.

Example Department of Social Welfare has recently carried out a socio-economic survey of a village. The information collected is related to the gender of the respondent and level of education (graduation). 1000 respondent were surveyed. The results are presented in the following table:

Gender	Educational Qualification		
	Undergraduate	Graduate	Total
Male	150	450	600
Female	150	250	400
Total	300	700	1000

A respondent has been selected randomly, what are the chances that -

- (a) The respondent will be Undergraduate (U)-

$$P(U) = 300/1000 = 0.3$$

(b) The respondent will be Graduate (G),

$$P(G) = 700/1000 = 0.7$$

(c) The respondent will be Female (F),

$$P(F) = 400/1000 = 0.4$$

These are the examples of Unconditional Probability. They are termed as unconditional because no condition is imposed on any event.

(d) The respondent will be Male-Graduate (MG)

$$P(\text{Male \& Graduate}) = P(MG) = 450/1000 = 0.45$$

(e) The respondent will be Undergraduate-Female (UF)

$$P(\text{Undergraduate and Female}) = P(UF) = 150/1000 = 0.15$$

The above two cases (d) and (e) are examples of Joint Probability.

(f) A randomly selected Female will be Graduate (G/F):

Here a condition has been imposed that randomly selected respondent has been Female. So, this is an example of Conditional Probability. In this case, we have to find out the probability of being Graduate, under the condition that the respondent should be Female. Hence consideration should be from a total of 400 (Female respondents only)

So, Probability of the respondent being Graduate, given Female-

$$P(G/F) = 250/400 = 0.625$$

The above concept can also be explained as under:

Probability of the respondent being Female,

$$P(F) = 400/1000 = 0.40$$

Probability of Female-Graduate

$$P(\text{Graduate \& Female}) = P(GF) = 250/1000 = 0.25$$

So,

$$P(G/F) = P(G \text{ and } F) / P(F) = 0.25/0.40 = 0.625$$

(g) A randomly selected Undergraduate will be Male (M/U)):

Probability of the respondent being Male, given Undergraduate,

$$P(M/U) = 150/300 = 0.50$$

Alternatively,

Probability of the respondent being Undergraduate,

$$P(U) = 300/1000 = 0.30$$

Probability of the respondent being Male-Undergraduate-

$$P(\text{Male} \& \text{Undergraduate}) = P(\text{MU}) = 150/1000 = 0.15$$

So,

$$P(\text{M/U}) = P(\text{M and U}) / P(\text{U}) = 0.15/0.30 = 0.50$$

HINTS & ASSUMPTIONS

Hint: Distinguish between conditional probability and joint probability by careful use of terms *given that* and *both ... and*: $P(A|B)$ is “the probability that A will occur *given that* B has occurred” and $P(AB)$ is “the probability that *both A and B* will occur.” And the *marginal probability* $P(A)$ is the “probability that A will occur, whether or not B happens.”

EXERCISES 4.6

Self-Check Exercises

- SC 4-9** According to a survey, the probability that a family owns two cars if its annual income is greater than \$35,000 is 0.75. Of the households surveyed, 60 percent had incomes over \$35,000 and 52 percent had two cars. What is the probability that a family has two cars and an income over \$35,000 a year?
- SC 4-10** Friendly’s Department Store has been the target of many shoplifters during the past month, but owing to increased security precautions, 250 shoplifters have been caught. Each shoplifter’s sex is noted; also noted is whether the perpetrator was a first-time or repeat offender. The data are summarized in the table.

Sex	First-Time Offender	Repeat Offender
Male	60	70
Female	44	76
	104	146

Assuming that an apprehended shoplifter is chosen at random, find

- The probability that the shoplifter is male.
- The probability that the shoplifter is a first-time offender, given that the shoplifter is male.
- The probability that the shoplifter is female, given that the shoplifter is a repeat offender.
- The probability that the shoplifter is female, given that the shoplifter is a first-time offender.
- The probability that the shoplifter is both male and a repeat offender.

Basic Concepts

- 4-32** Two events, A and B , are statistically dependent. If $P(A) = 0.39$, $P(B) = 0.21$, and $P(A \text{ or } B) = 0.47$, find the probability that
- Neither A nor B will occur.
 - Both A and B will occur.

- (c) B will occur, given that A has occurred.
- (d) A will occur, given that B has occurred.

4-33 Given that $P(A) = 3/14$, $P(B) = 1/6$, $P(C) = 1/3$, $P(AC) = 1/7$, and $P(B|C) = 5/21$, find the following probabilities: $P(A|C)$, $P(C|A)$, $P(BC)$, $P(C|B)$.

4-34 Assume that for two events A and B , $P(A) = 0.65$, $P(B) = 0.80$, $P(A|B) = P(A)$, and $P(B|A) = 0.85$. Is this a consistent assignment of probabilities? Explain.

Applications

4-35 At a soup kitchen, a social worker gathers the following data. Of those visiting the kitchen, 59 percent are men, 32 percent are alcoholics, and 21 percent are male alcoholics. What is the probability that a random male visitor to the kitchen is an alcoholic?

4-36 During a study of auto accidents, the Highway Safety Council found that 60 percent of all accidents occur at night, 52 percent are alcohol-related, and 37 percent occur at night and are alcohol-related.

- (a) What is the probability that an accident was alcohol-related, given that it occurred at night?

- (b) What is the probability that an accident occurred at night, given that it was alcohol-related?

4-37 If a hurricane forms in the eastern half of the Gulf of Mexico, there is a 76 percent chance that it will strike the western coast of Florida. From data gathered over the past 50 years, it has been determined that the probability of a hurricane's occurring in this area in any given year is 0.85.

- (a) What is the probability that a hurricane will occur in the eastern Gulf of Mexico and strike Florida this year?

- (b) If a hurricane in the eastern Gulf of Mexico is seeded (induced to rain by addition of chemicals from aircraft), its probability of striking Florida's west coast is reduced by one-fourth. If it is decided to seed any hurricane in the eastern gulf, what is the new value for the probability in part (a)?

4-38 Al Cascade, president of the Litre Corporation, is studying his company's chances of being awarded an important water purification system contract for the Tennessee Valley Authority. Accordingly, two events are of interest to him. First, Litre's major competitor, WTR, is conducting purification research, which it hopes to complete before the contract award deadline. Second, there are rumors of a TVA investigation of all recent contractors, of which Litre is one and WTR is not. If WTR finishes its research and there is no investigation, then Litre's probability of being awarded the contract is 0.67. If there is an investigation but WTR doesn't finish its research, the probability is 0.72. If both events occur, the probability is 0.58, and if neither occurs, the probability is 0.85. The occurrence of an investigation and WTR's completion of research in time are independent events.

- (a) Suppose that Al knows that the probability of WTR's completing its research in time is 0.80. How low must the probability of an investigation be so that the probability of Litre's being awarded the contract is at least 0.65?

- (b) Suppose that Al knows that the probability of an investigation is 0.70. How low must the probability of WTR's completing its research on time be so that the probability of Litre's being awarded the contract is at least 0.65?

- (c) Suppose that the probability of an investigation is 0.75 and the probability of WTR's completing its research in time is 0.85. What is the probability of Litre's being awarded the contract?

- 4-39 A company is considering upgrading its computer system, and a major portion of the upgrade is a new operating system. The company has asked an engineer for an evaluation of the operating system. Suppose the probability of a favorable evaluation is 0.65. If the probability the company will upgrade its system given a favorable evaluation is 0.85, what is the probability that the company will upgrade and receive a favorable evaluation?
- 4-40 The university's library has been randomly surveying patrons over the last month to see who is using the library and what services they have been using. Patrons are classified as undergraduate, graduate, or faculty. Services are classified as reference, periodicals, or books. The data for 350 people are given below. Assume a patron uses only one service per visit.

Patron	Reference	Periodicals	Books
Undergraduate	44	26	72
Graduate	24	61	20
Faculty	16	69	18
	84	156	110

Find the probability that a randomly chosen patron

- (a) Is a graduate student.
 - (b) Visited the periodicals section, given the patron is a graduate student.
 - (c) Is a faculty member, given a reference section visit.
 - (d) Is an undergraduate who visited the book section.
- 4-41 The southeast regional manager of General Express, a private parcel-delivery firm, is worried about the likelihood of strikes by some of his employees. He has learned that the probability of a strike by his pilots is 0.75 and the probability of a strike by his drivers is 0.65. Further, he knows that if the drivers strike, there is a 90 percent chance that the pilots will strike in sympathy.
- (a) What is the probability of both groups' striking?
 - (b) If the pilots strike, what is the probability that the drivers will strike in sympathy?
- 4-42 National Horticulture Board has been entrusted with the responsibility of sending good quality mangoes to overseas. For this purpose, an inspection is conducted on 10,000 boxes of mangoes from Malihabad and Hyderabad for exports. The inspection of boxes gave the following information:-

	Number of Boxes with		
	Number of Boxes	Damaged Fruit	Overripe Fruit
Malihabad	6000	200	840
Hyderabad	4000	365	295

- (a) What are the chances that a selected box will contain damaged or overripe fruit?
 - (b) A randomly selected box contains overripe fruit, what is the probability that it has came from Hyderabad?
- 4-43 Fragnance Soaps Pvt Ltd is a leading soap manufacturing company in India. "Active" is a well known brand of the company. Company conducted a survey to find out preference for this brand. The marketing research responses are as shown in the following table:

Prefer	Ahmedabad	Gwalior	Raipur	Lucknow
Yes	55	40	80	75
No	40	30	20	90
No Opinion	5	10	20	35

If a customer is selected at random, what is the probability?

- (a) That he or she prefers active?
- (b) The consumer prefers Active and is from Ahmadabad?
- (c) The consumer prefers Active given he is from Lucknow?
- (d) That he is from Raipur and has no opinion?

Worked-Out Answers to Self-Check Exercises

SC 4-9 Let I = income > \$35,000 C = 2 cars.

$$P(C \text{ and } I) = P(C|I)P(I) = (0.75)(0.6) = 0.45$$

SC 4-10 M/W = shoplifter is male/female; F/R = shoplifter is first-time/repeat offender.

- (a) $P(M) = (60 + 70)/250 = 0.520$
- (b) $P(F|M) = P(F \text{ and } M)/P(M) = (60/250)/(130/250) = 0.462$
- (c) $P(W|R) = P(W \text{ and } R)/P(R) = (76/250)/(146/250) = 0.521$
- (d) $P(W|F) = P(W \text{ and } F)/P(F) = (44/250)/(104/250) = 0.423$
- (e) $P(M \text{ and } R) = 70/250 = 0.280$

4.7 REVISING PRIOR ESTIMATES OF PROBABILITIES: BAYES' THEOREM

At the beginning of the baseball season, the fans of last year's pennant winner thought their team had a good chance of winning again. As the season progressed, however, injuries side lined the shortstop and the team's chief rival drafted a terrific home run hitter. The team began to lose. Late in the season, the fans realized that they must alter their prior probabilities of winning.

A similar situation often occurs in business. If a manager of a boutique finds that most of the purple and chartreuse ski jackets that she thought would sell so well are hanging on the rack, she must revise her prior probabilities and order a different color combination or have a sale.

In both these cases, certain probabilities were altered after the people involved got additional information. The new probabilities are known as revised, or *posterior*, probabilities. Because probabilities can be revised as more information is gained, probability theory is of great value in managerial decision making.

Posterior probabilities defined

The origin of the concept of obtaining posterior probabilities with limited information is attributable to the Reverend Thomas Bayes (1702–1761), and the basic formula for conditional probability under dependence

Bayes' theorem

$$P(B|A) = \frac{P(BA)}{P(A)}$$

[4-6]

is called *Bayes' Theorem*.

Bayes, an Englishman, was a Presbyterian minister and a competent mathematician. He pondered how he might prove the existence of God by examining whatever evidence the world about him provided. Attempting to show “that the Principal End of the Divine Providence . . . is the Happiness of His Creatures,” the Reverend Bayes used mathematics to study God. Unfortunately, the theological implications of his findings so alarmed the good Reverend Bayes that he refused to permit publication of his work during his lifetime. Nevertheless, his work outlived him, and modern decision theory is often called Bayesian decision theory in his honor.

Bayes’ theorem offers a powerful statistical method of evaluating new information and revising our prior estimates (based upon limited information only) of the probability that things are in one state or another. **If correctly used, it makes it unnecessary to gather masses of data over long periods of time in order to make good decisions based on probabilities.**

Value of Bayes’ theorem

Calculating Posterior Probabilities

Assume, as a first example of revising prior probabilities, that we have equal numbers of two types of deformed (biased or weighted) dice in a bowl. On half of them, ace (or one dot) comes up 40 percent of the time; therefore $P(\text{ace}) = 0.4$. On the other half, ace comes up 70 percent of the time; $P(\text{ace}) = 0.7$. Let us call the former type 1 and the latter type 2. One die is drawn, rolled once, and comes up ace. What is the probability that it is a type 1 die? Knowing the bowl contains the same number of both types of dice, we might incorrectly answer that the probability is one-half; but we can do better than this. To answer the question correctly, we set up Table 4-6.

Finding a new posterior estimate

The sum of the probabilities of the elementary events (drawing either a type 1 or a type 2 die) is 1.0 because there are only two types of dice. The probability of each type is 0.5. The two types constitute a mutually exclusive and collectively exhaustive list.

Revising probabilities based on one outcome

The sum of the $P(\text{ace} \mid \text{elementary event})$ column does *not* equal 1.0. The figures 0.4 and 0.7 simply represent the conditional probabilities of getting an ace, given type 1 and type 2 dice, respectively.

The fourth column shows the joint probability of ace and type 1 occurring together ($0.4 \times 0.5 = 0.20$), and the joint probability of ace and type 2 occurring together ($0.7 \times 0.5 = 0.35$). The sum of these joint probabilities (0.55) is the marginal probability of getting an ace. Notice that in each case, the joint probability was obtained by using the formula

$$P(AB) = P(A|B) \times P(B)$$

[4-7]

TABLE 4-6 FINDING THE MARGINAL PROBABILITY OF GETTING AN ACE

Elementary Event	Probability of Elementary Event	$P(\text{Ace} \mid \text{Elementary Event})$	$P(\text{Ace}, \text{Elementary Event})^*$
Type 1	0.5	0.4	$0.4 \times 0.5 = 0.20$
Type 2	0.5	0.7	$0.7 \times 0.5 = 0.35$
	1.0		$P(\text{ace}) = 0.55$

*A comma is used to separate joint events. We can join individual letters to indicate joint events without confusion (AB , for example), but joining whole words in this way could produce strange looking events ($\text{aceelementaryevent}$) in this table, and they could be confusing.

To find the probability that the die we have drawn is type 1, we use the formula for conditional probability under statistical dependence:

$$P(B|A) = \frac{P(BA)}{P(A)} \quad [4-6]$$

Converting to our problem, we have

$$P(\text{type 1|ace}) = \frac{P(\text{type 1, ace})}{P(\text{ace})}$$

or

$$P(\text{type 1|ace}) = \frac{0.20}{0.55} = 0.364$$

Thus, the probability that we have drawn a type 1 die is 0.364.

Let us compute the probability that the die is type 2:

$$P(\text{type 2|ace}) = \frac{P(\text{type 2, ace})}{P(\text{ace})} = \frac{0.35}{0.55} = 0.636$$

What have we accomplished with one additional piece of information made available to us? What inferences have we been able to draw from one roll of the die? Before we rolled this die, the best we could say was that there is a 0.5 chance it is a type 1 die and a 0.5 chance it is a type 2 die. However, after rolling the die, we have been able to *alter*, or revise, *our prior probability estimate*. Our new posterior estimate is that there is a higher probability (0.636) that the die we have in our hand is a type 2 than that it is a type 1 (only 0.364).

Conclusion after one roll

Posterior Probabilities with More Information

We may feel that one roll of the die is not sufficient to indicate its characteristics (whether it is type 1 or type 2). In this case, we can obtain additional information by rolling the die again. (Obtaining more information in most decision-making situations, of course, is more complicated and time-consuming.) Assume that the same die is rolled a second time and again comes up ace. What is the further revised probability that the die is type 1? To determine this answer, see Table 4-7.

Finding a new posterior estimate with more information

TABLE 4-7 FINDING THE MARGINAL PROBABILITY OF TWO ACES ON TWO SUCCESSIVE ROLLS

Elementary Event	Probability of Elementary Event	P(Ace Elementary Event)	P(2Aces Elementary Event)	P(2 Aces, Elementary Event)
Type 1	0.5	0.4	0.16	$0.16 \times 0.5 = 0.080$
Type 2	0.5	0.7	0.49	$0.49 \times 0.5 = 0.245$
	1.0			P(2 aces) = 0.325

We have one new column in this table. $P(2 \text{ aces} \mid \text{elementary event})$. This column gives the *joint* probability of two aces on two successive rolls if the die is type 1 and if it is type 2: $P(2 \text{ aces} \mid \text{type 1}) = 0.4 \times 0.4 = 0.16$, and $P(2 \text{ aces} \mid \text{type 2}) = 0.7 \times 0.7 = 0.49$. In the last column, we see the joint probabilities of two aces on two successive rolls and the elementary events (type 1 and type 2). That is, $P(2 \text{ aces, type 1})$ is equal to $P(2 \text{ aces} \mid \text{type 1})$ times the probability of type 1, or $0.16 \times 0.5 = 0.080$, and $P(2 \text{ aces, type 2})$ is equal to $P(2 \text{ aces} \mid \text{type 2})$ times the probability of type 2, or $0.49 \times 0.5 = 0.245$. The sum of these (0.325) is the marginal probability of two aces on two successive rolls.

We are now ready to compute the probability that the die we have drawn is type 1, given an ace on each of two successive rolls. Using the same general formula as before, we convert to

$$P(\text{type 1} \mid 2 \text{ aces}) = \frac{P(\text{type 1, 2 aces})}{P(2 \text{ aces})} = \frac{0.080}{0.325} = 0.246$$

Similarly,

$$P(\text{type 2} \mid 2 \text{ aces}) = \frac{P(\text{type 2, 2 aces})}{P(2 \text{ aces})} = \frac{0.245}{0.325} = 0.754$$

What have we accomplished with two rolls? When we first drew the die, all we knew was that there was a probability of 0.5 that it was type 1 and a probability of 0.5 that it was type 2. In other words, there was a 50–50 chance that it was either type 1 or type 2. After rolling the die once and getting an ace, we revised these original probabilities to the following:

Probability that it is type 1, given that an ace was rolled = 0.364

Probability that it is type 2, given that an ace was rolled = 0.636

After the second roll (another ace), we revised the probabilities again:

Probability that it is type 1, given that two aces were rolled = 0.246

Probability that it is type 2, given that two aces were rolled = 0.754

We have thus changed the original probabilities from 0.5 for each type to 0.246 for type 1 and 0.754 for type 2. This means that if a die turns up ace on two successive rolls, we can now assign a probability of 0.754 that it is type 2.

In both these experiments, we gained new information free of charge. We were able to roll the die twice, observe its behavior, and draw inferences from the behavior without any monetary cost. Obviously, there are few situations in which this is true, and managers must not only understand how to use new information to revise prior probabilities, but also be able to determine *how much that information is worth* to them before the fact. In many cases, the value of the information obtained may be considerably less than its cost.

A Problem with Three Pieces of Information

Consider the problem of a Little League baseball team that has been using an automatic pitching machine. If the machine is correctly set up—that is, properly adjusted—it will pitch strikes 85 percent of the time. If it is incorrectly set up, it will pitch strikes only 35 percent of the time. Past experience indicates that 75 percent of the setups of the machine are correctly done. After the machine has been set up at batting practice one day,

Example of posterior probability based on three trials

TABLE 4-8 POSTERIOR PROBABILITIES WITH THREE TRIALS

Event	P(Event) (1)	P(1 Strike Event) (2)	P(3 Strikes Event) (3)	P(Event, 3 Strikes) (4)
Correct	0.75	0.85	0.6141	$0.6141 \times 0.75 = 0.4606$
Incorrect	0.25	0.35	0.0429	$0.429 \times 0.25 = 0.0107$
	1.00			P(3 strikes) = 0.4713

it throws three strikes on the first three pitches. What is the revised probability that the setup has been done correctly? Table 4-8 illustrates how we can answer this question.

We can interpret the numbered table headings in Table 4-8 as follows:

1. P(event) describes the individual probabilities of correct and incorrect. $P(\text{correct}) = 0.75$ is given in the problem. Thus, we can compute

$$P(\text{incorrect}) = 1.00 - P(\text{correct}) = 1.00 - 0.75 = 0.25$$

2. $P(1 \text{ strike} | \text{event})$ represents the probability of a strike given that the setup is correct or incorrect. These probabilities are given in the problem.
3. $P(3 \text{ strikes} | \text{event})$ is the probability of getting three strikes on three successive pitches, given the event, that is, given a correct or incorrect setup. The probabilities are computed as follows:

$$P(3 \text{ strikes} | \text{correct}) = 0.85 \times 0.85 \times 0.85 = 0.6141$$

$$P(3 \text{ strikes} | \text{incorrect}) = 0.35 \times 0.35 \times 0.35 = 0.0429$$

4. $P(\text{event}, 3 \text{ strikes})$ is the probability of the joint occurrence of the event (correct or incorrect) and three strikes. We can compute the probability in the problem as follows:

$$P(\text{correct}, 3 \text{ strikes}) = 0.6141 \times 0.75 = 0.4606$$

$$P(\text{incorrect}, 3 \text{ strikes}) = 0.0429 \times 0.25 = 0.0107$$

Notice that if $A = \text{event}$ and $S = \text{strikes}$, these last two probabilities conform to the general mathematical formula for joint probabilities under conditions of dependence: $P(AS) = P(SA) = P(S | A) \times P(A)$, Equation 4-7.

After finishing the computation in Table 4-8, we are ready to determine the revised probability that the machine is correctly set up. We use the general formula

$$P(A|S) = \frac{P(AS)}{P(S)} \quad [4-6]$$

and convert it to the terms and numbers in this problem:

$$\begin{aligned} P(\text{correct} | 3 \text{ strikes}) &= \frac{P(\text{correct}, 3 \text{ strikes})}{P(3 \text{ strikes})} \\ &= \frac{0.4606}{0.4713} = 0.9773 \end{aligned}$$

The *posterior probability* that the machine is correctly set up is 0.9773, or 97.73 percent. We have thus revised our original probability of a correct setup from 75 to 97.73 percent, based on three strikes being thrown in three pitches.

TABLE 4-9 POSTERIOR PROBABILITIES WITH INCONSISTENT OUTCOMES

Event	P(Event)	P(S Event)	P(SBSSS Event)	P(Event, SBSSS)
Correct	0.75	0.85	$0.85 \times 0.15 \times 0.85 \times 0.85 \times 0.85 = 0.07830$	$0.07830 \times 0.75 = 0.05873$
Incorrect	0.25	0.35	$0.35 \times 0.65 \times 0.35 \times 0.35 \times 0.35 = 0.00975$	$0.00975 \times 0.25 = 0.00244$
		1.00		P(SBSSS) = 0.06117

$$\begin{aligned} P(\text{correct setup}|SBSSS) &= \frac{P(\text{correct setup}, SBSSS)}{P(SBSSS)} \\ &= \frac{0.05873}{0.06117} \\ &= 0.9601 \end{aligned}$$

Posterior Probabilities with Inconsistent Outcomes

In each of our problems so far, the behavior of the experiment was consistent: the die came up ace on two successive rolls, and the automatic machine three strikes on each of the first three pitches. In most situations, we would expect a less consistent distribution of outcomes. In the case of the pitching machine, for example, we might find the five pitches to be: strike, ball, strike, strike, strike. Calculating our posterior probability that the machine is correctly set up in this case is really no more difficult than it was with a set of perfectly consistent outcomes. Using the notation S = strike and B = ball, we have solved this example in Table 4-9.

An example with inconsistent outcomes

HINTS & ASSUMPTIONS

Posterior Probabilities under Bayes Theorem has an application utility, they provided revised estimates of priori probabilities (chances) to the decision maker using the additional information presented. This helps in more effective decision-making. So estimates of the probability, based on historical information, are revised using additional information.

Bayes' theorem is a formal procedure that lets decision makers combine classical probability theory with their best intuitive sense about what is likely to happen. Warning: The real value of Bayes' theorem is not in the algebra, but rather in the ability of informed managers to make good guesses about the future. Hint: In all situations in which Bayes' theorem will be used, first use all the historical data available to you, and then (and only then) add your own intuitive judgment to the process. Intuition used to make guesses about things that are already statistically well-described is misdirected.

EXERCISES 4.7

Self-Check Exercises

SC 4-11 Given: The probabilities of three events, A , B , and C , occurring are $P(A) = 0.35$, $P(B) = 0.45$, and $P(C) = 0.2$. Assuming that A , B , or C has occurred, the probabilities of another event, X , occurring are $P(X|A) = 0.8$, $P(X|B) = 0.65$, and $P(X|C) = 0.3$. Find $P(A|X)$, $P(B|X)$, and $P(C|X)$.

SC 4-12 A doctor has decided to prescribe two new drugs to 200 heart patients as follows: 50 get drug A, 50 get drug B, and 100 get both. The 200 patients were chosen so that each had an 80 percent chance of having a heart attack if given neither drug. Drug A reduces the probability of a heart attack by 35 percent, drug B reduces the probability by 20 percent, and the two drugs, when taken together, work independently. If a randomly selected patient in the program has a heart attack, what is the probability that the patient was given both drugs?

Basic Concept

4-44 Two related experiments are performed. The first has three possible, mutually exclusive outcomes: A , B , and C . The second has two possible, mutually exclusive outcomes: X and Y . We know $P(A) = 0.2$ and $P(B) = 0.65$. We also know the following conditional probabilities if the result of the second experiment is X : $P(X|A) = 0.75$, $P(X|B) = 0.60$, and $P(X|C) = 0.40$. Find $P(A|X)$, $P(B|X)$, and $P(C|X)$. What is the probability that the result of the second experiment is Y ?

Applications

4-45 Martin Coleman, credit manager for Beck's, knows that the company uses three methods to encourage collection of delinquent accounts. From past collection records, he learns that 70 percent of the accounts are called on personally, 20 percent are phoned, and 10 percent are sent a letter. The probabilities of collecting an overdue amount from an account with the three methods are 0.75, 0.60, and 0.65 respectively. Mr. Coleman has just received payment from a past-due account. What is the probability that this account

- (a) Was called on personally?
- (b) Received a phone call?
- (c) Received a letter?

4-46 A public-interest group was planning to make a court challenge to auto insurance rates in one of three cities: Atlanta, Baltimore, or Cleveland. The probability that it would choose Atlanta was 0.40; Baltimore, 0.35; and Cleveland, 0.25. The group also knew that it had a 60 percent chance of a favorable ruling if it chose Baltimore, 45 percent if it chose Atlanta, and 35 percent if it chose Cleveland. If the group did receive a favorable ruling, which city did it most likely choose?

4-47 EconOcon is planning its company picnic. The only thing that will cancel the picnic is a thunderstorm. The Weather Service has predicted dry conditions with probability 0.2, moist conditions with probability 0.45, and wet conditions with probability 0.35. If the probability of a thunderstorm given dry conditions is 0.3, given moist conditions is 0.6, and given wet conditions is 0.8, what is the probability of a thunderstorm? If we know the picnic was indeed canceled, what is the probability moist conditions were in effect?

4-48 An independent research group has been studying the chances that an accident at a nuclear power plant will result in radiation leakage. The group considers that the only possible types of accidents at a reactor are fire, mechanical failure, and human error, and that two or more accidents never occur together. It has performed studies that indicate that if there were a fire, a radiation leak would occur 20 percent of the time; if there were a mechanical failure, a radiation leak would occur 50 percent of the time; and if there were a human error,

a radiation leak would occur 10 percent of the time. Its studies have also shown that the probability of

- A fire and a radiation leak occurring together is 0.0010.
 - A mechanical failure and a radiation leak occurring together is 0.0015.
 - A human error and a radiation leak occurring together is 0.0012.
- What are the respective probabilities of a fire, mechanical failure, and human error?
 - What are the respective probabilities that a radiation leak was caused by a fire, mechanical failure, and human error?
 - What is the probability of a radiation leak?

4-49 A physical therapist at Enormous State University knows that the football team will play 40 percent of its games on artificial turf this season. He also knows that a football player's chances of incurring a knee injury are 50 percent higher if he is playing on artificial turf instead of grass. If a player's probability of knee injury on artificial turf is 0.42, what is the probability that

- A randomly selected football player incurs a knee injury?
- A randomly selected football player with a knee injury incurred the injury playing on grass?

4-50 The physical therapist from Exercise 4-48 is also interested in studying the relationship between foot injuries and position played. His data, gathered over a 3-year period, are summarized in the following table:

	Offensive Line	Defensive Line	Offensive Backfield	Defensive Backfield
Number of players	45	56	24	20
Number injured	32	38	11	9

Given that a randomly selected player incurred a foot injury, what is the probability that he plays in the (a) offensive line, (b) defensive line, (c) offensive backfield, and (d) defensive backfield?

4-51 A state Democratic official has decided that changes in the state unemployment rate will have a major effect on her party's chance of gaining or losing seats in the state senate. She has determined that if unemployment rises by 2 percent or more, the respective probabilities of losing more than 10 seats, losing 6 to 10 seats, gaining or losing 5 or fewer seats, gaining 6 to 10 seats, and gaining more than 10 seats are 0.25, 0.35, 0.15, 0.15, and 0.10, respectively. If unemployment changes by less than 2 percent, the respective probabilities are 0.10, 0.10, 0.15, 0.35, and 0.30. If unemployment falls by 2 percent or more, the respective probabilities are 0.05, 0.10, 0.10, 0.40, and 0.35. Currently this official believes that unemployment will rise by 2 percent or more with probability 0.25, change by less than 2 percent with probability 0.45, and fall by 2 percent or more with probability 0.30.

- If the Democrats gained seven seats, what is the probability that unemployment fell by 2 percent or more?
- If the Democrats lost one seat, what is the probability that unemployment changed by less than 2 percent?

4-52 T. C. Fox, marketing director for Metro-Goldmine Motion Pictures, believes that the studio's upcoming release has a 60 percent chance of being a hit, a 25 percent chance of being a moderate success, and a 15 percent chance of being a flop. To test the accuracy of his opinion,

T. C. has scheduled two test screenings. After each screening, the audience rates the film on a scale of 1 to 10, 10 being best. From his long experience in the industry, T. C. knows that 60 percent of the time, a hit picture will receive a rating of 7 or higher; 30 percent of the time, it will receive a rating of 4, 5, or 6; and 10 percent of the time, it will receive a rating of 3 or lower. For a moderately successful picture, the respective probabilities are 0.30, 0.45, and 0.25; for a flop, the respective probabilities are 0.15, 0.35, and 0.50.

- If the first test screening produces a score of 6, what is the probability that the film will be a hit?
- If the first test screening produces a score of 6 and the second screening yields a score of 2, what is the probability that the film will be a flop (assuming that the screening results are independent of each other)?

Worked-Out Answers to Self-Check Exercises

SC 4-11	Event	P(Event)	P(X Event)	P(X and Event)	P(Event X)
	A	0.35	0.80	0.2800	$0.2800/0.6325 = 0.4427$
	B	0.45	0.65	0.2925	$0.2925/0.6325 = 0.4625$
	C	0.20	0.30	0.0600	$0.0600/0.6325 = 0.0949$
$\mathbf{P(X) = 0.6325}$					

Thus, $P(A|X) = 0.4427$, $P(B|X) = 0.4625$, and $P(C|X) = 0.0949$.

SC 4-12 H = heart attack.

Event	P(Event)	P(H Event)	P(H and Event)	P(Event H)
A	0.25	$(0.8)(0.65) = 0.520$	0.130	$0.130/0.498 = 0.2610$
B	0.25	$(0.8)(0.80) = 0.640$	0.160	$0.160/0.498 = 0.3213$
$A \& B$	0.50	$(0.8)(0.65)(0.80) = 0.416$	0.208	$0.208/0.498 = 0.4177$
$\mathbf{P(X) = 0.498}$				

Thus, $P(A \& B | H) = 0.4177$.

STATISTICS AT WORK

Loveland Computers

Case 4: Probability “Aren’t you going to congratulate me, Uncle Walter?” Lee Azko asked the CEO of Loveland Computers as they waved goodbye to their new-found investment bankers who were boarding their corporate jet.

“Sure, Lee, it was pretty enough stuff. But you’ll find out that in business, there’s more to life than gathering data. You have to make decisions, too—and often you don’t have all the data you’d like because you’re trying to guess what *will* happen in the future, not what *did* happen in the past. Get in the car and I’ll explain.”

“When we first started Loveland Computers, it was pretty much a wholesaling business. We’d bring in the computers from Taiwan, Korea, or wherever, and just ship ‘em out the door with a label on them. Now that still works for some of the low-end products, but the higher-end stuff needs to be customized, so we run an assembly line here. Now I won’t call it a factory, because there isn’t a single thing that we

‘make’ here. We buy the cases from one place, the hard drives from somewhere else, and so on. Then we run the assembly line to make the machines just the way customers want them.”

“Why don’t you just have all the gizmos loaded on all the PCs, uncle?”

“Not a bad question, but here’s the reason we can’t do that. In this game, price is very important. And if you load a machine with something that a customer is never going to use—for example, going to the expense of adding a very large hard drive to a machine that’s going to be used in a local area network, where most of the data will be kept on a file server—you end up pricing yourself out of the market, or selling at a loss. We can’t afford to do either of those things. When we get back to the office, I want you to see Nancy Rainwater—she’s the head of Production. She needs some help figuring out this month’s schedule. This should give you some experience with real decision making.”

Nancy Rainwater had worked for Loveland Computers for 5 years. Although Nancy was short on book learning, growing up on a farm nearby, she had learned some important practical skills about managing a workforce and getting work done on time. Her rise through the ranks to Production Supervisor had been rapid. Nancy explained her problem to Lee as follows.

“We have to decide whether to close the production line on Martin Luther King Day on the 20th of the month. Most of the workers on the line have children who will be off school that day. Your uncle, Mr. Azko, won’t make it a paid vacation. But he might be open to closing the production line and letting people take the day off without pay if we can put in enough work days by the end of the month to meet our target production.”

“Well, that shouldn’t be too difficult to figure out—just count up the number of PCs produced on a typical day and divide that into the production target and see how many workdays you’ll need,” replied Lee with confidence.

“Well, I’ve already got that far. Not counting today, there are 19 workdays left until the end of the month, and I’ll need 17 days to complete the target production.”

“So let the workers take Martin Luther King Day off,” Lee concluded.

“But there’s more to it than that,” Nancy continued. “This is ‘colds and flu’ season. If too many people call in sick—and believe me that happens when there’s a ‘bug’ going around—I have to close the line for the day. I have records going back for a couple of years since I’ve been supervisor, and on an average winter day, there’s a 1 in 30 chance that we’ll have to close the line because of too many sick calls.

“And there’s always a chance that we’ll get a bad snowstorm—maybe even two—between now and the end of the month. Two years ago, two of the staff were in a terrible car wreck, trying to come to work on a day when the weather was real bad. So the company lawyer has told us to have a very tight ‘snow day’ policy. If the roads are dangerous, we close the line and lose that day’s production. I’m not allowed to schedule weekend work to make up—that costs us time-and-a-half on wages and costs get out of line.

“I’d feel a lot better about closing the line for the holiday if I could be reasonably certain that we’d get in enough workdays by the end of the month. But I guess you don’t have a crystal ball.”

“Well, not a crystal ball, exactly. But I do have some ideas,” Lee said, walking back toward the administrative offices, sketching something on a notepad. “By the way,” said the younger Azko, turning back toward Nancy Rainwater, “What’s *your* definition of ‘reasonably certain?’”

Study Questions: What was Lee sketching on the notepad? What type of calculation will Lee make and what additional information will be needed? What difference will it make if Nancy’s definition of “reasonably certain” means to meet the required production goal “75 percent of the time” or “99 percent of the time”?

CHAPTER REVIEW

Terms Introduced in Chapter 4

A Priori Probability Probability estimate made prior to receiving new information.

Bayes' Theorem The formula for conditional probability under statistical dependence.

Classical Probability The number of outcomes favorable to the occurrence of an event divided by the total number of possible outcomes.

Collectively Exhaustive Events A list of events that represents all the possible outcomes of an experiment.

Conditional Probability The probability of one event occurring, given that another event has occurred.

Event One or more of the possible outcomes of doing something, or one of the possible outcomes from conducting an experiment.

Experiment The activity that results in, or produces, an event.

Joint Probability The probability of two events occurring together or in succession.

Marginal Probability The unconditional probability of one event occurring; the probability of a single event.

Mutually Exclusive Events Events that cannot happen together.

Posterior Probability A probability that has been revised after additional information was obtained.

Probability The chance that something will happen.

Probability Tree A graphical representation showing the possible outcomes of a series of experiments and their respective probabilities.

Relative Frequency of Occurrence The proportion of times that an event occurs in the long run when conditions are stable, or the observed relative frequency of an event in a very large number of trials.

Sample Space The set of all possible outcomes of an experiment.

Statistical Dependence The condition when the probability of some event is dependent on, or affected by, the occurrence of some other event.

Statistical Independence The condition when the occurrence of one event has no effect on the probability of occurrence of another event.

Subjective Probability Probabilities based on the personal beliefs of the person making the probability estimate.

Venn Diagram A pictorial representation of probability concepts in which the sample space is represented as a rectangle and the events in the sample space as portions of that rectangle.

Equations Introduced in Chapter 4

$$4-1 \quad \text{Probability of an event} = \frac{\text{number of outcomes where the event occurs}}{\text{total number of possible outcomes}} \qquad \text{p. 158}$$

This is the definition of the *classical* probability that an event will occur.

$$P(A) = \text{probability of event } A \text{ happening} \qquad \text{p. 165}$$

A single probability refers to the probability of one particular event occurring, and it is called *marginal* probability.

$$P(A \text{ or } B) = \text{probability of either } A \text{ or } B \text{ happening} \quad \text{p. 167}$$

This notation represents the probability that one event *or* the other will occur.

$$4-2 \quad P(A \text{ or } B) = P(A) + P(B) - P(AB) \quad \text{p. 166}$$

The addition rule for events that are not mutually exclusive shows that the probability of *A* or *B* happening when *A* and *B* are not mutually exclusive is equal to the probability of event *A* happening plus the probability of event *B* happening minus the probability of *A* and *B* happening together, symbolized $P(AB)$.

$$4-3 \quad P(A \text{ or } B) = P(A) + P(B) \quad \text{p. 167}$$

The probability of either *A* or *B* happening when *A* and *B* are mutually exclusive equals the sum of the probability of event *A* happening and the probability of event *B* happening. This is the *addition rule for mutually exclusive events*.

$$4-4 \quad P(AB) = P(A) \times P(B) \quad \text{p. 172}$$

where

- $P(AB)$ = joint probability of events *A* and *B* occurring together or in succession
- $P(A)$ = marginal probability of event *A* happening
- $P(B)$ = marginal probability of event *B* happening

The *joint* probability of two or more *independent* events occurring together or in succession is the product of their marginal probabilities.

$$P(B | A) = \text{probability of event } B, \text{ given that event } A \text{ has happened} \quad \text{p. 176}$$

This notation shows *conditional* probability, the probability that a second event (*B*) will occur if a first event (*A*) has already happened.

$$4-5 \quad P(B | A) = P(B) \quad \text{p. 176}$$

For *statistically independent* events, the *conditional* probability of event *B*, given that event *A* has occurred, is simply the probability of event *B*. Independent events are those whose probabilities are in no way affected by the occurrence of each other.

$$4-6 \quad P(B|A) = \frac{P(BA)}{P(A)}$$

and

$$P(A|B) = \frac{P(AB)}{P(B)} \quad \text{p. 181}$$

For statistically *dependent* events, the *conditional* probability of event *B*, given that event *A* has occurred, is equal to the joint probability of events *A* and *B* divided by the marginal probability of event *A*.

$$4-7 \quad P(AB) = P(A|B) \times P(B)$$

and

$$P(BA) = P(B | A) \times P(A) \quad \text{p. 183}$$

Under conditions of statistical *dependence*, the *joint* probability of events *A* and *B* happening together or in succession is equal to the probability of event *A*, given that event *B* has already happened, multiplied by the probability that event *B* will happen.

Review and Application Exercises

- 4-53** Life insurance premiums are higher for older people, but auto insurance premiums are generally higher for younger people. What does this suggest about the risks and probabilities associated with these two areas of the insurance business?
- 4-54** “The chance of rain today is 80 percent.” Which of the following best explains this statement?
- It will rain 80 percent of the day today.
 - It will rain in 80 percent of the area to which this forecast applies today.
 - In the past, weather conditions of this sort have produced rain in this area 80 percent of the time.
- 4-55** “There is a 0.25 probability that a restaurant in the United States will go out of business this year.” When researchers make such statements, how have they arrived at their conclusions?
- 4-56** Using probability theory, explain the success of gambling and poker establishments.
- 4-57** Studies have shown that the chance of a new car being a “lemon” (one with multiple warranty problems) is greater for cars manufactured on Mondays and Fridays. Most consumers don’t know on which day their car was manufactured. Assuming a 5-day production week, for a consumer taking a car at random from a dealer’s lot,
- What is the chance of getting a car made on a Monday?
 - What is the chance of getting a car made on Monday or Friday?
 - What is the chance of getting a car made on Tuesday through Thursday?
 - What type of probability estimates are these?
- 4-58** Isaac T. Olduso, an engineer for Atlantic Aircraft, disagrees with his supervisor about the likelihood of landing-gear failure on the company’s new airliner. Isaac contends that the probability of landing-gear failure is 0.12, while his supervisor maintains that the probability is only 0.03. The two agree that if the landing gear fails, the airplane will crash with probability 0.55. Otherwise, the probability of a crash is only 0.06. A test flight is conducted, and the airplane crashes.
- Using Isaac’s figure, what is the probability that the airplane’s landing gear failed?
 - Repeat part (a) using the supervisor’s figure.
- 4-59** Congressman Bob Forehead has been thinking about the upcoming midterm elections and has prepared the following list of possible developments in his career during the midterm elections:
- He wins his party’s nomination for reelection.
 - He returns to his law practice.
 - He is nominated for vice president.
 - He loses his party’s nomination for reelection.
 - He wins reelection.
- Is each item on this list an “event” in the category of “Midterm Election Career Developments?”
 - Are all of the items qualifying as “events” in part (a) mutually exclusive? If not, are any mutually exclusive?
 - Are the events on the list collectively exhaustive?
- 4-60** Which of the following pairs of events are mutually exclusive?
- A defense department contractor loses a major contract, and the same contractor increases its work force by 50 percent.
 - A man is older than his uncle, and he is younger than his cousins.

- (c) A baseball team loses its last game of the year, and it wins the World Series.
 (d) A bank manager discovers that a teller has been embezzling, and she promotes the same teller.

4-61 The scheduling officer for a local police department is trying to decide whether to schedule additional patrol units in each of two neighborhoods. She knows that on any given day during the past year, the probabilities of major crimes and minor crimes being committed in the northern neighborhood were 0.478 and 0.602, respectively, and that the corresponding probabilities in the southern neighborhood were 0.350 and 0.523. Assume that major and minor crimes occur independently of each other and likewise that crimes in the two neighborhoods are independent of each other.

- (a) What is the probability that no crime of either type is committed in the northern neighborhood on a given day?
 (b) What is the probability that a crime of either type is committed in the southern neighborhood on a given day?
 (c) What is the probability that no crime of either type is committed in either neighborhood on a given day?

4-62 The Environmental Protection Agency is trying to assess the pollution effect of a paper mill that is to be built near Spokane, Washington. In studies of six similar plants built during the last year, the EPA determined the following pollution factors:

Plant	1	2	3	4	5	6
Sulfur dioxide emission in parts per million (ppm)	15	12	18	16	11	19

EPA defines excessive pollution as a sulfur dioxide emission of 18 ppm or greater.

- (a) Calculate the probability that the new plant will be an excessive sulfur dioxide polluter.
 (b) Classify this probability according to the three types discussed in the chapter: classical, relative frequency, and subjective.
 (c) How would you judge the accuracy of your result?

4-63 The American Cancer Society is planning to mail out questionnaires concerning breast cancer. From past experience with questionnaires, the Cancer Society knows that only 15 percent of the people receiving questionnaires will respond. It also knows that 1.3 percent of the questionnaires mailed out will have a mistake in address and never be delivered, that 2.8 percent will be lost or destroyed by the post office, that 19 percent will be mailed to people who have moved, and that only 48 percent of those who move leave a forwarding address.

- (a) Do the percentages in the problem represent classical, relative frequency, or subjective probability estimates?
 (b) Find the probability that the Cancer Society will get a reply from a given questionnaire.

4-64 McCormick and Tryon, Inc., is a "shark watcher," hired by firms fearing takeover by larger companies. This firm has found that one of its clients, Pare and Oyd Co., is being considered for takeover by two firms. The first, Engulf and Devour, considered 20 such companies last year and took over 7. The second, R. A. Venus Corp., considered 15 such companies last year and took over 6. What is the probability of Pare and Oyd's being taken over this year, assuming that

- (a) The acquisition rates of both Engulf and Devour and R. A. Venus are the same this year as they were last year?
 (b) This year's acquisition rates are independent of last year's?

In each case, assume that only one firm may take over Pare and Oyd.

- 4-65** As the administrator of a hospital, Cindy Turner wants to know what the probability is that a person checking into the hospital will require X-ray treatment and will also have hospital insurance that will cover the X-ray treatment. She knows that during the past 5 years, 23 percent of the people entering the hospital required X-rays, and that during the same period, 72 percent of the people checking into the hospital had insurance that covered X-ray treatments. What is the correct probability? Do any additional assumptions need to be made?
- 4-66** An air traffic controller at Dulles Airport must obey regulations that require her to divert one of two airplanes if the probability of the aircraft's colliding exceeds 0.025. The controller has two inbound aircraft scheduled to arrive 10 minutes apart on the same runway. She knows that Flight 100, scheduled to arrive first, has a history of being on time, 5 minutes late, and 10 minutes late 95, 3, and 2 percent of the time, respectively. Further, she knows that Flight 200, scheduled to arrive second, has a history of being on time, 5 minutes early, and 10 minutes early 97, 2, and 1 percent of the time, respectively. The flights' timings are independent of each other.
- Must the controller divert one of the planes, based on this information?
 - If she finds out that Flight 100 definitely will be 5 minutes late, must the controller divert one of the airplanes?
 - If the controller finds out that Flight 200 definitely will be 5 minutes early, must she divert one of the airplanes?
- 4-67** In a staff meeting called to address the problem of returned checks at the supermarket where you are interning as a financial analyst, the bank reports that 12 percent of all checks are returned for insufficient funds, and of those, in 50 percent of cases, there was cash given back to the customer. Overall, 10 percent of customers ask for cash back at the end of their transaction with the store. For 1,000 customer visits, how many transactions will involve:
- Insufficient funds?
 - Cash back to the customer?
 - Both insufficient funds and cash back?
 - Either insufficient funds or cash back?
- 4-68** Which of the following pairs of events are statistically independent?
- The times until failure of a calculator and of a second calculator marketed by a different firm.
 - The life-spans of the current U.S. and Russian presidents.
 - The amounts of settlements in asbestos poisoning cases in Maryland and New York.
 - The takeover of a company and a rise in the price of its stock.
 - The frequency of organ donation in a community and the predominant religious orientation of that community.
- 4-69** F. Liam Laytor, supervisor of customer relations for GLF Airlines, is studying his company's overbooking problem. He is concentrating on three late-night flights out of LaGuardia Airport in New York City. In the last year, 7, 8, and 5 percent of the passengers on the Atlanta, Kansas City, and Detroit flights, respectively, have been bumped. Further, 55, 20, and 25 percent of the late-night GLF passengers at LaGuardia take the Atlanta, Kansas City, and Detroit flights, respectively. What is the probability that a bumped passenger was scheduled to be on the
- Atlanta flight?
 - Kansas City flight?
 - Detroit flight?

- 4-70** An electronics manufacturer is considering expansion of its plant in the next 4 years. The decision depends on the increased production that will occur if either government or consumer sales increase. Specifically, the plant will be expanded if either (1) consumer sales increase 50 percent over the present sales level or (2) a major government contract is obtained. The company also believes that both these events will not happen in the same year. The planning director has obtained the following estimates:
- The probability of consumer sales increasing by 50 percent within 1, 2, 3, and 4 years is 0.05, 0.08, 0.12, and 0.16, respectively.
 - The probability of obtaining a major government contract within 1, 2, 3, and 4 years is 0.08, 0.15, 0.25, and 0.32, respectively.
- What is the probability that the plant will expand
- (a) Within the next year (in year 1)?
 - (b) Between 1 and 2 years from now (in year 2)?
 - (c) Between 2 and 3 years from now (in year 3)?
 - (d) Between 3 and 4 years from now (in year 4)?
 - (e) At all in the next 4 years (assume at most one expansion)?
- 4-71** Draw Venn diagrams to represent the following situations involving three events, A , B , and C , which are part of a sample space of events but do not include the whole sample space.
- (a) Each pair of events (A and B , A and C , and B and C) may occur together, but all three may not occur together.
 - (b) A and B are mutually exclusive, but not A and C nor B and C .
 - (c) A , B , and C are all mutually exclusive of one another.
 - (d) A and B are mutually exclusive, B and C are mutually exclusive, but A and C are not mutually exclusive.
- 4-72** Cartoonist Barry Bludeau sends his comics to his publisher via Union Postal Delivery. UPD uses rail and truck transportation in Mr. Bludeau's part of the country. In UPD's 20 years of operation, only 2 percent of the packages carried by rail and only 3.5 percent of the packages carried by truck have been lost. Mr. Bludeau calls the claims manager to inform him that a package containing a week of comics has been lost. If UPD sends 60 percent of the packages in that area by rail, which mode of transportation was more likely used to carry the lost comics? How does the solution change if UPD loses only 2 percent of its packages, regardless of the mode of transportation?
- 4-73** Determine the probability that
- (a) Both engines on a small airplane fail, given that each engine fails with probability 0.05 and that an engine is twice as likely to fail when it is the only engine working.
 - (b) An automobile is recalled for brake failure and has steering problems, given that 15 percent of that model were recalled for brake failure and 2 percent had steering problems.
 - (c) A citizen files his or her tax return and cheats on it, given that 70 percent of all citizens file returns and 25 percent of those who file cheat.
- 4-74** Two-fifths of clients at Show Me Realty come from an out-of-town referral network, the rest are local. The chances of selling a home on each showing are 0.075 and 0.053 for out-of-town and local clients, respectively. If a salesperson walks into Show Me's office and announces "It's a deal!" was the agent more likely to have conducted a showing for an out-of-town or local client?
- 4-75** A senior North Carolina senator knows he will soon vote on a controversial bill. To learn his constituents' attitudes about the bill, he met with groups in three cities in his state. An aide jotted down the opinions of 15 attendees at each meeting:

Opinion	City		
	Chapel Hill	Raleigh	Lumberton
Strongly oppose	2	2	4
Slightly oppose	2	4	3
Neutral	3	3	5
Slightly support	2	3	2
Strongly support	6	3	1
Total	15	15	15

- (a) What is the probability that someone from Chapel Hill is neutral about the bill? Strongly opposed?
- (b) What is the probability that someone in the three city groups strongly supports the bill?
- (c) What is the probability that someone from the Raleigh or Lumberton groups is neutral or slightly opposed?

4-76 The breakdown by political party of the 435 members of the U.S. House of Representatives before and after the 1992 Congressional elections was

	House Seats	
	Old	New
Democrats	268	259
Republicans	166	175
Independents	1	1

- (a) Determine the probability that a member selected at random before the 1992 election would be a Republican.
- (b) Determine the probability that a member selected at random after that election would not be a Republican.
- (c) Is it fair to conclude that the probability that a randomly selected Democratic incumbent was not re-elected was $9/268$? Explain.

4-77 A produce shipper has 10,000 boxes of bananas from Ecuador and Honduras. An inspection has determined the following information:

	# of Boxes	# of Boxes with	
		Damaged Fruit	Overripe Fruit
Ecuadoran	6,000	200	840
Honduran	4,000	365	295

- (a) What is the probability that a box selected at random will contain damaged fruit? Overripe fruit?
- (b) What is the probability that a randomly selected box is from Ecuador or Honduras?
- (c) Given that a randomly selected box contains overripe fruit, what is the probability that it came from Honduras?
- (d) If damaged fruit and overripe fruit are mutually exclusive, what is the probability that a box contains damaged or overripe fruit? What if they are not mutually exclusive?

4-78 Marcia Lerner will graduate in 3 months with a master's degree in business administration. Her school's placement office indicates that the probability of receiving a job offer as the result of any given on-campus interview is about 0.07 and is statistically independent from interview to interview.

- (a) What is the probability that Marcia will not get a job offer in any of her next three interviews?

- (b) If she has three interviews per month, what is the probability that she will have at least one job offer by the time she finishes school?
- (c) What is the probability that in her next five interviews she will get job offers on the third and fifth interviews only?

4-79 A standard set of pool balls contains 15 balls numbered from 1 to 15. Pegleg Woodhull, the famous blind poolplayer, is playing a game of 8-ball, in which the 8-ball must be, the last one hit into a pocket. He is allowed to touch the balls to determine their positions before taking a shot, but he does not know their numbers. Every shot Woodhull takes is successful.

- (a) What is the probability that he hits the 8-ball into a pocket, on his first shot, thus losing the game?
- (b) What is the probability that the 8-ball is one of the first three balls he hits?
- (c) What is the probability that Pegleg wins the game, that is, that the 8-ball is the last ball hit into a pocket?

4-80 BMT, Inc., is trying to decide which of two oil pumps to use in its new race car engine. One pump produces 75 pounds of pressure and the other 100. BMT knows the following probabilities associated with the pumps:

Probability of Engine Failure Due to		
	Seized Bearings	Ruptured Head Gasket
Pump A	0.08	0.03
Pump B	0.02	0.11

- (a) If seized bearings and ruptured head gaskets are mutually exclusive, which pump should BMT use?
- (b) If BMT devises a greatly improved “rupture-proof” head gasket, should it change its decision?

4-81 Sandy Irick is the public relations director for a large pharmaceutical firm that has been attacked in the popular press for distributing an allegedly unsafe vaccine. The vaccine protects against a virulent contagious disease that has a 0.04 probability of killing an infected person. Twenty-five percent of the population has been vaccinated.

A researcher has told her the following: The probability of any unvaccinated individual acquiring the disease is 0.30. Once vaccinated, the probability of acquiring the disease through normal means is zero. However, 2 percent of vaccinated people will show symptoms of the disease, and 3 percent of that group will die from it. Of people who are vaccinated and show no symptoms from the vaccination, 0.05 percent will die. Irick must draw some conclusions from these data for a staff meeting in 1 hour and a news conference later in the day.

- (a) If a person is vaccinated, what is the probability of dying from the vaccine? If he was not vaccinated, what is the probability of dying?
- (b) What is the probability of a randomly selected person dying from either the vaccine or the normally contracted disease?

4-82 The pressroom supervisor for a daily newspaper is being pressured to find ways to print the paper closer to distribution time, thus giving the editorial staff more leeway for last-minute changes. She has the option of running the presses at “normal” speed or at 110 percent of normal—“fast” speed. She estimates that they will run at the higher speed 60 percent of the time. The roll of paper (the newsprint “web”) is twice as likely to tear at the higher speed, which would mean temporarily stopping the presses,

- (a) If the web on a randomly selected printing run has a probability of 0.112 of tearing, what is the probability that the web will not tear at normal speed?

- (b) If the probability of tearing on fast speed is 0.20, what is the probability that a randomly selected torn web occurred on normal speed?
- 4-83** Refer to Exercise 4-83. The supervisor has noted that the web tore during each of the last four runs and that the speed of the press was not changed during these four runs. If the probabilities of tearing at fast and slow speeds were 0.14 and 0.07, respectively, what is the revised probability that the press was operating at fast speed during the last four runs?
- 4-84** A restaurant is experiencing discontentment among its customers. Historically it is known that there are three factors responsible for discontent amongst the customers viz. food quality, services quality, and interior décor. By conducting an analysis, it assesses the probabilities of discontentment with the three factors as 0.40, 0.35 and 0.25, respectively. By conducting a survey among customers, it also evaluates the probabilities of a customer going away discontented on account of these factors as 0.6, 0.8 and 0.5, respectively. The restaurant manager knows that a customer is discontented, what is the probability that it is due to service quality?
- 4-85** An economist believes that the chances of the Indian Rupee appreciating during period of high economic growth is 0.70, during moderate economic growth the chances of appreciation is 0.40, and during low economic growth it is 0.20. During any given time period the probability of high and moderate economic growth is 0.30 and 0.50 respectively. According to the RBI report the Rupee has been appreciating during the present period. What is the probability that the economy is experiencing a period of low economic growth?



Questions on Running Case: Academic Performance

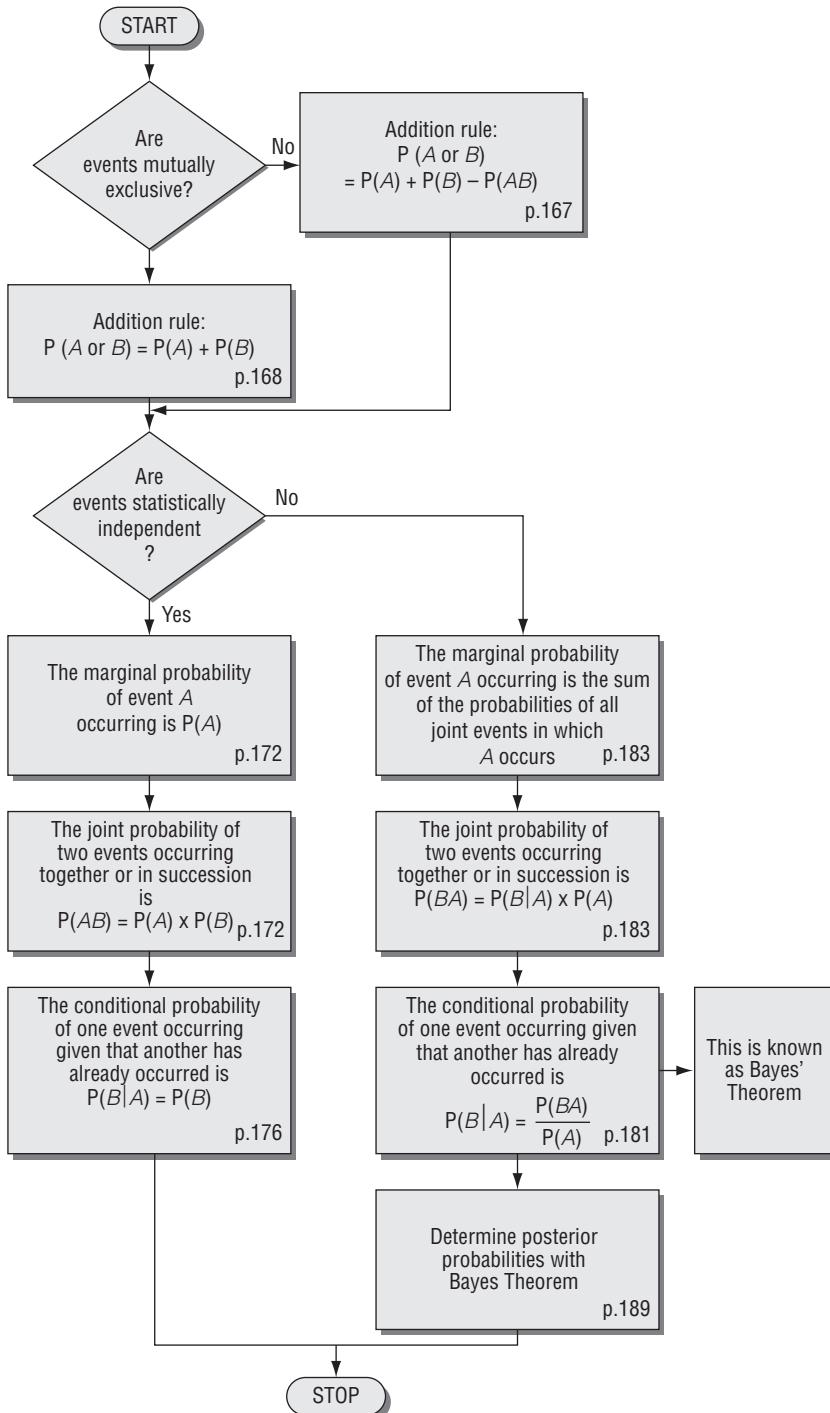
In the MBA-I Trimester of a college, XML Management School, there are 50 students. Their academic performance along with their gender and subject-stream has been noted down. The information is presented in the data sheet provide in Disk (Case_Academic Performance-Data.xls)

Answer the following questions:

1. If a student is randomly selected, what are the chances that she will be a male?
2. What is the probability that a randomly selected student has taken commerce stream in graduation?
3. What is the probability that a randomly selected student will be female and have taken professional stream in graduation?
4. What is the probability that a randomly selected female student has taken science stream in graduation?
5. What is the probability that a randomly selected arts student will be male?
6. What is the probability that a randomly selected student have secured at least 75% marks both in XII and graduation?
7. What is the probability that a randomly selected student has obtained less than 70% marks in XII provided he/she has more than 80% marks in X?
8. What is the probability that a randomly selected female student have secured more than 80 percentile in CAT if she has above 75% marks in graduation?
9. Are the events 'being male' and 'having science stream in graduation' independent?
10. A randomly selected student is found to be female, what are the chances that she has her CAT percentile in between 75 and 90?
11. A randomly selected student has got less than 65% marks in graduation, which event has more probability that the student would be a male or female?



Flow Chart: Probability I: Introductory Ideas



5 Probability Distributions

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To introduce the probability distributions most commonly used in decision making
 - To use the concept of expected value to make decisions
 - To show which probability distribution to use and how to find its values
 - To understand the limitations of each of the probability distributions you use
-

CHAPTER CONTENTS

- | | |
|---|---|
| 5.1 What is a Probability Distribution? 210 | ■ Statistics at Work 263 |
| 5.2 Random Variables 214 | ■ Terms Introduced in Chapter 5 265 |
| 5.3 Use of Expected Value in Decision Making 220 | ■ Equations Introduced in Chapter 5 265 |
| 5.4 The Binomial Distribution 225 | ■ Review and Application Exercises 266 |
| 5.5 The Poisson Distribution 238 | ■ Flow Chart: Probability Distributions 274 |
| 5.6 The Normal Distribution: A Distribution of a Continuous Random Variable 246 | |
| 5.7 Choosing the Correct Probability Distribution 263 | |

Modern filling machines are designed to work efficiently and with high reliability. Machines can fill toothpaste pumps to within 0.1 ounce of the desired level 80 percent of the time. A visitor to the plant, watching filled pumps being placed into cartons, asked, “What’s the chance that exactly half the pumps in a carton selected at random will be filled to within 0.1 ounce of the desired level?” Although we cannot make an exact forecast, the ideas about probability distributions discussed in this chapter enable us to give a pretty good answer to the question. ■

5.1 WHAT IS A PROBABILITY DISTRIBUTION?

In Chapter 2, we described frequency distributions as a useful way of summarizing variations in observed data. We prepared frequency distributions by listing all the possible outcomes of an experiment and then indicating the observed frequency of each possible outcome. *Probability distributions* are related to frequency distributions. **In fact, we can think of a probability distribution as a theoretical frequency distribution.** Now, what does that mean? A theoretical frequency distribution is a probability distribution that describes how outcomes are *expected* to vary. Because these distributions deal with expectations, they are useful models in making inferences and decisions under conditions of uncertainty. In later chapters, we will discuss the methods we use under these conditions.

Probability distributions and frequency distribution

Examples of Probability Distributions

To begin our study of probability distributions, let’s go back to the idea of a fair coin, which we introduced in Chapter 4. Suppose we toss a fair coin twice. Table 5-1 illustrates the possible outcomes from this two-toss experiment.

Experiment using a fair coin

Now suppose that we are interested in formulating a probability distribution of the number of tails that could possibly result when we toss the coin twice. We would begin by noting any outcome that did *not* contain a tail. With a fair coin, that is only the third outcome in Table 5-1: H, H. Then we would note the outcomes containing only one tail (the second and fourth outcomes in Table 5-1) and, finally, we would note that the first outcome contains two tails. In Table 5-2, we rearrange the outcomes of Table 5-1 to emphasize the number of tails contained in each outcome. We must be careful to note at this point that Table 5-2 is *not* the actual outcome of tossing a fair coin twice. Rather, it is a *theoretical* outcome, that is, it represents the way in which we would *expect* our two-toss experiment to behave over time.

TABLE 5-1 POSSIBLE OUTCOMES FROM TWO TOSSES OF A FAIR COIN

First Toss	Second Toss	Number of Tails on Two Tosses	Probability of the Four Possible Outcomes
T	T	2	$0.5 \times 0.5 = 0.25$
T	H	1	$0.5 \times 0.5 = 0.25$
H	H	0	$0.5 \times 0.5 = 0.25$
H	T	1	$0.5 \times 0.5 = 0.25$
			1.00

TABLE 5-2 PROBABILITY DISTRIBUTION OF THE POSSIBLE NUMBER OF TAILS FROM TWO TOSSES OF A FAIR COIN

Number of Tails, T	Tosses	Probability of This Outcome P(T)
0	(H, H)	0.25
1	(T, H) + (H, T)	0.50
2	(T, T)	0.25

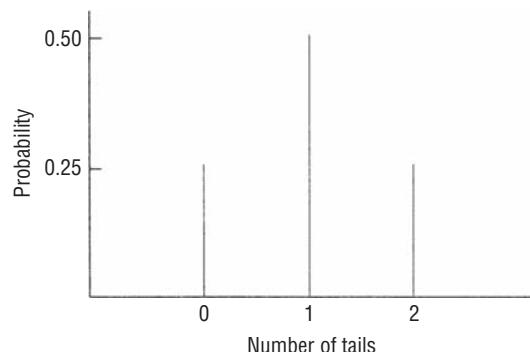


FIGURE 5-1 PROBABILITY DISTRIBUTION OF THE NUMBER OF TAILS IN TWO TOSSES OF A FAIR COIN

We can illustrate in graphic form the probability distribution in Table 5-2. To do this, we graph the number of tails we might see on two tosses against the probability that this number would happen. We show this graph in Figure 5-1.

Consider another example. A political candidate for local office is considering the votes she can get in a coming election.

Assume that votes can take on only four possible values. If the candidate's assessment is like this:

Number of votes	1,000	2,000	3,000	4,000	
Probability this will happen	0.1	0.3	0.4	0.2	Total 1.0

then the graph of the probability distribution representing her expectations will be like the one shown in Figure 5-2.

Before we move on to other aspects of probability distributions, we should point out that a **frequency distribution** is a listing of the observed frequencies of all the outcomes of an

Votoing Example
Difference between frequency distributions and probability distributions

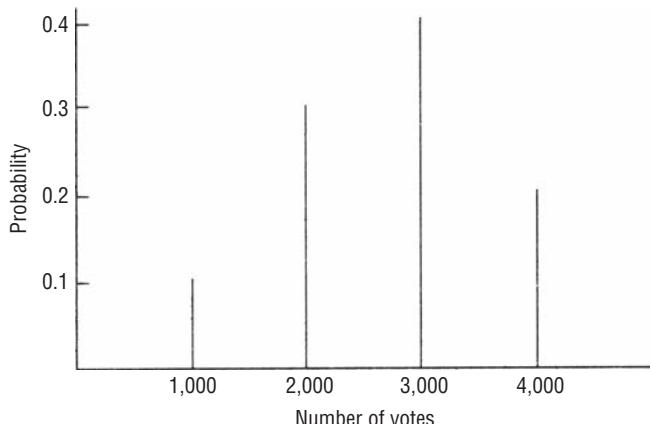


FIGURE 5-2 PROBABILITY DISTRIBUTION OF THE NUMBER OF VOTES

experiment that actually occurred when the experiment was done, whereas a probability distribution is a listing of the probabilities of all the possible outcomes that could result if the experiment were done. Also, as we can see in the two examples we presented in Figures 5-1 and 5-2, probability distributions can be based on theoretical considerations (the tosses of a coin) or on a subjective assessment of the likelihood of certain outcomes (the candidate's estimate). Probability distributions can also be based on experience. Insurance company actuaries determine insurance premiums, for example, by using long years of experience with death rates to establish probabilities of dying among various age groups.

Types of Probability Distributions

Probability distributions are classified as either *discrete* or *continuous*. A discrete probability can take on only a limited number of values, which can be listed. An example of a discrete probability distribution is shown in Figure 5-2, where we expressed the candidate's ideas about the coming election. There, votes could take on only four possible values (1,000, 2,000, 3,000, or 4,000). Similarly, the probability that you were born in a given month is also discrete because there are only 12 possible values (the 12 months of the year).

In a continuous probability distribution, on the other hand, the variable under consideration is allowed to take on any value within a given range, so we *cannot* list all the possible values. Suppose we were examining the level of effluent in a variety of streams, and we measured the level of effluent by parts of effluent per million parts of water. We would expect quite a continuous range of parts per million (ppm), all the way from very low levels in clear mountain streams to extremely high levels in polluted streams. In fact, it would be quite normal for the variable "parts per million" to take on an enormous number of values. We would call the distribution of this variable (ppm) a continuous distribution. Continuous distributions are convenient ways to represent discrete distributions that have many possible outcomes, all very close to each other.

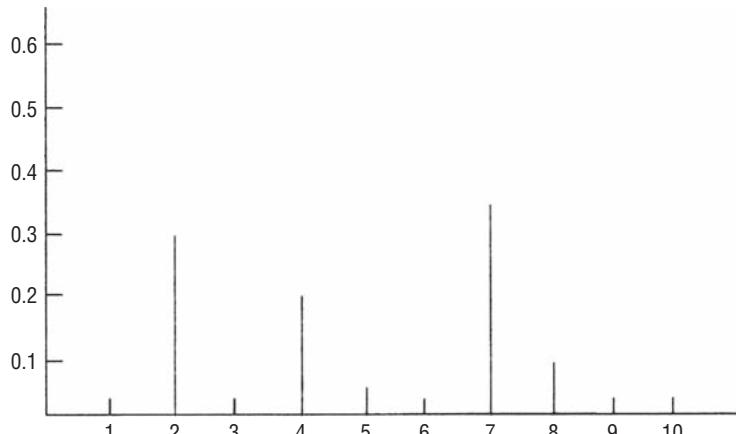
Discrete probability distributions

Continuous probability distributions

EXERCISES 5.1

Basic Concepts

- 5-1** Based on the following graph of a probability distribution, construct the corresponding table.



- 5-2** In the last chapter, we looked at the possible outcomes of tossing two dice, and we calculated some probabilities associated with various outcomes. Construct a table and a graph of the probability distribution representing the outcomes (in terms of total numbers of dots showing on both dice) for this experiment.
- 5-3** Which of the following statements regarding probability distributions are correct?
- A probability distribution provides information about the long-run or expected frequency of each outcome of an experiment.
 - The graph of a probability distribution has the possible outcomes of an experiment marked on the horizontal axis.
 - A probability distribution lists the probabilities that each outcome is random.
 - A probability distribution is always constructed from a set of observed frequencies like a frequency distribution.
 - A probability distribution may be based on subjective estimates of the likelihood of certain outcomes.

Applications

- 5-4** The regional chairman of the Muscular Dystrophy Association is trying to estimate the amount each caller will pledge during the annual MDA telethon. Using data gathered over the past 10 years, she has computed the following probabilities of various pledge amounts. Draw a graph illustrating this probability distribution.

Dollars pledged	25	50	75	100	125
Probability	0.45	0.25	0.15	0.10	0.05

- 5-5** Southport Autos offers a variety of luxury options on its cars. Because of the 6- to 8-week waiting period for customer orders, Ben Stoler, the dealer, stocks his cars with a variety of options. Currently, Mr. Stoler, who prides himself on being able to meet his customers' needs immediately, is worried because of an industrywide shortage of cars with V-8 engines. Stoler offers the following luxury combinations:

- | | | |
|-------------------------|-------------------|----------------------|
| 1. V-8 engine | electric sun roof | halogen headlights |
| 2. Leather interior | power door locks | stereo cassette deck |
| 3. Halogen headlights | V-8 engine | leather interior |
| 5. Stereo cassette deck | V-8 engine | power door locks |

Stoler thinks that combinations 2, 3, and 4 have an equal chance of being ordered, but that combination 1 is twice as likely to be ordered as any of these.

- What is the probability that any one customer ordering a luxury car will order one with a V-8 engine?
- Assume that two customers order luxury cars. Construct a table showing the probability distribution of the number of V-8 engines ordered.

- 5-6** Jim Rieck, a marketing analyst for Flatt and Mitney Aircraft, believes that the company's new Tigerhawk jet fighter has a 70 percent chance of being chosen to replace the U.S. Air Force's current jet fighter completely. However, there is one chance in five that the Air Force is going to buy only enough Tigerhawks to replace half of its 5,000 jet fighters. Finally, there is one chance in 10 that the Air Force will replace all of its jet fighters with Tigerhawks and will buy enough Tigerhawks to expand its jet fighter fleet by 10 percent. Construct a table and draw a graph of the probability distribution of sales of Tigerhawks to the Air Force.

5.2 RANDOM VARIABLES

A variable is random if it takes on different values as a result of the outcomes of a random experiment. A random variable can be either discrete or continuous. If a random variable is allowed to take on only a limited number of values, which can be listed, it is a *discrete random variable*. On the other hand, if it is allowed to assume any value within a given range, it is a *continuous random variable*.

You can think of a random variable as a value or magnitude that changes from occurrence to occurrence in no predictable sequence. A breast-cancer screening clinic, for example, has no way of knowing exactly how many women will be screened on any one day, so tomorrow's number of patients is a random variable. The values of a random variable are the numerical values corresponding to each possible outcome of the random experiment. If past daily records of the clinic indicate that the values of the random variable range from 100 to 115 patients daily, the random variable is a discrete random variable.

Table 5-3 illustrates the number of times each level has been reached during the last 100 days. Note that the table gives a frequency distribution. To the extent that we believe that the experience of the past 100 days has been typical, we can use this historical record to assign a probability to each possible number of patients and find a probability distribution. We have accomplished this in Table 5-4 by *normalizing* the observed frequency distribution.

TABLE 5-3 NUMBER OF WOMEN SCREENED DAILY DURING 100 DAYS

Number Screened	Number of Days This Level Was Observed
100	1
101	2
102	3
103	5
104	6
105	7
106	9
107	10
108	12
109	11
110	9
111	8
112	6
113	5
114	4
115	2
100	

Random variable defined

Example of discrete random variables

Creating a probability distribution

TABLE 5-4 PROBABILITY DISTRIBUTION FOR NUMBER OF WOMEN SCREENED

Number Screened (Value of the Random Variable)	Probability That the Random Variable Will Take on This Value
100	0.01
101	0.02
102	0.03
103	0.05
104	0.06
105	0.07
106	0.09
107	0.10
108	0.12
109	0.11
110	0.09
111	0.08
112	0.06
113	0.05
114	0.04
115	0.02
1.00	

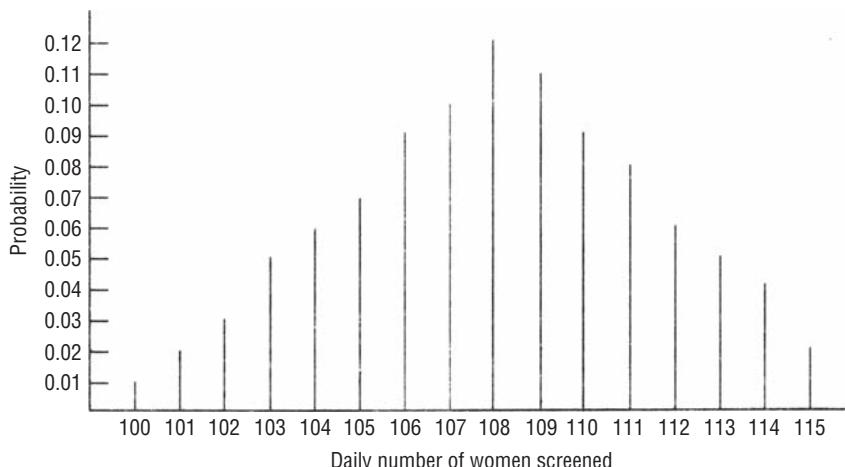


FIGURE 5-3 PROBABILITY DISTRIBUTION FOR THE DISCRETE RANDOM VARIABLE “DAILY NUMBER SCREENED”

(in this case, dividing each value in the right-hand column of Table 5-3 by 100, the total number of days for which the record has been kept). The probability distribution for the random variable “daily number screened” is illustrated graphically in Figure 5-3. Notice that the probability distribution for a random variable provides a probability for each possible value and that these probabilities must sum to 1. Table 5-4 shows that both these requirements have been met. Furthermore, both Table 5-4 and Figure 5-3 give us information about the long-run frequency of occurrence of daily patient screenings we would expect to observe if this random “experiment” were repeated.

The Expected Value of a Random Variable

Suppose you toss a coin 10 times and get 7 heads, like this:

Heads	Tails	Total
7	3	10

“Hmm, strange,” you say. You then ask a friend to try tossing the coin 20 times; she gets 15 heads and 5 tails. So now you have, in all, 22 heads and 8 tails out of 30 tosses.

What did you expect? Was it something closer to 15 heads and 15 tails (half and half)? Now suppose you turn the tossing over to a machine and get 792 heads and 208 tails out of 1,000 tosses of the same coin. You might now be suspicious of the coin because it didn’t live up to what you expected.

Expected value is a fundamental idea in the study of probability distributions. For many years, the concept has been put to considerable practical use by the insurance industry, and in the last 40 years, it has been widely used by many others who must make decisions under conditions of uncertainty.

To obtain the **expected value of a discrete random variable**, *Calculating expected value* we multiply each value that the random variable can assume by the probability of occurrence of that value and then sum these products. Table 5-5 illustrates this procedure for our clinic problem. The total in the table tells us that the expected value of the discrete random variable “number screened” is 108.02 women. What does this mean? It means that over a long period of time, the number of daily screenings should average about 108.02.

Remember that an-expected value of 108.02 does *not* mean that tomorrow exactly 108.02 women will visit the clinic.

The clinic director would base her decisions on the expected value of daily screenings because the expected value is a *weighted average of the outcomes she expects in the future*. Expected value *weights* each possible outcome by the frequency with which it is expected to occur. Thus, more common occurrences are given more weight than are less common ones. As conditions change over time, the director would recompute the expected value of daily screenings and use this new figure as a basis for decision making.

In our clinic example, the director used past patients' records as the basis for calculating the expected value of daily screenings. The expected value can also be derived from the director's subjective assessments of the probability that the random variable will take on certain values. In that case, the expected value represents nothing more than her personal convictions about the possible outcome.

Deriving expected value subjectively

In this section, we have worked with the probability distribution of a random variable in tabular form (Table 5-5) and in graphic form (Figure 5-3). In many situations, however, we will find it more convenient, in terms of the computations that must be done, to represent the probability distribution of a random variable in *algebraic* form. By doing this, we can make probability calculations by substituting numerical values directly into an algebraic formula. In the following sections, we shall illustrate some situations in which this is appropriate and methods for accomplishing it.

TABLE 5-5 CALCULATING THE EXPECTED VALUE OF THE DISCRETE RANDOM VARIABLE "DAILY NUMBER SCREENED"

Possible Values of the Random Variable	Probability That the Random Variable Will Take on These Values	
(1)	(2)	(1) × (2)
100	0.01	1.00
101	0.02	2.02
102	0.03	3.06
103	0.05	5.15
104	0.06	6.24
105	0.07	7.35
106	0.09	9.54
107	0.10	10.70
108	0.12	12.96
109	0.11	11.99
110	0.09	9.90
111	0.08	8.88
112	0.06	6.72
113	0.05	5.65
114	0.04	4.56
115	0.02	2.30

Expected value of the random variable "daily number screened" → **108.02**

HINTS & ASSUMPTIONS

The expected value of a discrete random variable is nothing more than the weighted average of each possible outcome, multiplied by the probability of that outcome happening, just like we did it in Chapter 3. Warning: The use of the term *expected* can be misleading. For example, if we calculated the expected value of number of women to be screened to be 11, we *don't* think exactly this many will show up tomorrow. We are saying that, absent any other information, 11 women is the best number we can come up with as a basis for planning how many nurses we'll need to screen them. Hint: If daily patterns in the data are discernible (more women on Monday than on Friday, for example) then build this into your decision. The same holds for monthly and seasonal patterns in the data.

EXERCISES 5.2

Self-Check Exercises

SC 5-1 Construct a probability distribution based on the following frequency distribution.

Outcome	102	105	108	111	114	117
Frequency	10	20	45	15	20	15

- (a) Draw a graph of the hypothetical probability distribution.
- (b) Compute the expected value of the outcome.

SC 5-2 Bob Walters, who frequently invests in the stock market, carefully studies any potential investment. He is currently examining the possibility of investing in the Trinity Power Company. Through studying past performance, Walters has broken the potential results of the investment into five possible outcomes with accompanying probabilities. The outcomes are annual rates of return on a single share of stock that currently costs \$150. Find the expected value of the return for investing in a single share of Trinity Power.

Return on investment (\$)	0.00	10.00	15.00	25.00	50.00
Probability	0.20	0.25	0.30	0.15	0.10

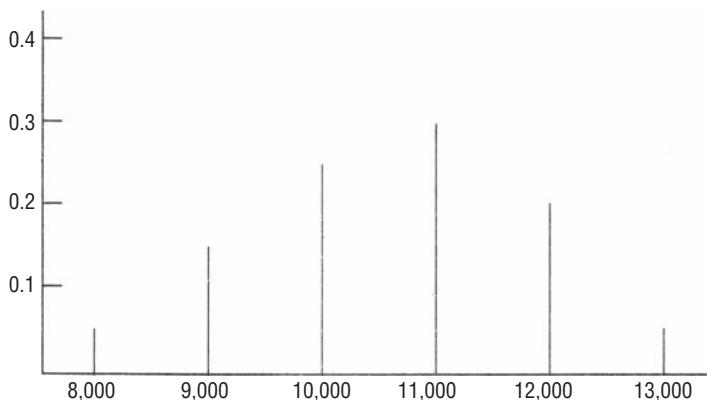
If Walters purchases stock whenever the expected rate of return exceeds 10 percent, will he purchase the stock, according to these data?

Basic Concepts

5-7 Construct a probability distribution based on the following frequency distribution:

Outcome	2	4	6	8	10	12	15
Frequency	24	22	16	12	7	3	1

- (a) Draw a graph of the hypothetical probability distribution.
 - (b) Compute the expected value of the outcome.
- 5-8** From the following graph of a probability distribution
- (a) Construct a table of the probability distribution.
 - (b) Find the expected value of the random variable.



- 5-9** The only information available to you regarding the probability distribution of a set of outcomes is the following list of frequencies:

X	0	15	30	45	60	75
Frequency	25	125	75	175	75	25

- (a) Construct a probability distribution for the set of outcomes.
(b) Find the expected value of an outcome.

Applications

- 5-10** Bill Johnson has just bought a VCR from Jim's Videotape Service at a cost of \$300. He now, has the option of buying an extended service warranty offering 5 years of coverage for \$100. After talking to friends and reading reports, Bill believes the following maintenance expenses could be incurred during the next five years:

Expense	0	50	100	150	200	250	300
Probability	0.35	0.25	0.15	0.10	0.08	0.05	0.02

Find the expected value of the anticipated maintenance costs. Should Bill pay \$100 for the warranty?

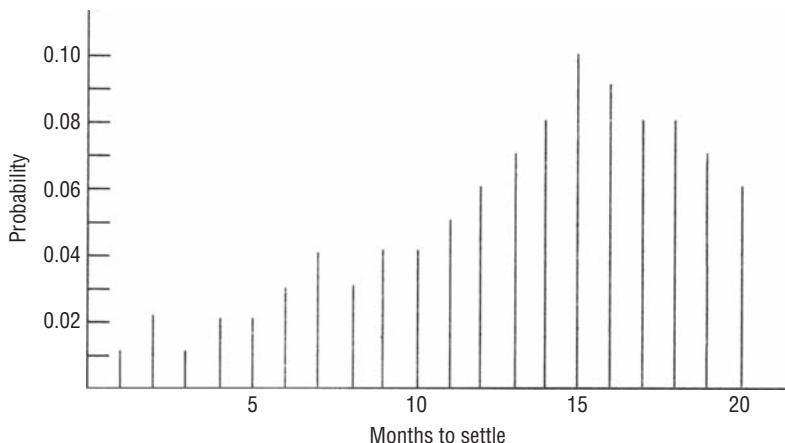
- 5-11** Steven T. Opsine, supervisor of traffic signals for the Fairfax County division of the Virginia State Highway Administration, must decide whether to install a traffic light at the reportedly dangerous intersection of Dolley Madison Blvd. and Lewinsville Rd. Toward this end, Mr. Opsine has collected data on accidents at the intersection:

Year	Number of Accidents											
	J	F	M	A	M	J	J	A	S	O	N	D
1995	10	8	10	6	9	12	2	10	10	0	7	10
1996	12	9	7	8	4	3	7	14	8	8	8	4

S.H.A. policy is to install a traffic light at an intersection at which the monthly expected number of accidents is higher than 7. According to this criterion, should Mr. Opsine recommend that a traffic light be installed at this intersection?

- 5-12** Alan Sarkid is the president of the Dinsdale Insurance Company and he is concerned about the high cost of claims that take a long time to settle. Consequently, he has asked his chief actuary,

Dr. Ivan Acke, to analyze the distribution of time until settlement. Dr. Acke has presented him with the following graph:



Dr. Acke also informed Mr. Sarkid of the expected amount of time to settle a claim. What is this figure?

5-13

The fire marshal of Baltimore County, Maryland, is compiling a report on single-family-dwelling fires. He has the following data on the number of such fires from the last 2 years:

Year	Number of Fires											
	J	F	M	A	M	J	J	A	S	O	N	D
1995	25	30	15	10	10	5	2	2	1	4	8	10
1996	20	25	10	8	5	2	4	0	5	8	10	15

Based on these data

- (a) What is the expected number of single-family-dwelling fires per month?
- (b) What is the expected number of single-family-dwelling fires per winter month (January, February, March)?

5-14

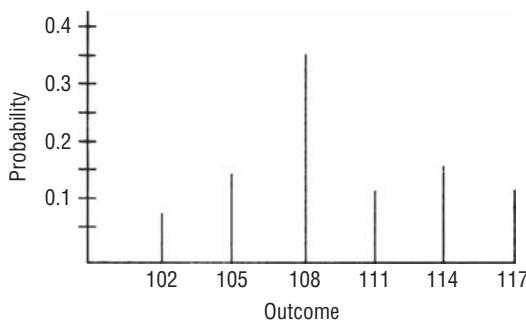
Ted Olson, the director of Overnight Delivery, Inc., has become concerned about the number of first-class letters lost by his firm. Because these letters are carried by both truck and airplane, Mr. Olson has broken down the lost letters for the last year into those lost from trucks and those lost from airplanes. His data are as follows:

Number Lost from	J	F	M	A	M	J	J	A	S	O	N	D
Truck	4	5	2	3	2	1	3	5	4	7	0	1
Airplane	5	6	0	2	1	3	4	2	4	7	4	0

Mr. Olson plans to investigate either the trucking or air division of the company, but not both. If he decides to investigate the division with the highest expected number of lost letters per month, which will he investigate?

Worked-Out Answers to Self-Check Exercises

SC 5-1 (a)



(b)

	Outcome (1)	Frequency (2)	P(Outcome) (3)	$(1) \times (3)$
	102	10	0.08	8.16
	105	20	0.16	16.80
	108	45	0.36	38.88
	111	15	0.12	13.32
	114	20	0.16	18.24
	117	15	0.12	14.04
	<u>125</u>		<u>1.00</u>	<u>109.44 = Expected outcome</u>

SC 5-2

	Return (1)	P(Return) (2)	$(1) \times (2)$
	0	0.20	0.00
	10	0.25	2.50
	15	0.30	4.50
	25	0.15	3.75
	50	0.10	5.00
	<u>1.00</u>		<u>15.75 = Expected return</u>

Bob will purchase the stock because the expected return of \$15.75 is greater than 10 percent of the \$150 purchase price.

5.3 USE OF EXPECTED VALUE IN DECISION MAKING

In the preceding section, we calculated the expected value of a random variable and noted that it can have significant value to decision makers. Now we need to take a moment to illustrate how decision makers combine the probabilities that a random variable will take on certain values with the monetary gain or loss that results when it does take on those values. Doing just this enables them to make intelligent decisions under uncertain conditions.

TABLE 5-6 SALES DURING 100 DAYS

Daily Sales	Number of Days Sold	Probability of Each Number Being Sold
10	15	0.15
11	20	0.20
12	40	0.40
13	25	0.25
	100	1.00

Combining Probabilities and Monetary Values

Let us look at the case of a fruit and vegetable wholesaler who sells strawberries. This product has a very limited useful life. If not sold on the day of delivery, it is worthless. One case of strawberries costs \$20, and the wholesaler receives \$50 for it. The wholesaler cannot specify the number of cases customers will call for on any one day, but her analysis of past records has produced the information in Table 5-6.

Wholesaler problem

Types of Losses Defined

Two types of losses are incurred by the wholesaler: (1) *obsolescence losses*, caused by stocking too much fruit on any one day and having to throw it away the next day; and (2) *opportunity losses*, caused by being out of strawberries any time that customers call for them. (Customers will not wait beyond the day a case is requested.)

Obsolescence and opportunity losses

Table 5-7 is a table of conditional losses. Each value in the table is conditional on a specific number of cases being stocked and a specific number being requested. The values in Table 5-7 include not only losses from decaying berries, but also those losses resulting from lost revenue when the wholesaler is unable to supply the requests she receives for the berries.

Table of conditional losses

Neither of these two types of losses is incurred when the number of cases stocked on any one day is the same as the number of cases requested. When that happens, the wholesaler sells all she has stocked and incurs no losses. This situation is indicated by a colored zero in the appropriate column. Figures above any zero represent losses arising from spoiled berries. In each case here, the number of cases stocked is greater than the number requested. For example, if the wholesaler stocks 12 cases but receives requests for only 10 cases, she loses \$40 (or \$20 per case for spoiled strawberries).

Obsolescence losses

TABLE 5-7 CONDITIONAL LOSS TABLE

Possible Requests for Strawberries	Possible Stock Options			
	10	11	12	13
10	\$0	\$20	\$40	\$60
11	30	0	20	40
12	60	30	0	20
13	90	60	30	0

TABLE 5-8 EXPECTED LOSS FROM STOCKING 10 CASES

Possible Requests	Conditional Loss		Probability of This Many Requests	=	Expected Loss
10	\$0	×	0.15	=	\$0.00
11	30	×	0.20	=	6.00
12	60	×	0.40	=	24.00
13	90	×	0.25	=	22.50
			<u>1.00</u>		<u>\$52.50</u>

Values **below** the colored zeros represent opportunity losses resulting from requests that cannot be filled. If only 10 cases are stocked on a day that 11 requests are received, the wholesaler suffers an opportunity loss of \$30 for the case she cannot sell (\$50 income per case that would have been received, minus \$20 cost, equals \$30).

Opportunity losses

Calculating Expected Losses

Examining each possible stock action, we can compute the expected loss. We do this by weighting each of the four possible loss figures in each column of Table 5-7 by the probabilities from Table 5-6. For a stock action of 10 cases, the expected loss is computed as in Table 5-8.

Meaning of expected loss

The conditional losses in Table 5-8 are taken from the second column of Table 5-7 for a stock action of 10 cases. The fourth column total in Table 5-8 shows us that if 10 cases are stocked each day, over a long period of time, the average or expected loss will be \$52.50 a day. There is no guarantee that tomorrow's loss will be exactly \$52.50.

Optimal solution

Tables 5-9 through 5-11 show the computations of the expected loss resulting from decisions to stock 11, 12, and 13 cases, respectively. **The optimal stock action is the one that minimize expected losses.** This action calls for the stocking of 12 cases each day, at which point the expected loss is minimized at \$17.50. We could just as easily have solved this problem by taking an alternative approach, that is, *maximizing expected gain* (\$50 received per case less \$20 cost per case) instead of minimizing expected loss. The answer, 12 cases, would have been the same.

TABLE 5-9 EXPECTED LOSS FROM STOCKING 11 CASES

Possible Requests	Conditional Loss		Probability of This Many Requests	=	Expected Loss
10	\$20	×	0.15	=	\$3.00
11	0	×	0.20	=	0.00
12	30	×	0.40	=	12.00
13	60	×	0.25	=	15.00
			<u>1.00</u>		<u>\$30.00</u>

TABLE 5-10 EXPECTED LOSS FROM STOCKING 12 CASES

Possible Requests	Conditional Loss		Probability of This Many Requests	=	Expected Loss
10	\$ 40	×	0.15	=	\$ 6.00
11	20	×	0.20	=	4.00
12	0	×	0.40	=	0.00
13	30	×	0.25	=	7.50
1.00 Minimum expected loss → \$17.50					

TABLE 5.11 EXPECTED LOSS FROM STOCKING 13 CASES

Possible Requests	Conditional Loss		Probability of This Many Requests	=	Expected Loss
10	\$ 60	×	0.15	=	\$9.00
11	40	×	0.20	=	8.00
12	20	×	0.40	=	8.00
13	0	×	0.25	=	0.00
1.00					\$25.00

In our brief treatment of expected value, we have made quite a few assumptions. To name only two, we've assumed that demand for the product can take on only four values, and that the berries are worth nothing one day later. Both these assumptions reduce the value of the answer we got. In Chapter 17, you will again encounter expected-value decision making, but there we will develop the ideas as a part of statistical decision theory (a broader use of statistical methods to make decisions), and we shall devote an entire chapter to expanding the basic ideas we have developed at this point.

HINTS & ASSUMPTIONS

Warning: In our illustrative exercise, we've allowed the random variable to take on only our values. This is unrealistic in the real world and we did it here only to make the explanation easier. Any manager facing this problem in her job would know that demand might be as low as zero on a given day (weather, holidays) and as high as perhaps 50 cases on another day. Hint: With demand ranging from zero to 50 cases, it's a computational nightmare to solve this problem by the method we just used. But don't panic, we will introduce another method in Chapter 17 that can do this easily.

EXERCISES 5.3

Self-Check Exercise

- SC 5-3** Mario, owner of Mario's Pizza Emporium, has a difficult decision on his hands. He has found that he always sells between one and four of his famous "everything but the kitchen sink" pizzas per night. These pizzas take so long to prepare, however, that Mario prepares all of them in advance and stores them in the refrigerator. Because the ingredients go bad within one day,

Mario always throws out any unsold pizzas at the end of each evening. The cost of preparing each pizza is \$7, and Mario sells each one for \$12. In addition to the usual costs, Mario also calculates that each “everything but” pizza that is ordered but he cannot deliver due to insufficient stock costs him \$5 in future business. How many “everything but” pizzas should Mario stock each night in order to minimize expected loss if the number of pizzas ordered has the following probability distribution?

Number of pizzas demanded	1	2	3	4
Probability	0.40	0.30	0.20	0.10

Applications

- 5-15** Harry Byrd, the director of publications for the Baltimore Orioles, is trying to decide how many programs to print for the team’s upcoming three-game series with the Oakland A’s. Each program costs 25¢ to print and sells for \$1.25. Any programs unsold at the end of the series must be discarded. Mr. Byrd has estimated the following probability distribution for program sales, using data from past program sales:

Programs sold	25,000	40,000	55,000	70,000
Probability	0.10	0.30	0.45	0.15

Mr. Byrd has decided to print either 25, 40, 55, or 70 thousand programs. Which number of programs will minimize the team’s expected losses?

- 5-16** Airport Rent-a-Car is a locally operated business in competition with several major firms. ARC is planning a new deal for prospective customers who want to rent a car for only one day and will return it to the airport. For \$35, the company will rent a small economy car to a customer, whose only expense is to fill the car with gas at day’s end. ARC is planning to buy number of small cars from the manufacturer at a reduced price of \$6,300. The big question is how many to buy. Company executives have decided the following distribution of demands per day for the service:

Number of cars rented	13	14	15	16	17	18
Probability	0.08	0.15	0.22	0.25	0.21	0.09

The company intends to offer the plan 6 days a week (312 days per year) and anticipates that its variable cost per car per day will be \$2.50. After the end of one year, the company expects to sell the cars and recapture 50 percent of the original cost. Disregarding the time value of money and any noncash expenses, use the expected-loss method to determine the optimal number of cars for ARC to buy.

- 5-17** We Care Air needs to make a decision about Flight 105. There are currently 3 seats reserved for last-minute customers, but the airline does not know if anyone will buy them. If they release the seats now, they know they will be able to sell them for \$250 each. Last-minute customers must pay \$475 per seat. The decision must be made now, and any number of seats may be released. We Care Air has the following probability distribution to help them:

Number of last-minute customers requesting seats	0	1	2	3
Probability	0.45	0.30	0.15	0.10

The company also counts a \$150 loss of goodwill for every last-minute customer who is turned away.

- How much revenue will be generated by releasing all 3 seats now?
- What is the company's expected net revenue (revenue less loss of goodwill) if 3 seats are released now?
- What is the company's expected net revenue if 2 seats are released now?
- How many seats should be released to maximize expected revenue?

Worked-Out Answer to Self-Check Exercise

SC 5-3

Loss Table					
	Pizzas Demanded				
	1	2	3	4	
Probability	0.4	0.3	0.2	0.1	
Pizzas Stocked					Expected Loss
1	0	10	20	30	10.0
2	7	0	10	20	6.8 ←
3	14	7	0	10	8.7
4	21	14	7	0	14.0

Mario should stock two “everything but” pizzas each night.

5.4 THE BINOMIAL DISTRIBUTION

One widely used probability distribution of a discrete random variable is the *binomial distribution*. It describes a variety of processes of interest to managers. The binomial distribution describes discrete, not continuous, data, resulting from an experiment known as a *Bernoulli process*, after the seventeenth-century Swiss mathematician Jacob Bernoulli. The tossing of a fair coin a fixed number of times is a Bernoulli process, and the outcomes of such tosses can be represented by the binomial probability distribution. The success or failure of interviewees on an aptitude test may also be described by a Bernoulli process. On the other hand, the frequency distribution of the lives of fluorescent lights in a factory would be measured on a continuous scale of hours and would not qualify as a binomial distribution.

The binomial distribution and Bernoulli processes

Use of the Bernoulli Process

We can use the outcomes of a fixed number of tosses of a fair coin as an example of a Bernoulli process. We can describe this process as follows:

Bernoulli process described

- Each trial (each toss, in this case) has only *two* possible outcomes: heads or tails, yes or no, success or failure.
- The probability of the outcome of any trial (toss) remains *fixed* over time. With a fair coin, the probability of heads remains 0.5 each toss regardless of the number of times the coin is tossed.
- The trials are *statistically independent*; that is, the outcome of one toss does not affect the outcome of any other toss.

Each Bernoulli process has its own characteristic probability. Take the situation in which historically seven-tenths of all people who applied for a certain type of job passed the job test. We would say that the characteristic probability here is 0.7, but we could describe our testing results as Bernoulli only if we felt certain that the proportion of those passing the test (0.7) remained constant over time. The other characteristics of the Bernoulli process would also have to be met, of course. Each test would have only two outcomes (success or failure), and the results of each test would have to be statistically independent.

Characteristic probability defined

In more formal language, the symbol p represents the probability of a success (in our example, 0.7), and the symbol q ($q = 1 - p$), the probability of a failure (0.3). To represent a certain number of successes, we will use the symbol r , and to symbolize the total number of trials, we use the symbol n . In the situations we will be discussing, the number of trials is fixed before the experiment is begun.

Using this language in a simple problem, we can calculate the chances of getting exactly two heads (in any order) on three tosses of a fair coin. Symbolically, we express the values as follows:

- p = characteristic probability or probability of success = 0.5
- $q = 1 - p$ = probability of failure = 0.5
- r = number of successes desired = 2
- n = number of trials undertaken = 3

We can solve the problem by using the *binomial formula*:

Binomial Formula

$$\text{Probability of } r \text{ successes in } n \text{ trials} = \frac{n!}{r!(n-r)!} p^r q^{n-r} \quad [5-1]$$

Although this formula may look somewhat complicated, it can be used quite easily. The symbol ! means *factorial*, which is computed as follows: 3! means $3 \times 2 \times 1$, or 6. To calculate 5!, we multiply $5 \times 4 \times 3 \times 2 \times 1 = 120$. Mathematicians define 0! as equal to 1. Using the binomial formula to solve our problem, we discover

$$\begin{aligned} \text{Probability of 2 successes in 3 trials} &= \frac{3!}{2!(3-2)!} (0.5)^2 (0.5)^1 \\ &= \frac{3 \times 2 \times 1}{(2 \times 1)(1 \times 1)} (0.5)^2 (0.5) \\ &= \frac{6}{2} (0.25)(0.5) \\ &= 0.375 \end{aligned}$$

Thus, there is a 0.375 probability of getting two heads on three tosses of a fair coin.

By now you've probably recognized that we can use the binomial distribution to determine the probabilities for the toothpaste pump problem we introduced at the beginning of this chapter. Recall that historically, eight-tenths of the pumps were correctly filled (successes). If we want to compute

the probability of getting exactly three of six pumps (half a carton) correctly filled, we can define our symbols this way:

$$\begin{aligned} p &= 0.8 \\ q &= 0.2 \\ r &= 3 \\ n &= 6 \end{aligned}$$

and then use the binomial formula as follows:

$$\text{Probability of } r \text{ successes in } n \text{ trials} = \frac{n!}{r!(n-r)!} p^r q^{n-r} \quad [5-1]$$

$$\begin{aligned} \text{Probability of 3 out of 6 pumps correctly filled} &= \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(3 \times 2 \times 1)} (0.8)^3 (0.2)^3 \\ &= \frac{720}{6 \times 6} (0.512)(0.008) \\ &= (20)(0.512)(0.008) \\ &= 0.08192 \end{aligned}$$

Of course, we *could* have solved these two problems using the probability trees we developed in Chapter 4, but for larger problems, trees become quite cumbersome. In fact, using the binomial formula (Equation 5-1) is no easy task when we have to compute the value of something like 19 factorial. For this reason, binomial probability tables have been developed, and we shall use them shortly.

Binomial tables are available

Some Graphic Illustrations of the Binomial Distribution

To this point, we have dealt with the binomial distribution only in terms of the binomial formula, but the binomial, like any other distribution, can be expressed graphically as well.

To illustrate several of these distributions, consider a situation at Kerr Pharmacy, where employees are often late. Five workers are in the pharmacy. The owner has studied the situation over a period of time and has determined that there is a 0.4 chance of any one employee being late and that they arrive independently of one another. How would we draw a binomial probability distribution illustrating the probabilities of 0, 1, 2, 3, 4, or 5 workers being late simultaneously? To do this, we would need to use the binomial formula, where

$$\begin{aligned} p &= 0.4 \\ q &= 0.6 \\ n &= 5^* \end{aligned}$$

and to make a separate computation for each r , from 0 through 5. Remember that, mathematically, any number to the zero power is defined as being equal to 1. Beginning with our binomial formula:

$$\text{Probability of } r \text{ late arrivals out of } n \text{ workers} = \frac{n!}{r!(n-r)!} p^r q^{n-r} \quad [5-1]$$

*When we define n , we look at the number of workers. The fact that there is a possibility that none will be late does not alter our choice of $n = 5$.

For $r = 0$, we get

$$\begin{aligned} P(0) &= \frac{5!}{0!(5-0)!}(0.4)^0(0.6)^5 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(1)(5 \times 4 \times 3 \times 2 \times 1)}(1)(0.6)^5 \\ &= \frac{120}{120}(1)(0.07776) \\ &= (1)(1)(0.07776) \\ &= 0.07776 \end{aligned}$$

For $r = 1$, we get

$$\begin{aligned} P(1) &= \frac{5!}{1!(5-1)!}(0.4)^1(0.6)^4 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(1)(4 \times 3 \times 2 \times 1)}(0.4)(0.6)^4 \\ &= \frac{120}{24}(0.4)(0.1296) \\ &= (5)(0.4)(0.1296) \\ &= 0.2592 \end{aligned}$$

For $r = 2$, we get

$$\begin{aligned} P(2) &= \frac{5!}{2!(5-2)!}(0.4)^2(0.6)^3 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)}(0.4)^2(0.6)^3 \\ &= \frac{120}{12}(0.16)(0.216) \\ &= (10)(0.03456) \\ &= 0.3456 \end{aligned}$$

For $r = 3$, we get

$$\begin{aligned} P(3) &= \frac{5!}{3!(5-3)!}(0.4)^3(0.6)^2 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(2 \times 1)}(0.4)^3(0.6)^2 \\ &= (10)(0.064)(0.36) \\ &= 0.2304 \end{aligned}$$

Using the formula to derive the binomial probability distribution

For $r = 4$, we get

$$\begin{aligned} P(4) &= \frac{5!}{4!(5-4)!}(0.4)^4(0.6)^1 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(1)}(0.4)^4(0.6) \\ &= (5)(0.0256)(0.6) \\ &= 0.0768 \end{aligned}$$

Finally, for $r = 5$, we get

$$\begin{aligned} P(5) &= \frac{5!}{5!(5-5)!}(0.4)^5(0.6)^0 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1)(1)}(0.4)^5(1) \\ &= (1)(0.01024)(1) \\ &= 0.01024 \end{aligned}$$

The binomial distribution for this example is shown graphically in Figure 5-4.

Without doing all the calculations involved, we can illustrate the general appearance of a family of binomial probability distributions. In Figure 5-5, for example, each distribution represents $n = 5$. In each case, the p and q have been changed and are noted beside each distribution. The probabilities in Figure 5-5 sum to slightly less than 1.0000 because of rounding. From Figure 5-5, we can make the following generalizations:

General appearance of binomial distributions

1. When p is small (0.1), the binomial distribution is skewed to the right.
2. As p increases (to 0.3, for example), the skewness is less noticeable.
3. When $p = 0.5$, the binomial distribution is symmetrical.
4. When p is larger than 0.5, the distribution is skewed to the left.

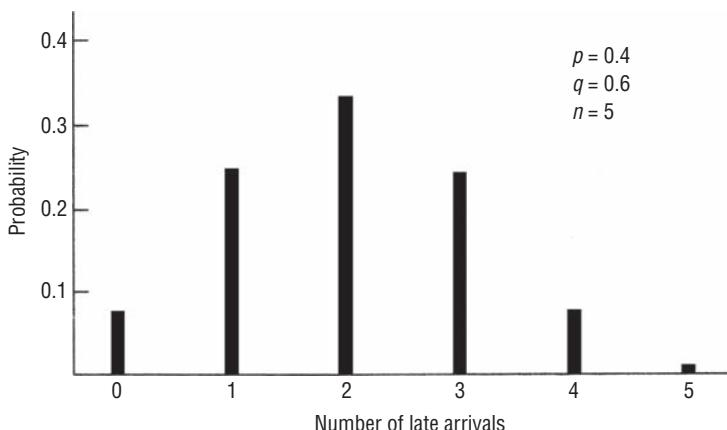
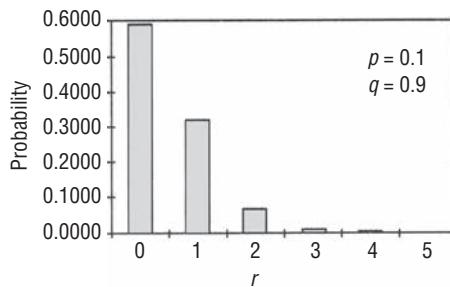
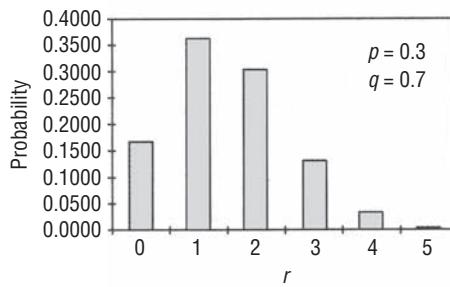


FIGURE 5-4 BINOMIAL PROBABILITY DISTRIBUTION OF LATE ARRIVALS

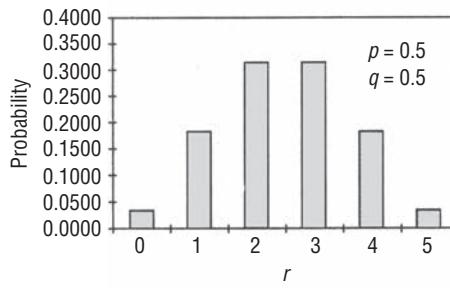
$n = 5, p = 0.1$	
r	Probability
0	0.5905
1	0.3280
2	0.0729
3	0.0081
4	0.0004
5	0.0000
	0.9999



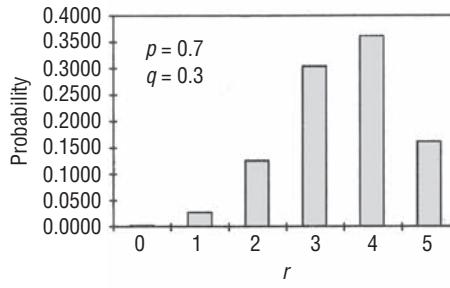
$n = 5, p = 0.3$	
r	Probability
0	0.1681
1	0.3601
2	0.3087
3	0.1323
4	0.0283
5	0.0024
	0.9999



$n = 5, p = 0.5$	
r	Probability
0	0.0312
1	0.1562
2	0.3125
3	0.3125
4	0.1562
5	0.0312
	0.9998



$n = 5, p = 0.7$	
r	Probability
0	0.0024
1	0.0283
2	0.1323
3	0.3087
4	0.3601
5	0.1681
	0.9999



$n = 5, p = 0.9$	
r	Probability
0	0.0000
1	0.0004
2	0.0081
3	0.0729
4	0.3280
5	0.5905
	0.9999

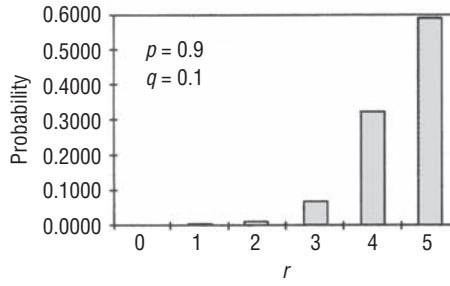


FIGURE 5-5 FAMILY OF BINOMIAL PROBABILITY DISTRIBUTIONS WITH CONSTANT $n = 5$ AND VARIOUS p AND q VALUES

5. The probabilities for 0.3, for example, are the same as those for 0.7 except that the values of p and q are *reversed*. This is true for any pair of complementary p and q values (0.3 and 0.7, 0.4 and 0.6, and 0.2 and 0.8).

Let us examine graphically what happens to the binomial distribution when p stays constant but n is increased. Figure 5-6 illustrates the general shape of a family of binomial distributions with a constant p of 0.4 and n 's from 5 to 30. As n increases, the vertical lines not only become more numerous but also tend to bunch up together to form a *bell shape*. We shall have more to say about this bell shape shortly.

Using the Binomial Tables

Earlier we recognized that it is tedious to calculate probabilities using the binomial formula when n is a large number. Fortunately, we can use Appendix Table 3 to determine binomial probabilities quickly.

Solving problems using the binomial tables

To illustrate the use of the binomial tables, consider this problem. What is the probability that 8 of the 15 registered Democrats on Prince Street will fail to vote in the coming primary if the probability of any individual's not voting is 0.30 and if people decide independently of each other whether or not to vote? First, we represent the elements in this problem in binomial distribution notation:

$$\begin{aligned} n &= 15 && \text{number of registered Democrats} \\ p &= 0.30 && \text{probability that any one individual won't vote} \\ r &= 8 && \text{number of individuals who will fail to vote} \end{aligned}$$

Then, because the problem involves 15 trials, we must find the table corresponding to $n = 15$. Because the probability of an individual's not voting is 0.30, we look through the binomial tables until we find the column headed 0.30. We then move down that column until we are opposite the $r = 8$ row, where we read the answer 0.0348. This is the probability of eight registered voters not voting.

How to use the binomial tables

Suppose the problem had asked us to find the probability of eight or more registered voters not voting? We would have looked under the 0.30 column and added up the probabilities there from 8 to the bottom of the column like this:

8	0.0348
9	0.0116
10	0.0030
11	0.0006
12	0.0001
13	0.0000
	0.0501

The answer is that there is a 0.0501 probability of eight or more registered voters not voting.

Suppose now that the problem asked us to find the probability of *fewer* than eight non-voters. Again, we would have begun with the 0.30 column, but this time we would add the probabilities from 0 (the top of the $n = 15$ column) *down* to 7 (the highest value less than 8), like this:

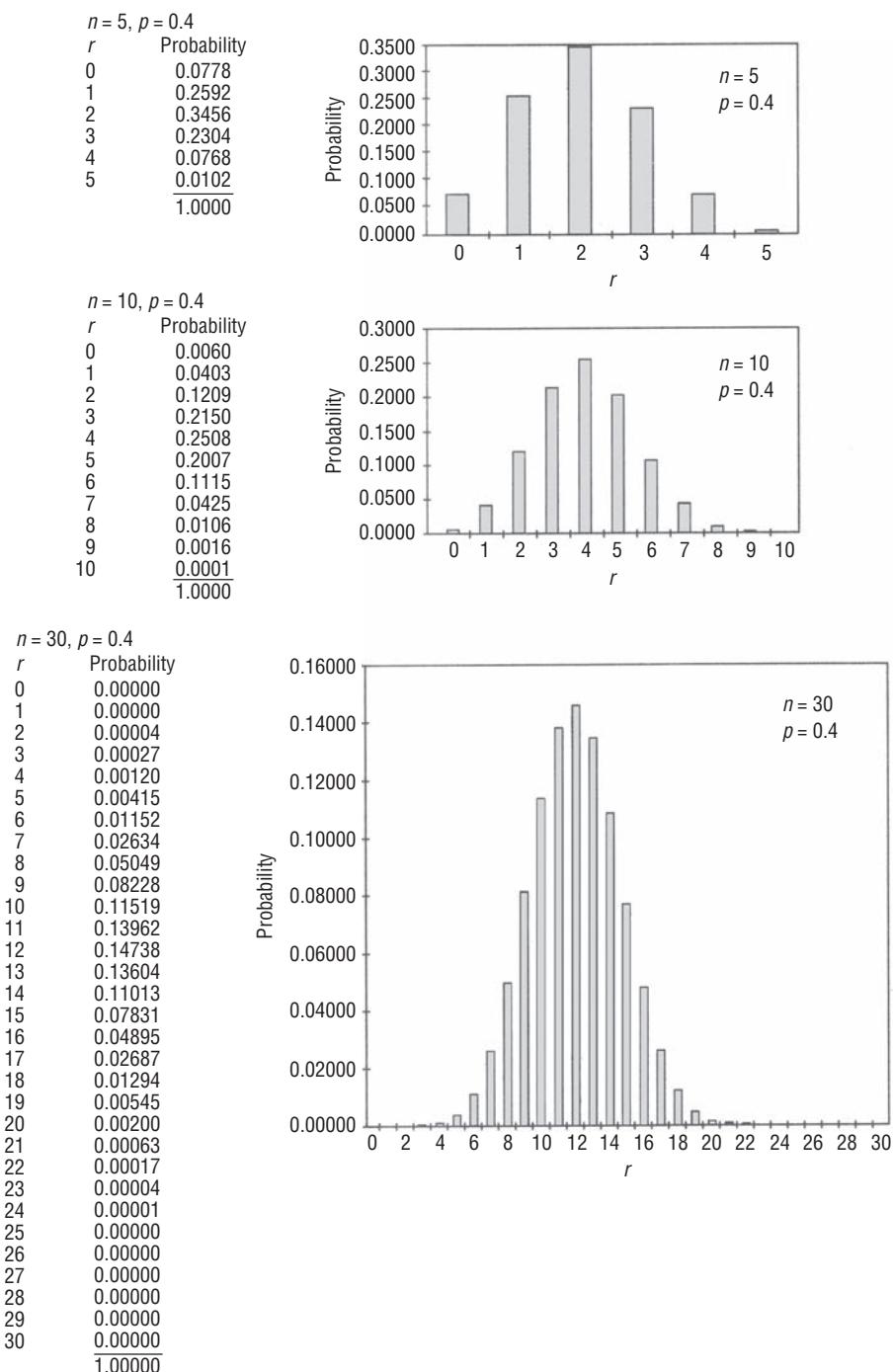


FIGURE 5-6 FAMILY OF BINOMIAL PROBABILITY DISTRIBUTIONS WITH CONSTANT $p = 0.4$ AND $n = 5, 10$, AND 30

0	0.0047
1	0.0305
2	0.0916
3	0.1700
4	0.2186
5	0.2061
6	0.1472
7	0.0811
	<hr/>
	0.9498

The answer is that there is a 0.9498 probability of fewer than eight nonvoters.

Because r (the number of nonvoters) is *either* 8 or more, *or else* fewer than 8, it must be true that

$$P(r \geq 8) + P(r < 8) = 1$$

But according to the values we just calculated,

$$P(r \geq 8) + P(r < 8) = 0.0501 + 0.9498 = 0.9999$$

The slight difference between 1 and 0.9999 is due to rounding errors resulting from the fact that the binomial table gives the probabilities to only 4 decimal places of accuracy.

You will see that the binomial table probabilities at the tops of the columns of figures go only up to 0.50. How do you solve problems with probabilities larger than 0.5? Simply go back through the binomial tables and look this time at the probability values at the *bottoms* of the columns; these go from 0.50 through 0.99.

Measures of Central Tendency and Dispersion for the Binomial Distribution

Earlier in this chapter, we encountered the concept of the expected value or mean of a probability distribution. The binomial distribution has an expected value or mean (μ) and a standard deviation (σ), and we should be able to compute both these statistical measures. Intuitively, we can reason that if a certain machine produces good parts with a $p = 0.5$, then, over time, the mean of the distribution of the number of good parts in the output would be 0.5 times the total output. If there is a 0.5 chance of tossing a head with a fair coin, over a large number of tosses, the mean of the binomial distribution of the number of heads would be 0.5 times the total number of tosses.

Computing the mean and the standard deviation

Symbolically, we can represent the mean of a binomial distribution as

Mean of a Binomial Distribution

$$\mu = np$$

[5-2]

The mean

where

- n = number of trials
- p = probability of success

And we can calculate the standard deviation of a binomial distribution by using the formula

Standard Deviation of a Binomial Distribution

$$\sigma = \sqrt{npq} \quad [5-3]$$

The standard deviation

where

- n = number of trials
- p = probability of success
- q = probability of failure = $1 - p$

To see how to use Equations 5-2 and 5-3, take the case of a packaging machine that produces 20 percent defective packages. If we take a random sample of 10 packages, we can compute the mean and the standard deviation of the binomial distribution of that process like this:

$$\mu = np \quad [5-2]$$

$$= (10)(0.2)$$

$$= 2 \leftarrow \text{Mean}$$

$$\sigma = \sqrt{npq} \quad [5-3]$$

$$= \sqrt{(10)(0.2)(0.8)}$$

$$= \sqrt{1.6}$$

$$= 1.265 \leftarrow \text{Standard deviation}$$

Meeting the Conditions for Using the Bernoulli Process

We need to be careful in the use of the binomial probability to make certain that the three conditions necessary for a Bernoulli process introduced earlier are met, particularly conditions 2 and 3. Condition 2 requires the probability of the outcome of any trial to remain fixed over time. In many industrial processes, however, it is extremely difficult to guarantee that this is indeed the case. Each time an industrial machine produces a part, for instance, there is some infinitesimal wear on the machine. If this wear accumulates beyond a reasonable point, the proportion of acceptable parts produced by the machine will be altered and condition 2 for the use of the binomial distribution may be violated. This problem is not present in a coin-toss experiment, but it is an integral consideration in all real applications of the binomial probability distribution.

Condition 3 requires that the trials of a Bernoulli process be statistically independent, that is, the outcome of one trial cannot affect in any way the outcome of any other trial. Here, too, we can encounter some problems in real applications. Consider an interviewing process in which high-potential candidates are being screened for top positions. If the interviewer has talked with five unacceptable candidates in

Problems in applying the binomial distribution to real-life situations

a row, he may not view the sixth with complete impartiality. The trials, therefore, might not be statistically independent.

HINTS & ASSUMPTIONS

Warning: One of the requirements for using a Bernoulli process is that the probability of the outcome must be fixed over time. This is a very difficult condition to meet in practice. Even a fully automatic machine making parts will experience some wear as the number of parts increases and this will affect the probability of producing acceptable parts. Still another condition for its use is that the trials (manufacture of parts in our machine example) be independent. This too is a condition that is hard to meet. If our machine produces a long series of bad parts, this could affect the position (or sharpness) of the metal-cutting tool in the machine. Here, as in every other situation, going from the textbook to the real world is often difficult, and smart managers use their experience and intuition to know when a Bernoulli process is appropriate.

EXERCISES 5.4

Self-Check Exercises

SC 5-4 For a binomial distribution with $n = 12$ and $p = 0.45$, use Appendix Table 3 to find

- (a) $P(r = 8)$.
- (b) $P(r > 4)$.
- (c) $P(r \leq 10)$.

SC 5-5 Find the mean and standard deviation of the following binomial distributions:

- (a) $n = 16, p = 0.40$.
- (b) $n = 10, p = 0.75$.
- (c) $n = 22, p = 0.15$.
- (d) $n = 350, p = 0.90$.
- (e) $n = 78, p = 0.05$.

SC 5-6 The latest nationwide political poll indicates that for Americans who are randomly selected, the probability that they are conservative is 0.55, the probability that they are liberal is 0.30, and the probability that they are middle-of-the-road is 0.15. Assuming that these probabilities are accurate, answer the following questions pertaining to a randomly chosen group of 10 Americans. (Do not use Appendix Table 3.)

- (a) What is the probability that four are liberal?
- (b) What is the probability that none are conservative?
- (c) What is the probability that two are middle-of-the-road?
- (d) What is the probability that at least eight are liberal?

Basic Concepts

5-18 For a binomial distribution with $n = 7$ and $p = 0.2$, find

- (a) $P(r = 5)$.
- (b) $P(r > 2)$.

- (c) $P(r < 8)$.
- (d) $P(r \geq 4)$.

5-19 For a binomial distribution with $n = 15$ and $p = 0.2$, use Appendix Table 3 to find

- (a) $P(r = 6)$.
- (b) $P(r \geq 11)$.
- (c) $P(r \leq 4)$.

5-20 Find the mean and standard deviation of the following binomial distributions:

- (a) $n = 15, p = 0.20$.
- (b) $n = 8, p = 0.42$.
- (c) $n = 72, p = 0.06$.
- (d) $n = 29, p = 0.49$.
- (e) $n = 642, p = 0.21$.

5-21 For $n = 8$ trials, compute the probability that $r \geq 1$ for each of the following values of p :

- (a) $p = 0.1$.
- (b) $p = 0.3$.
- (c) $p = 0.6$.
- (d) $p = 0.4$.

Applications

5-22 Harley Davidson, director of quality control for the Kyoto Motor company, is conducting his monthly spot check of automatic transmissions. In this procedure, 10 transmissions are removed from the pool of components and are checked for manufacturing defects. Historically, only 2 percent of the transmissions have such flaws. (Assume that flaws occur independently in different transmissions.)

- (a) What is the probability that Harley's sample contains more than two transmissions with manufacturing flaws? (Do not use the tables.)
- (b) What is the probability that none of the selected transmissions has any manufacturing flaws? (Do not use the tables.)

5-23 Diane Bruns is the mayor of a large city. Lately, she has become concerned about the possibility that large numbers of people who are drawing unemployment checks are secretly employed. Her assistants estimate that 40 percent of unemployment beneficiaries fall into this category, but Ms. Bruns is not convinced. She asks one of her aides to conduct a quiet investigation of 10 randomly selected unemployment beneficiaries.

- (a) If the mayor's assistants are correct, what is the probability that more than eight of the individuals investigated have jobs? (Do not use the tables.)
- (b) If the mayor's assistants are correct, what is the probability that one or three of the investigated individuals have jobs? (Do not use the tables.)

5-24 A month later, Mayor Bruns (from Exercise 5-23) picks up the morning edition of the city's leading newspaper, the *Sun-American*, and reads an exposé of unemployment fraud. In this article, the newspaper claims that out of every 15 unemployment beneficiaries, the probability that four or more have jobs is 0.9095, and the expected number of employed beneficiaries exceeds 7. You are a special assistant to Mayor Bruns, who must respond to these claims at an afternoon press conference. She asks you to find the answers to the following two questions:

- (a) Are the claims of the *Sun-American* consistent with each other?
- (b) Does the first claim conflict with the opinion of the mayor's assistants?

- 5-25** A recent study of how Americans spend their leisure time surveyed workers employed more than 5 years. They determined the probability an employee has 2 weeks of vacation time to be 0.45, 1 week of vacation time to be 0.10, and 3 or more weeks to be 0.20. Suppose 20 workers are selected at random. Answer the following questions without Appendix Table 3.
- What is the probability that 8 have 2 weeks of vacation time?
 - What is the probability that only one worker has 1 week of vacation time?
 - What is the probability that at most 2 of the workers have 3 or more weeks of vacation time?
 - What is the probability that at least 2 workers have 1 week of vacation time?
- 5-26** Harry Ohme is in charge of the electronics section of a large department store. He has noticed that the probability that a customer who is just browsing will buy something is 0.3. Suppose that 15 customers browse in the electronics section each hour. Use Appendix Table 3 in the back of the book to answer the following questions:
- What is the probability that at least one browsing customer will buy something during a specified hour?
 - What is the probability that at least four browsing customers will buy something during a specified hour?
 - What is the probability that no browsing customers will buy anything during a specified hour?
 - What is the probability that no more than four browsing customers will buy something during a specified hour?

Worked-Out Answers to Self-Check Exercises

SC 5-4 Binomial ($n = 12, p = 0.45$).

- $P(r = 8) = 0.0762$
- $P(r > 4) = 1 - P(r \leq 4) = 1 - (0.0008 + 0.0075 + 0.0339 + 0.0923 + 0.1700) = 0.6955$
- $P(r \leq 10) = 1 - P(r \geq 11) = 1 - (0.0010 + 0.0001) = 0.9989$

SC 5-5	n	p	$\mu = np$	$\sigma = \sqrt{npq}$
(a)	16	0.40	6.4	1.960
(b)	10	0.75	7.5	1.369
(c)	22	0.15	3.3	1.675
(d)	350	0.90	315.0	5.612
(e)	78	0.05	3.9	1.925

SC 5-6 (a) $n = 10, p = 0.30, P(r = 4) = \left(\frac{10!}{4!6!}\right)(0.30)^4(0.70)^6 = 0.2001$

(b) $n = 10, p = 0.55, P(r = 0) = \left(\frac{10!}{0!10!}\right)(0.55)^0(0.45)^{10} = 0.0003$

$$(c) \quad n = 10, p = 0.15, P(r = 2) = \left(\frac{10!}{2!8!} \right) (0.15)^2 (0.85)^8 = 0.2759$$

$$(d) \quad n = 10, p = 0.30, P(r - 8) = P(r = 8) + P(r = 9) + P(r = 10)$$

$$= \left(\frac{10!}{8!2!} \right) (0.30)^8 (0.70)^2 + \left(\frac{10!}{9!1!} \right) (0.30)^9 (0.70)^1 + \left(\frac{10!}{10!0!} \right) (0.30)^{10} (0.70)^0 \\ = 0.00145 + 0.00014 + 0.00001 = 0.0016$$

5.5 THE POISSON DISTRIBUTION

There are many discrete probability distributions, but our discussion will focus on only two: the *binomial*, which we have just concluded, and the *Poisson*, which is the subject of this section. The Poisson distribution is named for Siméon Denis Poisson (1781–1840), a French mathematician who developed the distribution from studies during the latter part of his lifetime.

The Poisson distribution is used to describe a number of processes, including the distribution of telephone calls going through a switchboard system, the demand (needs) of patients for service at a health institution, the arrivals of trucks and cars at a tollbooth, and the number of accidents at an intersection. These examples all have a common element: They can be described by a discrete random variable that takes on integer (whole) values (0, 1, 2, 3, 4, 5, and so on). The number of patients who arrive at a physician's office in a given interval of time will be 0, 1, 2, 3, 4, 5, or some other whole number. Similarly, if you count the number of cars arriving at a tollbooth on the New Jersey Turnpike during some 10-minute period, the number will be 0, 1, 2, 3, 4, 5, and so on.

Examples of Poisson distributions

Characteristics of Processes That Produce a Poisson Probability Distribution

The number of vehicles passing through a single turnpike toll-booth at rush hour serves as an illustration of Poisson probability distribution characteristics:

Conditions leading to a Poisson probability distribution

1. The average (mean) number of vehicles that arrive per rush hour can be estimated from past traffic data.
2. If we divide the rush hour into periods (intervals) of one second each, we will find these statements to be true:
 - (a) The probability that exactly one vehicle will arrive at the single booth per second is a very small number and is constant for every one-second interval.
 - (b) The probability that two or more vehicles will arrive within a one-second interval is so small that we can assign it a zero value.
 - (c) The number of vehicles that arrive in a given one-second interval is independent of the time at which that one-second interval occurs during the rush hour.
 - (d) The number of arrivals in any one-second interval is not dependent on the number of arrivals in any other one-second interval.

Now, we can generalize from these four conditions described for our tollbooth example and apply them to other processes. If these new processes meet the same four conditions, then we can use a Poisson probability distribution to describe them.

Calculating Poisson Probabilities Using Appendix Table 4a

The Poisson probability distribution, as we have explained, is concerned with certain processes that can be described by a discrete random variable. The letter X usually represents that discrete random variable, and X can take on integer values (0, 1, 2, 3, 4, 5, and so on). We use capital X to represent the random variable and lowercase x to represent a specific value that capital X can take. The probability of exactly x occurrences in a Poisson distribution is calculated with the formula

Poisson Formula
$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$ [5-4]

Poisson distribution formula

Look more closely at each part of this formula:

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

An example using the Poisson formula

Suppose that we are investigating the safety of a dangerous intersection. Past police records indicate a mean of five accidents per month at this intersection. The number of accidents is distributed according to a Poisson distribution, and the Highway Safety

Division wants us to calculate the probability in any month of exactly 0, 1, 2, 3, or 4 accidents. We can use Appendix Table 4a to avoid having to calculate e 's to negative powers. Applying the formula

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!} \quad [5-4]$$

we can calculate the probability of no accidents:

$$\begin{aligned} P(0) &= \frac{(5)^0(e^{-5})}{0!} \\ &= \frac{(1)(0.00674)}{1} \\ &= 0.00674 \end{aligned}$$

For exactly one accident:

$$\begin{aligned} P(1) &= \frac{(5)^1 \times (e^{-5})}{1!} \\ &= \frac{(5)(0.00674)}{1} \\ &= 0.03370 \end{aligned}$$

For exactly two accidents:

$$\begin{aligned} P(2) &= \frac{(5)^2 (e^{-5})}{2!} \\ &= \frac{(25)(0.00674)}{2 \times 1} \\ &= 0.08425 \end{aligned}$$

For exactly three accidents:

$$\begin{aligned} P(3) &= \frac{(5)^3 (e^{-5})}{3!} \\ &= \frac{(125)(0.00674)}{3 \times 2 \times 1} \\ &= \frac{0.8425}{6} \\ &= 0.14042 \end{aligned}$$

Finally, for exactly four accidents:

$$\begin{aligned} P(4) &= \frac{(5)^4 (e^{-5})}{4!} \\ &= \frac{(625)(0.00674)}{4 \times 3 \times 2 \times 1} \\ &= \frac{4.2125}{24} \\ &= 0.17552 \end{aligned}$$

Our calculations will answer several questions. Perhaps we want to know the probability of 0, 1, or 2 accidents in any month. We find this by adding the probabilities of exactly 0, 1, and 2 accidents like this:

Using these results

$$\begin{aligned} P(0) &= 0.00674 \\ P(1) &= 0.03370 \\ P(2) &= 0.08425 \\ P(0 \text{ or } 1 \text{ or } 2) &= \underline{\underline{0.12469}} \end{aligned}$$

We will take action to improve the intersection if the probability of more than three accidents per month exceeds 0.65. Should we act? To solve this problem, we need to calculate the probability of having

0, 1, 2, or 3 accidents and then subtract the sum from 1.0 to get the probability for more than 3 accidents. We begin like this:

$$\begin{aligned} P(0) &= 0.00674 \\ P(1) &= 0.03370 \\ P(2) &= 0.08425 \\ P(3) &= \underline{0.14042} \\ P(3 \text{ or fewer}) &= \mathbf{0.26511} \end{aligned}$$

Because the Poisson probability of three or fewer accidents is 0.26511, the probability of more than three must be 0.73489, $(1.00000 - 0.26511)$. Because 0.73489 exceeds 0.65, steps should be taken to improve the intersection.

We could continue calculating the probabilities for more than four accidents and eventually produce a Poisson probability distribution of the number of accidents per month at this intersection. Table 5-12 illustrates such a distribution. To produce this table, we have used Equation 5-4. Try doing the calculations yourself for the probabilities beyond exactly four accidents. Figure 5-7 illustrates graphically the Poisson probability distribution of the number of accidents.

Constructing a Poisson probability distribution

Looking Up Poisson Probabilities Using Appendix Table 4b

Fortunately, hand calculations of Poisson probabilities are not necessary. Appendix Table 4b produces the same result as hand calculation but avoids the tedious work.

TABLE 5-12 POISSON PROBABILITY DISTRIBUTION OF ACCIDENTS PER MONTH

$x = \text{Number of Accidents}$	$P(x) = \text{Probability of Exactly That Number}$
0	0.00674
1	0.03370
2	0.08425
3	0.14042
4	0.17552
5	0.17552
6	0.14627
7	0.10448
8	0.06530
9	0.03628
10	0.01814
11	0.00824
	0.99486 ← Probability of 0 through 11 accidents
12 or more	0.00514 ← Probability of 12 or more $(1.0 - 0.99486)$
	1.00000

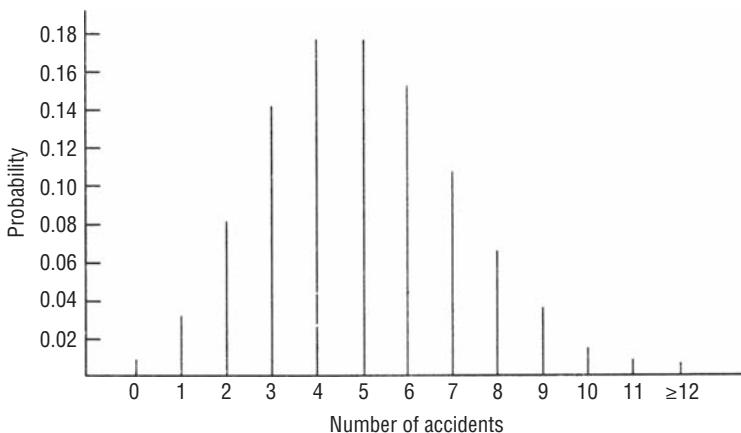


FIGURE 5-7 POISSON PROBABILITY DISTRIBUTION OF THE NUMBER OF ACCIDENTS

Look again at our intersection problem first introduced on page 239. There we calculated the probability of four accidents this way:

$$\begin{aligned} P(x) &= \frac{\lambda^x \times e^{-\lambda}}{x!} \\ P(4) &= \frac{(5)^4 (e^{-5})}{4!} \\ &= \frac{(625)(0.00674)}{4 \times 3 \times 2 \times 1} \\ &= 0.17552 \end{aligned} \quad [5-4]$$

To use Appendix Table 4b all we need to know are the values for x and λ , in this instance 4 and 5, respectively. Now look in Appendix Table 4b. First find the column headed 5; then come down the column until you are opposite 4, and read the answer directly, 0.1755. That's much less work isn't it?

Using Appendix Table 4b to look up Poisson probabilities

One more example will make sure we've mastered this new method. On page 241, we calculated the Poisson probability of 0, 1, or 2 accidents as being 0.12469. Finding this same result using Appendix Table 4b requires that we again look for the column headed 5, then come down that column, and add up the values we find beside 0, 1, and 2 like this:

- 0.0067 (Probability of 0 accidents)
- 0.0337 (Probability of 1 accident)
- 0.0842 (Probability of 2 accidents)
- 0.1246** (Probability of 0, 1, or 2 accidents)

Once again, the slight differences in the two answers are due to rounding errors.

Poisson Distribution as an Approximation of the Binomial Distribution

Sometimes, if we wish to avoid the tedious job of calculating binomial probability distributions, we can use the Poisson instead. The Poisson distribution can be a reasonable approximation of the binomial, but only under certain conditions. These conditions occur when n is large and p is small, that is, when the number of trials is large and the binomial probability of success is small. **The rule most often used by statisticians is that the Poisson is a good approximation of the binomial when n is greater than or equal to 20 and p is less than or equal to 0.05.** In cases that meet these conditions, we can substitute the mean of the binomial distribution (np) in place of the mean of the Poisson distribution (λ) so that the formula becomes

Using a modification of the Poisson formula to approximate binomial probabilities

Poisson Probability Distribution as an Approximation of the Binomial

$$P(x) = \frac{(np)^x \times e^{-np}}{x!} \quad [5-5]$$

Let us use both the binomial probability formula (5-1) and the Poisson approximation formula (5-5) on the same problem to determine the extent to which the Poisson is a good approximation of the binomial. Say that we have a hospital with 20 kidney dialysis machines and that the chance of any one of them malfunctioning during any day is 0.02. What is the probability that exactly three machines will be out of service on the same day? Table 5-13 shows the answers to this question. As we can see, the difference between the two probability distributions is slight (only about a 10 percent error, in this example).

Comparing the Poisson and binomial formulas

TABLE 5-13 COMPARISON OF POISSON AND BINOMIAL PROBABILITY APPROACHES TO THE KIDNEY DIALYSIS SITUATION

Poisson Approach	Binomial Approach
$P(x) = \frac{(np)^x \times e^{-np}}{x!}$ [5-5]	$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$ [5-1]
$P(3) = \frac{(20 \times 0.02)^3 e^{-(20 \times 0.02)}}{3!}$ $= \frac{(0.4)^3 e^{-0.4}}{3 \times 2 \times 1}$ $= \frac{(0.064)(0.67032)}{6}$ $= 0.00715$	$P(3) = \frac{20!}{3!(20-3)!} (0.02)^3 (0.98)^{17}$ $= 0.0065$

HINTS & ASSUMPTIONS

Statisticians look for situations where one distribution (Poisson, for example) whose probabilities are relatively easy to calculate can be substituted for another (binomial) whose probabilities are somewhat cumbersome to calculate. Even though a slight bit of accuracy is often lost in doing this, the time trade-off is favorable. When we do this, we assume that the Poisson distribution is a good approximation of the binomial distribution, but we qualify our assumption by requiring n to be greater than or equal to 20 and p to be less than or equal to 0.05. Assumptions based on such proven statistical values will not get us into trouble.

EXERCISES 5.5**Self-Check Exercises**

SC 5-7 Given $\lambda = 4.2$, for a Poisson distribution, find

- (a) $P(x \leq 2)$.
- (b) $P(x \geq 5)$.
- (c) $P(x = 8)$.

SC 5-8 Given a binomial distribution with $n = 30$ trials and $p = 0.04$, use the Poisson approximation to the binomial to find

- (a) $P(r = 25)$.
- (b) $P(r = 3)$.
- (c) $P(r = 5)$.

Basic Concepts

5-27 Given a binomial distribution with $n = 28$ trials and $p = 0.025$, use the Poisson approximation to the binomial to find

- (a) $P(r \geq 3)$.
- (b) $P(r < 5)$.
- (c) $P(r = 9)$.

5-28 If the prices of new cars increase an average of four times every 3 years, find the probability of

- (a) No price hikes in a randomly selected period of 3 years.
- (b) Two price hikes.
- (c) Four price hikes.
- (d) Five or more.

5-29 Given a binomial distribution with $n = 25$ and $p = 0.032$, use the Poisson approximation to the binomial to find

- (a) $P(r = 3)$
- (b) $P(r = 5)$
- (c) $P(r \leq 2)$

5-30 Given $\lambda = 6.1$ for a Poisson distribution, find

- (a) $P(x \leq 3)$
- (b) $P(x \geq 2)$

- (c) $P(x = 6)$
- (d) $P(1 \leq x \leq 4)$

Applications

- 5-31** Concert pianist Donna Prima has become quite upset at the number of coughs occurring in the audience just before she begins to play. On her latest tour, Donna estimates that on average eight coughs occur just before the start of her performance. Ms. Prima has sworn to her conductor that if she hears more than five coughs at tonight's performance, she will refuse to play. What is the probability that she will play tonight?
- 5-32** Guy Ford, production supervisor for the Winstead Company's Charlottesville plant, is worried about an elderly employee's ability to keep up the minimum work pace. In addition to the normal daily breaks, this employee stops for short rest periods an average of 4.1 times per hour. The rest period is a fairly consistent 3 minutes each time. Ford has decided that if the probability of the employee resting for 12 minutes (not including normal breaks) or more per hour is greater than 0.5, he will move the employee to a different job. Should he do so?
- 5-33** On average, five birds hit the Washington Monument and are killed each week. Bill Garcy, an official of the National Parks Service, has requested that Congress allocate funds for equipment to scare birds away from the monument. A Congressional subcommittee has replied that funds cannot be allocated unless the probability of more than three birds being killed in week exceeds 0.7. Will the funds be allocated?
- 5-34** Southwestern Electronics has developed a new calculator that performs a series of functions not yet performed by any other calculator. The marketing department is planning to demonstrate this calculator to a group of potential customers, but it is worried about some initial problems, which have resulted in 4 percent of the new calculators developing mathematical inconsistencies. The marketing VP is planning on randomly selecting a group of calculators for this demonstration and is worried about the chances of selecting a calculator that could start malfunctioning. He believes that whether or not a calculator malfunctions is a Bernoulli process, and he is convinced that the probability of a malfunction is really about 0.04.
- (a) Assuming that the VP selects exactly 50 calculators to use in the demonstration, and using the Poisson distribution as an approximation of the binomial, what is the chance of getting at least three calculators that malfunction?
 - (b) No calculators malfunctioning?
- 5-35** The Orange County Dispute Settlement Center handles various kinds of disputes, but most are marital disputes. In fact, 96 percent of the disputes handled by the DSC are of a marital nature.
- (a) What is the probability that, out of 80 disputes handled by the DSC, exactly seven are nonmarital?
 - (b) None are nonmarital?
- 5-36** The U.S. Bureau of Printing and Engraving is responsible for printing this country's paper money. The BPE has an impressively small frequency of printing errors; only 0.5 percent of all bills are too flawed for circulation. What is the probability that out of a batch of 1,000 bills
- (a) None are too flawed for circulation?
 - (b) Ten are too flawed for circulation?
 - (c) Fifteen are too flawed for circulation?

Worked-Out Answers to Self-Check Exercises

SC 5-7 $\lambda = 4.2$, $e^{-4.2} = 0.0150$.

$$(a) P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$= \frac{(4.2)^0 e^{-4.2}}{0!} + \frac{(4.2)^1 e^{-4.2}}{1!} + \frac{(4.2)^2 e^{-4.2}}{2!}$$

$$= 0.0150 + 0.0630 + 0.1323 = 0.2103$$

$$(b) P(x \geq 5) = 1 - P(x \leq 4) = 1 - P(x = 4) - P(x = 3) - P(x \leq 2)$$

$$= 1 - \frac{(4.2)^4 e^{-4.2}}{4!} - \frac{(4.2)^3 e^{-4.2}}{3!} - 0.2103$$

$$= 1 - 0.1944 - 0.1852 - 0.2103 = 0.4101$$

$$(c) P(x = 8) = \frac{(4.2)^8 e^{-4.2}}{8!} = 0.0360$$

SC 5-8 Binomial, $n = 30$, $p = 0.04$; $\lambda = np = 1.2$; $e^{-1.2} = 0.30119$.

$$(a) P(r = 25) = \frac{(1.2)^{25} e^{-1.2}}{25!} = 0.0000$$

$$(b) P(r = 3) = \frac{(1.2)^3 e^{-1.2}}{3!} = 0.0867$$

$$(c) P(r = 5) = \frac{(1.2)^5 e^{-1.2}}{5!} = 0.0062$$

5.6 THE NORMAL DISTRIBUTION: A DISTRIBUTION OF A CONTINUOUS RANDOM VARIABLE

So far in this chapter, we have been concerned with discrete probability distributions. In this section, we shall turn to cases in which the variable can take on *any* value within a given range and in which the probability distribution is continuous.

Continuous distribution defined

A very important continuous probability distribution is the *normal distribution*. Several mathematicians were instrumental in its development, including the eighteenth-century mathematician–astronomer Karl Gauss. In honor of his work, the normal probability distribution is often called the Gaussian distribution.

There are two basic reasons why the normal distribution occupies such a prominent place in statistics. First, it has some properties that make it applicable to a great many situations in which it is necessary to make inferences by taking samples. In Chapter 6, we will find that the normal distribution is a useful sampling distribution. Second, the normal distribution comes close to fitting the actual observed frequency distributions of many phenomena, including human characteristics (weights, heights, and IQs), outputs from physical processes (dimensions and yields), and other measures of interest to managers in both the public and private sectors.

Importance of the normal distribution

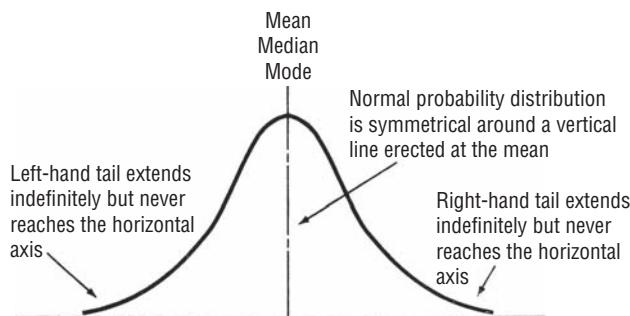


FIGURE 5-8 FREQUENCY CURVE FOR THE NORMAL PROBABILITY DISTRIBUTION

Characteristics of the Normal Probability Distribution

Look for a moment at Figure 5-8. This diagram suggests several important features of a normal probability distribution:

1. The curve has a single peak; thus, it is unimodal. It has the bell shape that we described earlier.
2. The mean of a normally distributed population lies at the center of its normal curve.
3. Because of the symmetry of the normal probability distribution, the median and the mode of the distribution are also at the center; thus, for a normal curve, the mean, median, and mode are the same value.
4. The two tails of the normal probability distribution extend indefinitely and never touch the horizontal axis. (Graphically, of course, this is impossible to show.)

Most real-life populations do not extend forever in both directions, but for such populations the normal distribution is a convenient approximation. There is no single normal curve, but rather a family of normal curves. To define a particular normal probability distribution, we need only two parameters: the mean (μ) and the standard deviation (σ). In Table 5-14, each of the populations is described only by its mean and its standard deviation, and each has a particular normal curve.

Significance of the two parameters that describe a normal distribution

Figure 5-9 shows three normal probability distributions, each of which has the same mean but a different standard deviation. Although these curves differ in appearance, all three are “normal curves.”

TABLE 5-14 DIFFERENT NORMAL PROBABILITY DISTRIBUTIONS

Nature of the Population	Its Mean	Its Standard Deviation
Annual earnings of employees at one plant	\$17,000/year	\$1,000
Length of standard 8' building lumber	8'	0.05"
Air pollution in one community	2,500 particles per million	750 particles per million
Per capita income in a single developing country	\$1,400	\$300
Violent crimes per year in a given city	8,000	900

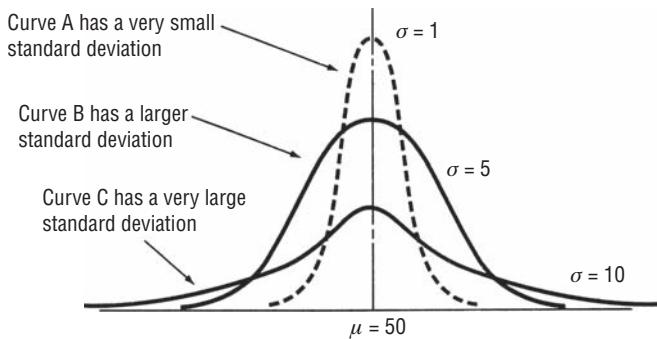


FIGURE 5-9 NORMAL PROBABILITY DISTRIBUTIONS WITH IDENTICAL MEANS BUT DIFFERENT STANDARD DEVIATIONS

Figure 5-10 illustrates a “family” of normal curves, all with the same standard deviation, but each with a different mean.

Finally, Figure 5-11 shows three different normal probability distributions, each with a different mean *and* a different standard deviation. The normal probability distributions illustrated in Figures 5-9,

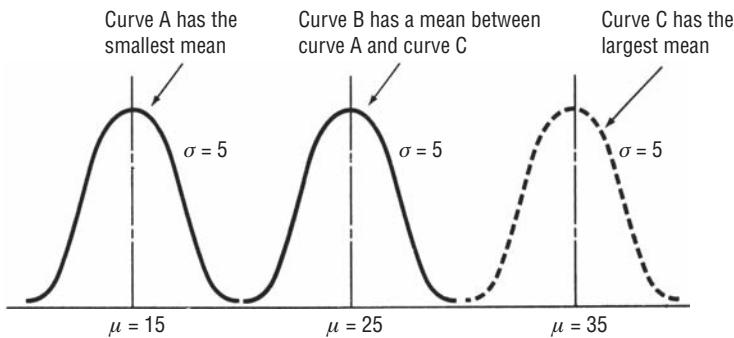


FIGURE 5-10 NORMAL PROBABILITY DISTRIBUTION WITH DIFFERENT MEANS BUT THE SAME STANDARD DEVIATION

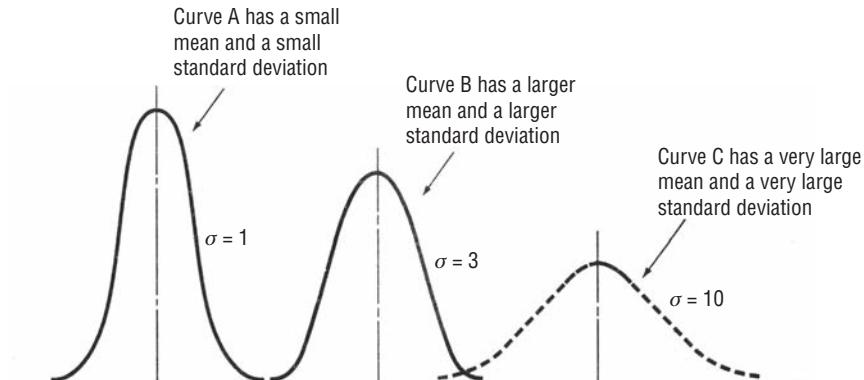


FIGURE 5-11 THREE NORMAL PROBABILITY DISTRIBUTIONS, EACH WITH A DIFFERENT MEAN AND A DIFFERENT STANDARD DEVIATION

5-10, and 5-11 demonstrate that the normal curve can describe a large number of populations, differentiated only by the mean and/or the standard deviation.

Areas under the Normal Curve

No matter what the values of μ and σ are for a normal probability distribution, the total area under the normal curve is 1.00, so that we may think of areas under the curve as probabilities.

Measuring the area under a normal curve

Mathematically, it is true that

1. Approximately 68 percent of all the values in a normally distributed population lie within ± 1 standard deviation from the mean.
2. Approximately 95.5 percent of all the values in a normally distributed population lie within ± 2 standard deviations from the mean.
3. Approximately 99.7 percent of all the values in a normally distributed population lie within ± 3 standard deviations from the mean.

These three statements are shown graphically in Figure 5-12.

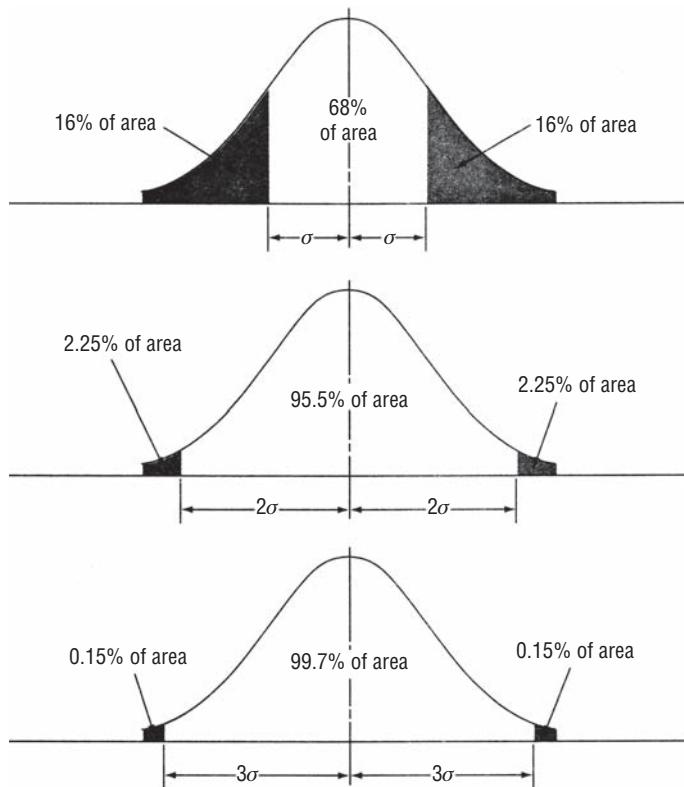


FIGURE 5-12 RELATIONSHIP BETWEEN THE AREA UNDER THE CURVE FOR A NORMAL PROBABILITY DISTRIBUTION AND THE DISTANCE FROM THE MEAN MEASURED IN STANDARD DEVIATIONS

Figure 5-12 shows three different ways of measuring the area under the normal curve. However, very few of the applications we shall make of the normal probability distribution involve intervals of *exactly* 1, 2, or 3 standard deviations (plus and minus) from the mean. What should we do about all these other cases? Fortunately, we can refer to statistical tables constructed for precisely these situations. They indicate portions of the area under the normal curve that are contained within any number of standard deviations (plus and minus) from the mean.

It is not possible or necessary to have a different table for every possible normal curve. Instead, we can use a table of the *standard normal probability distribution* (a normal distribution with $(\mu=0$ and $\sigma=1$) to find areas under any normal curve. With this table, we can determine the area, or probability, that the normally distributed random variable will lie within certain distances from the mean. These distances are defined in terms of standard deviations.

Standard normal probability distribution

We can better understand the concept of the standard normal probability distribution by examining the special relationship of the standard deviation to the normal curve. Look at Figure 5-13. Here we have illustrated two normal probability distributions, each with a different mean and a different standard deviation. Both area *a* and area *b*, the shaded areas under the curves, contain the *same* proportion of the total area under the normal curve. Why? Because both these areas are defined as being the area between the mean and one standard deviation to the right of the mean. *All* intervals containing the same number of standard deviations from the mean will contain the same proportion of the total area under the curve for *any* normal probability distribution. This makes possible the use of only one standard normal probability distribution table.

Let's find out what proportion of the total area under the curve is represented by colored areas in Figure 5-13. In Figure 5-12, we saw that an interval of one standard deviation (plus *and* minus) from the mean contained about 68 percent of the total area under the curve. In Figure 5-13, however, we are interested only in the area between the mean and 1 standard deviation to the *right* of the mean (plus, *not* plus and minus). This area must be half of 68 percent, or 34 percent, for both distributions.

Finding the percentage of the total area under the curve

One more example will reinforce our point. Look at the two normal probability distributions in Figure 5-14. Each of these has a different mean and a different standard deviation. The colored area under *both* curves, however, contains the same proportion of the total area under the curve. Why? Because both colored areas fall within 2 standard deviations (plus and minus) from the mean. Two

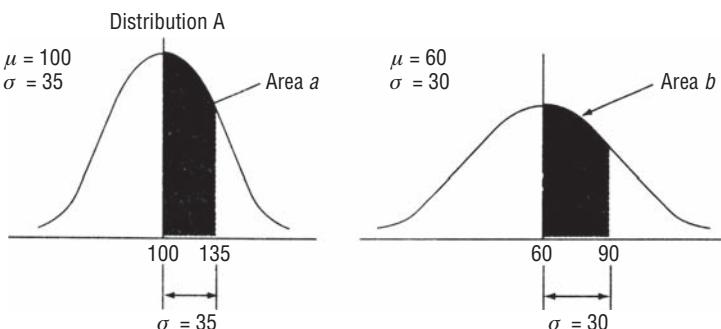


FIGURE 5-13 TWO INTERVALS, EACH ONE STANDARD DEVIATION TO THE RIGHT OF THE MEAN

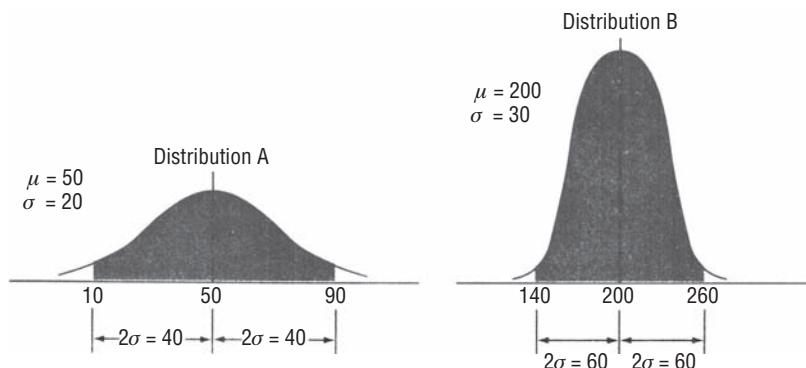


FIGURE 5-14 TWO INTERVALS, EACH ± 2 STANDARD DEVIATIONS FROM THE MEAN

standard deviations (plus and minus) from the mean include the same proportion of the total area under *any* normal probability distribution. In this case, we can refer to Figure 5-12 again and see that the colored areas in both distributions in Figure 5-14 contain about 95.5 percent of the total area under the curve.

Using the Standard Normal Probability Distribution Table

Appendix Table 1 shows the area under the normal curve between the mean and any value of the normally distributed random variable. Notice in this table the location of the column labeled z . The value for z is derived from the formula

Standardizing a Normal Random Variable	Formula for measuring distances under the normal curve
$z = \frac{x - \mu}{\sigma}$	[5-6]

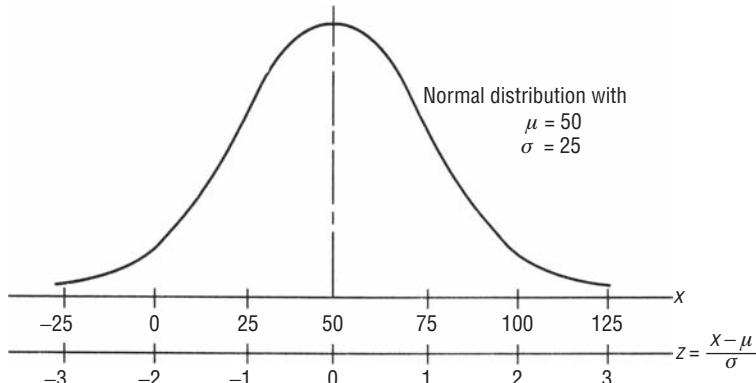


FIGURE 5-15 NORMAL DISTRIBUTION ILLUSTRATING COMPARABILITY OF Z VALUES AND STANDARD DEVIATIONS

where

- x = value of the random variable with which we are concerned
- μ = mean of the distribution of this random variable
- σ = standard deviation of this distribution
- z = number of standard deviations from x to the mean of this distribution

Why do we use z rather than “the number of standard deviations”? Normally distributed random variables take on many *different units of measure*: dollars, inches, parts per million, pounds, time. Because we shall use one table, Table 1 in the Appendix, we talk in terms of *standard units* (which really means standard deviations), and we denote them by the symbol z .

We can illustrate this graphically. In Figure 5-15, we see that the use of z is just a change of the scale of measurement on the horizontal axis.

The Standard Normal Probability Distribution Table, Appendix Table 1, is organized in terms of standard units, or z values. It gives the values for only *half* the area under the normal curve, beginning with 0.0 at the mean. Because the normal probability distribution is symmetrical (return to Figure 5-8 to review this point), the values true for one half of the curve are true for the other. We can use this one table for problems involving both sides of the normal curve. Working a few examples will help us to feel comfortable with the table.

Using z values

Standard Normal Probability Distribution Table

Data for Examples We have a training program designed to upgrade the supervisory skills of production-line supervisors. Because the program is self-administered, supervisors require different numbers of hours to complete the program. A study of past participants indicates that the mean length of time spent on the program is 500 hours and that this normally distributed random variable has a standard deviation of 100 hours.

Using the table to find probabilities (examples)

Example 1 What is the probability that a participant selected at random will require more than 500 hours to complete the program?

Solution In Figure 5-16, we see that half of the area under the curve is located on either side of the mean of 500 hours. Thus, we can deduce that the probability that the random variable will take on a value higher than 500 is the colored half, or 0.5.

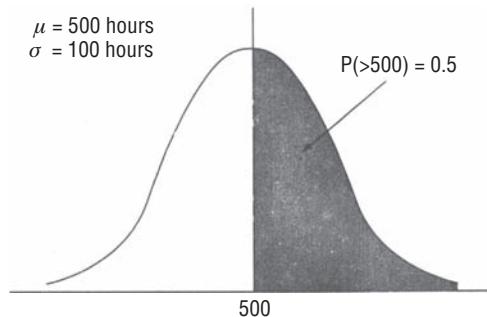


FIGURE 5-16 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL MORE THAN 500 HOURS IN COLOR

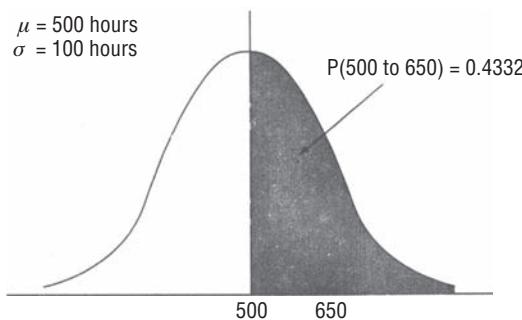


FIGURE 5-17 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL 500 TO 650 HOURS IN COLOR

Example 2 What is the probability that a candidate selected at random will take between 500 and 650 hours to complete the training program?

Solution We have shown this situation graphically in Figure 5-17. The probability that will answer this question is represented by the colored area between the mean (500 hours) and the x value in which we are interested (650 hours). Using Equation 5-6, we get a z value of

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} & [5-6] \\ &= \frac{650 - 500}{100} \\ &= \frac{150}{100} \\ &= 1.5 \text{ standard deviations} \end{aligned}$$

If we look up $z = 1.5$ in Appendix Table 1, we find a probability of 0.4332. Thus, the chance that a candidate selected at random would require between 500 and 650 hours to complete the training program is slightly higher than 0.4.

Example 3 What is the probability that a candidate selected at random will take more than 700 hours to complete the program?

Solution This situation is different from our previous examples. Look at Figure 5-18. We are interested in the colored area to the right of the value “700 hours.” How can we solve this problem? We can begin by using Equation 5-6:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} & [5-6] \\ &= \frac{700 - 500}{100} \\ &= \frac{200}{100} \\ &= 2 \text{ standard deviations} \end{aligned}$$

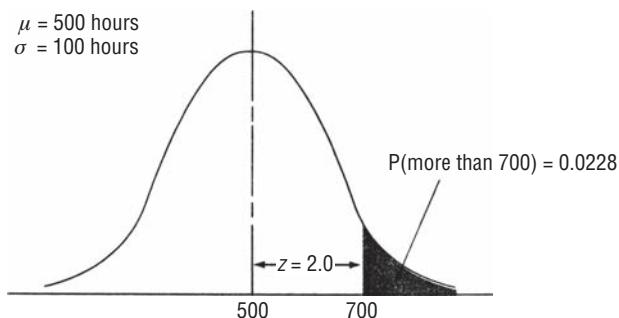


FIGURE 5-18 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL ABOVE 700 HOURS IN COLOR

Looking in Appendix Table 1 for a z value of 2.0, we find a probability of 0.4772. That represents the probability the program will require *between* 500 and 700 hours. However, we want the probability it will take *more* than 700 hours (the colored area in Figure 5-18). Because the right half of the curve (between the mean and the right-hand tail) represents a probability of 0.5, we can get our answer (the area to the right of the 700-hour point) if we subtract 0.4772 from 0.5; $0.5000 - 0.4772 = 0.0228$. Therefore, there are just over 2 chances in 100 that a participant chosen at random would take more than 700 hours to complete the course.

Example 4 Suppose the training-program director wants to know the probability that a participant chosen at random would require between 550 and 650 hours to complete the required work.

Solution This probability is represented by the colored area in Figure 5-19. This time, our answer will require two steps. First, we calculate a z value for the 650-hour point, as follows:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{650 - 500}{100} \end{aligned} \quad [5-6]$$

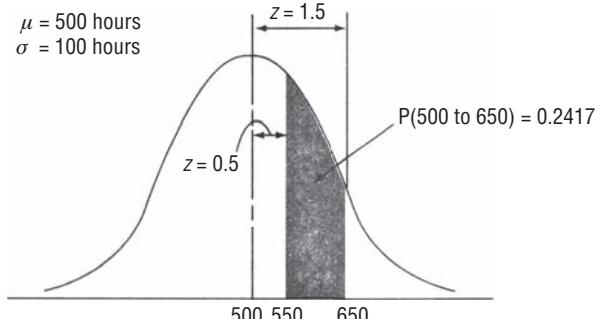


FIGURE 5-19 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL BETWEEN 550 AND 650 HOURS IN COLOR

$$\begin{aligned}
 &= \frac{150}{100} \\
 &= 1.5 \text{ standard deviations}
 \end{aligned}$$

When we look up a z of 1.5 in Appendix Table 1, we see a probability value of 0.4332 (the probability that the random variable will fall between the mean and 650 hours). Now for step 2. We calculate a z value for our 550-hour point like this:

$$\begin{aligned}
 z &= \frac{x - \mu}{\sigma} & [5-6] \\
 &= \frac{550 - 500}{100} \\
 &= \frac{50}{100} \\
 &= 0.5 \text{ standard deviation}
 \end{aligned}$$

In Appendix Table 1, the z value of 0.5 has a probability of 0.1915 (the chance that the random variable will fall between the mean and 550 hours). To answer our question, we must subtract as follows:

$$\begin{array}{rcl}
 0.4332 & \text{(Probability that the random variable will lie between the mean and 650 hours)} \\
 - 0.1915 & \text{(Probability that the random variable will lie between the mean and 550 hours)} \\
 \hline
 \mathbf{0.2417} & \leftarrow \text{(Probability that the random variable will lie between 550 and 650 hours)}
 \end{array}$$

Thus, the chance of a candidate selected at random taking between 550 and 650 hours to complete the program is a bit less than 1 in 4.

Example 5 What is the probability that a candidate selected at random will require fewer than 580 hours to complete the program?

Solution This situation is illustrated in Figure 5-20. Using Equation 5-6 to get the appropriate z value for 580 hours, we have

$$z = \frac{x - \mu}{\sigma} \quad [5-6]$$

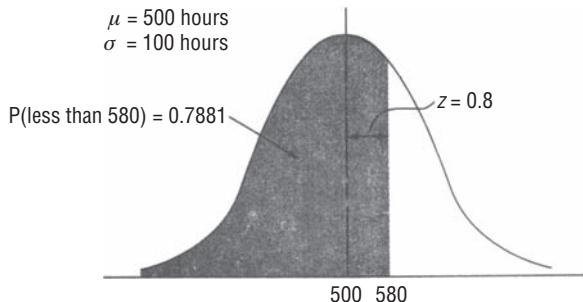


FIGURE 5-20 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL LESS THAN 580 HOURS IN COLOR

$$\begin{aligned}
 &= \frac{580 - 500}{100} \\
 &= \frac{80}{100} \\
 &= 0.8 \text{ standard deviation}
 \end{aligned}$$

Looking in Appendix Table 1 for a z value of 0.8, we find a probability of 0.2881—the probability that the random variable will lie between the mean and 580 hours. We must add to this the probability that the random variable will be between the left-hand tail and the mean. Because the distribution is symmetrical with half the area on each side of the mean, we know this value must be 0.5. As a final step, then, we add the two probabilities:

$$\begin{aligned}
 0.2881 &\quad (\text{Probability that the random variable will lie between the mean and 580 hours}) \\
 + 0.5000 &\quad (\text{Probability that the random variable will lie between the left-hand tail and the mean}) \\
 \mathbf{0.7881} &\leftarrow (\text{Probability that the random variable will lie between the left-hand tail and 580 hours})
 \end{aligned}$$

Thus, the chances of a candidate requiring less than 580 hours to complete the programme slightly higher than 75 percent.

Example 6 What is the probability that a candidate chosen at random will take between 420 and 570 hours to complete the program?

Solution Figure 5-21 illustrates the interval in question, from 420 to 570 hours. Again, the solution requires two steps. First, we calculate a z value for the 570-hour point:

$$\begin{aligned}
 z &= \frac{x - \mu}{\sigma} & [5-6] \\
 &= \frac{570 - 500}{100} \\
 &= \frac{70}{100} \\
 &= 0.7 \text{ standard deviation}
 \end{aligned}$$

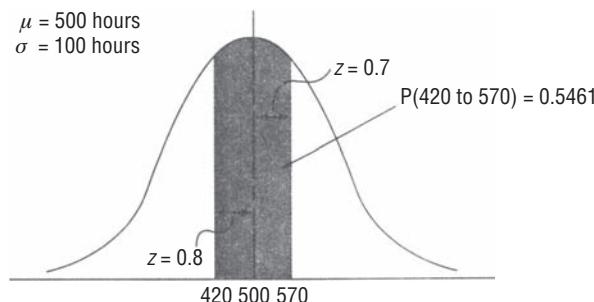


FIGURE 5-21 DISTRIBUTION OF THE TIME REQUIRED TO COMPLETE THE TRAINING PROGRAM, WITH THE INTERVAL BETWEEN 420 AND 570 HOURS IN COLOR

We look up the z value of 0.7 in Appendix Table 1 and find a probability value of 0.2580. Second, we calculate the z value for the 420-hour point:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{420 - 500}{100} \\ &= \frac{-80}{100} \\ &= -0.8 \text{ standard deviation} \end{aligned} \quad [5-6]$$

Because the distribution is symmetrical, we can disregard the sign and look for a z value of 0.8. The probability associated with this z value is 0.2881. We find our answer by adding these two values as follows:

$$\begin{array}{rl} 0.2580 & \text{(Probability that the random variable will lie between the mean and 570 hours)} \\ +0.2881 & \text{(Probability that the random variable will lie between the mean and 420 hours)} \\ \hline 0.5461 & \leftarrow \text{(Probability that the random variable will lie between 420 and 570 hours)} \end{array}$$

Shortcomings of the Normal Probability Distribution

Earlier in this section, we noted that the tails of the normal distribution approach but never touch the horizontal axis. This implies that there is *some* probability (although it may be very small) that the random variable can take on enormous values. It is possible for the right-hand tail of a normal curve to assign a minute probability of a person's weighing 2,000 pounds. Of course, no one would believe that such a person exists. (A weight of one ton or more would lie about 50 standard deviations to the right of the mean and would have a probability that began with 250 zeros to the right of the decimal point!) **We do not lose much accuracy by ignoring values far out in the tails. But in exchange for the convenience of using this theoretical model, we must accept the fact that it can assign impossible empirical values.**

Theory and practice

The Normal Distribution as an Approximation of the Binomial Distribution

Although the normal distribution is continuous, it is interesting to note that it can sometimes be used to approximate discrete distributions. To see how we can use it to approximate the binomial distribution, suppose we would like to know the probability of getting 5, 6, 7, or 8 heads in 10 tosses of a fair coin. We could use Appendix Table 3 to find this probability, as follows:

Sometimes the normal is used to approximate the binomial

$$\begin{aligned} P(r = 5, 6, 7 \text{ or } 8) &= P(r = 5) + P(r = 6) + P(r = 7) + P(r = 8) \\ &= 0.2461 + 0.2051 + 0.1172 + 0.0439 \\ &= 0.6123 \end{aligned}$$

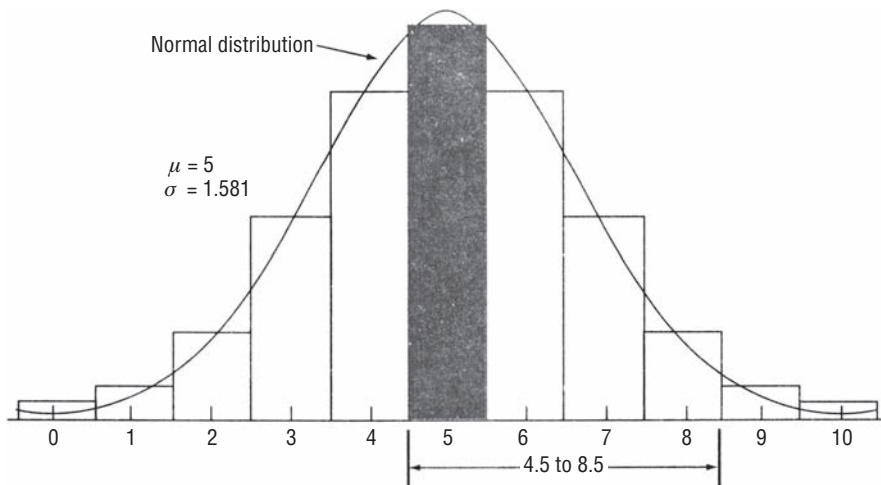


FIGURE 5-22 BINOMIAL DISTRIBUTION WITH $n = 10$ AND $p = \frac{1}{2}$, WITH A SUPERIMPOSED NORMAL DISTRIBUTION WITH $\mu = 5$ AND $\sigma = 1.581$

Figure 5-22 shows the binomial distribution for $n = 10$ and $p = \frac{1}{2}$ with a normal distribution superimposed on it with the same mean ($\mu = np = 10(\frac{1}{2}) = 5$) and the same standard deviation ($\sigma = \sqrt{npq} = \sqrt{10(\frac{1}{2})(\frac{1}{2})} = \sqrt{2.5} = 1.581$).

Look at the area under the normal curve between $5 - \frac{1}{2}$ and $5 + \frac{1}{2}$. We see that the area is approximately the same size as the area of the colored bar representing the binomial probability of getting five heads. The two $\frac{1}{2}$'s that we add to and subtract from 5 are called *continuity correction factors* and are used to improve the accuracy of the approximation.

Using the continuity correction factors, we see that the binomial probability of 5, 6, 7, or 8 heads can be approximated by the area under the normal curve between 4.5 and 8.5. Compute that probability by finding the z values corresponding to 4.5 and 8.5.

$$\text{At } x = 4.5 < z = \frac{x - \mu}{\sigma} \quad [5-6]$$

$$= \frac{4.5 - 5}{1.581} \\ = -0.32 \text{ standard deviation}$$

$$\text{At } x = 8.5 < z = \frac{x - \mu}{\sigma} \quad [5-6]$$

$$= \frac{8.5 - 5}{1.581} \\ = 2.1 \text{ standard deviations}$$

Two distributions with the same means and standard deviations

Continuity correction factors

Now, from Appendix Table 1, we find

0.1255	(Probability that z will be between -0.32 and 0 (and, correspondingly, that x will be between 4.5 and 5))
+0.4864	(Probability that z will be between 0 and 2.21 (and, correspondingly, that x will be between 5 and 8.5))
0.6119	(Probability that x will be between 4.5 and 8.5)

Comparing the binomial probability of 0.6123 (Appendix Table 3) with this normal approximation of 0.6119 , we see that the error in the approximation is less than .1 percent.

The error in estimating is slight

The normal approximation to the binomial distribution is very convenient because it enables us to solve the problem without extensive tables of the binomial distribution. (You might note that Appendix Table 3, which gives binomial probabilities for values of n up to 20 , is already 9 pages long.) **We should note that some care needs to be taken in using this approximation, but it is quite good whenever both np and nq are at least 5.**

Care must be taken

HINTS & ASSUMPTIONS

Warning: The normal distribution is the probability distribution most often used in statistics. Statisticians fear that too often, the data being analyzed are not well-described by a normal distribution. Fortunately there is a test to help you decide whether this is indeed the case, and we'll introduce it in Chapter 11 when we've laid a bit more foundation. Hint: Students who have trouble calculating probabilities using the normal distribution tend to do better when they actually sketch the distribution in question, indicate the mean and standard deviation, and then show the limits of the random variable in question (we use color but pencil shading is just as good). Visualizing the situation this way makes decisions easier (and answers more accurate).

EXERCISES 5.6

Self-Check Exercises

SC 5-9 Use the normal approximation to compute the binomial probabilities in parts (a)–(d) below:

- (a) $n = 30, p = 0.35$, between 10 and 15 successes, inclusive.
- (b) $n = 42, p = 0.62$, 30 or more successes.
- (c) $n = 15, p = 0.40$, at most 7 successes.
- (d) $n = 51, p = 0.42$, between 17 and 25 successes, inclusive.

SC 5-10 Dennis Hogan is the supervisor for the Conowingo Hydroelectric Dam. Mr. Hogan knows that the dam's turbines generate electricity at the peak rate only when at least $1,000,000$ gallons of water pass through the dam each day. He also knows, from experience, that the daily flow is normally distributed, with the mean equal to the previous day's flow and a standard deviation of $200,000$ gallons. Yesterday, $850,000$ gallons flowed through the dam. What is the probability that the turbines will generate at peak rate today?

Basic Concepts

- 5-37** Given that a random variable, X , has a normal distribution with mean 6.4 and standard deviation 2.7, find
- $P(4.0 < x < 5.0)$.
 - $P(x > 2.0)$.
 - $P(x < 7.2)$.
 - $P((x < 3.0) \text{ or } (x > 9.0))$.
- 5-38** Given that a random variable, X , has a binomial distribution with $n = 50$ trials and $p = 0.25$, use the normal approximation to the binomial to find
- $P(x > 10)$.
 - $P(x < 18)$.
 - $P(x > 21)$.
 - $P(9 < x < 14)$.
- 5-39** In a normal distribution with a standard deviation of 5.0, the probability that an observation selected at random exceeds 21 is 0.14.
- Find the mean of the distribution.
 - Find the value below which 4 percent of the values in the distribution lie.
- 5-40** Use the normal approximation to compute the binomial probabilities in parts (a)–(e) below.
- $n = 35, p = 0.15$, between 7 and 10 successes inclusive.
 - $n = 29, p = 0.25$, at least 9 successes.
 - $n = 84, p = 0.42$, at most 40 successes.
 - $n = 63, p = 0.11, 10$ or more successes.
 - $n = 18, p = 0.67$, between 9 and 12 successes inclusive.

Applications

- 5-41** The manager of a small postal substation is trying to quantify the variation in the weekly demand for mailing tubes. She has decided to assume that this demand is normally distributed. She knows that on average 100 tubes are purchased weekly and that 90 percent of the time, weekly demand is below 115.
- What is the standard deviation of this distribution?
 - The manager wants to stock enough mailing tubes each week so that the probability of running out of tubes is no higher than 0.05. What is the lowest such stock level?
- 5-42** The Gilbert Machinery Company has received a big order to produce electric motors for a manufacturing company. In order to fit in its bearing, the drive shaft of the motor must have a diameter of 5.1 ± 0.05 (inches). The company's purchasing agent realizes that there is a large stock of steel rods in inventory with a mean diameter of $5.07''$ and a standard deviation of $0.07''$. What is the probability of a steel rod from inventory fitting the bearing?
- 5-43** The manager of a Spiffy Lube auto lubrication shop is trying to revise his policy on ordering grease gun cartridges. Currently, he orders 110 cartridges per week, but he runs out of cartridges 1 out of every 4 weeks. He knows that, on average, the shop uses 95 cartridges per week. He is also willing to assume that demand for cartridges is normally distributed.
- What is the standard deviation of this distribution?
 - If the manager wants to order enough cartridges so that his probability of running out during any week is no greater than 0.2, how many cartridges should he order per week?

- 5-44** Jarrid Medical, Inc., is developing a compact kidney dialysis machine, but its chief engineer, Mike Crowe, is having trouble controlling the variability of the rate at which fluid moves through the device. Medical standards require that the hourly flow be 4 liters, plus or minus 0.1 liter, 80 percent of the time. Mr. Crowe, in testing the prototype, has found that 68 percent of the time, the hourly flow is within 0.08 liter of 4.02 liters. Does the prototype satisfy the medical standards?
- 5-45** Sgt. Wellborn Fitte, the U.S. Army's quartermaster at Fort Riley, Kansas, prides himself on being able to find a uniform to fit virtually any recruit. Currently, Sgt. Fitte is revising his stock requirements for fatigue caps. Based on experience, Sgt. Fitte has decided that hat size among recruits varies in such a way that it can be approximated by a normal distribution with a mean of 7". Recently, though, he has revised his estimate of the standard deviation from 0.75 to 0.875. Present stock policy is to have on hand hats in every size (increments of $\frac{1}{8}$ ") from $6\frac{1}{4}$ " to $7\frac{3}{4}$ ". Assuming that a recruit is fit if his or her hat size is within this range, find the probability that a recruit is fit using
- The old estimate of the standard deviation.
 - The new estimate of the standard deviation.
- 5-46** Glenn Howell, VP of personnel for the Standard Insurance Company, has developed a new training program that is entirely self-paced. New employees work various stages at their own pace; completion occurs when the material is learned. Howell's program has been especially effective in speeding up the training process, as an employee's salary during training is only 67 percent of that earned upon completion of the program. In the last several years, average completion time of the program was 44 days, and the standard deviation was 12 days.
- Find the probability an employee will finish the program in 33 to 42 days.
 - What is the probability of finishing the program in fewer than 30 days?
 - Fewer than 25 or more than 60 days?
- 5-47** On the basis of past experience, automobile inspectors in Pennsylvania have noticed that 5 percent of all cars coming in for their annual inspection fail to pass. Using the normal approximation to the binomial, find the probability that between 7 and 18 of the next 200 cars to enter the Lancaster inspection station will fail the inspection.
- 5-48** R. V. Poppin, the concession stand manager for the local hockey rink, just had 2 cancellations on his crew. This means that if more than 72,000 people come to tonight's hockey game, the lines for hot dogs will constitute a disgrace to Mr. Poppin and will harm business at future games. Mr. Poppin knows from experience that the number of people who come to the game is normally distributed with mean 67,000 and standard deviation 4,000 people.
- What is the probability that there will be more than 72,000 people?
 - Suppose Mr. Poppin can hire two temporary employees to make sure business won't be harmed in the future at an additional cost of \$200. If he believes the future harm to business of having more than 72,000 fans at the game would be \$5,000, should he hire the employees? Explain. (Assume there will be no harm if 72,000 or fewer fans show up, and that the harm due to too many fans doesn't depend on how many more than 72,000 show up.)
- 5-49** Maurine Lewis, an editor for a large publishing company, calculates that it requires 11 months on average to complete the publication process from manuscript to finished book, with a standard deviation of 2.4 months. She believes that the normal distribution well describes the distribution of publication times. Out of 19 books she will handle this year, approximately how many will complete the process in less than a year?

5-50 The Quickie Sales Corporation has just been given two conflicting estimates of sales for the upcoming quarter. Estimate I says that sales (in millions of dollars) will be normally distributed with $\mu = 325$ and $\sigma = 60$. Estimate II says that sales will be normally distributed with $\mu = 300$ and $\sigma = 50$. The board of directors finds that each estimate appears to be equally believable a priori. In order to determine which estimate should be used for future predictions, the board of directors has decided to meet again at the end of the quarter to use updated sales information to make a statement about the credibility of each estimate.

- Assuming that Estimate I is accurate, what is the probability that Quickie will have quarterly sales in excess of \$350 million?
- Rework part (a) assuming that Estimate II is correct.
- At the end of the quarter, the board of directors finds that Quickie Sales Corp. has had sales in excess of \$350 million. Given this updated information, what is the probability that Estimate I was originally the accurate one? (*Hint:* Remember Bayes' theorem.)
- Rework part (c) for Estimate II.

5-51 The Nobb Door Company manufactures doors for recreational vehicles. It has two conflicting objectives: It wants to build doors as small as possible to save on material costs, but to preserve its good reputation with the public, it feels obligated to manufacture doors that are tall enough for 95 percent of the adult population in the United States to pass through without stooping. In order to determine the height at which to manufacture doors, Nobb is willing to assume that the height of adults in America is normally distributed with mean 73 inches and standard deviation 6 inches. How tall should Nobb's doors be?

Worked-Out Answers to Self-Check Exercises

SC 5-9

$$(a) \mu = np = 30(0.35) = 10.5 \quad \sigma = \sqrt{npq} = \sqrt{30(0.35)(0.65)} = 2.612$$

$$P(10 \leq r \leq 15) = P\left(\frac{9.5 - 10.5}{2.612} \leq z \leq \frac{15.5 - 10.5}{2.612}\right)$$

$$= P(-0.38 \leq z \leq 1.91) = 0.1480 + 0.4719 = 0.6199$$

$$(b) \mu = np = 42(0.62) = 26.04 \quad \sigma = \sqrt{npq} = \sqrt{42(0.62)(0.38)} = 3.146$$

$$P(r \geq 30) = P\left(z \geq \frac{29.5 - 26.04}{3.146}\right) = P(z \geq 1.10) = 0.5 - 0.3643 = 0.1357$$

$$(c) \mu = np = 15(0.40) = 6 \quad \sigma = \sqrt{npq} = \sqrt{15(0.40)(0.60)} = 1.895$$

$$P(r \leq 7) = P\left(z \leq \frac{7.5 - 6}{1.897}\right) = P(z \leq 0.79) = 0.5 + 0.2852 = 0.7852$$

$$(d) \mu = np = 51(0.42) = 21.42 \quad \sigma = \sqrt{npq} = \sqrt{51(0.42)(0.58)} = 3.525$$

$$P(17 \leq r \leq 25) = P\left(\frac{16.5 - 21.42}{3.525} \leq z \leq \frac{25.5 - 21.42}{3.525}\right)$$

$$P(-1.40 \leq z \leq 1.16) = 0.4192 + 0.3770 = 0.7962$$

SC 5-10 For today, $\mu = 850,000$, $\sigma = 200,000$

$$\begin{aligned} P(x \geq 1,000,000) &= P\left(z \geq \frac{1,000,000 - 850,000}{200,000}\right) = P(z \geq 0.75) \\ &= 0.5 - 0.2734 = 0.2266 \end{aligned}$$

5.7 CHOOSING THE CORRECT PROBABILITY DISTRIBUTION

If we plan to use a probability to describe a situation, we must be careful to choose the right one. We need to be certain that we are not using the *Poisson* probability distribution when it is the *binomial* that more nearly describes the situation we are studying. Remember that the binomial distribution is applied when the number of trials is fixed before the experiment begins, and each trial is independent and can result in only two mutually exclusive outcomes (success/failure, either/or, yes/no). Like the binomial, the Poisson distribution applies when each trial is independent. But although the probabilities in a Poisson distribution approach zero after the first few values, the number of possible values is infinite. The results are not limited to two mutually exclusive outcomes. Under some conditions, the Poisson distribution can be used as an approximation of the binomial, but not always. All the assumptions that form the basis of a distribution must be met if our use of that distribution is to produce meaningful results.

Even though the normal probability distribution is the only continuous distribution we have discussed in this chapter, we should realize that there are other useful continuous distributions. In the chapters to come, we shall study three additional continuous distributions: Student's *t*, χ^2 , and *F*. Each of these is of interest to decision makers who solve problems using statistics.

EXERCISES 5.7

- 5-52** Which probability distribution is most likely the appropriate one to use for the following variables: binomial, Poisson, or normal?
- The life span of a female born in 1977.
 - The number of autos passing through a tollbooth.
 - The number of defective radios in a lot of 100.
 - The water level in a reservoir.
- 5-53** What characteristics of a situation help to determine which is the appropriate distribution to use?
- 5-54** Explain in your own words the difference between discrete and continuous random variables. What difference do such classifications make in determining the probabilities of future events?
- 5-55** In practice, managers see many different types of distributions. Often, the nature of these distributions is not as apparent as are some of the examples provided in this book. What alternatives are open to students, teachers, and researchers who want to use probability distributions in their work but who are not sure exactly which distributions are appropriate for given situations?

STATISTICS AT WORK

Loveland Computers

Case 5: Probability Distributions “So, Nancy Rainwater tells me she’s ‘reasonably certain’ about her decision on how she’s going to schedule the production line.” Walter Azko was beginning to feel that hiring Lee Azko as an assistant was one of his better investments. “But don’t get too comfortable,

I've got another problem I want you to work on. Tomorrow, I want you to spend some time with Jeff Cohen—he's the head of purchasing here."

Jeff Cohen would be the first to say that he was surprised to find himself as the head of purchasing for a computer company. An accountant by training, he had first run into Walter Azko when he was assigned by his CPA firm to help Walter prepare the annual financial statements for his importing company. Because Walter traveled frequently and was always trying out new product lines, the financial records were a mess of invoices and check stubs for manufacturers, brokers, and shippers. Jeff's brief assignment turned into a permanent position, and when Loveland Computers was formed, he somewhat reluctantly agreed to handle purchasing, as long as Walter negotiated the deals. For Jeff, the best part of the job was that he could indulge his taste for oriental art.

Lee Azko found Jeff in a corner office that looked like a surgery room prepared for an operation: There was not so much as a paper clip on his desk, and the bookshelves contained neat rows of color-coded binders. "Let me explain my problem to you, Lee," Cohen launched in immediately. "We import our midrange line fully assembled from Singapore. Because it's a high-value product, it makes sense to pay to have it airfreighted to us. The best part of that is that we don't have to keep much inventory here in Colorado and we're not paying to have hundreds of thousands of dollars' worth of computers to sit on docks and on boats for several weeks. The computers are boxed and wrapped on pallets in a shape that just fits in the cargo hold of an MD-11 freighter. So it makes sense for us to order the midrange in lots of 200 units."

"I understand," said Lee, making a mental note that each shipment was worth about a quarter of a million dollars. "I've seen them arrive at the inbound dock."

"About half of the computers are sent on to customers without even being taken out of the box. But the rest need some assembly work on Nancy Rainwater's production line. We need to add a modem—you know, the device that lets a computer 'talk' to another machine through regular telephone lines. The modem comes on one board and just snaps into a slot. There's not much to it. I can get modems locally from several different electronics firms. But for each lot of computers, I have to decide how many modems to order. And I don't know how many customers will want a modem. If I order too many, I end up with unused inventory that just adds to my costs. The overstock eventually gets used up for customers who call in after the purchase and want a modem as an 'add on.' But if I order too few, I have to use a lot of staff time to round up a few extras, and, of course, none of the suppliers wants to give me a price break on a small lot."

"Well, you've got the records," Lee replied. "Why don't you just order the 'average' number of modems needed for each lot?"

"Because although the *average* number of modems per lot has stayed the same over the last few years, the *actual* number requested by customers on any single lot jumps around a bit. Take a look at these numbers," Jeff said as he walked across to the bookcase and pulled out a folder. "It's much worse for me to end up with too few modems in stock when a shipment of midranges comes through the production line than to have too many. So I suppose I tend to order above the average. It just seems that there ought to be a way to figure out how many to order so that we can be reasonably sure that we can operate the line without running out."

"Well, there's only one question remaining," said Lee. "You have to tell me how many times—out of 100 lots of computers—you can tolerate being wrong in your guess. Would a 95 percent success rate work for you?"

Study Questions: What calculations is Lee going to make? Why does Lee need to know Jeff Cohen's desired "success" rate for this prediction? What does Lee know about the underlying distribution of the parameter "number of modems per lot"? Finally, what additional information will Lee need?

CHAPTER REVIEW

Terms Introduced in Chapter 5

Bernoulli Process A process in which each trial has only two possible outcomes, the probability of the outcome of any trial remains fixed over time, and the trials are statistically independent.

Binomial Distribution A discrete distribution describing the results of an experiment known as a Bernoulli process.

Continuity Correction Factor Corrections used to improve the accuracy of the approximation of a binomial distribution by a normal distribution.

Continuous Probability Distribution A probability distribution in which the variable is allowed to take on any value within a given range.

Continuous Random Variable A random variable allowed to take on any value within a given range.

Discrete Probability Distribution A probability distribution in which the variable is allowed to take on only a limited number of values, which can be listed.

Discrete Random Variable A random variable that is allowed to take on only a limited number of values, which can be listed.

Expected Value A weighted average of the outcomes of an experiment.

Expected Value of a Random Variable The sum of the products of each value of the random variable with that value's probability of occurrence.

Normal Distribution A distribution of a continuous random variable with a single-peaked, bell-shaped curve. The mean lies at the center of the distribution, and the curve is symmetrical around a vertical line erected at the mean. The two tails extend indefinitely, never touching the horizontal axis.

Poisson Distribution A discrete distribution in which the probability of the occurrence of an event within a very small time period is a very small number, the probability that two or more such events will occur within the same time interval is effectively 0, and the probability of the occurrence of the event within one time period is independent of where that time period is.

Probability Distribution A list of the outcomes of an experiment with the probabilities we would expect to see associated with these outcomes.

Random Variable A variable that takes on different values as a result of the outcomes of a random experiment.

Standard Normal Probability Distribution A normal probability distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Equations Introduced in Chapter 5

- 5-1 Probability of r successes in n trials =
$$\frac{n!}{r!(n-r)!} p^r q^{n-r}$$
 p. 226
 where
- r = number of successes desired
 - n = number of trials undertaken
 - p = probability of success (characteristic probability)
 - q = probability of failure ($q = 1 - p$)

This *binomial formula* enables us to calculate algebraically the probability of r successes. We can apply it to any Bernoulli process, where each trial has only two possible outcomes (a success or a failure), the probability of success remains the same trial after trial, and the trials are statistically independent.

5-2

$$\mu = np$$

p. 233

The *mean of a binomial distribution* is equal to the number of trials multiplied by the probability of success.

5-3

$$\sigma = \sqrt{npq}$$

p. 234

The *standard deviation of a binomial distribution* is equal to the square root of the product of the number of trials, the probability of a success, and the probability of a failure (found by taking $q = 1 - p$).

5-4

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

p. 239

This formula enables us to calculate the probability of a discrete random variable occurring in a *Poisson distribution*. The formula states that the probability of *exactly* x occurrences is equal to λ , or lambda (the mean number of occurrences per interval of time in a Poisson distribution), raised to the x th power and multiplied by e , or 2.71828 (the base of the natural logarithm system), raised to the negative lambda power, and the product divided by x factorial. Appendix Tables 4a and 4b can be used for computing Poisson probabilities.

5-5

$$P(x) = \frac{(np)^x \times e^{-np}}{x!}$$

p. 243

If we substitute in Equation 5-4 the mean of the binomial distribution (np) in place of the mean of the Poisson distribution (λ), we can use the Poisson probability distribution as a reasonable approximation of the binomial. The approximation is good when n is greater than or equal to 20 and p is less than or equal to 0.05.

5-6

$$z = \frac{x - \mu}{\sigma}$$

p. 251

where

- x = value of the random variable with which we are concerned
- μ = mean of the distribution of this random variable
- σ = standard deviation of this distribution
- z = number of standard deviations from x to the mean of this distribution

Once we have derived z using this formula, we can use the Standard Normal Probability Distribution Table (which gives the values for areas under half the normal curve, beginning with 0.0 at the mean) and determine the probability that the random variable with which we are concerned is within that distance from the mean of this distribution.

Review and Application Exercises

5-56

- In the past 20 years, on average, only 3 percent of all checks written to the American Heart Association have bounced. This month, the A.H.A. received 200 checks. What is the probability that
- Exactly 10 of these checks bounced?
 - Exactly 5 of these checks bounced?

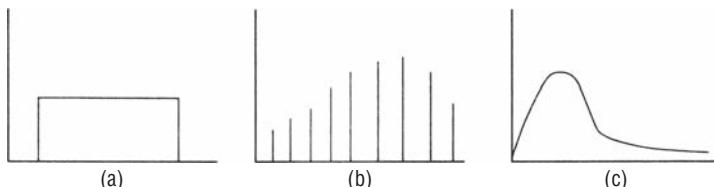
- 5-57** An inspector for the U.S. Department of Agriculture is about to visit a large meat-packing company. She knows that, on average, 2 percent of all sides of beef inspected by the USDA are contaminated. She also knows that if she finds that more than 5 percent of the meat-packing company's beef is contaminated, the company will be closed for at least 1 month. Out of curiosity, she wants to compute the probability that this company will be shut down as a result of her inspection. Should she assume her inspection of the company's sides of beef is a Bernoulli process? Explain.

- 5-58** The regional office of the Environmental Protection Agency annually hires second-year law students as summer interns to help the agency prepare court cases. The agency is under a budget and wishes to keep its costs at a minimum. However, hiring student interns is less costly than hiring full-time employees. Accordingly, the agency wishes to hire the maximum number of students without overstaffing. On the average, it takes two interns all summer to research a case. The interns turn their work over to staff attorneys, who prosecute the cases in the fall when the circuit court convenes. The legal staff coordinator has to place his budget request in June of the preceding summer for the number of positions he wishes to maintain. It is therefore impossible for him to know with certainty how many cases will be researched in the following summer. The data from preceding summers are as follows:

Year	1987	1988	1989	1990	1991
Number of cases	6	4	8	7	5
Year	1992	1993	1994	1995	1996
Number of cases	6	4	5	4	5

Using these data as his probability distribution for the number of cases, the legal staff coordinator wishes to hire enough interns to research the expected number of cases that will arise. How many intern positions should be requested in the budget?

- 5-59** Label the following probability distributions as discrete or continuous:



- 5-60** Which probability distribution would you use to find binomial probabilities in the following situations: binomial, Poisson, or normal?
- 112 trials, probability of success 0.06.
 - 15 trials, probability of success 0.4.
 - 650 trials, probability of success 0.02.
 - 59 trials, probability of success 0.1.

- 5-61** The French bread made at La Fleur de Farine costs \$8 per dozen baguettes to produce. Fresh bread sells at a premium, \$16 per dozen baguettes, but it has a short shelf life. If La Fleur de Farine bakes more bread than its customers demand on any given day, the leftover day-old

bread goes for croutons in local restaurants at a discounted \$7 per dozen baguettes. Conversely, producing less bread than customers demand leads to lost sales. La Fleur de Farine bakes its French bread in batches of 350 dozen baguettes. The daily demand for bread is a random variable, taking the values two, three, four, or five batches, with probabilities 0.2, 0.25, 0.4, and 0.15, respectively. If La Fleur de Farine wishes to maximize expected profits, how much bread should it bake each morning?

5-62 Reginald Dunfey, president of British World Airlines, is fiercely proud of his company's on-time percentage; only 2 percent of all BWA flights arrive more than 10 minutes early or late. In his upcoming speech to the BWA board of directors, Mr. Dunfey wants to include the probability that none of the 200 flights scheduled for the following week will be more than 10 minutes early or late. What is the probability? What is the probability that exactly 10 flights will be more than 10 minutes early or late?

5-63 Marvin Thornbury, an attorney working for the Legal Aid Society, estimates that, on average, seven of the daily arrivals to the L.A.S. office are people who were (in their opinion) unfairly evicted. Further, he estimates that, on average, five of the daily arrivals are people whose landlords have raised their rent illegally.

- (a) What is the probability that six of the daily arrivals report an unfair eviction?
- (b) What is the probability that eight daily arrivals have suffered from an illegal rent increase?

5-64 The City Bank of Durham has recently begun a new credit program. Customers meeting certain credit requirements can obtain a credit card accepted by participating area merchants that carries a discount. Past numbers show that 25 percent of all applicants for this card are rejected. Given that credit acceptance or rejection is a Bernoulli process, out of 15 applicants, what is the probability that

- (a) Exactly four will be rejected?
- (b) Exactly eight?
- (c) Fewer than three?
- (d) More than five?

5-65 Anita Daybride is a Red Cross worker aiding earthquake victims in rural Colombia. Ms. Daybride knows that typhus is one of the most prevalent post-earthquake diseases: 44 percent of earthquake victims in rural areas contract the disease. If Anita treats 12 earthquake victims, what is the probability that

- (a) Six or more have typhus?
- (b) Seven or fewer?
- (c) Nine or more?

5-66 On average, 12 percent of those enrolled in the Federal Aviation Administration's air traffic controller training program will have to repeat the course. If the current class size at the Leesburg, Virginia, training center is 15, what is the probability that

- (a) Fewer than 6 will have to repeat the course?
- (b) Exactly 10 will pass the course?
- (c) More than 12 will pass the course?

5-67 The Virginia Department of Health and Welfare publishes a pamphlet, *A Guide to Selecting Your Doctor*. Free copies are available to individuals, institutions, and organizations that are willing to pay the postage. Most of the copies have gone to a small number of groups who, in turn, have disseminated the literature. Mailings for 5 years have been as follows:

	Year				
	1992	1993	1994	1995	1996
Virginia Medical Association	7,000	3,000	—	2,000	4,000
Octogenarian Clubs	1,000	1,500	1,000	700	1,000
Virginia Federation of Women's Clubs	4,000	2,000	3,000	1,000	—
Medical College of Virginia	—	—	3,000	2,000	3,000
U.S. Department of Health, Education, and Welfare	1,000	—	1,000	—	1,000

In addition, an average of 2,000 copies per year were mailed or given to walk-in customers. Assistant secretary Susan Fleming, who has to estimate the number of pamphlets to print for 1997, knows that a revised edition of the pamphlet will be published in 1998. She feels that the demand in 1997 will most likely resemble that of 1994. She has constructed this assessment of the probabilities:

	Year				
	1992	1993	1994	1995	1996
Probability that 1997 will resemble this year	0.10	0.25	0.45	0.10	0.10

- (a) Construct a table of the probability distribution of demand for the pamphlet, and draw a graph representing that distribution.
- (b) Assuming Fleming's assessment of the probabilities was correct, how many pamphlets should she order to be certain that there will be enough for 1997?

5-68

Production levels for Giles Fashion vary greatly according to consumer acceptance of the latest styles. Therefore, the company's weekly orders of wool cloth are difficult to predict in advance. On the basis of 5 years of data, the following probability distribution for the company's weekly demand for wool has been computed:

Amount of wool (lb)	2,500	3,500	4,500	5,500
Probability	0.30	0.45	0.20	0.05

From these data, the raw-materials purchaser computed the expected number of pounds required. Recently, she noticed that the company's sales were lower in the last year than in years before. Extrapolating, she observed that the company will be lucky if its weekly demand averages 2,500 this year.

- (a) What was the expected weekly demand for wool based on the distribution from past data?
- (b) If each pound of wool generates \$5 in revenue and costs \$4 to purchase, ship, and handle, how much would Giles Fashion stand to gain or lose each week if it orders wool based on the past expected value and the company's demand is only 2,500?

5-69

Heidi Tanner is the manager of an exclusive shop that sells women's leather clothing and accessories. At the beginning of the fall/winter season, Ms. Tanner must decide how many full-length leather coats to order. These coats cost her \$100 each and will sell for \$200 each. Any coats left over at the end of the season will have to be sold at a 20 percent discount in order to make room for spring/summer inventory. From past experience, Heidi knows that demand for the coats has the following probability distribution:

Number of coats demanded	8	10	12	14	16
Probability	0.10	0.20	0.25	0.30	0.15

She also knows that any leftover coats can be sold at discount.

- (a) If Heidi decides to order 14 coats, what is her expected profit?
- (b) How would the answer to part (a) change if the leftover coats were sold at a 40 percent discount?

5-70 The Executive Camera Company provides full expenses for its sales force. When attempting to budget automobile expenses for its employees, the financial department uses mileage figures to estimate gas, tire, and repair expenses. Distances driven average 5,650 miles a month, and have a standard deviation of 120. The financial department wants its expense estimate and subsequent budget to be adequately high and, therefore, does not want to use any of the data from drivers who drove fewer than 5,500 miles. What percentage of Executive's sales force drove 5,500 miles or more?

5-71 Mission Bank is considering changing the day for scheduled maintenance for the automatic teller machine (ATM) in the lobby. The average number of people using it between 8 and 9 A.M. is 30, except on Fridays, when the average is 45. The management decision must balance the efficient use of maintenance staff while minimizing customer inconvenience.

- (a) Does knowledge of the two average figures affect the manager's expected value (for inconvenienced customers)?
- (b) Taking the data for all days together, the relative probability of inconveniencing 45 customers is quite small. Should the manager expect many inconvenienced customers if the maintenance day is changed to Friday?

5-72 The purchasing agent in charge of procuring automobiles for the state of Minnesota's inter-agency motor pool was considering two different models. Both were 4-door, 4-cylinder cars with comparable service warranties. The decision was to choose the automobile that achieved the best mileage per gallon. The state had done some tests of its own, which produced the following results for the two automobiles in question:

	Average MPG	Standard Deviation
Automobile A	42	4
Automobile B	38	7

The purchasing agent was uncomfortable with the standard deviations, so she set her own decision criterion for the car that would be more likely to get more than 45 miles per gallon.

- (a) Using the data provided in combination with the purchasing agent's decision criterion, which car should she choose?
- (b) If the purchasing agent's criterion was to reject the automobile that more likely obtained less than 39 mpg, which car should she buy?

5-73 In its third year, attendance in the Liberty Football League averaged 16,050 fans per game, and had a standard deviation of 2,500.

- (a) According to these data, what is the probability that the number of fans at any given game was greater than 20,000?
- (b) Fewer than 10,000?
- (c) Between 14,000 and 17,500?

- 5-74** Ted Hughes, the mayor of Chapelboro, wants to do something to reduce the number of accidents in the town involving motorists and bicyclists. Currently, the probability distribution of the number of such accidents per week is as follows:

Number of accidents	0	1	2	3	4	5
Probability	0.05	0.10	0.20	0.40	0.15	0.10

The mayor has two choices of action: He can install additional lighting on the town's streets or he can expand the number of bike lanes in the town. The respective revised probability distributions for the two options are as follows:

Number of accidents	0	1	2	3	4	5
Probability (lights)	0.10	0.20	0.30	0.25	0.10	0.05
Probability (lanes)	0.20	0.20	0.20	0.30	0.05	0.05

Which plan should the mayor approve if he wants to produce the largest possible reduction in

- (a) Expected number of accidents per week?
- (b) Probability of more than three accidents per week?
- (c) Probability of three or more accidents per week?

- 5-75** Copy Chums of Boulder leases office copying machines and resells returned machines at a discount. Leases are normally distributed, with a mean of 24 months and a standard deviation of 7.5 months.

- (a) What is the probability that a copier will still be on lease after 28 months?
- (b) What is the probability that a copier will be returned within one year?

- 5-76** Sensurex Productions, Incorporated, has recently patented and developed an ultrasensitive smoke detector for use in both residential and commercial buildings. Whenever a detectable amount of smoke is in the air, a wailing siren is set off. In recent tests conducted in a $20' \times 15' \times 8'$ room, the smoke levels that activated the smoke detector averaged 320 parts per million (ppm) of smoke in the room, and had a standard deviation of 25 ppm.

- (a) If a cigarette introduces 82 ppm into the atmosphere of a $20' \times 15' \times 8'$ room, what is the probability that four people smoking cigarettes simultaneously will set off the alarm?
- (b) Three people?

- 5-77** Rework Exercise 5-65 using the normal approximation. Compare the approximate and exact answers.

- 5-78** Try to use the normal approximation for Exercise 5-66. Notice that np is only 1.8. Comment on the accuracy of the approximation.

- 5-79** Randall Finan supervises the packaging of college textbooks for Newsome-Cluett Publishers. He knows that the number of cardboard boxes he will need depends partly on the size of the books. All Newsome-Cluett books use the same size paper but may have differing numbers of pages. After pulling shipment records for the last 5 years, Finan derived the following set of probabilities:

# of pages	100	300	500	700	900	1100
Probability	0.05	0.10	0.25	0.25	0.20	0.15

- (a) If Finan bases his box purchase on an expected length of 600 pages, will he have enough boxes?
- (b) If all 700-page books are edited down to 500 pages, what expected number of pages should he use?

5-80 D'Addario Rose Co. is planning rose production for Valentine's Day. Each rose costs \$0.35 to raise and sells wholesale for \$0.70. Any roses left over after Valentine's Day can be sold the next day for \$0.10 wholesale. D'Addario has the following probability distribution based on previous years:

Roses sold	15,000	20,000	25,000	30,000
Probability	0.10	0.30	0.40	0.20

How many roses should D'Addario produce to minimize the firm's expected losses?

5-81 A certain business school has 400 students in its MBA program. One hundred sixteen of the students are married. Without using Appendix Table 3, determine

- (a) The probability that exactly 2 of 3 randomly selected students are married.
- (b) The probability that exactly 4 of 13 students chosen at random are married.

5-82 Kenan Football Stadium has 4 light towers with 25 high-intensity floodlights mounted on each. Sometimes an entire light tower will go dark. Smitty Moyer, head of maintenance, wonders what the distribution of light tower failures is. He knows that any individual tower has a probability of 0.11 of failing during a football game and that the towers fail independently of one another.

Construct a graph, like Figure 5-4, of a binomial probability distribution showing the probabilities of exactly 0, 1, 2, 3, or 4 towers going dark during the same game.

5-83 Smitty Moyer (see Exercise 5-82) knows that the probability that any one of the 25 individual floodlights in a light tower fails during a football game is 0.05. The individual floodlights in a tower fail independently of each other.

- (a) Using both the binomial and the Poisson approximation, determine the probability that seven floodlights from a given tower will fail during the same game.
- (b) Using both methods, determine the probability that two will fail.

5-84 Ansel Fearrington wants to borrow \$75,000 from his bank for a new tractor for his farm. The loan officer doesn't have any data specifically on the bank's history of equipment loans, but he does tell Ansel that over the years, the bank has received about 1460 loan applications per year and that the probability of approval was, on average, about 0.8.

- (a) Ansel is curious about the average and standard deviation of the number of loans approved per year. Find these figures for him.
- (b) Suppose that after careful research the loan officer tells Ansel the correct figures actually are 1,327 applications per year with an approval probability of 0.77. What are the mean and standard deviation now?

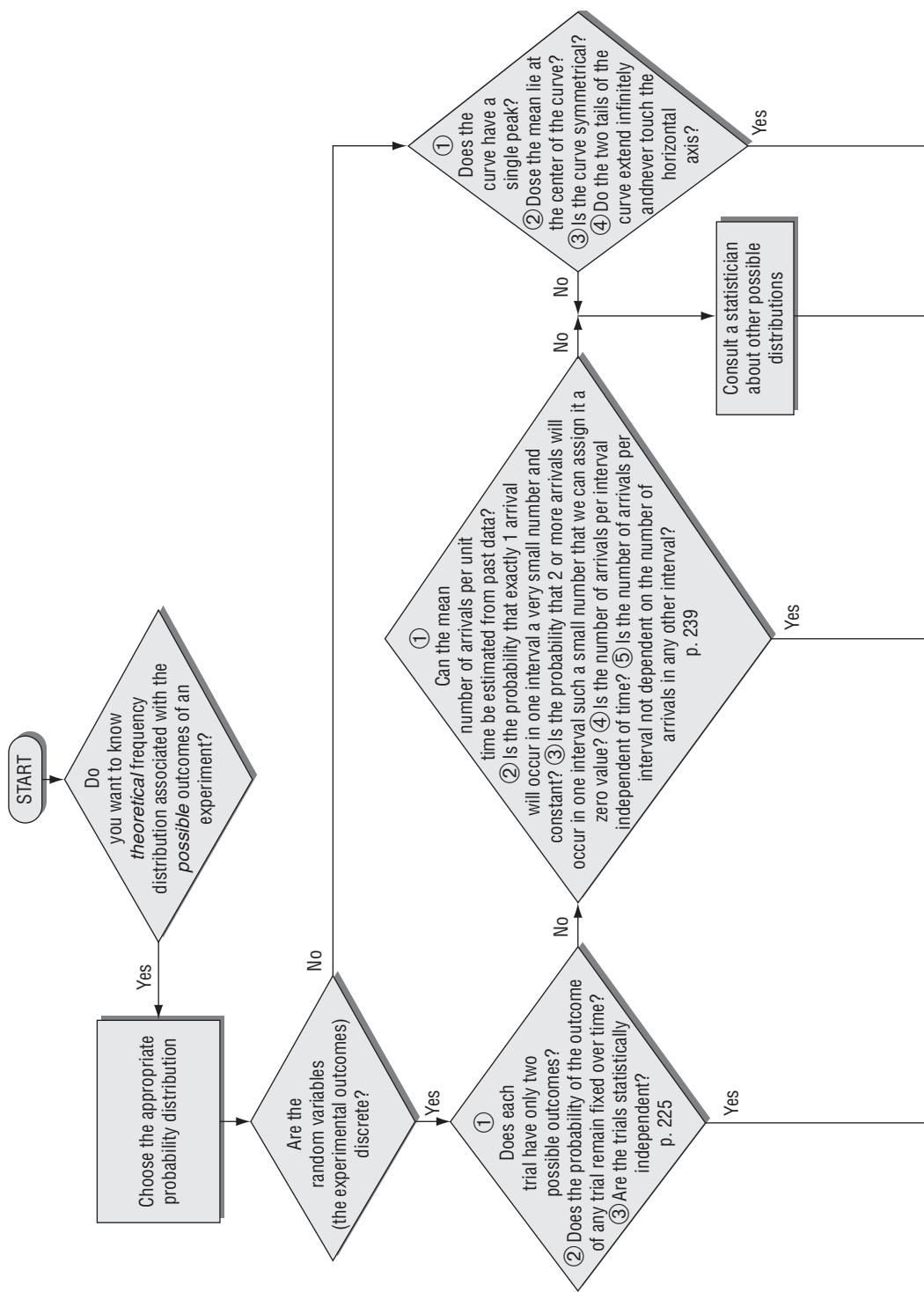
5-85 Ansel Fearrington (see Exercise 5-84) learns that the loan officer has been fired for failing to follow bank lending guidelines. The bank now announces that all financially sound loan applications will be approved. Ansel guesses that three out of every five applications are unsound.

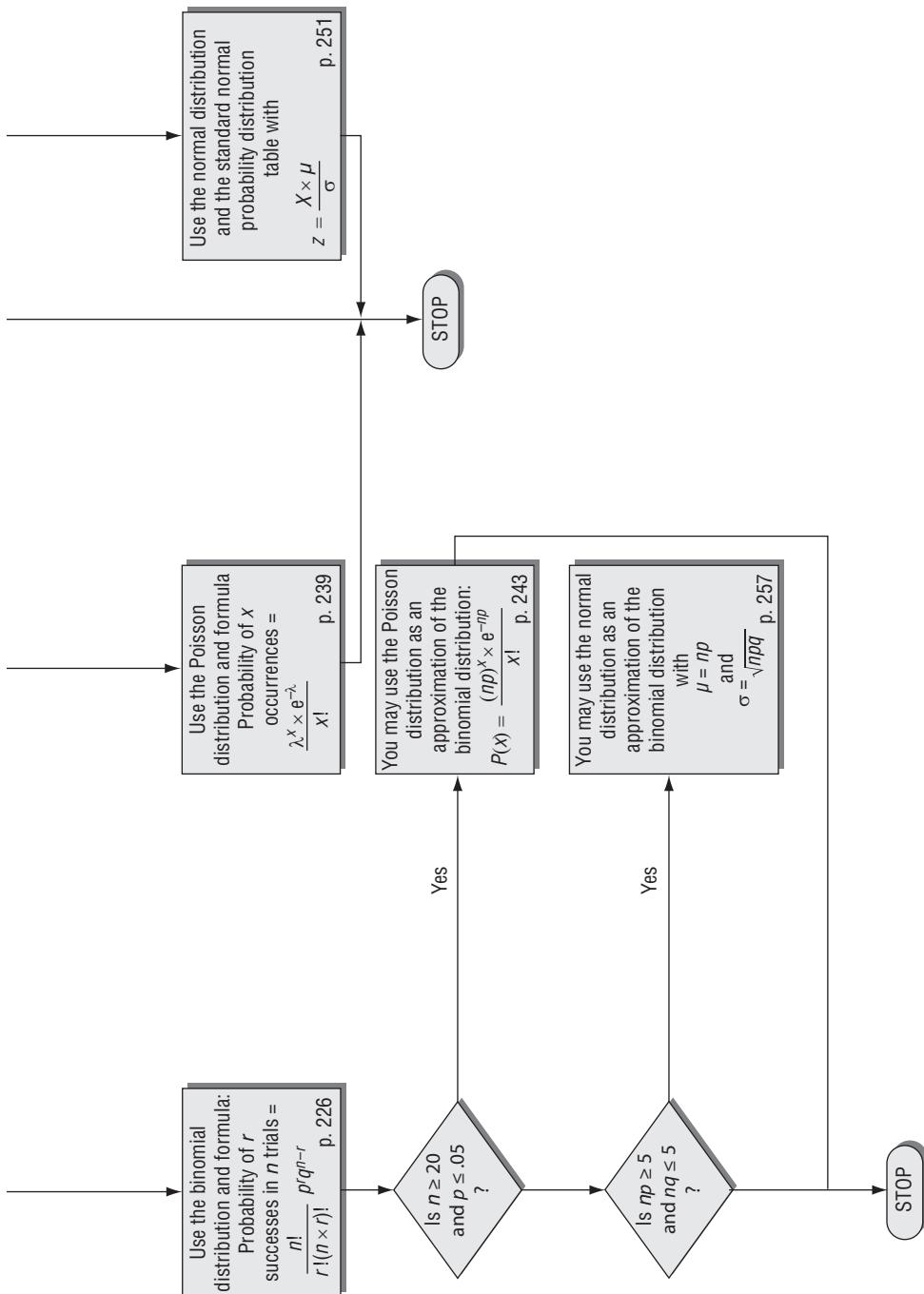
- (a) If Ansel is right, what is the probability that exactly 6 of the next 10 applications will be approved?
- (b) What is the probability that more than 3 will be approved?
- (c) What is the probability that more than 2 but fewer than 6 will be approved?

- 5-86** Krista Engel is campaign manager for a candidate for U.S. Senator. Staff consensus is that the candidate has the support of 40 percent of registered voters. A random sample of 300 registered voters shows that 34 percent would vote for Krista's candidate. If 40 percent of voters really are allied with her candidate, what is the probability that a sample of 300 voters would indicate 34 percent or fewer on her side? Is it likely that the 40 percent estimate is correct?
- 5-87** Krista Engel (see Exercise 5-86) has learned that her candidate's major opponent, who has the support of 50 percent of registered voters, will likely lose the support of $\frac{1}{4}$ of those voters because of his recent support of clear-cutting of timber in national forests, a policy to which Krista's candidate is opposed.

If Krista's candidate now has the support of 34 percent of registered voters, and all the dissatisfied voters then switch to Krista's candidate, what is the probability that a new survey of 250 registered voters would show her candidate to have the support of 51 to 55 percent of the voters?

Flow Chart: Probability Distribution





This page is intentionally left blank.

Sampling and Sampling Distributions

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To take a sample from an entire population and use it to describe the population
 - To make sure the samples you do take are an accurate representation of the population from which they came
 - To introduce the concepts of sampling distributions
 - To understand the trade-offs between the cost of taking larger samples and the additional accuracy this gives to decisions made from them
 - To introduce experimental design: sampling procedures to gather the most information for the least cost
-

CHAPTER CONTENTS

6.1	Introduction to Sampling	278
6.2	Random Sampling	281
6.3	Non-Random Sampling	289
6.4	Design of Experiments	292
6.5	Introduction to Sampling Distributions	296
6.6	Sampling Distributions in More Detail	300
6.7	An Operational Consideration in Sampling: The Relationship between Sample Size and Standard Error	313

■	Statistics at Work	319
■	Terms Introduced in Chapter 6	320
■	Equations Introduced in Chapter 6	322
■	Review and Application Exercises	323
■	Flow Chart: Sampling and Sampling Distributions	326

Although there are over 200 million TV viewers in the United States and somewhat over half that many TV sets, only about 1,000 of those sets are sampled to determine what programs Americans watch. Why select only about 1,000 sets out of 100 million? Because time and the average cost of an interview prohibit the rating companies from trying to reach millions of people. And since polls are reasonably accurate, interviewing everybody is unnecessary. In this chapter, we examine questions such as these: How many people should be interviewed? How should they be selected? How do we know when our sample accurately reflects the entire population? ■

6.1 INTRODUCTION TO SAMPLING

Shoppers often sample a small piece of cheese before purchasing any. They decide from one piece what the larger chunk will taste like. A chemist does the same thing when he takes a sample of alcohol from a still, determines that it is 90 proof, and infers that all the alcohol in the still is 90 proof. If the chemist tests all the alcohol or the shoppers taste all the cheese, there will be none to sell. Testing all of the product often destroys it and is unnecessary. To determine the characteristics of the whole, we have to sample only a portion.

Reasons for sampling

Suppose that, as the personnel director of a large bank, you need to write a report describing all the employees who have voluntarily left the company in the last 10 years. You would have a difficult task locating all these thousands of people. They are not easily accessible as a group—many have died, moved from the community, left the country, or acquired a new name by marriage. How do you write the report? The best idea is to locate a representative sample and interview them in order to generalize about the entire group.

Time is also a factor when managers need information quickly in order to adjust an operation or change a policy. Consider an automatic machine that sorts thousands of pieces of mail daily. Why wait for an entire day's output to check whether the machine is working accurately (whether the *population characteristics* are those required by the postal service)? Instead, samples can be taken at specific intervals, and if necessary, the machine can be adjusted right away.

Census or sample

Sometimes it is possible and practical to examine every person or item in the population we wish to describe. We call this a *complete enumeration*, or *census*. We use sampling when it is not possible to count or measure every item in the population.

Examples of populations and samples

Statisticians use the word *population* to refer not only to people but to all items that have been chosen for study. In the cases we have just mentioned, the populations are all the cheese in the chunk, all the whiskey in the vat, all the employees of the large bank who voluntarily left in the last 10 years, and all mail sorted by the automatic machine since the previous sample check. **Statisticians use the word *sample* to describe a portion chosen from the population.**

Function of statistics and parameters

Statistics and Parameters

Mathematically, we can describe samples and populations by using measures such as the mean, median, mode, and standard deviation, which we introduced in Chapter 3. When these terms describe the characteristics of a sample, they are called *statistics*.

When they describe the characteristics of a population, they are called *parameters*. A statistic is a characteristic of a sample; a parameter is a characteristic of a population.

Suppose that the mean height in inches of all tenth graders in the United States is 60 inches. In this case, 60 inches is a characteristic of the population “all tenth graders” and can be called a *population parameter*. On the other hand, if we say that the mean height in Ms. Jones’s tenth-grade class in Bennettsville is 60 inches, we are using 60 inches to describe a characteristic of the sample “Ms. Jones’s tenth graders.” In that case, 60 inches would be a *sample statistic*. If we are convinced that the mean height of Ms. Jones’s tenth graders is an accurate estimate of the mean height of all tenth graders in the United States, we could use the sample statistic “mean height of Ms. Jones’s tenth graders” to estimate the population parameter “mean height of all U.S. tenth graders” without having to measure all the millions of tenth graders in the United States.

To be consistent, statisticians use lowercase Roman letters to denote sample statistics and Greek or capital letters for population parameters. Table 6-1 lists these symbols and summarizes the definitions we have studied so far in this chapter.

Using statistics to estimate parameters

N , μ , σ , and n , \bar{x} , s : standard symbols

Types of Sampling

There are two methods of selecting samples from populations: *nonrandom* or *judgment* sampling, and *random* or *probability* sampling. In probability sampling, all the items in the population have a chance of being chosen in the sample. In judgment sampling, personal knowledge and opinion are used to identify the items from the population that are to be included in the sample. A sample selected by judgment sampling is based on someone’s expertise about the population. A forest ranger, for example, would have a judgment sample if he decided ahead of time which parts of a large forested area he would walk through to estimate the total board feet of lumber that could be cut. Sometimes a judgment sample is used as a pilot or trial sample to decide how to take a random sample later. The rigorous statistical analysis that can be done with probability samples cannot be done with judgment samples. They are more convenient and can be used successfully even if we are unable to measure their validity. But if a study uses judgment sampling and loses a significant degree of representativeness, it will have purchased convenience at too high a price.

Judgment and probability sampling

TABLE 6-1 DIFFERENCES BETWEEN POPULATIONS AND SAMPLES

	Population	Sample
DEFINITION	Collection of items being considered	Part or portion of the population chosen for study
CHARACTERISTICS	“Parameters”	“Statistics”
SYMBOLS	Population size = N Population mean = μ Population standard deviation = σ	Sample size = n Sample mean = \bar{x} Sample standard deviation = s

Biased Samples

The Congress is debating some gun control laws. You are asked to conduct an opinion survey. Because hunters are the ones that are most affected by the gun control laws, you went to a hunting lodge and interviewed the members there. Then you reported that in a survey done by you, about 97 percent of the respondents were in favor of repealing all gun control laws.

A week later, the Congress took up another bill: “Should working pregnant women be given a maternity leave of one year with full pay to take care of newborn babies?” Because this issue affects women most, this time you went to all the high-rise office complexes in your city and interviewed several working women of child-bearing age. Again you reported that in a survey done by you, about 93 percent of the respondents were in favor of the one-year maternity leave with full pay.

couple of biased polls

In both of these situations you picked a biased sample by choosing people who would have very strong feelings on one side of the issue. How can we be sure that pollsters we listen to and read about don’t make the same mistake you did? The answer is that unless the pollsters have a strong reputation for statistically accurate polling, we can’t. However, we can be alert to the risks we take when we don’t ask for more information or do more research into their competence.

EXERCISES 6.1

Basic Concepts

- 6-1 What is the major drawback of judgment sampling?
- 6-2 Are judgment sampling and probability sampling necessarily mutually exclusive? Explain.
- 6-3 List the advantages of sampling over complete enumeration, or census.
- 6-4 What are some disadvantages of probability sampling versus judgment sampling?

Applications

- 6-5 Farlington Savings and Loan is considering a merger with Sentry Bank, but needs shareholder approval before the merger can be accomplished. At its annual meeting, to which all shareholders are invited, the president of FS&L asks the shareholders whether they approve of the deal. Eighty-five percent approve. Is this percentage a sample statistic or a population parameter?
- 6-6 Jean Mason, who was hired by Former Industries to determine employee attitudes toward the upcoming union vote, met with some difficulty after reporting her findings to management. Mason’s study was based on statistical sampling, and from the beginning data, it was clear (or so Jean thought) that the employees were favoring a unionized shop. Jean’s report was shrugged off with the comment, “This is no good. Nobody can make statements about employee sentiments when she talks to only a little over 15 percent of our employees. Everyone knows you have to check 50 percent to have any idea of what the outcome of the union vote will be. We didn’t hire you to make guesses.” Is there any defense for Jean’s position?
- 6-7 A consumer protection organization is conducting a census of people who were injured by a particular brand of space heater. Each victim is asked questions about the behavior of the heater just before its malfunction; this information generally is available only from the victim, because the heater in question tends to incinerate itself upon malfunction. Early in the census, it is discovered that several of the victims were elderly and have died. Is any census of the victims now possible? Explain.

6.2 RANDOM SAMPLING

In a random or probability sample, we know what the chances are that an element of the population will or will not be included in the sample. As a result, we can assess objectively the estimates of the population characteristics that result from our sample; that is, we can describe mathematically how objective our estimates are. Let us begin our explanation of this process by introducing four methods of random sampling:

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

Simple Random Sampling

Simple random sampling selects samples by methods that allow *each possible sample to have an equal probability of being picked* and *each item in the entire population to have an equal chance of being included in the sample*. We can illustrate these requirements with an example. Suppose we have a population of four students in a seminar and we want samples of two students at a time for interviewing purposes. Table 6-2 illustrates all of the possible combinations of samples of two students in a population size of four, the probability of each sample being picked, and the probability that each student will be in a sample.

An example of simple random sampling

Our example illustrated in Table 6-2 uses a *finite* population of four students. By *finite*, we mean that the population has stated or limited size, that is to say, there is a whole number (N) that tells

Defining finite and with replacement

TABLE 6-2 CHANCES OF SELECTING SAMPLES OF TWO STUDENTS FROM A POPULATION OF FOUR STUDENTS

Students <i>A</i> , <i>B</i> , <i>C</i> , and <i>D</i>	
Possible samples of two people: <i>AB</i> , <i>AC</i> , <i>AD</i> , <i>BC</i> , <i>BD</i> , <i>CD</i>	
Probability of drawing this sample of two people must be	
$P(AB) = \frac{1}{6}$	
$P(AC) = \frac{1}{6}$	
$P(AD) = \frac{1}{6}$	(There are only six possible samples of two people)
$P(BC) = \frac{1}{6}$	
$P(BD) = \frac{1}{6}$	
$P(CD) = \frac{1}{6}$	
Probability of this student in the sample must be	
$P(A) = \frac{1}{2}$	[In Chapter 4, we saw that the marginal probability is equal to the sum of the joint probabilities of the events within which the event is contained: $P(A) = P(AB) + P(AC) + P(AD) = \frac{1}{2}$
$P(B) = \frac{1}{2}$	
$P(C) = \frac{1}{2}$	
$P(D) = \frac{1}{2}$	

us how many items there are in the population. Certainly, if we sample without “replacing” the student, we shall soon exhaust our small population group. Notice, too, that if we *sample with replacement* (that is, if we replace the sampled student immediately after he or she is picked and before the second student is chosen), the same person could appear twice in the sample.

We have used this example only to help us think about sampling from an infinite population. An *infinite population* is a population in which it is theoretically impossible to observe all the elements. Although many populations appear to be exceedingly large, no truly infinite population of physical objects actually exists. After all, given unlimited resources and time, we could enumerate any finite population, even the grains of sand on the beaches of North America. As a practical matter, then, we will use the term *infinite population* when we are talking about a population that could not be enumerated in a reasonable period of time. In this way, we will use the theoretical concept of infinite population as an approximation of a large finite population, just as we earlier used the theoretical concept of continuous random variable as an approximation of a discrete random variable that could take on many closely spaced values.

An infinite population

How to Do Random Sampling The easiest way to select a sample randomly is to use random numbers. These numbers can be generated either by a computer programmed to scramble numbers or by a table of random numbers, which should properly be called a *table of random digits*.

Table 6-3 illustrates a portion of such a table. Here we have 1,150 random digits in sets of 10 digits. These numbers have been generated by a completely random process. The probability that any one digit from 0 through 9 will appear is the same as that for any other digit, and the probability of one sequence of digits occurring is the same as that for any other sequence of the same length.

To see how to use this table, suppose that we have 100 employees in a company and wish to interview a randomly chosen sample of 10. We could get such a random sample by assigning every employee a number of 00 to 99, consulting Table 6-3, and picking a systematic method of selecting two-digit numbers. In this case, let's do the following:

Using a table of random digits

1. Go from the top to the bottom of the columns beginning with the left-hand column, and read only the first two digits in each row. Notice that our first number using this method would be 15, the second 09, the third 41, and so on.
2. If we reach the bottom of the last column on the right and are still short of our desired 10 two-digit numbers of 99 and under, we can go back to the beginning (the top of the left-hand column) and start reading the third, and fourth digits of each number. These would begin 81, 28, and 12.

Another way to select our employees would be to write the name of each one on a slip of paper and deposit the slips in a box. After mixing them thoroughly, we could draw 10 slips at random.

Using slips of paper

This method works well with a small group of people but presents problems if the people in the population number in the thousands. There is the added problem, too, of not being certain that the slips of paper are mixed well. In the draft lottery of 1970, for example, when capsules were drawn from a bowl to determine by birthdays the order for selecting draftees for the armed services, December birthdays appeared more often than the probabilities would have suggested. As it turned out, the December capsules had been placed in the bowl last, and the capsules had not been mixed properly. Thus, December capsules had the highest probability of being drawn.

TABLE 6.3 1,150 RANDOM DIGITS*

1581922396	2068577984	8262130892	8374856049	4637567488
0928105582	7295088579	9586111652	7055508767	6472382934
4112077556	3440672486	1882412963	0684012006	0933147914
7457477468	5435810788	9670852913	1291265730	4890031305
0099520858	3090908872	2039593181	5973470495	9776135501
7245174840	2275698645	8416549348	4676463101	2229367983
6749420382	4832630032	5670984959	5432114610	2966095680
5503161011	7413686599	1198757695	0414294470	0140121598
7164238934	7666127259	5263097712	5133648980	4011966963
3593969525	0272759769	0385998136	9999089966	7544056852
4192054466	0700014629	5169439659	8408705169	1074373131
9697426117	6488888550	4031652526	8123543276	0927534537
2007950579	9564268448	3457416988	1531027886	7016633739
4584768758	2389278610	3859431781	3643768456	4141314518
3840145867	9120831830	7228567652	1267173884	4020651657
0190453442	4800088084	1165628559	5407921254	3768932478
6766554338	5585265145	5089052204	9780623691	2195448096
6315116284	9172824179	5544814339	0016943666	3828538786
3908771938	4035554324	0840126299	4942059208	1475623997
5570024586	9324732596	1186563397	4425143189	3216653251
2999997185	0135968938	7678931194	1351031403	6002561840
7864375912	8383232768	1892857070	2323673751	3188881718
7065492027	6349104233	3382569662	4579426926	1513082455

*Based on first 834 serial numbers of selective service lottery as reported by The New York Times, October 30, 1940, p. 12.
 © 1940 by The New York Times Company. Reprinted by permission.

Systematic Sampling

In *systematic sampling*, elements are selected from the population at a uniform interval that is measured in time, order, or space. If we wanted to interview every twentieth student on a college campus, we would choose a random starting point in the first 20 names in the student directory and then pick every twentieth name thereafter.

Systematic sampling differs from simple random sampling in that each *element* has an equal chance of being selected but each *sample* does *not* have an equal chance of being selected. This would have been the case if, in our earlier example, we had assigned numbers between 00 and 99 to our employees and then had begun to choose a sample of 10 by picking every tenth number beginning

Characteristics of systematic sampling

1, 11, 21, 31, and so forth. Employees numbered 2, 3, 4, and 5 would have had *no* chance of being selected together.

In systematic sampling, there is the problem of introducing an error into the sample process. Suppose we were sampling paper waste produced by households, and we decided to sample 100 households every Monday. Chances are high that our sample would not be representative, because Monday's trash would very likely include the Sunday newspaper. Thus, the amount of waste would be biased upward by our choice of this sampling procedure.

Systematic sampling has advantages, too, however. Even though systematic sampling may be inappropriate when the elements lie in a sequential pattern, this method may require less time and sometimes results in lower costs than the simple random-sample method.

Shortcomings of systematic sampling

Stratified Sampling

To use *stratified sampling*, we divide the population into relatively homogeneous groups, called *strata*. Then we use one of two approaches. Either we select at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole or we draw an equal number of elements from each stratum and give weight to the results according to the stratum's proportion of total population. With either approach, stratified sampling guarantees that every element in the population has a chance of being selected.

Two ways to take stratified samples

Stratified sampling is appropriate when the population is already divided into groups of different sizes and we wish to acknowledge this fact. Suppose that a physician's patients are divided into four groups according to age, as shown in Table 6-4. The physician wants to find out how many hours his patients sleep. To obtain an estimate of this characteristic of the population, he could take a random sample from each of the four age groups and give weight to the samples according to the percentage of patients in that group. This would be an example of a stratified sample.

When to use stratified sampling

TABLE 6-4 COMPOSITION OF PATIENTS BY AGE

Age Group	Percentage of Total
Birth–19 years	30
20–39 years	40
40–59 years	20
60 years and older	10

The advantage of stratified samples is that when they are properly designed, they more accurately reflect characteristics of the population from which they were chosen than do other kinds of samples.

Cluster Sampling

In *cluster sampling*, we divide the population into groups, or *clusters*, and then select a random sample of these clusters. We assume that these individual clusters are representative of the population as a whole. If a market research team is attempting to determine by sampling the average number of television sets per household in a large city, they could use a city map to divide the territory into blocks and then choose a certain number of blocks (clusters) for interviewing. Every household in each of these blocks would be interviewed. A well-designed cluster sampling procedure can produce a more precise sample at considerably less cost than that of simple random sampling.

With both stratified and cluster sampling, the population is divided into well-defined groups. We use *stratified* sampling when each group has small variation within itself but there is a wide variation between the groups. We use *cluster* sampling in the case where there is considerable variation within each group but the groups are essentially homogeneous.

Comparison of stratified and cluster sampling

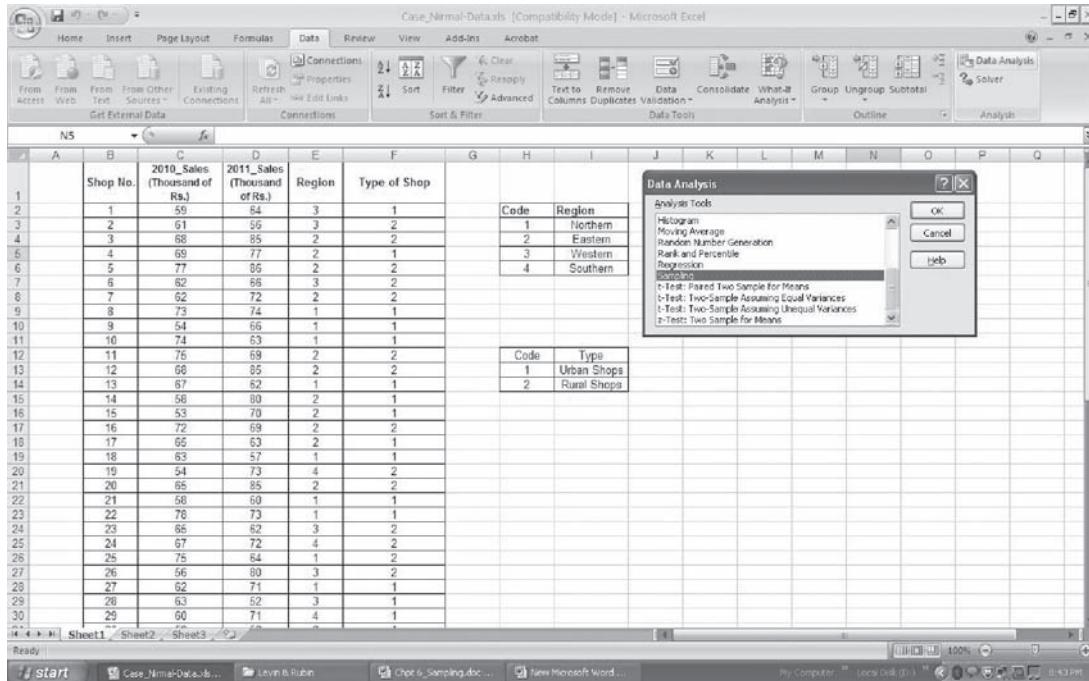
Basis Of Statistical Inference: Simple Random Sampling

Systematic sampling, stratified sampling, and cluster sampling attempt to approximate simple random sampling. All are methods that have been developed for their precision, economy, or physical ease. Even so, assume for the rest of the examples and problems in this book that we obtain our data using simple random sampling. The principles of simple random sampling are the foundation for making inferences about populations from information contained in samples. Because the inferential statistics that have been developed for simple random sampling, their extensions, and their modifications are conceptually quite simple but somewhat involved mathematically, if you have been involved in simple random sampling, you will have a good grasp of the concepts even if you must leave the technical details to the professional statistician.

Why we assume random sampling

Drawing a Random Sample Using MS Excel

MS-Excel can be used to draw a random sample from a list of population elements. For drawing a random sample go to **Data > Data Analysis > Sampling**.



When the **Sampling** dialogue-box opens, enter the range of population elements into **Input Range**, check **Random** option button under **Sampling Method** and enter desired **sample-size**. Pressing **OK** will give you the desired random sample.

The screenshot shows a Microsoft Excel window titled "Case_Normal-Data.xls [Compatibility Mode] - Microsoft Excel". The "Sampling" dialog box is open in the foreground, overlaid on the main spreadsheet area. The dialog box has the following settings:

- Input Range:** \$C\$1:\$C\$61
- Labels:** Checked
- Sampling Method:** Random (radio button selected)
- Number of Samples:** 15
- Output options:** New Worksheet By: (radio button selected)

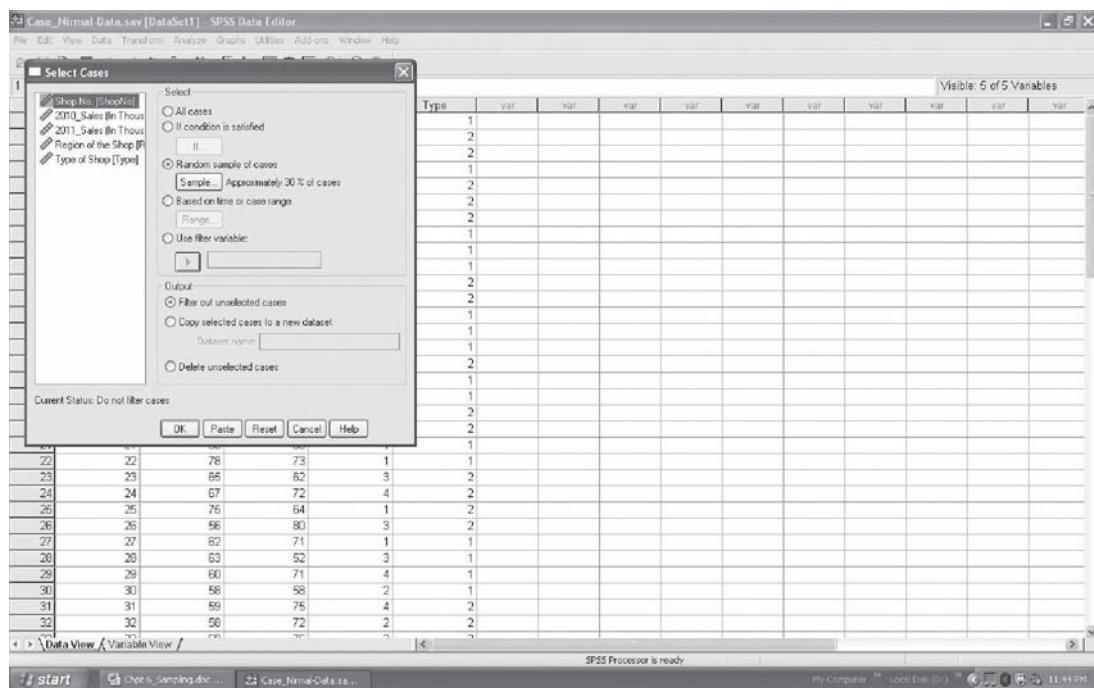
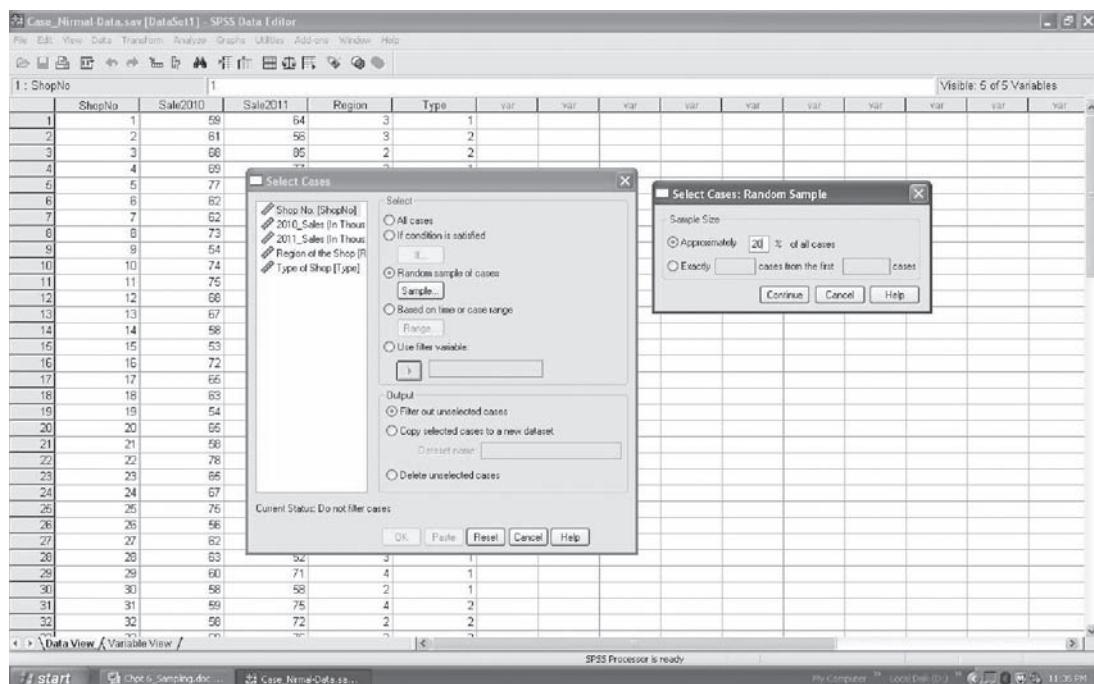
The main spreadsheet contains data from row 38 to 61, columns A through F. The data consists of two columns of numbers: column A (37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62) and column C (59, 73, 64, 74, 51, 64, 54, 70, 63, 65, 60, 67, 61, 59, 62, 61, 74, 59, 89, 83). The "Sampling" dialog box is positioned in the upper right quadrant of the screen.

The screenshot shows the same Microsoft Excel window as the previous one, but now the sampled data is visible in a new worksheet. The "Sampling" dialog box is no longer open. The new worksheet, titled "Sampling", contains 15 rows of data, labeled "Sample" from 1 to 15. The data is identical to the original population data, but only 15 rows are present. The "Sampling" dialog box is no longer open.

Sample	A	B	C	D	E	F
1	37		59	70	2	1
2	38		73	75	1	2
3	39		64	76	3	2
4	40		74	58	1	1
5	41		51	75	3	1
6	42		64	65	1	1
7	43		54	71	2	1
8	44		70	69	1	1
9	45		63	78	1	1
10	46		65	65	1	2
11	47		60	60	2	1
12	48		67	75	2	2
13	49		61	79	3	2
14	50		59	85	2	2
15	51		62	75	1	1
16	52		62	68	2	2
17	53		63	55	1	2
18	54		57	67	3	2
19	55		70	67	1	1
20	56		62	59	3	1
21	57		61	74	1	1
22	58		67	59	1	1
23	59		89	75	2	2
24	60		83	75	4	1

Drawing a Random Sample Using SPSS

SPSS can also be used to draw a random sample from a list of population elements. For drawing a random sample go to **Data > Select Cases**. Click on **Random sample of cases** and press **Sample**. In the resulting sub-dialogue box **Select Cases: Random Sample**, there are two options, that can be used to instruct appropriate number of samples. One is Approximately _____ % of all cases and the second is Exactly _____ cases from the first _____ cases. Then press **Continue** and upon coming back to main dialogue, press **OK**.



	ShopNo	Sale2010	Sale2011	Region	Type	filter, \$	var1	var2	var3	var4	var5	var6	var7
1	1	59	64	3	1	1							
2	2	61	56	3	2	0							
3	3	68	65	2	2	0							
4	4	69	77	2	1	1							
5	5	77	86	2	2	1							
6	6	82	68	3	2	0							
7	7	62	72	2	2	0							
8	8	73	74	1	1	0							
9	9	54	66	1	1	1							
10	10	74	63	1	1	1							
11	11	75	69	2	2	1							
12	12	68	95	2	2	1							
13	13	67	62	1	1	0							
14	14	58	80	2	1	1							
15	15	53	70	2	1	0							
16	16	72	69	2	2	0							
17	17	65	63	2	1	0							
18	18	63	57	1	1	1							
19	19	54	73	4	2	0							
20	20	65	85	2	2	1							
21	21	56	60	1	1	1							
22	22	78	73	1	1	0							
23	23	85	62	3	2	1							
24	24	67	72	4	2	1							
25	25	76	64	1	2	0							
26	26	56	80	3	2	0							
27	27	62	71	1	1	1							
28	28	63	52	3	1	0							
29	29	60	71	4	1	1							
30	30	58	58	2	1	0							
31	31	59	75	4	2	1							
32	32	58	72	2	2	1							

6.3 NON-RANDOM SAMPLING

Non-random sampling designs do not provide unit in the population a known chance of being selected in the sample. The selection procedure is partially subjective. These sampling designs do not provide representative sample because of lack of objectivity but they can be more frequently applied in business because here complete list of population elements is not required for sampling. So, inspite of being less effective, non-random sampling designs are widely used in business scenario. Some popular non-random sampling designs are as follows:

Convenience Sampling It is based on convenience of the researcher. Researcher selects the sample which is most convenient to him/her. No planning is required for the sampling. It is least effective and should be used only for introductory purpose and not for conclusive purpose.

Judgement Sampling Researcher exercises his/her judgement to draw a sample which he/she thinks is representative of the population or otherwise appropriate. It is also known as Purposive Sampling. It is better than the previous one but personal bias limits the applicability of this sampling scheme.

Quota Sampling It consists of fixation of certain quotas on the basis of certain parameter (s) so as to make sample representative of the population under study. It is one of the most commonly used non-random sampling. It is much similar to stratified random sampling but does not require complete list of population elements.

Shopping-Mall Intercept Sampling This sample involves drawing samples (establishing malls) in market-places, shopping malls, fairs in different socioeconomic locations, so to make

sample representative of the population. This scheme is very popular because of convenience and representativeness.

Snowball Sampling Here, initial respondents are selected randomly then additional respondents are selected by their referrals and so on. This scheme is also known as Multiplicity Sampling. It is useful for rare population.

HINTS & ASSUMPTIONS

Warning: Even when precautions are taken, many so-called random samples are still not random. When you try to take a random sample of mall shoppers, you get a biased sample because many people are not willing to take the time to stop to talk to the interviewer. Nowadays, when telephone pollers try to take a random sample, often they don't get through to people with call-screening devices on their phones. There *are* ways to counter these problems in random sampling, but often the "fix" is more complicated and/or costly than the sampling organization wants to face.

EXERCISES 6.2

Self-Check Exercises

SC 6-1 If we have a population of 10,000 and we wish to sample 20 randomly, use the random digits table (Table 6-3) to select 20 individuals from the 10,000. List the numbers of the elements selected, based on the random digits table.

SC 6-2 A Senate study on the issue of self-rule for the District of Columbia involved surveying 2,000 people from the population of the city regarding their opinions on a number of issues related to self-rule. Washington, D.C., is a city in which many neighborhoods are poor and many neighborhoods are rich, with very few neighborhoods falling between the extremes. The researchers who were administering the survey had reasons to believe that the opinions expressed on the various questions would be highly dependent on income. Which method was more appropriate, stratified sampling or cluster sampling? Explain briefly.

Basic Concepts

6-8 In the examples below, probability distributions for three natural subgroups of a larger population are shown. For which situation would you recommend stratified sampling?



6-9 We wish to sample 15 pages from this textbook. Use the random digits table (Table 6-3) to select 15 pages at random and count the number of words in *italics* on each page. Report your results.

- 6-10** Using a calendar, systematically sample every eighteenth day of the year, beginning with January 6.
- 6-11** A population is made up of groups that have wide variation within each group but little variation from group to group. The appropriate type of sampling for this population is
(a) Stratified.
(b) Systematic.
(c) Cluster.
(d) Judgment.
- 6-12** Consult Table 6-3. What is the probability that a 4 will appear as the leftmost digit in each set of 10 digits? That a 7 will appear? 2? How many times would you expect to see each of these digits in the leftmost position? How many times is each found in that position? Can you explain any differences in the number found and the number expected?

Applications

- 6-13** The local cable television company is planning to add one channel to its basic service. There are five channels to choose from, and the company would like some input from its subscribers. There are about 20,000 subscribers, and the company knows that 35 percent of these are college students, 45 percent are white-collar workers, 15 percent are blue-collar workers, and 5 percent are other. However, the company believes there is much variation within these groups. Which of the following sampling methods is more appropriate: random, systematic, stratified, or cluster sampling?
- 6-14** A nonprofit organization is conducting a door-to-door opinion poll on municipal day-care centers. The organization has devised a scheme for random sampling of houses, and plans to conduct the poll on weekdays from noon to 5 P.M. Will this scheme produce a random sample?
- 6-15** Bob Peterson, public relations manager for Piedmont Power and Light, has implemented an institutional advertising campaign to promote energy consciousness among its customers. Peterson, anxious to know whether the campaign has been effective, plans to conduct a telephone survey of area residents. He plans to look in the telephone book and select random numbers with addresses that correspond to the company's service area. Will Peterson's sample be a random one?
- 6-16** At the U.S. Mint in Philadelphia, 10 machines stamp out pennies in lots of 50. These lots are arranged sequentially on a single conveyor belt, which passes an inspection station. An inspector decides to use systematic sampling in inspecting the pennies and is trying to decide whether to inspect every fifth or every seventh lot of pennies. Which is better? Why?
- 6-17** The state occupational safety board has decided to do a study of work-related accidents within the state, to examine some of the variables involved in the accidents, such as the type of job, the cause of the accident, the extent of the injury, the time of day, and whether the employer was negligent. It has been decided that 250 of the 2,500 work-related accidents reported last year in the state will be sampled. The accident reports are filed by date in a filing cabinet. Marsha Gulley, a department employee, has proposed that the study use a systematic sampling technique and select every tenth report in the file for the sample. Would her plan of systematic sampling be appropriate here? Explain.
- 6-18** Bob Bennett, product manager for Clipper Mowers Company is interested in looking at the kinds of lawn mowers used throughout the country. Assistant product manager Mary Wilson has recommended a stratified random-sampling process in which the cities and communities

studied are separated into substrata, depending on the size and nature of the community. Mary Wilson proposes the following classification

Category	Type of Community
Urban	Inner city (population 100,000+)
Suburban	Outlying areas of cities or smaller communities (pop. 20,000 to 100,000)
Rural	Small communities (fewer than 20,000 residents)

Is stratified random sampling appropriate here?

Worked-Out Answers to Self-Check Exercises

SC 6-1 Starting at the top of the third column and choosing the last 4 digits of the numbers in that column gives the following sample (reading across rows):

892	1652	2963	2913	3181	9348	4959
7695	7712	8136	9659	2526	6988	1781
7652	8559	2204	4339	6299	3397	

SC 6-2 Stratified sampling is more appropriate in this case because there appear to be two very dissimilar groups, which probably have smaller variation within each group than between groups.

6.4 DESIGN OF EXPERIMENTS

We encountered the term *experiment* in Chapter 4, “Probability I.” There we defined an *event* as one or more of the possible outcomes of doing something, and an *experiment* as an activity that would produce such events. In a coin-toss experiment, the possible events are heads and tails.

Events and experiments revisited

Planning Experiments

If we are to conduct experiments that produce meaningful results in the form of usable conclusions, the way in which these experiments are designed is of the utmost importance. Sections 6.1 and 6.2 discussed ways of ensuring that random sampling was indeed being done. The way in which sampling is conducted is only a part of the total design of an experiment. In fact, the design of experiments is itself the subject of quite a number of books, some of them rather formidable in both scope and volume.

Sampling is only one part

Phases of Experimental Design

To get a better feel for the complexity of experimental design without actually getting involved with the complex details, take an example from the many that confront us every day, and follow that example through from beginning to end.

A claim is made

The statement is made that a Crankmaster battery will start your car's engine better than Battery X. Crankmaster might design its experiment in the following way.

Objective This is our beginning point. Crankmaster wants to test its battery against the leading competitor. Although it is possible to design an experiment that would test the two batteries on several characteristics (life, size, cranking power, weight, and cost, to name but a few), Crankmaster has decided to limit this experiment to cranking power.

Objectives are set

What Is to Be Measured This is often called the response variable. If Crankmaster is to design an experiment that compares the cranking power of its battery to that of another, it must define how cranking power is to be measured. Again, there are quite a few ways in which this can be done. For example, Crankmaster could measure the time it took for the batteries to run down completely while cranking engines, the total number of engine starts it took to run down the batteries, or the number of months in use that the two batteries could be expected to last. Crankmaster decides that the response variable in its experiment will be the time it takes for batteries to run down completely while cranking engines.

The response variable is selected

How Large a Sample Size Crankmaster wants to be sure that it chooses a sample size large enough to support claims it makes for its battery, without fear of being challenged; however, it knows that the more batteries it tests, the higher the cost of conducting the experiment. As we shall point out in Section 6 of this chapter, there is a diminishing return in sampling; and although sampling more items does, in fact, improve accuracy, the benefit may not be worth the cost. Not wishing to choose a sample size that is too expensive to contend with, Crankmaster decides that comparing 10 batteries from each of the two companies (itself and its competitor) will suffice.

How many to test

Conducting the Experiment Crankmaster must be careful to conduct its experiment under controlled conditions; that is, it has to be sure that it is measuring *cranking power*; and that the other variables (such as temperature, age of engine, and condition of battery cables, to name only a few) are held as nearly constant as practicable. In an effort to accomplish just this, Crankmaster's statistical group uses new cars of the same make and model, conducts the tests at the same outside air temperature, and is careful to be quite precise in measuring the time variable. Crankmaster gathers experimental data on the performance of the 20 batteries in this manner.

Experimental conditions are kept constant

Analyzing the Data Data on the 20 individual battery tests are subjected to hypothesis testing in the same way that we shall see in Chapter 9, "Testing Hypotheses: Two-Sample Tests." Crankmaster is interested in whether there is a significant difference between the cranking power of its battery and that of its competitor. It turns out that the difference between the mean cranking life of Crankmaster's battery and that of its competitor *is* significant. Crankmaster incorporates the result of this experiment into its advertising.

Data are analyzed

Reacting to Experimental Claims

How should we, as consumers, react to Crankmaster's new battery-life claims in its latest advertising? Should we conclude from the tests it has run that the Crankmaster battery *is* superior to the competitive battery? If we stop for a moment to consider the nature of the experiment, we may not be so quick to come to such a conclusion.

How should the consumer react?

How do we know that the ages and conditions of the cars' engines in the experiment *were* identical? And are we absolutely sure that the battery's cables were identical in size and resistance to current? And what about air temperatures during the tests? Were they the same? These are the normal kinds of questions that we should ask.

Are we sure?

How should we react to the statement, if it is made, that "we subjected the experimental results to extensive statistical testing"? The answer to that will have to wait until Chapter 9, where we can determine whether such a difference in battery lives is too large to be attributed to chance. At this point, we, as consumers, need to be appropriately skeptical.

Other Options Open

Of course, Crankmaster would have had the same concerns we did, and in all likelihood would *not* have made significant advertising claims solely on the basis of the experimental design we have just described. One possible course of action to avoid criticism is to *ensure* that all variables except the one being measured have indeed been controlled. Despite the care taken to produce such controlled conditions, it turns out that these overcontrolled experiments do not really solve our problem. Normally, instead of investing resources in attempts to *eliminate* experimental variations, we choose a *completely different route*. The next few paragraphs show how we can accomplish this.

Another route for Crankmaster

Factorial Experiments

In the Crankmaster situation, we had two batteries (let's refer to them now as A and B) and three test conditions that were of some concern to us: temperature, age of the engine, and condition of the battery cable. Let's introduce the notion of *factorial experiments* by using this notation:

H = hot temperature	N = new engine	G = good cable
C = cold temperature	O = old engine	W = worn cable

Handling all test conditions at the same time

Of course, in most experiments, we could find more than two temperature conditions and, for that matter, more than two categories for engine condition and battery-cable condition. But it's better to introduce the idea of factorial experiments using a somewhat simplified example.

How many combinations?

Now, because there are two batteries, two temperature possibilities, two engine condition possibilities, and two battery-cable possibilities, there are $2 \times 2 \times 2 \times 2 = 16$ possible combinations of factors. If we wanted to write these sixteen possibilities down, they would look like Table 6-5.

TABLE 6-5 SIXTEEN POSSIBLE COMBINATIONS OF FACTORS FOR BATTERY TEST

Test	Battery	Temperature	Engine Condition	Cable Condition
1	A	H	N	G
2	A	H	N	W
3	A	H	O	G
4	A	H	O	W
5	A	C	N	G
6	A	C	N	W
7	A	C	O	G
8	A	C	O	W
9	B	H	N	G
10	B	H	N	W
11	B	H	O	G
12	B	H	O	W
13	B	C	N	G
14	B	C	N	W
15	B	C	O	G
16	B	C	O	W

Having set up all the possible combinations of factors involved in this experiment, we could now conduct the 16 tests in the table, if we did this, we would have conducted a complete factorial experiment, because each of the two *levels* of each of the four *factors* would have been used once with each possible combination of other levels of other factors. Designing the experiment this way would permit us to use techniques we shall introduce in Chapter 11, "Chi-Square and Analysis of Variance," to test the effect of each of the factors.

Levels and factors to be handled

We need to point out, before we leave this section, that in an actual experiment we would hardly conduct the tests in the order in which they appear in the table. They were arranged in that order to facilitate your counting the combinations and determining that all possible combinations were indeed represented. In actual practice, we would randomize the order of the tests, perhaps by putting 16 numbers in a hat and drawing out the order of the experiment in that simple manner.

Randomizing

Being More Efficient in Experimental Design

As you saw from our four-factor experiment, 16 tests were required to compare all levels with all factors. If we were to compare the same two batteries, but this time with five levels of temperature, four measures of engine condition, and three measures of battery-cable condition, it would take $2 \times 5 \times 4 \times 3 = 120$ tests for a complete factorial experiment.

A bit of efficiency

Fortunately, statisticians have been able to help us reduce the number of tests in cases like this. To illustrate how this works, look at a consumer-products company that wants to test market a new

toothpaste in four different cities with four different kinds of packages and with four different advertising programs. In such a case, a complete factorial experiment would take $4 \times 4 \times 4 = 64$ tests. However, if we do some clever planning, we can actually do it with far fewer tests—16, to be precise.

Let's use the notation:

A = City 1	I = Package 1	1 = Ad program 1
B = City 2	II = Package 2	2 = Ad program 2
C = City 3	III = Package 3	3 = Ad program 3
D = City 4	IV = Package 4	4 = Ad program 4

		Advertising program			
		1	2	3	4
Package	I	C	B	D	A
	II	B	C	A	D
	III	D	A	B	C
	IV	A	D	C	B

FIGURE 6-1 A LATIN SQUARE

Now we arrange the cities, packages, and advertising programs in a design called a *Latin square* (Figure 6-1).

In the experimental design represented by the Latin square, we would need only 16 tests instead of 64 as originally calculated. Each combination of city, package, and advertising program would be represented in the 16 tests. The actual statistical analysis of the data obtained from such a Latin square experimental design would require a form of analysis of variance a bit beyond the scope of this book.

6.5 INTRODUCTION TO SAMPLING DISTRIBUTIONS

In Chapter 3, we introduced methods by which we can use sample data to calculate statistics such as the mean and the standard deviation. So far in this chapter, we have examined how samples can be taken from populations. If we apply what we have learned and take several samples from a population, the statistics we would compute for each sample need not be the same and most probably would vary from sample to sample.

Statistics differ among samples from the same population

Suppose our samples each consist of ten 25-year-old women from a city with a population of 100,000 (an infinite population, according to our usage). By computing the mean height and standard deviation of that height for each of these samples, we would quickly see that the mean of each sample and the standard deviation of each sample would be different. **A probability distribution of all the possible means of the samples is a distribution of the sample means.** Statisticians call this a *sampling distribution of the mean*.

Sampling distribution defined

We could also have a sampling distribution of a proportion. Assume that we have determined the proportion of beetle-infested pine trees in samples of 100 trees taken from a very large forest. We have taken a large number of those 100-item samples. If we plot a probability distribution of the possible proportions of infested trees in all these samples, we would see a distribution of the sample proportions. In statistics, this is called a *sampling distribution of the proportion*. (Notice that the term *proportion* refers to the proportion that is infested.)

Describing Sampling Distributions

Any probability distribution (and, therefore, any sampling distribution) can be partially described by its mean and standard deviation. Table 6-6 illustrates several populations. Beside each, we have indicated

TABLE 6-6 EXAMPLES OF POPULATIONS, SAMPLES, SAMPLE STATISTICS, AND SAMPLING DISTRIBUTIONS

Population	Sample	Sample Statistic	Sampling Distribution
Water in a river	10-gallon containers of water	Mean number of parts of mercury per million parts of water	Sampling distribution of the mean
All professional basketball teams	Groups of 5 players	Median height	Sampling distribution of the median
All parts produced by a manufacturing process	50 parts	Proportion defective	Sampling distribution of the proportion

the sample taken from that population, the sample statistic we have measured, and the sampling distribution that would be associated with that statistic.

Now, how would we describe each of the sampling distributions in Table 6-6? In the first example, the sampling distribution of the mean can be partially described by its mean and standard deviation. The sampling distribution of the median in the second example can be partially described by the mean and standard deviation of the distribution of the medians. And in the third, the sampling distribution of the proportion can be partially described by the mean and standard deviation of the distribution of the proportions.

Concept of Standard Error

Rather than say “standard deviation of the distribution of sample means” to describe a distribution of sample means, statisticians refer to the *standard error of the mean*. Similarly, the “standard deviation of the distribution of sample proportions” is shortened to the *standard error of the proportion*. The term *standard error* is used because it conveys a specific meaning. An example will help explain the reason for the name. Suppose we wish to learn something about the height of freshmen at a large state university. We could take a series of samples and calculate the mean height for each sample. It is highly unlikely that all of these sample means would be the same; we expect to see some variability in our observed means. This variability in the sample statistics results from *sampling error* due to chance; that is, there are differences between each sample and the population, and among the several samples, owing solely to the elements we happened to choose for the samples.

**Explanation of the term
standard error**

The standard deviation of the distribution of sample means measures the extent to which we expect the means from the different samples to vary because of this chance error in the sampling process. Thus, **the standard deviation of the distribution of a sample statistic is known as the standard error of the statistic**.

The standard error indicates not only the size of the chance error that has been made, but also the accuracy we are likely to get if we use a sample statistic to estimate a population parameter. A distribution of sample means that is less spread out (that has a small standard error) is a better estimator of the population mean than a distribution of sample means that is widely dispersed and has a larger standard error.

Use of the standard error

Table 6-7 indicates the proper use of the term *standard error*. In Chapter 7, we shall discuss how to estimate population parameters using sample statistics.

TABLE 6-7 CONVENTIONAL TERMINOLOGY USED TO REFER TO SAMPLE STATISTICS

When We Wish to Refer to the	We Use the Conventional Term
Standard deviation of the distribution of sample means	Standard error of the mean
Standard deviation of the distribution of sample proportions	Standard error of the proportion
Standard deviation of the distribution of sample medians	Standard error of the median
Standard deviation of the distribution of sample ranges	Standard error of the range

One Use of the Standard Error

A school that trains private pilots for their instrument examination advertised that “our graduates score higher on the instrument written examination than graduates of other schools.” To the unsuspecting reader, this seems perfectly clear. If you want to score higher on your instrument written examination, then this school is your best bet.

In fact, however, whenever we are using tests, we have to deal with standard error. Specifically, we need some measure of the precision of the test instrument, usually represented by standard error. This would tell us how large a difference in one school’s grades would have to be for it to be statistically significant. Unfortunately, the advertisement did not offer data; it merely asserted that “our graduates do better.”

HINTS & ASSUMPTIONS

Understanding sampling distributions allows statisticians to take samples that are both meaningful and cost-effective. Because large samples are very expensive to gather, decision makers should always aim for the smallest sample that gives reliable results. In describing distributions, statisticians have their own shorthand, and when they use the term *standard error* to describe a distribution, they are referring to the distribution’s standard deviation. Instead of saying “the standard deviation of the distribution of sample means” they say “the standard error of the mean.” Hint: The standard error indicates how spread-out (dispersed) the means of the samples are. Warning: Although the *standard error of the mean* and the *population standard deviation* are related to each other, as we shall soon see, it is important to remember that they are not the same thing.

EXERCISES 6.3

Self-Check Exercises

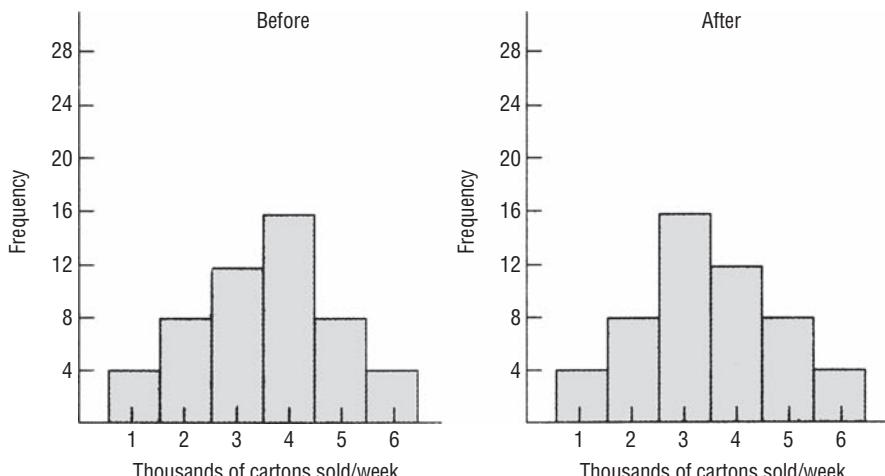
- SC 6-3** A machine that fills bottles is known to have a mean filling amount of 125 grams and a standard deviation of 20 grams. A quality control manager took a random sample of filled bottles and found the sample mean to be 130. The quality control manager assumed the sample must not be representative. Is the conclusion correct?
- SC 6-4** The president of the American Dental Association wants to determine the average number of times that each dentist’s patients floss per day. Toward this end, he asks each of 100 randomly selected dentists to poll 50 of their patients at random and submit the mean number of flossings per day to the ADA. These numbers are computed and submitted to the president. Has he been given a sample from the population of patients or from some other distribution?

Basic Concepts

- 6-19** Suppose you are sampling from a population with a mean of 2.15. What sample size will guarantee that
- The sample mean is 2.15?
 - The standard error of the mean is zero?
- 6-20** The term *error*, in standard error of the mean, refers to what type of error?

Applications

- 6-21** You recently purchased a box of raisin bran and measured the number of raisins. The company claims that the number of raisins per box is 2.0 cups on average, with a standard deviation of 0.2 cup. Your box contained only 1.9 cups. Could the company's claim be correct?
- 6-22** North Carolina Electric and Gas has determined that the cost per 100 sq ft. for the residential population electrical service is \$0,314 on average, with a standard deviation of \$0.07. Two different samples are selected at random, and the means are \$0.30 and \$0.35, respectively. The assistant in charge of data collection concludes that the second sample is the better one because it is better to overestimate than underestimate the true mean. Comment. Is one of the means "better" in some ways, given the true population mean?
- 6-23** A woman working for Nielsen ratings service interviews passersby on a New York street and records each subject's estimate of average time spent viewing prime-time television per night. These interviews continue for 20 days, and at the end of each day, the interviewer computes the mean time spent viewing among all those interviewed during the day. At the conclusion of all interviews, she constructs a frequency distribution for these daily means. Is this a sampling distribution of the mean? Explain.
- 6-24** Charlotte Anne Serrus, a marketing analyst for the Florris Tobacco Company, wants to assess the damage done to FTC's sales by the appearance of a new competitor. Accordingly, she has compiled weekly sales figures from one-year periods before and after the competitor's appearance. Charlotte has graphed the corresponding frequency distributions as follows:



Based on these graphs, what has been the effect of the competitor's appearance on average weekly sales?

- 6-25** In times of declining SAT scores and problems of functional illiteracy, the admissions committee of a prestigious university is concerned with keeping high standards of admission. Each year, after decisions on acceptance are made, the committee publishes and distributes statistics on students admitted, giving, for example, the average SAT score. On the report containing the statistics are the words "Standard Error of the Mean." The secretary who types the report knows that for several years, the average SAT score was about 1,200 and has assumed that the standard error of the mean was how much the committee allowed an admitted student's score to deviate from the mean. Is the assumption correct? Explain.
- 6-26** A mail-order distribution firm is interested in the level of customer satisfaction. The CEO has randomly selected 50 regional managers to survey customers. Each manager randomly selects 5 supervisors to randomly survey 30 customers. The surveys are conducted and results are computed and sent to the CEO. What type of distribution did the sample come from?

Worked-Out Answers to Self-Check Exercises

- SC 6-3** No. The mean of a sample usually does not exactly equal the population mean because of sampling error.
- SC 6-4** The information gathered concerns mean flossings per day for groups of 50 patients, not for single patients, so it is a sample from the sampling distribution of the mean of samples of size 50 drawn from the patient population. It is not a sample from the patient population.

6.6 SAMPLING DISTRIBUTIONS IN MORE DETAIL

In Section 6.4, we introduced the idea of a sampling distribution. We examined the reasons why sampling from a population and developing a distribution of these sample statistics would produce a sampling distribution, and we introduced the concept of standard error. Now we will study these concepts further, so that we will not only be able to understand them conceptually, but also be able to handle them computationally.

Conceptual Basis for Sampling Distributions

Figure 6-2 will help us examine sampling distributions without delving too deeply into statistical theory. We have divided this illustration into three parts. Figure 6-2(a) illustrates a *population distribution*. Assume that this population is all the filter screens in a large industrial pollution-control system and that this distribution is the operating hours before a screen becomes clogged. The distribution of operating hours has a mean μ (*mu*) and a standard deviation σ (*sigma*).

Deriving the sampling distribution of the mean

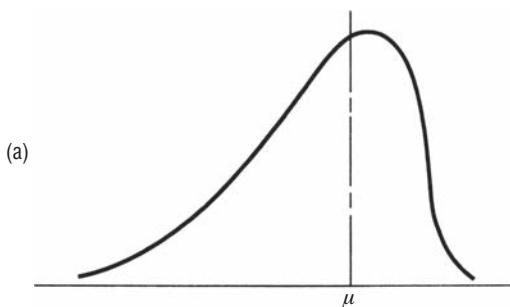
Suppose that somehow we are able to take all the possible samples of 10 screens from the population distribution (actually, there would be far too many for us to consider). Next we would calculate the mean and the standard deviation for each one of these *samples*, as represented in Figure 6-2(b). As a result, each sample would have its own mean, \bar{x} (*x bar*), and its own standard deviation, s . All the individual sample means would *not* be the same as the population mean. They would tend to be near the population mean, but only rarely would they be exactly that value.

As a last step, we would produce a distribution of all the means from every sample that could be taken. This distribution, called the *sampling distribution of the mean*, is illustrated in Figure 6-2(c). This distribution of the sample means (the sampling distribution) would have its own mean, $\mu_{\bar{x}}$ (*mu sub x bar*), and its own standard deviation, or standard error, $\sigma_{\bar{x}}$ (*sigma sub x bar*).

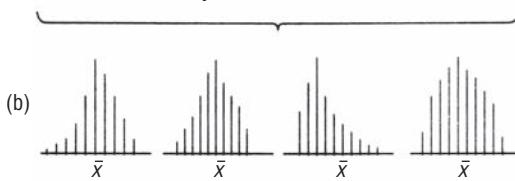
In statistical terminology, the sampling distribution we would obtain by taking all the samples of a given size is a *theoretical sampling distribution*. Figure 6-2(c) describes such an example.

The sampling distribution of the mean

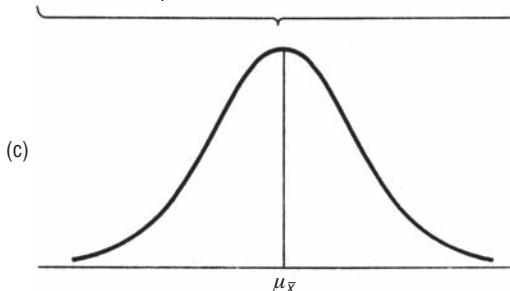
Function of theoretical sampling distributions



If somehow we were able to take *all* the possible samples of a given size from this population *distribution*, they would be represented graphically by these four samples below. Although we have shown only four such samples, there would actually be an enormous number of them.



Now, if we were able to take the means from all the *sample distributions* and produce a distribution of these sample means, it would look like this:



The population distribution:

This distribution is the distribution of the operating hours of *all* the filter screens. It has:

μ = the mean of this distribution

σ = the standard deviation of this distribution

The sample frequency distribution:

These only *represent* the enormous number of sample distributions possible. *Each* sample distribution is a discrete distribution and has:

← \bar{x} = its own mean, called “ x bar”

s = its own standard deviation

The sampling distribution of the mean:

This distribution is the distribution of all the sample means and has:

← $\mu_{\bar{x}}$ = mean of the sampling distribution of the means, called “ μ sub \bar{x} bar”

← $\sigma_{\bar{x}}$ = standard error of the mean (standard deviation of the sampling distribution of the mean), called “ σ sub \bar{x} bar”

FIGURE 6-2 CONCEPTUAL POPULATION DISTRIBUTION, SAMPLE DISTRIBUTIONS, AND SAMPLING DISTRIBUTION

In practice, the size and character of most populations prohibit decision makers from taking all the possible samples from a population distribution. Fortunately, statisticians have developed formulas for estimating the characteristics of these theoretical sampling distributions, making it unnecessary for us to collect large numbers of samples. In most cases, decision makers take only one sample from the population, calculate statistics for that sample, and from those statistics infer something about the parameters of the entire population. We shall illustrate this shortly.

In each example of sampling distributions in the remainder of this chapter, we shall use the sampling distribution of the mean. We could study the sampling distributions of the median, range, or proportion, but we will stay with the mean for the continuity it will add to the explanation. Once you develop an understanding of how to deal computationally with the sampling distribution of the mean, you will be able to apply it to the distribution of any other sample statistic.

Why we use the sampling distribution of the mean

Sampling from Normal Populations

Suppose we draw samples from a normally distributed population with a mean of 100 and standard deviation of 25, and that we start by drawing samples of five items each and by calculating their means. The first mean might be 95, the second 106, the third 101, and so on. Obviously, there is just as much chance for the sample mean to be above the population mean of 100 as there is for it to be below 100. Because we are *averaging* five items to get each sample mean, very large values in the sample would be averaged down and very small values up. We would reason that we would get less spread among the sample means than we would among the individual items in the original population. That is the same as saying that the standard error of the mean, or standard deviation of the sampling distribution of the mean, would be less than the standard deviation of the *individual* items in the population. Figure 6-3 illustrates this point graphically.

Sampling distribution of the mean from normally distributed populations

Now suppose we increase our sample size from 5 to 20. This would not change the standard deviation of the items in the original population. But with samples of 20, we have increased the effect of averaging in each sample and would expect even *less* dispersion among the sample means. Figure 6-4 illustrates this point.

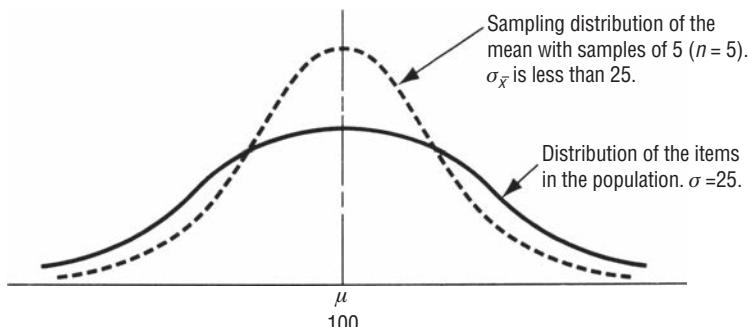


FIGURE 6-3 RELATIONSHIP BETWEEN THE POPULATION DISTRIBUTION AND THE SAMPLING DISTRIBUTION OF THE MEAN FOR A NORMAL POPULATION

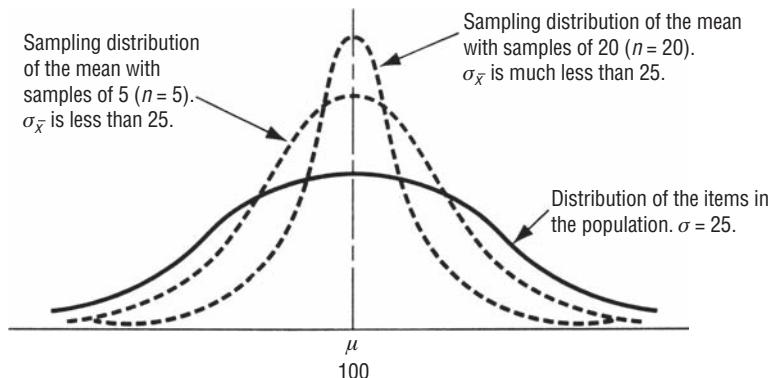


FIGURE 6-4 RELATIONSHIP BETWEEN THE POPULATION DISTRIBUTION AND SAMPLING DISTRIBUTION OF THE MEAN WITH INCREASING n 's

The sampling distribution of a mean of a sample taken from a normally distributed population demonstrates the important properties summarized in Table 6-8. An example will further illustrate these properties. A bank calculates that its individual savings accounts are normally distributed with a mean of \$2,000 and a standard deviation of \$600. If the bank takes a random sample of 100 accounts, what is the probability that the sample mean will lie between \$1,900 and \$2,050? This is a question about the sampling distribution of the mean; therefore, we must first calculate the standard error of the mean. In this case, we shall use the equation for the standard error of the mean designed for situations in which the population is infinite (later, we shall introduce an equation for finite populations):

Properties of the sampling distribution of the mean

Standard Error of the Mean for Infinite Populations

$$\text{Standard error of the mean} \longrightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [6-1]$$

where

- σ = population standard deviation
- n = sample size

TABLE 6-8 PROPERTIES OF THE SAMPLING DISTRIBUTION OF THE MEAN WHEN THE POPULATION IS NORMALLY DISTRIBUTED

Property	Illustrated Symbolically
The sampling distribution has a mean equal to the population mean	$\mu_{\bar{x}} = \mu$
The sampling distribution has a standard deviation (a standard error) equal to the population standard deviation divided by the square root of the sample size	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Applying this to our example, we get

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\$600}{\sqrt{100}} \\ &= \frac{\$600}{10} \\ &= \$60 \leftarrow \text{Standard error of the mean}\end{aligned}$$

Finding the standard error of the mean for infinite populations

Next, we need to use the table of z values (Appendix Table 1) and Equation 5-6, which enables us to use the Standard Normal Probability Distribution Table. With these, we can determine the probability that the sample mean will lie between \$1,900 and \$2,050.

$$z = \frac{x - \mu}{\sigma} \quad [5-6]$$

Equation 5-6 tells us that to convert any normal random variable to a standard normal random variable, we must subtract the mean of the variable being standardized and divide by the standard error (the standard deviation of that variable). Thus, in this particular case, Equation 5-6 becomes

Standardizing the Sample Mean

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad [6-2]$$

Sample mean Population mean
 Standard error of the mean = $\frac{\sigma}{\sqrt{n}}$

Now we are ready to compute the two z values as follows:

For $\bar{x} = \$1,900$:

Converting the sample mean to a z value

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \\ &= \frac{\$1,900 - \$2,000}{\$60} \\ &= -\frac{100}{60} \\ &= -1.67 \leftarrow \text{Standard deviations from the mean of a standard normal probability distribution}\end{aligned} \quad [6-2]$$

For $\bar{x} = \$2,050$

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \\ &= \frac{\$2,050 - \$2,000}{\$60} \\ &= \frac{50}{60} \\ &= 0.83 \leftarrow \text{Standard deviation from the mean of a standard normal probability distribution}\end{aligned} \quad [6-2]$$

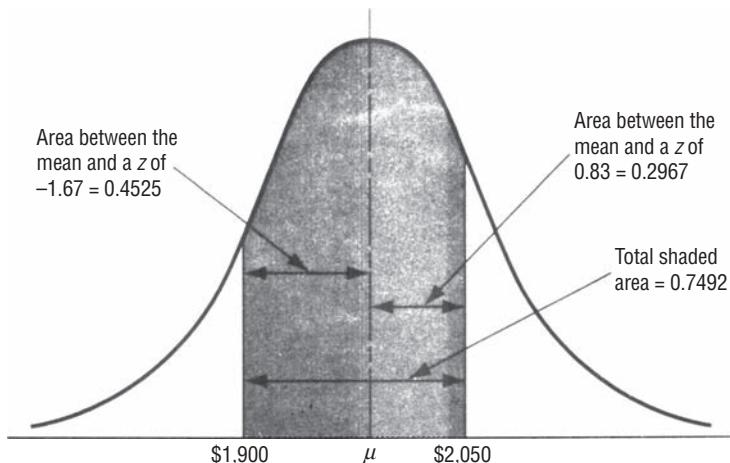


FIGURE 6-5 PROBABILITY OF SAMPLE MEAN LYING BETWEEN \$1,900 AND \$2,050

Appendix Table 1 gives us an area of 0.4525 corresponding to a z value of -1.67 , and it gives an area of .2967 for a z value of $.83$. If we add these two together, we get 0.7492 as the total probability that the sample mean will lie between \$1,900 and \$2,050. We have shown this problem graphically in Figure 6-5.

Sampling from Nonnormal Populations

In the preceding section, we concluded that when the population is normally distributed, the sampling distribution of the mean is also normal. Yet decision makers must deal with many populations that are not normally distributed. How does the sampling distribution of the mean react when the population from which the samples are drawn is not normal? An illustration will help us answer this question.

Consider the data in Table 6-9, concerning five motorcycle owners and the lives of their tires. Because only five people are involved, the population is too small to be approximated by a normal distribution. We'll take all of the possible samples of the owners in groups of three, compute the sample means (\bar{x}), list them, and compute the mean of the sampling distribution ($\mu_{\bar{x}}$). We have done this in Table 6-10. These calculations show that even in a case in which the population is not normally distributed, $\mu_{\bar{x}}$, the mean of the sampling distribution, is *still* equal to the population mean, μ .

The mean of the sampling distribution of the mean equals the population mean

Now look at Figure 6-6. Figure 6-6(a) is the population distribution of tire lives for the five motorcycle owners, a distribution that is anything, but normal in shape. In Figure 6-6(b), we show the sampling distribution of the mean for a sample size of three, taking the information from Table 6-10. Notice the difference between the probability distributions in Figures 6-6(a) and 6-6(b). In Figure 6-6(b), the distribution looks a little more like the bell shape of the normal distribution.

TABLE 6-9 EXPERIENCE OF FIVE MOTORCYCLE OWNERS WITH LIFE OF TIRES

OWNER	Carl	Debbie	Elizabeth	Frank	George	
TIRE LIFE (MONTHS)	3	3	7	9	14	Total: 36 months
	$\text{Mean} = \frac{36}{5} = 7.2 \text{ months}$					

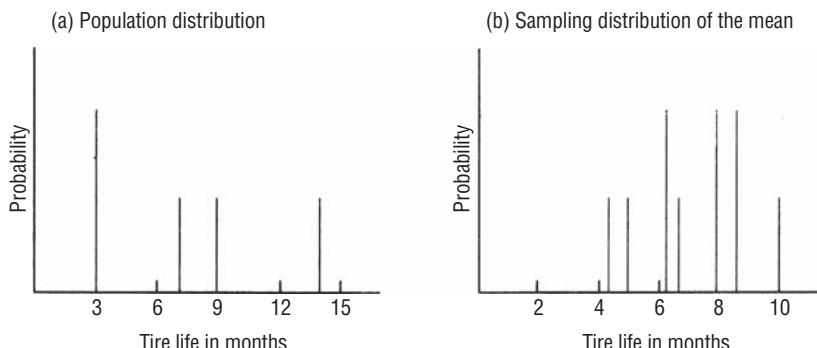
TABLES 6-10 CALCULATION OF SAMPLE MEAN TIRE LIFE WITH $n = 3$

Samples of Three	Sample Data (Tire Lives)	Sample Mean
EFG*	7 + 9 + 14	10
DFG	3 + 9 + 14	$8\frac{2}{3}$
DEG	3 + 7 + 14	8
DEF	3 + 7 + 9	$6\frac{1}{3}$
CFG	3 + 9 + 14	$8\frac{2}{3}$
CEG	3 + 7 + 14	8
CEF	3 + 7 + 9	$6\frac{1}{3}$
CDF	3 + 3 + 9	5
CDE	3 + 3 + 7	$4\frac{1}{3}$
CDG	3 + 3 + 14	$6\frac{2}{3}$
		72 months
		$\mu_{\bar{x}} = \frac{72}{10}$
		$= 7.2 \text{ months}$

*Names abbreviated by first initial

If we had a long time and much space, we could repeat this example and enlarge the population size to 40. Then we could take samples of *different* sizes. Next we would plot the sampling distributions of the mean that would occur for the *different* sizes. Doing this would show quite dramatically how quickly the sampling distribution of the mean approaches normality, *regardless* of the shape of the population distribution. Figure 6-7 simulates this process graphically without all the calculations.

Increase in the size of samples leads to a more normal sampling distribution

**FIGURE 6-6** POPULATION DISTRIBUTION AND SAMPLING DISTRIBUTION OF THE MEAN TIRE LIFE

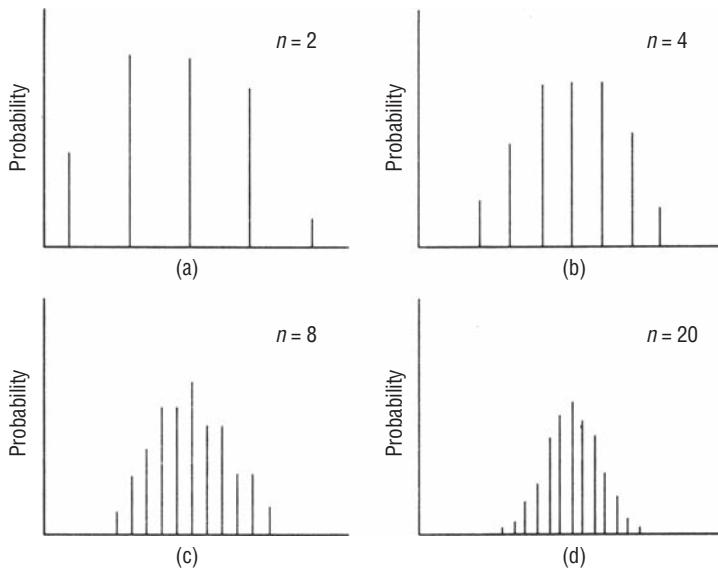


FIGURE 6-7 SIMULATED EFFECT OF INCREASES IN THE SAMPLE SIZE ON THE APPEARANCE OF THE SAMPLING DISTRIBUTION

The Central Limit Theorem

The example in Table 6-10 and the four probability distributions in Figure 6-7 should suggest several things to you. First, **the mean of the sampling distribution of the mean will equal the population mean** regardless of the sample size, even if the population is not normal. Second, as the sample size increases, **the sampling distribution of the mean will approach normality**, regardless of the shape of the population distribution.

This relationship between the shape of the population distribution and the shape of the sampling distribution of the mean is called the *central limit theorem*. The central limit theorem is perhaps the most important theorem in all of statistical inference. It **assures us that the sampling distribution of the mean approaches normal as the sample size increases**. There are theoretical situations in which the central limit theorem fails to hold, but they are almost never encountered in practical decision making. Actually, a sample does not have to be very large for the sampling distribution of the mean to approach normal. Statisticians use the normal distribution as an approximation to the sampling distribution whenever the sample size is at least 30, but the sampling distribution of the mean can be nearly normal with samples of even half that size. **The significance of the central limit theorem is that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.** Putting this ability to work is the subject of much of the material in the subsequent chapters of this book.

Let's illustrate the use of the central limit theorem. The distribution of annual earnings of all bank tellers with five years'

Results of increasing sample size

Significance of the central limit theorem

Using the central limit theorem

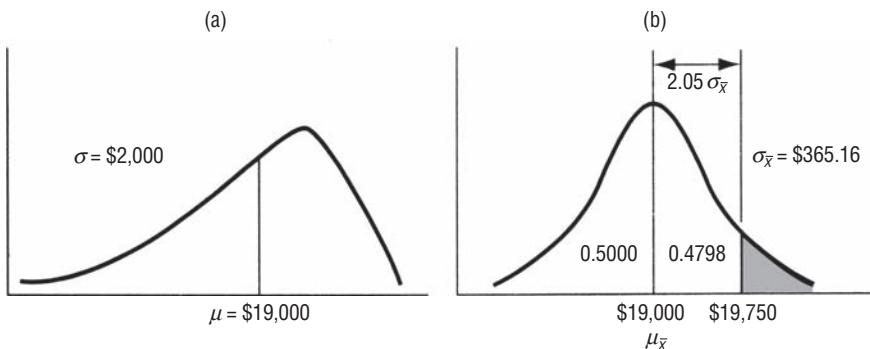


FIGURE 6-8 POPULATION DISTRIBUTION AND SAMPLING DISTRIBUTION FOR BANK TELLERS' EARNINGS

experience is skewed negatively, as shown in Figure 6-8(a). This distribution has a mean of \$19,000 and a standard deviation of \$2,000. If we draw a random sample of 30 tellers, what is the probability that their earnings will average more than \$19,750 annually? In Figure 6-8(b), we show the sampling distribution of the mean that would result, and we have colored the area representing “earnings over \$19,750.”

Our first task is to calculate the standard error of the mean from the population standard deviation, as follows

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} & [6-1] \\ &= \frac{\$2,000}{\sqrt{30}} \\ &= \frac{\$2,000}{5.477} \\ &= \$365.16 \leftarrow \text{Standard error of the mean}\end{aligned}$$

Because we are dealing with a sampling distribution, we must now use Equation 6-2 and the Standard Normal Probability Distribution (Appendix Table 1).

For $x = \$19,750$:

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} & [6-2] \\ &= \frac{\$19,750 - \$19,000}{\$365.16} \\ &= \frac{\$750.00}{\$365.16} \\ &= 2.05 \leftarrow \text{Standard deviations from the mean of a standard normal probability distribution}\end{aligned}$$

This gives us an area of 0.4798 for a z value of 2.05. We show this area in Figure 6-8 as the area between the mean and \$19,750. Since half, or 0.5000, of the area under the curve lies between the mean

and the right-hand tail, the colored area must be

$$\begin{array}{r} 0.5000 \text{ (Area between the mean and the right-hand tail)} \\ -0.4798 \text{ (Area between the mean \$19,750)} \\ \hline 0.0202 \leftarrow \text{(Area between the right-hand tail and \$19,750)} \end{array}$$

Thus, we have determined that there is slightly more than a 2 percent chance of average earnings being more than \$19,750 annually in a group of 30 tellers.

Sampling Distribution of Proportion

In many situations, the issue of interest is categorical in nature, which can be classified as occurrence or non-occurrence. In these situations, researcher is interested in estimating proportion of occurrence. Since, information from complete population is not available, sample proportion is used to estimate the ‘true’ proportion.

$$\text{Sample Proportion } \hat{p} = \frac{x}{n}$$

where, x is number of occurrences out of a total of the sample size of ‘ n ’

‘ x ’ will follow binomial distribution with probability of occurrence as p .

According to Binomial Distribution:

$$\text{Mean of } x: \mu_x = np$$

$$\text{Standard deviation of } x: \sigma_x = \sqrt{npq}$$

$$\text{where } q = 1 - p$$

If we consider sample proportion $\hat{p} = \frac{x}{n}$, then sampling distribution sample statistic (sample proportion) \hat{p} will have

$$\text{Mean of } \hat{p} = \frac{x}{n}: \mu_{\hat{p}} = \frac{np}{n} = p$$

$$\text{Standard error of } \hat{p}: \sigma_{\hat{p}} = \frac{\sqrt{npq}}{n}$$

$$= \sqrt{\frac{pq}{n}}$$

If sample size ‘ n ’ is large, considering normal distribution as an approximation of the Binomial Distribution.

So, sampling distribution of $\hat{p} = \frac{x}{n}$, will have normal distribution with

$$\text{Mean: } \mu_{\hat{p}} = p$$

$$\text{Standard error: } \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

$$\text{hence } Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

In case of finite population (N)

Sampling distribution of $\hat{p} = \frac{x}{n}$ will be normal distribution with

$$\text{Mean } \mu_{\hat{p}} = p$$

$$\text{Standard error } \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

HINTS & ASSUMPTIONS

The central limit theorem is one of the most powerful concepts in statistics. What it really says is that the distribution of sample means tends to be a normal distribution. This is true *regardless* of the shape of the population distribution from which the samples were taken. Hint: Go back and look at Figures 6-6 and 6-7 on pages 318–319. Watch again how fast the distribution of sample means taken from the clearly nonnormal population in Figure 6-6 begins to look like a normal distribution in Figure 6-7 once we start to increase the sample size. And it really doesn't make any difference what the distribution of the population looks like; this will *always* happen. We could prove this to you, but first you'd have to go back and take several advanced mathematics courses to understand the proof.

EXERCISES 6.4**Self-Check Exercises**

- SC 6-5** In a sample of 25 observations from a normal distribution with mean 98.6 and standard deviation 17.2
- What is $P(92 < \bar{x} < 102)$?
 - Find the corresponding probability given a sample of 36.
- SC 6-6** Mary Bartel, an auditor for a large credit card company, knows that, on average, the monthly balance of any given customer is \$112, and the standard deviation is \$56. If Mary audits 50 randomly selected accounts, what is the probability that the sample average monthly balance is
- Below \$100?
 - Between \$100 and \$130?

Basic Concepts

- 6-27** In a sample of 16 observations from a normal distribution with a mean of 150 and a variance of 256, what is
- $P(\bar{x} < 160)$?
 - $P(\bar{x} > 142)$?
- If, instead of 16 observations, 9 observations are taken, find
- $P(\bar{x} < 160)$.
 - $P(\bar{x} > 142)$.
- 6-28** In a sample of 19 observations from a normal distribution with mean 18 and standard deviation 4.8
- What is $P(16 < \bar{x} < 20)$?
 - What is $P(16 \leq \bar{x} \leq 20)$?
 - Suppose the sample size is 48. What is the new probability in part (a)?
- 6-29** In a normal distribution with mean 56 and standard deviation 21, how large a sample must be taken so that there will be at least a 90 percent chance that its mean is greater than 52?
- 6-30** In a normal distribution with mean 375 and standard deviation 48, how large a sample must be taken so that the probability will be at least 0.95 that the sample mean falls between 370 and 380?

Applications

- 6-31** An astronomer at the Mount Palomar Observatory notes that during the Geminid meteor shower, an average of 50 meteors appears each hour, with a variance of 9 meteors squared. The Geminid meteor shower will occur next week.
- If the astronomer watches the shower for 4 hours, what is the probability that at least 48 meteors per hour will appear?
 - If the astronomer watches for an additional hour, will this probability rise or fall? Why?
- 6-32** The average cost of a studio condominium in the Cedar Lakes development is \$62,000 and the standard deviation is \$4,200.
- What is the probability that a condominium in this development will cost at least \$65,000?
 - Is the probability that the average cost of a sample of two condominiums will be at least \$65,000 greater or less than the probability of one condominium's costing that much? By how much?
- 6-33** Robertson Employment Service customarily gives standard intelligence and aptitude tests to all people who seek employment through the firm. The firm has collected data for several years and has found that the distribution of scores is not normal, but is skewed to the left with a mean of 86 and a standard deviation of 16. What is the probability that in a sample of 75 applicants who take the test, the mean score will be less than 84 or greater than 90?
- 6-34** An oil refinery has backup monitors to keep track of the refinery flows continuously and to prevent machine malfunctions from disrupting the process. One particular monitor has an average life of 4,300 hours and a standard deviation of 730 hours. In addition to the primary monitor, the refinery has set up two standby units, which are duplicates of the primary one. In the case of malfunction of one of the monitors, another will automatically take over in its place. The operating life of each monitor is independent of the others.
- What is the probability that a given set of monitors will last at least 13,000 hours?
 - At most 12,630 hours?
- 6-35** A recent study by the EPA has determined that the amount of contaminants in Minnesota lakes (in parts per million) is normally distributed with mean 64 ppm and variance 17.6. Suppose 35 lakes are randomly selected and sampled. What is the probability that the sample average amount of contaminants is
- Above 72 ppm?
 - Between 64 and 72 ppm?
 - Exactly 64 ppm?
 - Above 94 ppm?
 - If, in our sample, we found $\bar{x} = 100$ ppm, would you feel confident in the study conducted by the EPA? Explain briefly.
- 6-36** Calvin Ensor, president of General Telephone Corp., is upset at the number of telephones produced by GTC that have faulty receivers. On average, 110 telephones per day are being returned because of this problem, and the standard deviation is 64. Mr. Ensor has decided that unless he can be at least 80 percent certain that, on average, no more than 120 phones per day will be returned during the next 48 days, he will order the process overhauled. Will the overhaul be ordered?
- 6-37** Clara Voyant, whose job is predicting the future for her venture capital company, has just received the statistics describing her company's performance on 1,800 investments last year.

Clara knows that, in general, investments generate profits that have a normal distribution with mean \$7,500 and standard deviation \$3,300. Even before she looked at the specific results from each of the 1,800 investments from last year, Clara was able to make some accurate predictions by using her knowledge of sampling distributions. Follow her analysis by finding the probability that the sample mean of last year's investments

- (a) Exceeded \$7,700.
- (b) Was less than \$7,400.
- (c) Was greater than \$7,275, but less than \$7,650.

6-38 Farmer Braun, who sells grain to Germany, owns 60 acres of wheat fields. Based on past experience, he knows that the yield from each individual acre is normally distributed with mean 120 bushels and standard deviation 12 bushels. Help Farmer Braun plan for his next year's crop by finding

- (a) The expected mean of the yields from Farmer Braun's 60 acres of wheat.
- (b) The standard deviation of the sample mean of the yields from Farmer Braun's 60 acres.
- (c) The probability that the mean yield per acre will exceed 123.8 bushels.
- (d) The probability that the mean yield per acre will fall between 117 and 122 bushels.

6-39 A ferry carries 25 passengers. The weight of each passenger has a normal distribution with mean 168 pounds and variance 361 pounds squared. Safety regulations state that for this particular ferry, the total weight of passengers on the boat should not exceed 4,250 pounds more than 5 percent of the time. As a service to the ferry owners, find

- (a) The probability that the total weight of passengers on the ferry will exceed 4,250 pounds.
- (b) The 95th percentile of the distribution of the total weight of passengers on the ferry.

Is the ferry complying with safety regulations?

Worked-Out Answers to Self-Check Exercises

SC 6-5 (a) $n = 25 \quad \mu = 98.6 \quad \sigma = 17.2 \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 17.2/\sqrt{25} = 3.44$

$$\begin{aligned} P(92 < \bar{x} < 102) &= P\left(\frac{92 - 98.6}{3.44} < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{102 - 98.6}{3.44}\right) \\ &= P(-1.92 < z < 0.99) = 0.4726 + 0.3389 = 0.8115 \end{aligned}$$

(b) $n = 36 \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 17.2/\sqrt{36} = 2.87$

$$\begin{aligned} P(92 < \bar{x} < 102) &= P\left(\frac{92 - 98.6}{2.87} < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{102 - 98.6}{2.87}\right) \\ &= P(-2.30 < z < 1.18) = 0.4893 + 0.3810 = 0.8703 \end{aligned}$$

SC 6-6 The sample size of 50 is large enough to use the central limit theorem.

$$\mu = 112 \quad \sigma = 56 \quad n = 50 \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 56/\sqrt{50} = 7.920$$

$$(a) P(\bar{x} < 100) = P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{100 - 112}{7.920}\right) = P(z < -1.52) = 0.5 - 0.4357 = 0.0643$$

$$\begin{aligned}
 \text{(b)} \quad P(100 < \bar{x} < 130) &= P\left(\frac{100 - 112}{7.920} < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{130 - 112}{7.920}\right) \\
 &= P(-1.52 < z < 2.27) = 0.4357 + 0.4884 + 0.9241
 \end{aligned}$$

6.7 AN OPERATIONAL CONSIDERATION IN SAMPLING: THE RELATIONSHIP BETWEEN SAMPLE SIZE AND STANDARD ERROR

We saw earlier in this chapter that the standard error, $\sigma_{\bar{x}}$ is a measure of dispersion of the sample means around the population mean. If the dispersion decreases (if $\sigma_{\bar{x}}$ becomes smaller), then

Precision of the sample mean

the values taken by the sample mean tend to cluster *more* closely around μ . Conversely, if the dispersion increases (if $\sigma_{\bar{x}}$ becomes larger), the values taken by the sample mean tend to cluster *less* closely around μ . We can think of this relationship this way: **As the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean.** Statisticians describe this phenomenon in another way: As the standard error decreases, the *precision* with which the sample mean can be used to estimate the population mean increases.

If we refer to Equation 6-1, we can see that as n increases, $\sigma_{\bar{x}}$ decreases. This happens because in Equation 6-1 a larger denominator on the right side would produce smaller $\sigma_{\bar{x}}$ on the left side. Two examples will show this relationship; both assume the same population standard deviation σ of 100.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [6-1]$$

When $n = 10$:

$$\begin{aligned}
 \sigma_{\bar{x}} &= \frac{100}{\sqrt{10}} \\
 &= \frac{100}{3.162} \\
 &= 31.63 \leftarrow \text{Standard error of the mean}
 \end{aligned}$$

And when $n = 100$:

$$\begin{aligned}
 \sigma_{\bar{x}} &= \frac{100}{\sqrt{100}} \\
 &= \frac{100}{10} \\
 &= 10 \leftarrow \text{Standard error of the mean}
 \end{aligned}$$

What have we shown? As we increased our sample size from 10 to 100 (a tenfold increase), the standard error dropped from 31.63 to 10, which is only about one-third of its former value. **Our examples show that, because $\sigma_{\bar{x}}$ varies inversely with the square root of n , there is diminishing return in sampling.**

Increasing the sample size:

Diminishing returns

It is true that sampling more items will decrease the standard error, but this benefit may not be worth the cost. A statistician would say, “The increased precision is not worth the additional sampling cost.” In a statistical sense, it seldom pays to take excessively large samples. Managers should always assess *both* the worth and the cost of the additional precision they will obtain from a larger sample before they commit resources to take it.

The Finite Population Multiplier

To this point in our discussion of sampling distributions, we have used Equation 6-1 to calculate the standard error of the mean:

Modifying Equation 6-1

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [6-1]$$

This equation is designed for situations in which the population is infinite, or in which we sample from a finite population with replacement (that is, after each item is sampled, it is put back into the population before the next item is chosen, so that the same item can possibly be chosen more than once). If you will refer back to page 303, where we introduced Equation 6-1, you will recall our parenthesized note, which said, “Later we shall introduce an equation for finite populations.” Introducing that equation is the purpose of this section.

Many of the populations decision makers examine are finite, that is, of stated or limited size. Examples of these include the employees in a given company, the clients of a city social-services agency, the students in a specific class, and a day’s production in a given manufacturing plant. Not one of these populations is infinite, so we need to modify Equation 6-1 to deal with them. The formula designed to find the standard error of the mean when the population is *finite*, and we sample *without replacements* is

Finding the standard error of the mean for finite populations

Standard Error of the Mean for Finite Populations

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad [6-3]$$

where

- N = size of the population
- n = size of the sample

This new term on the right-hand side, which we multiply by our original standard error, is called the *finite population multiplier*:

Finite Population Multiplier

$$\text{Finite population multiplier} = \sqrt{\frac{N-n}{N-1}} \quad [6-4]$$

A few examples will help us become familiar with interpreting and using Equation 6-3. Suppose we are interested in a population of 20 textile companies of the same size, all of which are experiencing excessive labor turnover. Our study indicates that the standard deviation of the distribution of annual turnover is 75 employees. If we sample five of these textile companies, without replacement, and wish to compute the standard error of the mean, we would use Equation 6-3 as follows:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \\ &= \frac{75}{\sqrt{5}} \times \sqrt{\frac{20-5}{20-1}} \\ &= (33.54)(0.888) \\ &= 29.8 \leftarrow \text{Standard error of the mean of a finite population}\end{aligned}\quad [6-3]$$

In this example, a finite population multiplier of 0.888 reduced the standard error from 33.54 to 29.8.

In cases in which the population is very large in relation to the size of the sample, this finite population multiplier is close to 1 and has little effect on the calculation of the standard error. Say that we have a population of 1,000 items and that we have taken a sample of 20 items. If we use Equation 6-4 to calculate the finite population multiplier, the result would be

$$\begin{aligned}\text{Finite population multiplier} &= \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{1,000-20}{1,000-1}} \\ &= \sqrt{0.981} \\ &= 0.99\end{aligned}\quad [6-4]$$

Using this multiplier of 0.99 would have little effect on the calculation of the standard error of the mean.

This last example shows that when we sample a small fraction of the entire population (that is, when the population size N is very large relative to the sample size n), the finite population multiplier takes on a value close to 1.0. Statisticians refer to the fraction n/N as the *sampling fraction*, because it is the fraction of the population N that is contained in the sample.

When the sampling fraction is small, the standard error of the mean for finite populations is so close to the standard error of the mean for infinite populations that we might as well use the same formula for both, namely, Equation 6-1: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The generally accepted rule is: **When the sampling fraction is less than 0.05, the finite population multiplier need not be used.**

When we use Equation 6-1, σ is constant, and so the measure of sampling precision, $\sigma_{\bar{x}}$ depends only on the sample size n and not on the proportion of the population sampled. That is, to make $\sigma_{\bar{x}}$ smaller, it is necessary only to make n larger. **Thus, it turns out that it is the absolute size of the sample that determines sampling precision, not the fraction of the population sampled.**

Sometimes the finite population multiplier is close to 1

Sampling fraction defined

Sample size determines sampling precision

HINTS & ASSUMPTIONS

Although the *law of diminishing return* comes from economics, it has a definite place in statistics too. It says that there is diminishing return in sampling. Specifically, although sampling more items will decrease the standard error (the standard deviation of the distribution of sample means), the increased precision may not be worth the cost. Hint: Look again at Equation 6-1 on page 303. Because n is in the denominator, when we increase it (take larger samples) the standard error ($\sigma_{\bar{x}}$) decreases. Now look at page 313. When we increased the sample size from 10 to 100 (a tenfold increase) the standard error fell only from 31.63 to 10 (about a two-thirds decrease). Maybe it wasn't smart to spend so much money increasing the sample size to get this result. That's exactly why statisticians (and smart managers) focus on the concept of the "right" sample size. Another hint: In dealing with the finite population multiplier, remember that even though we can count them, some finite populations are so large that they are treated as if they were infinite. An example of this would be the number of TV households in the United States.

Sample size determination

Determination of appropriate sample size depends upon two criteria-

- Degree of precision or extent of the permissible error (e)
- Degree of confidence placed with the sample results ($1 - \alpha$).

Estimating Population Mean,

$$\text{Sample mean} = \bar{x}$$

$$\text{Population mean} = \mu$$

$$e = (\bar{x} - \mu)$$

$$Z_{\alpha} = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}}$$

$$Z_{\alpha} = \frac{e}{\sigma/\sqrt{n}}$$

$$e = Z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{Z_{\alpha} \cdot \sigma}{e} \right)^2$$

Estimating Population Proportion,

$$\text{Sample proportion} = p$$

$$\text{Population proportion} = P.$$

$$e = |(p - P)|$$

$$Z_{\alpha} = \frac{(p - P)}{\sqrt{\frac{P(1-P)}{n}}} = \frac{e}{\sqrt{\frac{P(1-P)}{n}}}$$

$$e = Z_{\alpha} \times \sqrt{\frac{P(1-P)}{n}}$$

$$\begin{aligned} \text{Taking max. value of } P(1 - P) &= \frac{1}{4} \\ &= .25 \\ n &= \frac{Z_{\alpha}^2 \times 0.25}{e^2} \end{aligned}$$

EXERCISES 6.5

Self-Check Exercises

- SC 6-7** From a population of 125 items with a mean of 105 and a standard deviation of 17, 64 items were chosen.
- What is the standard error of the mean?
 - What is the $P(107.5 < \bar{x} < 109)$?
- SC 6-8** Jonida Martinez, researcher for the Columbian Coffee Corporation, is interested in determining the rate of coffee usage per household in the United States. She believes that yearly consumption per household is normally distributed with an unknown mean μ and a standard deviation of about 1.25 pounds.
- If Martinez takes a sample of 36 households and records their consumption of coffee for one year, what is the probability that the sample mean is within one-half pound of the population mean?
 - How large a sample must she take in order to be 98 percent certain that the sample mean is within one-half pound of the population mean?

Basic Concepts

- 6-40** From a population of 75 items with a mean of 364 and a variance of 18, 32 items were randomly selected without replacement.
- What is the standard error of the mean?
 - What is the $P(363 \leq \bar{x} \leq 366)$?
 - What would your answer to part (a) be if we sampled with replacement?
- 6-41** Given a population of size $N = 80$ with a mean of 22 and a standard deviation of 3.2, what is the probability that a sample of 25 will have a mean between 21 and 23.5?
- 6-42** For a population of size $N = 80$ with a mean of 8.2 and a standard deviation of 2.1, find the standard error of the mean for the following sample sizes:
- $n = 16$
 - $n = 25$
 - $n = 49$

Applications

- 6-43** Tread-On-Us has designed a new tire, and they don't know what the average amount of tread life is going to be. They do know that tread life is normally distributed with a standard deviation of 216.4 miles.

- (a) If the company samples 800 tires and records their tread life, what is the probability the sample mean is between the true mean and 300 miles over the true mean?
- (b) How large a sample must be taken to be 95 percent sure the sample mean will be within 100 miles of the true mean?

6-44 An underwater salvage team is preparing to explore a site off the coast of Florida where an entire flotilla of 45 Spanish galleons sank. From historical records, the team expects these wrecks to generate an average of \$225,000 in revenue when explored, and a standard deviation of \$39,000. The team's financier, however, remains skeptical, and has stated that if the exploration expenses of \$2.1 million are not recouped from the first nine wrecks, he will cancel the remainder of the exploration. What is the probability that the exploration continues past the first nine wrecks?

6-45 An X-ray technician is taking readings from her machine to ensure that it adheres to federal safety guidelines. She knows that the standard deviation of the amount of radiation emitted by the machine is 150 millirems, but she wants to take readings until the standard error of the sampling distribution is no higher than 25 millirems. How many readings should she take?

6-46 Sara Gordon is heading a fund-raising drive for Milford College. She wishes to concentrate on the current tenth-reunion class, and hopes to get contributions from 36 percent of the 250 members of that class. Past data indicate that those who contribute to the tenth-year reunion gift will donate 4 percent of their annual salaries. Sara believes that the reunion class members' annual salaries have an average of \$32,000 and a standard deviation of \$9,600. If her expectations are met (36 percent of the class donate 4 percent of their salaries), what is the probability that the tenth-reunion gift will be between \$110,000 and \$120,000?

6-47 Davis Aircraft Co. is developing a new wing de-icer system, which it has installed on 30 commercial airliners. The system is designed so that the percentage of ice removed is normally distributed with mean 96 and standard deviation 7. The FAA will do a spot check of six of the airplanes with the new system, and will approve the system if at least 98 percent of the ice is removed on average. What is the probability that the system receives FAA approval?

6-48 Food Place, a chain of 145 supermarkets, has been bought out by a larger nationwide supermarket chain. Before the deal is finalized, the larger chain wants to have some assurance that Food Place will be a consistent moneymaker. The larger chain has decided to look at the financial records for 36 of the Food Place stores. Food Place management claims that each store's profits have an approximately normal distribution with the same mean and a standard deviation of \$1,200. If the Food Place management is correct, what is the probability that the sample mean for the 36 stores will fall within \$200 of the actual mean?

6-49 Miss Joanne Happ, chief executive officer of Southwestern Life & Surety Corp., wants to undertake a survey of the huge number of insurance policies that her company has underwritten. Miss Happ's firm makes a yearly profit on each policy that is distributed with mean \$310 and standard deviation \$150. Her personal accuracy requirements dictate that the survey must be large enough to reduce the standard error to no more than 1.5 percent of the population mean. How large should her sample be?

6-50 In a study of reading habits among management students, it is desired to estimate average time spent by management students reading in library per week. From the past experience it is known that population standard deviation of the reading time is 90 minutes. How

large a sample would be required, if the researcher wants to be able to assert with 95% confidence that sample mean time would differ from the actual mean time by atmost half an hour?

6-51 Indian Oil Company has recently launched a public relation campaign to persuade its subscribers to reduce the wasteful use of the fuel. The Company's marketing research director believes that about 40% of the subscribers are aware of the campaign. He wishes to find out how large a sample would be needed to be 95% confident that true proportion is within 3% of the sample proportion.

6-52 An automobile insurance company wants to estimate from a sample study what proportion of its policy holders are interested in buying a new car within the next financial year. The total number of the policy holders is 6000. How large a sample is required to be able to assert with 95% confidence that proportion of policy holders interested in buying obtained from the sample would differ from true proportion by at most 4 percent?

Worked-Out Answers to Self-Check Exercises

SC 6-7 $N = 125 \quad \mu = 105 \quad s = 17 \quad n = 64$

$$(a) \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{17}{8} \times \sqrt{\frac{61}{124}} = 1.4904$$

$$(b) (107.5 < \bar{x} < 109) = P\left(\frac{107.5 - 105}{1.4904} < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{109 - 105}{1.4904}\right) \\ = P(1.68 < z < 2.68) = 0.4963 - 0.4535 = 0.0428$$

SC 6-8 (a) $\sigma = 1.25 \quad n = 36 \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.25/\sqrt{36} = 0.2083$

$$P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) = P\left(\frac{-0.5}{0.2083} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq \frac{0.5}{0.2083}\right) \\ = P(-2.4 \leq z \leq 2.4) = 0.4918 + 0.4918 = 0.9836$$

$$(b) 0.98 = P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) = P\left(\frac{-0.5}{1.25/\sqrt{n}} \leq z \leq \frac{0.5}{1.25/\sqrt{n}}\right) \\ = P(-2.33 \leq z \leq 2.33)$$

$$\text{Hence, } 2.33 = \frac{0.5}{1.25/\sqrt{n}} = 0.4\sqrt{n} \quad \text{and} \quad n = (2.33/0.4)^2 = 33.93.$$

She should sample at least 34 households.

STATISTICS AT WORK

Loveland Computers

Case 6: Samplings and Sampling Distributions After less than a week on the job as an administrative assistant to Loveland Computers' CEO, Lee Azko was feeling almost overwhelmed with the range

of projects that seemed to demand attention. But, there was no use denying, it sure felt good to put into practice some of the techniques that had been taught in school. And the next day on the job brought a new set of challenges.

"I guess those folks in production must like you," Walter Azko greeted Lee by the coffee machine. "I hope you're all done with purchasing because production has a quality control problem it needs help with. Go and see Nancy Rainwater again."

Lee went down to the assembly line but was greeted by an unfamiliar face. Tyronza Wilson introduced himself. "Nancy said you'd be down. I'm in charge of checking the components we use when we assemble high-end computers for customers. For most of the components, the suppliers are so reliable that we just assume they're going to work. In the very rare case there's a failure, we catch it at the end of the line, where we run the computers overnight on a test program to 'burn them in.' That means, we don't want to be surprised by a part that fails when it's been on the job for only a few hours.

"Recently, we've been having a problem with the 3-gigabyte hard drives. You know, everyone used to be happy with one or two gigabytes of storage, but new programs with fancy graphics eat up a great deal of disk space and many of the customers are specifying the large drive for their computers. To move large amounts of data, *access time* becomes very important—that's a measure of the average time that it takes to retrieve a standard amount of data from a hard drive. Because access-time performance is important to our customers, I can't just assume that every hard drive is going to work within specifications. If we wait to test access time at the end of the line and find we have a drive that's too slow, we have to completely rebuild the computer with a new drive and drive controller. That's a lot of expensive rework that we should avoid.

"But it'd be even more expensive to test every one of them at the beginning of the process—the only way I can measure the access time of each drive is to hook it up to a computer and run a diagnostic program. All told, that takes the best part of a quarter of an hour. I don't have the staff or the machines to test every one, and it's rather pointless because the vast majority of them will pass inspection.

"There's more demand than supply for the high-capacity hard drives right now, so we've been buying them all over the place. As a result, there seem to be 'good shipments' and 'bad shipments.' If the average access time of a shipment is too long, we return them to the supplier and reject their invoice. That saves us paying for something we can't use, but if I reject too many shipments, it leaves us short of disk drives to complete our orders.

"Obviously we need some kind of sampling scheme here—we need to measure the access time on a sample of each shipment and then make our decision about the lot. But I'm not sure how many we should test."

"Well, I think you have a good handle on the situation," said Lee, taking out a notepad. "Let me begin by asking you a few questions."

Study Questions: What types of sampling schemes will Lee consider and what factors will influence the choice of scheme? What questions should Lee have for Tyronza?

CHAPTER REVIEW

Terms Introduced in Chapter 6

Census The measurement or examination of every element in the population.

Central Limit Theorem A result assuring that the sampling distribution of the mean approaches normality as the sample size increases, regardless of the shape of the population distribution from which the sample is selected.

Clusters Within a population, groups that are essentially similar to each other, although the groups themselves have wide internal variation.

Cluster Sampling A method of random sampling in which the population is divided into groups, or clusters of elements, and then a random sample of these clusters is selected.

Factorial Experiment An experiment in which each factor involved is used once with each other factor. In a complete factorial experiment, every level of each factor is used with each level of every other factor.

Finite Population A population having a stated or limited size.

Finite Population Multiplier A factor used to correct the standard error of the mean for studying a population of finite size that is small in relation to the size of the sample.

Infinite Population A population in which it is theoretically impossible to observe all the elements.

Judgment Sampling A method of selecting a sample from a population in which personal knowledge or expertise is used to identify the items from the population that are to be included in the sample.

Latin Square An efficient experimental design that makes it unnecessary to use a complete factorial experiment.

Parameters Values that describe the characteristics of a population.

Precision The degree of accuracy with which the sample mean can estimate the population mean, as revealed by the standard error of the mean.

Random or Probability Sampling A method of selecting a sample from a population in which all the items in the population have an equal chance of being chosen in the sample.

Sample A portion of the elements in a population chosen for direct examination or measurement.

Sampling Distribution of the Mean A probability distribution of all the possible means of samples of a given size, n , from a population.

Sampling Distribution of a Statistic For a given population, a probability distribution of all the possible values a statistic may take on for a given sample size.

Sampling Error Error or variation among sample statistics due to chance, that is, differences between each sample and the population, and among several samples, which are due solely to the elements we happen to choose for the sample.

Sampling Fraction The fraction or proportion of the population contained in a sample.

Sampling with Replacement A sampling procedure in which sampled items are returned to the population after being picked, so that some members of the population can appear in the sample more than once.

Sampling without Replacement A sampling procedure in which sampled items are not returned to the population after being picked, so that no member of the population can appear in the sample more than once.

Simple Random Sampling Methods of selecting samples that allow each possible sample an equal probability of being picked and each item in the entire population an equal chance of being included in the sample.

Standard Error The standard deviation of the sampling distribution of a statistic.

Standard Error of the Mean The standard deviation of the sampling distribution of the mean; a measure of the extent to which we expect the means from different samples to vary from the population mean, owing to the chance error in the sampling process.

Statistical Inference The process of making inferences about populations from information contained in samples.

Statistics Measures describing the characteristics of a sample.

Strata Groups within a population formed in such a way that each group is relatively homogeneous, but wider variability exists among the separate groups.

Stratified Sampling A method of random sampling in which the population is divided into homogeneous groups, or strata, and elements within each stratum are selected at random according to one of two rules: (1) A specified number of elements is drawn from each stratum corresponding to the proportion of that stratum in the population, or (2) equal numbers of elements are drawn from each stratum, and the results are weighted according to the stratum's proportion of the total population.

Systematic Sampling A method of sampling in which elements to be sampled are selected from the population at a uniform interval that is measured in time, order, or space.

Equations Introduced in Chapter 6

$$6-1 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{p. 303}$$

Use this formula to derive the standard error of the mean when the population is infinite, that is, when the elements of the population cannot be enumerated in a reasonable period of time, or when we sample with replacement. This equation states that the sampling distribution has a standard deviation, which we also call a standard error, equal to the population standard deviation divided by the square root of the sample size.

$$6-2 \quad z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{p. 304}$$

A modified version of Equation 5-6, this formula allows us to determine the distance of the *sample mean* \bar{x} from the population mean μ , when we divide the difference by the standard error of the mean $\sigma_{\bar{x}}$. Once we have derived a z value, we can use the Standard Normal Probability Distribution Table and compute the probability that the sample mean will be that distance from the population mean. Because of the central limit theorem, we can use this formula for nonnormal distributions if the sample size is at least 30.

$$6-3 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad \text{p. 314}$$

where

- N = size of the population
- n = size of the sample

This is the formula for finding the *standard error of the mean* when the population is *finite*, that is, of stated or limited size, and the sampling is done *without* replacement.

$$6-4 \quad \text{Finite population multiplier} = \sqrt{\frac{N-n}{N-1}} \quad \text{p. 314}$$

In Equation 6-3, the term $\sqrt{(N-n)/(N-1)}$, which we multiply by the standard error from Equation (6-1), is called the *finite population multiplier*. When the population is small in relation to the size of the sample, the finite population multiplier reduces the size of the standard error. Any decrease in the standard error increases the precision with which the sample mean can be used to estimate the population mean.

Review and Application Exercises

- 6-50** Crash Davis is the line supervisor for the Benicia, California, plant of a manufacturer of in-line skates. Close fit is important for in-line skating gear, so Crash tests each day's production by selecting a size 13 pair from the line and skating to get his afternoon cappuccino down the street. Crash points out that he selects each pair "at random." Is this, in fact, a random sample of the day's production, or is it judgmental?
- 6-51** Jim Ford, advertising manager for a retail department store chain, is responsible for choosing the final advertisements from sample layouts designed by his staff. He has been in the retail advertising business for years and has been responsible for the chain's advertising for quite some time. His assistant, however, having learned the latest advertising effectiveness measurement techniques while at a New York agency, wants to do effectiveness tests for each advertisement considered, using random samples of consumers in the store's retail trading district. These tests will be quite costly. Jim is sure that his experience enables him to decide on appropriate ads, so there has been some disagreement between the two. Can you defend either position?
- 6-52** Burt Purdue, manager of the Sea Island Development Company, wants to find out residents' feelings toward the development's recreation facilities and the improvements they would like to see implemented. The development includes residents of various ages and income levels, but a large proportion are middle-class residents between the ages of 30 and 50. As yet, Burt is unsure whether there are differences among age groups or income levels in their desire for recreation facilities. Would stratified random sampling be appropriate here?
- 6-53** A camera manufacturer is attempting to find out what employees feel are the major problems with the company and what improvements are needed. To assess the opinions of the 37 departments, management is considering a sampling plan. It has been recommended to the personnel director that management adopt a cluster sampling plan. Management would choose six departments and interview all the employees. Upon collecting and assessing the data gathered from these employees, the company could then make changes and plan for areas of job improvement. Is a cluster sampling plan appropriate in this situation?
- 6-54** By reviewing sales since opening 6 months ago, a restaurant owner found that the average bill for a couple was \$26, and the standard deviation was \$5.65. How large would a sample of customers have to be for the probability to be at least 95.44 percent that the mean cost per meal for the sample would fall between \$25 and \$27?
- 6-55** The end of March in 1992 saw the following state-by-state unemployment rates in the United States.

State	Unemployment Rate (%)	State	Unemployment Rate (%)
Alabama	7.5	Montana	7.3
Alaska	10.1	Nebraska	2.8
Arizona	8.4	Nevada	6.8
Arkansas	7.0	New Hampshire	7.5
California	8.7	New Jersey	7.5
Colorado	6.3	New Mexico	7.6
Connecticut	7.4	New York	8.5
Delaware	6.4	North Carolina	6.4
District of Columbia	8.2	North Dakota	5.3
Florida	8.1	Ohio	7.8

(continued)

(contd.)

State	Unemployment Rate (%)	State	Unemployment Rate (%)
Georgia	6.3	Oklahoma	6.8
Hawaii	3.5	Oregon	8.6
Idaho	7.8	Pennsylvania	7.6
Illinois	8.2	Rhode Island	8.9
Indiana	6.3	South Carolina	7.1
Iowa	5.3	South Dakota	4.0
Kansas	3.6	Tennessee	7.0
Kentucky	7.0	Texas	7.4
Louisiana	6.9	Utah	5.0
Maine	8.4	Vermont	7.1
Maryland	7.4	Virginia	6.8
Massachusetts	10.0	Washington	8.3
Michigan	10.0	West Virginia	12.9
Minnesota	6.3	Wisconsin	5.7
Mississippi	8.1	Wyoming	7.5
Missouri	5.6		

Source: Sharon R. Cohany, "Current Labor Statistics: Employment Data," *Monthly Labor Review* 115 (6), (June 1992): 80–82.

- (a) Compute the population mean and standard deviation of the unemployment rates.
- (b) Using the states of Alabama, Kansas, Michigan, Nebraska, and North Carolina as a random sample (taken without replacement), determine the sample mean, \bar{x} .
- (c) What are the mean ($\mu_{\bar{x}}$) and standard deviation ($\sigma_{\bar{x}}$) of the sampling distribution of \bar{x} , the sample mean of all samples of size $n = 5$, taken without replacement?
- (d) Consider the sampling distribution of \bar{x} for samples of size $n = 5$, taken without replacement. Is it reasonable to assume that this distribution is normal or approximately so? Explain.
- (e) Notwithstanding your answer to part (d), assume that the sampling distribution of \bar{x} for samples of size $n = 5$, taken without replacement, is approximately normal. What is the probability that the mean of such a random sample will lie between 5.9 and 6.5?

6-56 Joan Fargo, president of Fargo-Lanna Ltd., wants to offer videotaped courses for employees during the lunch hour, and wants to get some idea of the courses that employees would like to see offered. Accordingly, she has devised a ballot that an employee can fill out in 5 minutes, listing his or her preferences among the possible courses. The ballots, which cost very little to print, will be distributed with paychecks, and the results will be tabulated by the as yet unassigned clerical staff of a recently dissolved group within the company. Ms. Fargo plans to poll all employees. Are there any reasons to poll a sample of the employees rather than the entire population?

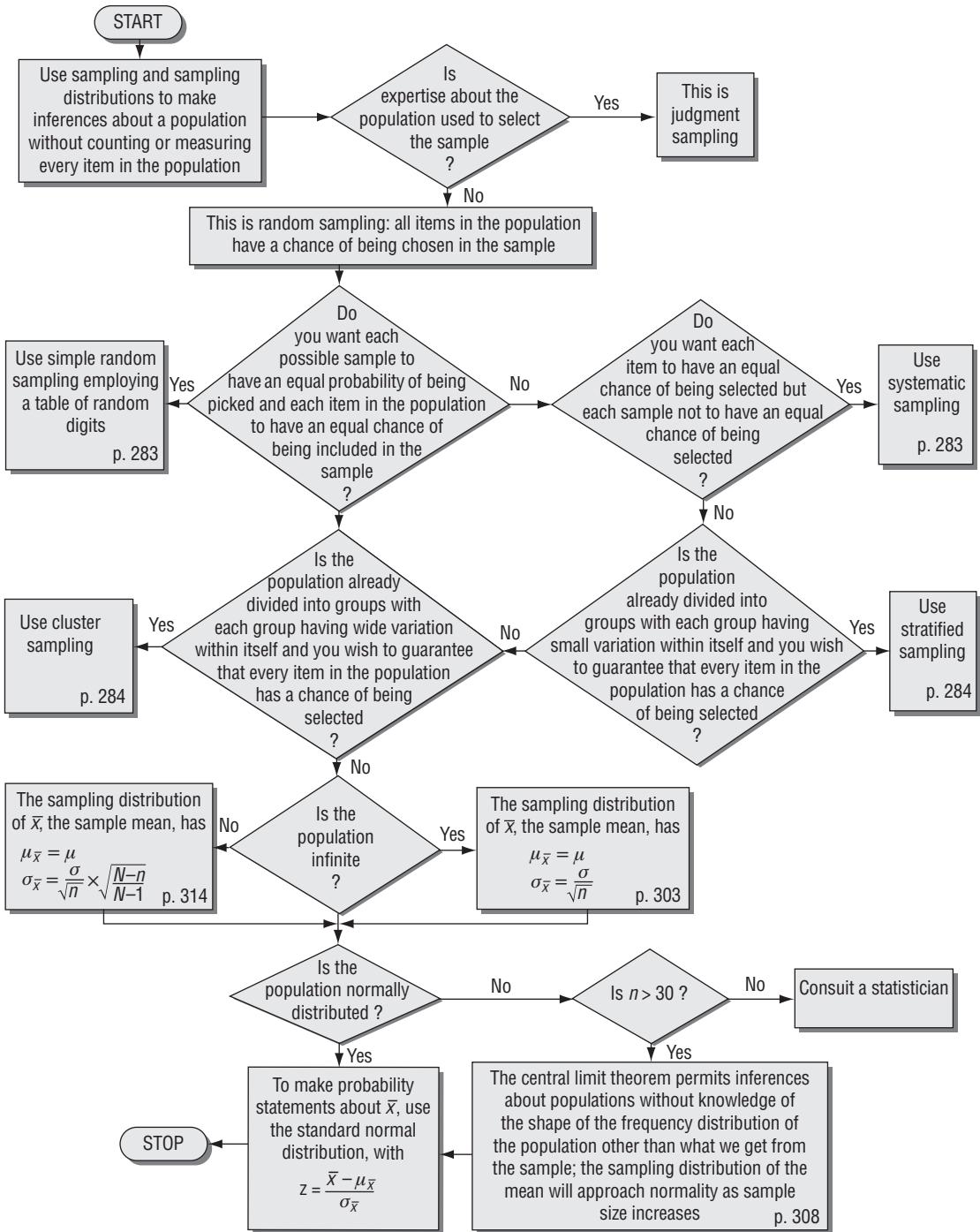
6-57 A drug manufacturer knows that for a certain antibiotic, the average number of doses ordered for a patient is 20. Steve Simmons, a salesman for the company, after looking at one day's prescription orders for the drug in his territory, announced that the sample mean for this drug should be lower. He said, "For any sample, the mean should be lower, since the sampling mean always understates the population mean because of sample variation." Is there any truth to what Simmons said?

6-58 Several weeks later at a sales meeting, Steve Simmons again demonstrated his expertise in statistics. He had drawn a graph and presented it to the group, saying, "This is a sampling

distribution of means. It is a normal curve and represents a distribution of all observations in each possible sample combination.” Is Simmons right? Explain.

- 6-59** Low-Cal Foods Company uses estimates of the level of activity for various market segments to determine the nutritional composition of its diet food products. Low-Cal is considering the introduction of a liquid diet food for older women, since this segment has special weight problems not met by the competitor’s diet foods. To determine the desired calorie content of this new product, Dr. Nell Watson, researcher for the company, conducted tests on a sample of women to determine calorie consumption per day. Her results showed that the average number of calories expended per day for older women is 1,328 and the standard deviation is 275. Dr. Watson estimates that the benefits she obtains with a sample size of 25 are worth \$1,720. She expects that reducing the standard error by half its current value will double the benefit. If it costs \$16 for every woman in the sample, should Watson reduce her standard error?
- 6-60** The U.S. Customs Agency routinely checks all passengers arriving from foreign countries as they enter the United States. The department reports that the number of people per day found to be carrying contraband material as they enter the United States through John F. Kennedy airport in New York averages 42 and has a standard deviation of 11. What is the probability that in five days at the airport, the average number of passengers found carrying contraband will exceed 50?
- 6-61** HAL Corporation manufactures large computer systems and has always prided itself on the reliability of its System 666 central processing units. In fact, past experience has shown that the monthly downtime of System 666 CPUs averages 41 minutes, and has a standard deviation of 8 minutes. The computer center at a large state university maintains an installation built around six System 666 CPUs. James Kitchen, the director of the computer center, feels that a satisfactory level of service is provided to the university community if the average downtime of the six CPUs is less than 50 minutes per month. In any given month, what is the probability that Kitchen will be satisfied with the level of service?
- 6-62** Members of the Organization for Consumer Action send more than 250 volunteers a day all over the state to increase support for a consumer protection bill that is currently before the state legislature. Usually, each volunteer will visit a household and talk briefly with the resident in the hope that the resident will sign a petition to be given to the state legislature. The number of signatures a volunteer obtains for the petition each day averages 5.8 and has a standard deviation of 0.8. What is the probability a sample of 20 volunteers will result in an average between 5.5 and 6.2 signatures per day?
- 6-63** Jill Johnson, product manager for Southern Electric’s smoke alarm, is concerned over recent complaints from consumer groups about the short life of the device. She has decided to gather evidence to counteract the complaints by testing a sample of the alarms. For the test, it costs \$4 per unit in the sample. Precision is desirable for presenting persuasive statistical evidence to consumer groups, so Johnson figures the benefits she will receive for various sample sizes are determined by the formula Benefits = $\$5,249/\sigma_{\bar{x}}$. If Johnson wants to increase her sample until the cost equals the benefit, how many units should she sample? The population standard deviation is 265.
- 6-64** Seventy data clerks at the Department of Motor Vehicles make an average of 18 errors per day, normally distributed with a standard deviation of 4. A field auditor can check the work of 15 clerks per day. What is the probability that the average number of errors in a group of 15 clerks checked on one day is
- Fewer than 15.5?
 - Greater than 20?

Flow Chart: Sampling and Sampling Distributions



7

Estimation

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn how to estimate certain characteristics of a population from samples
 - To learn the strengths and shortcomings of point estimates and interval estimates
 - To calculate how accurate our estimates really are
 - To learn how to use the t distribution to make interval estimates in some cases when the normal distribution cannot be used
 - To calculate the sample size required for any desired level of precision in estimation
-

CHAPTER CONTENTS

- | | |
|---|---|
| 7.1 Introduction 328 | ■ Statistics at Work 370 |
| 7.2 Point Estimates 331 | ■ Terms Introduced in Chapter 7 371 |
| 7.3 Interval Estimates: Basic Concepts 336 | ■ Equations Introduced in Chapter 7 371 |
| 7.4 Interval Estimates and Confidence Intervals 341 | ■ Review and Application Exercises 372 |
| 7.5 Calculating Interval Estimates of the Mean from Large Samples 344 | ■ Flow Chart: Estimation 377 |
| 7.6 Calculating Interval Estimates of the Proportion from Large Samples 349 | |
| 7.7 Interval Estimates Using the t Distribution 353 | |
| 7.8 Determining the Sample Size in Estimation 364 | |

As part of the budgeting process for next year, the manager of the Far Point electric generating plant must estimate the coal he will need for this year. Last year, the plant almost ran out, so he is reluctant to budget for that same amount again. The plant manager, however, does feel that past usage data will help him *estimate* the number of tons of coal to order. A random sample of 10 plant operating weeks chosen over the last 5 years yielded a mean usage of 11,400 tons a week, and a sample standard deviation of 700 tons a week. With the data he has and the methods we shall discuss in this chapter, the plant manager can make a sensible estimate of the amount to order this year, including some idea of the accuracy of the estimate he has made. ■

7.1 INTRODUCTION

Everyone makes estimates. When you are ready to cross a street, you estimate the speed of any car that is approaching, the distance between you and that car, and your own speed. Having made these quick estimates, you decide whether to wait, walk, or run.

All managers must make quick estimates too. The outcome of these estimates can affect their organizations as seriously as the outcome of your decision as to whether to cross the street.

Reasons for estimates

University department heads make estimates of next fall's enrollment in statistics. Credit managers estimate whether a purchaser will eventually pay his bills. Prospective home buyers make estimates concerning the behavior of interest rates in the mortgage market. All these people make estimates without worry about whether they are scientific but with the hope that the estimates bear a reasonable resemblance to the outcome.

Managers use estimates because in all but the most trivial decisions, they must make rational decisions without complete information and with a great deal of uncertainty about what the future will bring. As educated citizens and professionals, you will be able to make more useful estimates by applying the techniques described in this and subsequent chapters.

The material on probability theory covered in Chapters 4, 5, and 6 forms the foundation for *statistical inference*, the branch of statistics concerned with using probability concepts to deal with uncertainty in decision making. Statistical inference is based on *estimation*, which we shall introduce in this chapter, and *hypothesis testing*, which is the subject of Chapters 8, 9, and 10. In both estimation and hypothesis testing, we shall be making inferences about characteristics of populations from information contained in samples.

Making statistical inferences

How do managers use sample statistics to estimate population parameters? The department head attempts to estimate enrollments next fall from current enrollments in the same courses. The credit manager attempts to estimate the creditworthiness of prospective customers from a sample of their past payment habits. The home buyer attempts to estimate the future course of interest rates by observing the current behavior of those rates. In each case, somebody is trying to infer something about a population from information taken from a sample.

Using samples

This chapter introduces methods that enable us to estimate with reasonable accuracy the *population proportion* (the proportion of the population that possesses a given characteristic) and the *population mean*. To calculate the exact proportion or the exact mean would be an impossible goal. Even so, we will be able to make an estimate, make a statement about the error that

Estimating population parameters

will probably accompany this estimate, and implement some controls to avoid as much of the error as possible. As decision makers, we will be forced at times to rely on blind hunches. Yet in other situations, in which information is available and we apply statistical concepts, we can do better than that.

Types of Estimates

We can make two types of estimates about a population: a *point estimate defined* and an *interval estimate*. A **point estimate is a single number that is used to estimate an unknown population parameter**. If, while watching the first members of a football team come onto the field, you say, “Why, I bet their line must average 250 pounds,” you have made a point estimate. A department head would make a point estimate if she said, “Our current data indicate that this course will have 350 students in the fall.”

A point estimate is often insufficient, because it is either right *Shortcoming of point estimates* or wrong. If you are told only that her point estimate of enrollment is wrong, you do not know *how* wrong it is, and you cannot be certain of the estimate’s reliability. If you learn that it is off by only 10 students, you would accept 350 students as a good estimate of future enrollment. But if the estimate is off by 90 students, you would reject it as an estimate of future enrollment. Therefore, a point estimate is much more useful if it is accompanied by an estimate of the error that might be involved.

An **interval estimate is a range of values used to estimate a population parameter**. It indicates the error in two ways: *Interval estimate defined* by the extent of its range and by the probability of the true population parameter lying within that range. In this case, the department head would say something like, “I estimate that the true enrollment in this course in the fall will be between 330 and 380 and that it is very likely that the exact enrollment will fall within this interval.” She has a better idea of the reliability of her estimate. If the course is taught in sections of about 100 students each, and if she had tentatively scheduled five sections, then on the basis of her estimate, she can now cancel one of those sections and offer an elective instead.

Estimator and Estimates

Any sample statistic that is used to estimate a population parameter is called an *estimator*, that is, **an estimator is a sample statistic used to estimate a population parameter**. The sample mean \bar{x} can be an estimator of the population mean μ , and the sample proportion can be used as an estimator of the population proportion. We can also use the sample range as an estimator of the population range.

When we have observed a specific numerical value of our *Estimator defined* estimator, we call that value an *estimate*. In other words, **an estimate is a specific observed value of a statistic**. We form an estimate by taking a sample and computing the value taken by our estimator in that sample. Suppose that we calculate the mean odometer reading (mileage) from a sample of used taxis and find it to be 98,000 miles. If we use this specific value to estimate the mileage for a whole fleet of used taxis, the value 98,000 miles would be an estimate. Table 7-1 illustrates several populations, population parameters, estimators, and estimates.

TABLE 7-1 POPULATIONS, POPULATION PARAMETERS, ESTIMATORS, AND ESTIMATES

Population in Which We Are Interested	Population Parameter We Wish to Estimate	Sample Statistic We Will Use as an Estimator	Estimate We Make
Employees in a furniture factory	Mean turnover per year	Mean turnover for a period of 1 month	8.9% turnover per year
Applicants for Town Manager of Chapel Hill	Mean formal education (years)	Mean formal education of every fifth applicant	17.9 years of formal education
Teenagers in a given community	Proportion who have criminal records	Poportion of a sample of 50 teenagers who have criminal records	0.02, or 2%, have criminal records

Criteria of a Good Estimator

Some statistics are better estimators than others. Fortunately, we can evaluate the quality of a statistic as an estimator by using four criteria:

Qualities of a good estimator

- 1. Unbiasedness.** This is a desirable property for a good estimator to have. The term *unbiasedness* refers to the fact that a sample mean is an *unbiased estimator* of a population mean because **the mean of the sampling distribution of sample means taken from the same population is equal to the population mean itself**. We can say that a statistic is an unbiased estimator if, on average, it tends to assume values that are above the population parameter being estimated as frequently and to the same extent as it tends to assume values that are below the population parameter being estimated.
- 2. Efficiency.** Another desirable property of a good estimator is that it be efficient. *Efficiency* refers to the size of the standard error of the statistic. If we compare two statistics from a sample of the same size and try to decide which one is the more efficient estimator, we would pick the statistic that has the smaller standard error, or standard deviation of the sampling distribution. Suppose we choose a sample of a given size and must decide whether to use the sample mean or the sample median to estimate the population mean. If we calculate the standard error of the sample mean and find it to be 1.05 and then calculate the standard error of the sample median and find it to be 1.6, we would say that the sample mean is a *more efficient estimator* of the population mean *because its standard error is smaller*. It makes sense that an estimator with a smaller standard error (with less variation) will have more chance of producing an estimate nearer to the population parameter under consideration.
- 3. Consistency.** A statistic is a *consistent estimator* of a population parameter if *as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter*. If an estimator is consistent, it becomes more reliable with large samples. Thus, if you are wondering whether to increase the sample size to get more information about a population parameter, find out first whether your statistic is a consistent estimator. If it is not, you will waste time and money by taking larger samples.
- 4. Sufficiency.** An estimator is *sufficient* if it makes so much use of the information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

We present these criteria here to make you aware of the care that statisticians must use in picking an estimator.

A given sample statistic is not always the best estimator of its analogous population parameter. Consider a symmetrically distributed population in which the values of the median and the mean coincide. In this instance, the sample mean would be an *unbiased* estimator of population median. Also, the sample mean would be a *consistent* estimator of the population median because, as the sample size increases, the value of the sample mean would tend to come very close to the population median. And the sample mean would be a more *efficient* estimator of the population median than the sample median itself because in large samples, the sample mean has a smaller standard error than the sample median. At the same time, the sample median in a symmetrically distributed population would be an unbiased and consistent estimator of the population mean but *not the most efficient* estimator because in large samples, its standard error is larger than that of the sample mean.

Finding the best estimator**EXERCISES 7.1**

- 7-1** What two basic tools are used in making statistical inferences?
- 7-2** Why do decision makers often measure samples rather than entire populations? What is the disadvantage?
- 7-3** Explain a shortcoming that occurs in a point estimate but not in an interval estimate. What measure is included with an interval estimate to compensate for this?
- 7-4** What is an estimator? How does an estimate differ from an estimator?
- 7-5** List and describe briefly the criteria of a good estimator.
- 7-6** What role does consistency play in determining sample size?

7.2 POINT ESTIMATES

The sample mean \bar{x} is the best estimator of the population mean μ . It is unbiased, consistent, the most efficient estimator, and, as long as the sample is sufficiently large, its sampling distribution can be approximated by the normal distribution.

Using the sample mean to estimate the population mean

If we know the sampling distribution of \bar{x} , we can make statements about any estimate we may make from sampling information. Let's look at a medical-supplies company that produces disposable hypodermic syringes. Each syringe is wrapped in a sterile package and then jumble-packed in a large corrugated carton. Jumble packing causes the cartons to contain differing numbers of syringes. Because the syringes are sold on a per unit basis, the company needs an estimate of the number of syringes per carton for billing purposes. We have taken a sample of 35 cartons at random and recorded the number of syringes in each carton. Table 7-2 illustrates our results. Using the results of Chapter 3, we can obtain the sample mean \bar{x} by finding the sum of all our results, $\sum x$, and dividing this total by n , the number of cartons we have sampled:

$$\bar{x} = \frac{\sum x}{n} \quad [3-2]$$

Using this equation to solve our problem, we get

$$\begin{aligned}\bar{x} &= \frac{3,570}{35} \\ &= 102 \text{ syringes}\end{aligned}$$

TABLE 7-2 RESULTS OF A SAMPLE OF 35 CARTONS OF HYPODERMIC SYRINGES (SYRINGES PER CARTON)

101	103	112	102	98	97	93
105	100	97	107	93	94	97
97	100	110	106	110	103	99
93	98	106	100	112	105	100
114	97	110	102	98	112	99

Thus, using the sample mean \bar{x} as our estimator, the point estimate of the population mean μ is 102 syringes per carton. The manufactured price of a disposable hypodermic syringe is quite small (about 25¢), so both the buyer and seller would accept the use of this point estimate as the basis for billing, and the manufacturer can save the time and expense of counting each syringe that goes into a carton.

Point Estimate of the Population Variance and Standard Deviation

Suppose the management of the medical-supplies company wants to estimate the variance and/or standard deviation of the distribution of the number of packaged syringes per carton. The most frequently used estimator of the population standard deviation σ is the sample standard deviation s . We can calculate the sample standard deviation as in Table 7-3 and discover that it is 6.01 syringes.

Using the sample standard deviation to estimate the population standard deviation

TABLE 7-3 CALCULATION OF SAMPLE VARIANCE AND STANDARD DEVIATION FOR SYRINGES PER CARTON

Values of x (Needles per Carton) (1)	x^2 (2)	Sample Mean x (3)	$(x - \bar{x})$ (4) = (1) - (3)	$(x - \bar{x})^2$ (5) = (4) ²
101	10,201	102	-1	1
105	11,025	102	3	9
97	9,409	102	-5	25
93	8,649	102	-9	81
114	12,996	102	12	144
103	10,609	102	1	1
100	10,000	102	-2	4
100	10,000	102	-2	4
98	9,604	102	-4	16
97	9,409	102	-5	25
112	12,544	102	10	100
97	9,409	102	-5	25

TABLE 7-3 CALCULATION OF SAMPLE VARIANCE AND STANDARD DEVIATION FOR SYRINGES PER CARTON (Contd.)

110	12,100	102	8	64
106	11,236	102	4	16
110	12,100	102	8	64
102	10,404	102	0	0
107	11,449	102	5	25
106	11,236	102	4	16
100	10,000	102	-2	4
102	10,404	102	0	0
98	9,604	102	-4	16
93	8,649	102	-9	81
110	12,100	102	8	64
112	12,544	102	10	100
98	9,604	102	-4	16
97	9,409	102	-5	25
94	8,836	102	-8	64
103	10,609	102	1	1
105	11,025	102	3	9
112	12,544	102	10	100
93	8,649	102	-9	81
97	9,409	102	-5	25
99	9,801	102	-3	9
100	10,000	102	-2	4
99	9,801	102	-3	9
3,570	365,368			
		Sum of all the squared differences	$\Sigma(x - \bar{x})^2 \rightarrow$	1,228
[3-17]	$s^2 = \frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}$ $= \frac{365,368}{34} - \frac{35(102)^2}{34}$ $= \frac{1,228}{34}$ $= 36.12$	Sum of the squared differences divided by 34, the number of items in the sample - 1 (sample variance)	$\frac{\Sigma(x - \bar{x})^2}{n-1} \rightarrow 36.12$	
[3-18]	$s = \sqrt{s^2}$ $= \sqrt{36.12}$ $= 6.01 \text{ syringes}$	Sample standard deviation s	$\sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}} \rightarrow 6.01 \text{ syringes}$	

If, instead of considering

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1} \quad [3-17]$$

as our sample variance, we had considered

Why is n – 1 the divisor?

$$s^2 = \frac{\sum(x - \bar{x})^2}{n}$$

the result would have some *bias* as an estimator of the population variance; specifically, it would tend to be too low. Using a divisor of $n - 1$ gives us an unbiased estimator of σ^2 . Thus, we will use s^2 (as defined in Equation 3-17) and s (as defined in Equation 3-18) to estimate σ^2 and σ .

Point Estimate of the Population Proportion

The proportion of units that have a particular characteristic in a given population is symbolized p . If we know the proportion of units in a sample that have that same characteristic (symbolized \bar{p}), we can use this \bar{p} as an estimator of p . It can be shown that \bar{p} has all the desirable properties we discussed earlier; it is unbiased, consistent, efficient, and sufficient.

Using the sample proportion to estimate the population proportion

Continuing our example of the manufacturer of medical supplies, we shall try to estimate the population proportion from the sample proportion. Suppose management wishes to estimate the number of cartons that will arrive damaged, owing to poor handling in shipment after the cartons leave the factory. We can check a sample of 50 cartons from their shipping point to the arrival at their destination and then record the presence or absence of damage. If, in this case, we find that the proportion of damaged cartons in the sample is 0.08, we would say that

$$\bar{p} = 0.08 \leftarrow \text{Sample proportion damaged}$$

Because the sample proportion \bar{p} is a convenient estimator of the population proportion p , we can estimate that the proportion of damaged cartons in the population will also be 0.08.

HINTS & ASSUMPTIONS

Putting all of the definitions aside, the reason we study estimators is so we can learn about populations by sampling, without counting every item in the population. Of course, there is no free lunch here either, and when we give up counting everything, we lose some accuracy. Managers would like to know the accuracy that *is* achieved when we sample, and using the ideas in this chapter, we can tell them. Hint: Determining the best sample size is not just a statistical decision. Statisticians can tell you how the standard error behaves as you increase or decrease the sample size, and market researchers can tell you what the cost of taking more or larger samples will be. But it's you who must use your judgment to combine these two inputs to make a sound *managerial* decision.

EXERCISES 7.2

Self-Check Exercises

- SC 7-1** The Greensboro Coliseum is considering expanding its seating capacity and needs to know both the average number of people who attend events there and the variability in this number. The following are the attendances (in thousands) at nine randomly selected sporting events. Find point estimates of the mean and the variance of the population from which the sample was drawn.

8.8 14.0 21.3 7.9 12.5 20.6 16.3 14.1 13.0

- SC 7-2** The Pizza Distribution Authority (PDA) has developed quite a business in Carrboro by delivering pizza orders promptly. PDA guarantees that its pizzas will be delivered in 30 minutes or less from the time the order was placed, and if the delivery is late, the pizza is free. The time that it takes to deliver each pizza order that is on time is recorded in the Official Pizza Time Book (OPTB), and the delivery time for those pizzas that are delivered late is recorded as 30 minutes in the OPTB. Twelve random entries from the OPTB are listed.

15.3 29.5 30.0 10.1 30.0 19.6
10.8 12.2 14.8 30.0 22.1 18.3

- Find the mean for the sample.
- From what population was this sample drawn?
- Can this sample be used to estimate the average time that it takes for PDA to deliver a pizza? Explain.

Applications

- 7-7** Joe Jackson, a meteorologist for local television station WDUL, would like to report the average rainfall for today on this evening's newscast. The following are the rainfall measurements (in inches) for today's date for 16 randomly chosen past years. Determine the sample mean rainfall.

0.47 0.27 0.13 0.54 0.00 0.08 0.75 0.06
0.00 1.05 0.34 0.26 0.17 0.42 0.50 0.86

- 7-8** The National Bank of Lincoln is trying to determine the number of tellers available during the lunch rush on Fridays. The bank has collected data on the number of people who entered the bank during the last 3 months on Friday from 11 A.M. to 1 P.M. Using the data below, find point estimates of the mean and standard deviation of the population from which the sample was drawn.

242 275 289 306 342 385 279 245 269 305 294 328

- 7-9** Electric Pizza was considering national distribution of its regionally successful product and was compiling pro forma sales data. The average monthly sales figures (in thousands of dollars) from its 30 current distributors are listed. Treating them as (a) a sample and (b) a population, compute the standard deviation.

7.3 5.8 4.5 8.5 5.2 4.1
2.8 3.8 6.5 3.4 9.8 6.5

6.7	7.7	5.8	6.8	8.0	3.9
6.9	3.7	6.6	7.5	8.7	6.9
2.1	5.0	7.5	5.8	6.4	5.2

- 7-10 In a sample of 400 textile workers, 184 expressed extreme dissatisfaction regarding a prospective plan to modify working conditions. Because this dissatisfaction was strong enough to allow management to interpret plan reaction as being highly negative, they were curious about the proportion of total workers harboring this sentiment. Give a point estimate of this proportion.
- 7-11 The Friends of the Psychics network charges \$3 per minute to learn the secrets that can turn your life around. The network charges for whole minutes only and rounds up to benefit the company. Thus, a 2 minute 10 second call costs \$9. Below is a list of 15 randomly selected charges.

3 9 15 21 42 30 6 9 6 15 21 24 32 9 12

- (a) Find the mean of the sample.
- (b) Find a point estimate of the variance of the population.
- (c) Can this sample be used to estimate the average length of a call? If so, what is your estimate? If not, what can we estimate using this sample?

Worked-Out Answers to Self-Check Exercises

SC 7-1 $\sum x^2 = 2003.65 \quad \sum x = 128.5 \quad n = 9$

$$\bar{x} = \frac{\sum x}{n} = \frac{128.5}{9} = 14.2778 \text{ thousands of people}$$

$$s^2 = \frac{1}{n-1} (\sum x^2 - n\bar{x}^2) = \frac{2003.65 - 9(14.2778)^2}{8} \\ = 21.119 \text{ (1,000s of people)}^2$$

SC 7-2 (a) $\bar{x} = \frac{\sum x}{n} = \frac{242.7}{12} = 20.225 \text{ minutes.}$

- (b) The population of times recorded in the OPTB.
- (c) No, it cannot. Because every delivery time over 30 minutes is recorded as 30 minutes, use of these will consistently underestimate the average of the delivery time.

7.3 INTERVAL ESTIMATES: BASIC CONCEPTS

The purpose of gathering samples is to learn more about a population. We can compute this information from the sample data as either *point* estimates, which we have just discussed, or as *interval* estimates, the subject of the rest of this chapter. **An interval estimate describes a range of values within which a population parameter is likely to lie.**

Suppose the marketing research director needs an estimate of the average life in months of car batteries his company manufactures. We select a random sample of 200 batteries, record the car owners' names and addresses as listed in store records, and interview these owners about the battery life they have experienced. Our sample of

Start with the point estimate

200 users has a mean battery life of 36 months. If we use the point estimate of the sample mean \bar{x} as the best estimator of the population mean μ , we would report that the mean life of the company's batteries is 36 months.

But the director also asks for a statement about the uncertainty that will be likely to accompany this estimate, that is, a statement about the range within which the unknown population mean is likely to lie. To provide such a statement, we need to find *the standard error of the mean*.

We learned from Chapter 6 that if we select and plot a large number of sample means from a population, the distribution of these means will approximate a normal curve. Furthermore, the mean of the sample means will be the same as the population mean. Our sample size of 200 is large enough that we can apply the central limit theorem, as we have done graphically in Figure 7-1. To measure the spread, or dispersion, in our distribution of sample means, we can use the following formula* and calculate the standard error of the mean:

$$\text{Standard error of the mean for an infinite population} \rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \leftarrow \text{Standard deviation of the population} \quad [6-1]$$

Suppose we have already estimated the standard deviation of the population of the batteries and reported that it is 10 months. Using this standard deviation and the first equation from Chapter 6, we can calculate the standard error of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [6-1]$$

$$\begin{aligned} &= \frac{10}{\sqrt{200}} \\ &= \frac{10}{14.14} \end{aligned}$$

$$= 0.707 \text{ month} \leftarrow \text{One standard error of the mean}$$

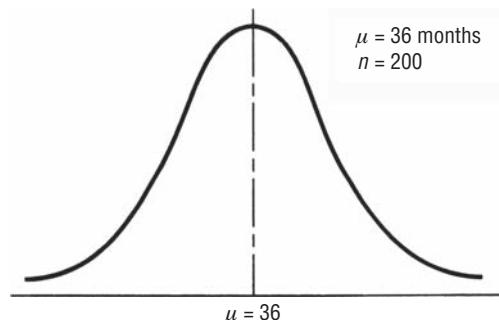


FIGURE 7-1 SAMPLING DISTRIBUTION OF THE MEAN FOR SAMPLES OF 200 BATTERIES

We could now report to the director that our estimate of the life of the company's batteries is 36 months, and the standard error that accompanies this estimate is 0.707. In other words, the actual mean life for all the batteries *may* lie somewhere in the interval estimate of 35.293 to 36.707 months. This is helpful but insufficient information for the director. Next, we need to calculate the chance that the actual life will lie in this interval *or* in other intervals of different widths that we might choose, $\pm 2\sigma$ (2×0.707), $\pm 3\sigma$ (3×0.707), and so on.

Making an interval estimate

* We have not used the finite population multiplier to calculate the standard error of the mean because the population of batteries is large enough to be considered infinite.

Probability of the True Population Parameter Falling within the Interval Estimate

To begin to solve this problem, we should review relevant parts of Chapter 5. There we worked with the normal probability distribution and learned that specific portions of the area under the normal curve are located between plus and minus any given number of standard deviations from the mean. In Figure 5-12, we saw how to relate these portions to specific probabilities.

Fortunately, we can apply these properties to the standard error of the mean and make the following statement about the range of values used to make an interval estimate for our battery problem.

Finding the chance the mean will fall in this interval estimate

The probability is 0.955 that the mean of a sample size of 200 will be within ± 2 standard errors of the population mean. Stated differently, 95.5 percent of all the sample means are within ± 2 standard errors from μ , and hence μ is **within ± 2 standard errors of 95.5 percent of all the sample means**. Theoretically, if we select 1,000 samples at random from a given population and then construct an interval of ± 2 standard errors around the mean of each of these samples, about 955 of these intervals will include the population mean. Similarly, the probability is 0.683 that the mean of the sample will be within ± 1 standard error of the population mean, and so forth. This theoretical concept is basic to our study of interval construction and statistical inference. In Figure 7-2, we have illustrated the concept graphically, showing five such intervals. Only the interval constructed around the sample mean \bar{x}_4 does not contain the population mean. In words, statisticians would describe the interval estimates represented in Figure 7-2 by saying, “The population mean μ will be located within ± 2 standard errors from the sample mean 95.5 percent of the time.”

As far as any particular interval in Figure 7-2 is concerned, it either contains the population mean or it does not, because the population mean is a fixed parameter. Because we know that in 95.5 percent of all samples, the interval will contain the population mean, we say that we are 95.5 percent confident that the interval contains the population mean.

Applying this to the battery example, we can now report to the director. Our best estimate of the life of the company’s batteries is 36 months, and we are 68.3 percent confident that the life lies in the interval from 35.293 to 36.707 months ($36 \pm 1\sigma_{\bar{x}}$). Similarly, we are 95.5 percent confident that the life falls within the interval of 34.586 to 37.414 months ($36 \pm 2\sigma_{\bar{x}}$), and we are 99.7 percent confident that battery life falls within the interval of 33.879 to 38.121 months ($36 \pm 3\sigma_{\bar{x}}$).

A more useful estimate of battery life

HINTS & ASSUMPTIONS

Every time you make an estimate there is an implied error in it. For people to understand this error, it’s common practice to describe it with a statement like “Our best estimate of the life of this set of tires is 40,000 miles and we are 90 percent sure that the life will be between 35,000 and 45,000 miles.” But if your boss demanded to know the precise average life of a set of tires, and if she were not into sampling, you’d have to watch hundreds of thousands of sets of tires being worn out and then calculate how long they lasted on average. Warning: Even then you’d be sampling because it’s impossible to watch and measure every set of tires that’s being used. It’s a lot less expensive and a lot faster to use sampling to find the answer. And if you understand estimates, you can tell your boss what risks she is taking in using a sample to estimate real tire life.

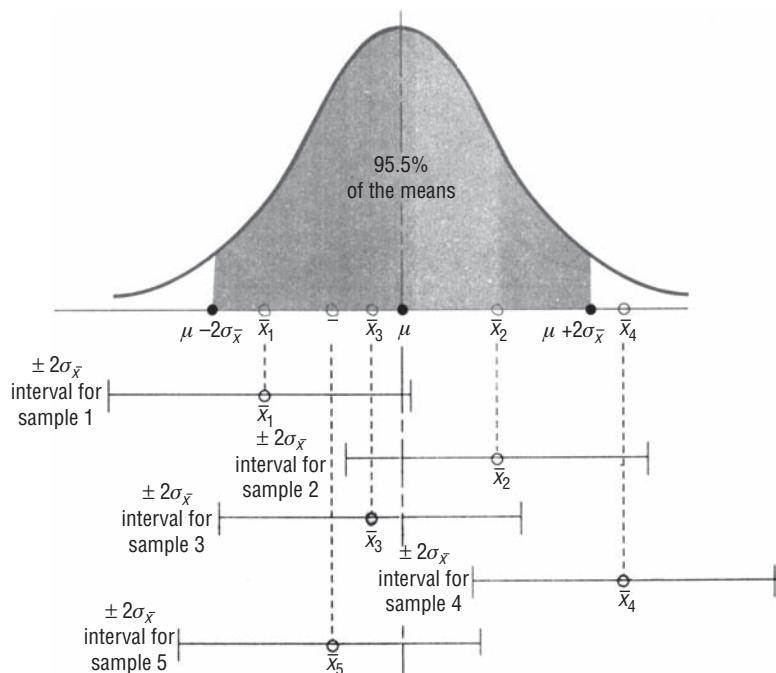


FIGURE 7-2 A NUMBER OF INTERVALS CONSTRUCTED AROUND SAMPLE MEANS; ALL EXCEPT ONE INCLUDE THE POPULATION MEAN

EXERCISES 7.3

Self-Check Exercises

- SC 7-3** For a population with a known variance of 185, a sample of 64 individuals leads to 217 as an estimate of the mean.
- Find the standard error of the mean.
 - Establish an interval estimate that should include the population mean 68.3 percent of the time.
- SC 7-4** Eunice Gunterwal is a frugal undergraduate at State U. Who is interested in purchasing a used car. She randomly selected 125 want ads and found that the average price of a car in this sample was \$3,250. Eunice knows that the standard deviation of used-car prices in this city is \$615.
- Establish an interval estimate for the average price of a car so that Eunice can be 68.3 percent certain that the population mean lies within this interval.
 - Establish an interval estimate for the average price of a car so that Miss Gunterwal can be 95.5 percent certain that the population mean lies within this interval.

Basic Concepts

- 7-12** From a population known to have a standard deviation of 1.4, a sample of 60 individuals is taken. The mean for this sample is found to be 6.2.

- (a) Find the standard error of the mean.
 (b) Establish an interval estimate around the sample mean, using one standard error of the mean.
- 7-13** From a population with known standard deviation of 1.65, a sample of 32 items resulted in 34.8 as an estimate of the mean.
 (a) Find the standard error of the mean.
 (b) Compute an interval estimate that should include the population mean 99.7 percent of the time.

Applications

- 7-14** The University of North Carolina is conducting a study on the average weight of the many bricks that make up the University's walkways. Workers are sent to dig up and weigh a sample of 421 bricks and the average brick weight of this sample was 14.2 lb. It is a well-known fact that the standard deviation of brick weight is 0.8 lb.
 (a) Find the standard error of the mean.
 (b) What is the interval around the sample mean that will include the population mean 95.5 percent of the time?
- 7-15** Because the owner of the Bard's Nook, a recently opened restaurant, has had difficulty estimating the quantity of food to be prepared each evening, he decided to determine the mean number of customers served each night. He selected a sample of 30 nights, which resulted in a mean of 71. The population standard deviation has been established as 3.76.
 (a) Give an interval estimate that has a 68.3 percent probability of including the population mean.
 (b) Give an interval estimate that has a 99.7 percent chance of including the population mean.
- 7-16** The manager of the Neuse River Bridge is concerned about the number of cars "running" the toll gates and is considering altering the toll-collection procedure if such alteration would be cost-effective. She randomly sampled 75 hours to determine the rate of violation. The resulting average violations per hour was 7. If the population standard deviation is known to be 0.9, estimate an interval that has a 95.5 percent chance of containing the true mean.
- 7-17** Gwen Taylor, apartment manager for WillowWood Apartments, wants to inform potential renters about how much electricity they can expect to use during August. She randomly selects 61 residents and discovers their average electricity usage in August to be 894 kilowatt hours (kwh). Gwen believes the variance in usage is about 131 (kwh)².
 (a) Establish an interval estimate for the average August electricity usage so Gwen can be 68.3 percent certain the true population mean lies within this interval.
 (b) Repeat part (a) with a 99.7 percent certainty.
 (c) If the price per kwh is \$0.12, within what interval can Gwen be 68.3 percent certain that the average August cost for electricity will lie?
- 7-18** The school board of Forsight County considers its most important task to be keeping the average class size in Forsight County schools less than the average class size in neighboring Hindsight County. Miss Dee Marks, the school superintendent for Forsight County, has just received reliable information indicating that the average class size in Hindsight County this year is 30.3 students. She does not yet have the figures for all 621 classes in her own school system, so Dee is forced to rely upon the 76 classes that have reported class sizes, yielding an average class size of 29.8 students. Dee knows that the class size of Forsight County

classes has a distribution with an unknown mean and standard deviation equal to 8.3 students. Assuming that the sample of 76 that Miss Marks possesses is randomly chosen from the population of all Forsight County class sizes:

- (a) Find an interval that Dee can be 95.5 percent certain will contain the true mean.
- (b) Do you think that Dee has met her goal?

Worked-Out Answers to Self-Check Exercises

SC 7-3 $\sigma^2 = 185 \quad \sigma = \sqrt{185} = 13.60 \quad n = 64 \quad \bar{x} = 217$

- (a) $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 13.60 / \sqrt{64} = 1.70$
- (b) $\bar{x} \pm \sigma_{\bar{x}} = 217 \pm 1.70 = (215.3, 218.7)$

SC 7-4 $\sigma = 615 \quad n = 125 \quad \bar{x} = 3,250 \quad \sigma_{\bar{x}} = \sigma / \sqrt{n} = 615 / \sqrt{125} = 55.01$

- (a) $\bar{x} \pm \sigma_{\bar{x}} = 3,250 \pm 55.01 = (\$3,194.99, \$3,305.01)$
- (b) $\bar{x} + 2\sigma_{\bar{x}} = 3,250 + 2(55.01) = 3,250 + 110.02 = (\$3,139.98, \$3,360.02)$

7.4 INTERVAL ESTIMATES AND CONFIDENCE INTERVALS

In using interval estimates, we are not confined to ± 1 , 2 , and 3 standard errors. According to Appendix Table 1, for example, ± 1.64 standard errors includes about 90 percent of the area under the curve; it includes 0.4495 of the area on either side of the mean in a normal distribution. Similarly, ± 2.58 standard errors includes about 99 percent of the area, or 49.51 percent on each side of the mean.

In statistics, the probability that we associate with an interval estimate is called the confidence level. This probability indicates how confident we are that the interval estimate will include the population parameter. A higher probability means more confidence. In estimation, the most commonly used confidence levels are 90 percent, 95 percent, and 99 percent, but we are free to apply *any* confidence level. In Figure 7-2, for example, we used a 95.5 percent confidence level.

The confidence interval is the range of the estimate we are making. If we report that we are 90 percent confident that the mean of the population of incomes of people in a certain community will lie between \$8,000 and \$24,000, then the range \$8,000–\$24,000 is our confidence interval. Often, however, we will express the confidence interval in standard errors rather than in numerical values. Thus, we will often express confidence intervals like this: $\bar{x} \pm 1.64 \sigma_{\bar{x}}$, where

$$\begin{aligned}\bar{x} + 1.64\sigma_{\bar{x}} &= \text{upper limit of the confidence interval} \\ \bar{x} - 1.64\sigma_{\bar{x}} &= \text{lower limit of the confidence interval}\end{aligned}$$

Thus, *confidence limits* are the upper and lower limits of the confidence interval. In this case, $\bar{x} + 1.64\sigma_{\bar{x}}$ is called the *upper confidence limit (UCL)* and $\bar{x} - 1.64\sigma_{\bar{x}}$ is the *lower confidence limit (LCL)*.

Relationship between Confidence Level and Confidence Interval

You may think that we should use a high confidence level, such as 99 percent, in all estimation problems. After all, a high confidence level seems to signify a high degree of accuracy in the estimate. In

TABLE 7-4 ILLUSTRATION OF THE RELATIONSHIP BETWEEN CONFIDENCE LEVEL AND CONFIDENCE INTERVAL

Customer's Question	Store Manager's Response	Implied Confidence Level	Implied Confidence Interval
Will I get my washing machine within 1 year?	I am absolutely certain of that.	Better than 99%	1 year
Will you deliver the washing machine within 1 month?	I am almost positive it will be delivered this month.	At least 95%	1 month
Will you deliver the washing machine within 1 week?	I am pretty certain it will go out within this week.	About 80%	1 week
Will I get my washing machine tomorrow?	I am not certain we can get it to you then.	About 40%	1 day
Will my washing machine get home before I do?	There is little chance it will beat you home.	Near 1%	1 hour

practice, however, high confidence levels will produce large confidence intervals, and such large intervals are not precise; they give very fuzzy estimates.

Consider an appliance store customer who inquires about the delivery of a new washing machine. In Table 7-4 are several of the questions the customer might ask and the likely responses. This table indicates the direct relationship that exists between the confidence level and the confidence interval for any estimate. As the customer sets a tighter and tighter confidence interval, the store manager agrees to a lower and lower confidence level. Notice, too, that when the confidence interval is too wide, as is the case with a one-year delivery, the estimate may have very little real value, even though the store manager attaches a 99 percent confidence level to that estimate. Similarly, if the confidence interval is too narrow ("Will my washing machine get home before I do?"), the estimate is associated with such a low confidence level (1 percent) that we question its value.

Using Sampling and Confidence Interval Estimation

In our discussion of the basic concepts of interval estimation, particularly in Figure 7-2, we described samples being drawn repeatedly from a given population in order to estimate a population parameter. We also mentioned selecting a large number of sample means from a population. In practice, however, it is often difficult or expensive to take more than one sample from a population. Based on just one sample, we estimate the population parameter. We must be careful, then, about interpreting the results of such a process.

Suppose we calculate from one sample in our battery example the following confidence interval and confidence level: "We are 95 percent confident that the mean battery life of the population lies within 30 and 42 months." **This statement does not mean that the chance is 0.95 that the mean life of all our batteries falls within the interval established from this one sample. Instead, it means that if we select many random samples of the same size and calculate a confidence interval for each of these samples, then in about 95 percent of these cases, the population mean will lie within that interval.**

Estimating from only one sample

HINTS & ASSUMPTIONS

Warning: There is no free lunch in dealing with confidence levels and confidence intervals. When you want more of one, you have to take less of the other. Hint: To understand this important relationship, go back to Table 7-4. If you want the estimate of the time of delivery to be *perfectly* accurate (100 percent), you have to sacrifice tightness in the confidence interval and accept a very wide delivery promise ("sometime this year"). On the other hand, if you aren't concerned with the accuracy of the estimate, you could get a delivery person to say "I'm 1 percent sure I can get it to you within an hour." You *can't* have both at the same time.

EXERCISES 7.4

Self-Check Exercise

- SC 7-5** Given the following confidence levels, express the lower and upper limits of the confidence interval for these levels in terms of \bar{x} and $\sigma_{\bar{x}}$.
- 54 percent.
 - 75 percent.
 - 94 percent.
 - 98 percent.

Basic Concepts

- 7-19** Define the confidence level for an interval estimate.
- 7-20** Define the confidence interval.
- 7-21** Suppose you wish to use a confidence level of 80 percent. Give the upper limit of the confidence interval in terms of the sample mean, \bar{x} , and the standard error, $\sigma_{\bar{x}}$.
- 7-22** In what way may an estimate be less meaningful because of
 - A high confidence level?
 - A narrow confidence interval?
- 7-23** Suppose a sample of 50 is taken from a population with standard deviation 27 and that the sample mean is 86.
 - Establish an interval estimate for the population mean that is 95.5 percent certain to include the true population mean.
 - Suppose, instead, that the sample size was 5,000. Establish an interval for the population mean that is 95.5 percent certain to include the true population mean.
 - Why might estimate (a) be preferred to estimate (b)? Why might (b) be preferred to (a)?
- 7-24** Is the confidence level for an estimate based on the interval constructed from a single sample?
- 7-25** Given the following confidence levels, express the lower and upper limits of the confidence interval for these levels in terms of \bar{x} and $\sigma_{\bar{x}}$.
 - 60 percent.
 - 70 percent.
 - 92 percent.
 - 96 percent.

Applications

- 7-26** Steve Klippers, owner of Steve's Barbershop, has built quite a reputation among the residents of Cullowhee. As each customer enters his barbershop, Steve yells out the number of minutes that the customer can expect to wait before getting his haircut. The only statistician in town, after being frustrated by Steve's inaccurate point estimates, has determined that the actual waiting time for any customer is normally distributed with mean equal to Steve's estimate in minutes and standard deviation equal to 5 minutes divided by the customer's position in the waiting line. Help Steve's customers develop 95 percent probability intervals for the following situations:
- The customer is second in line and Steve's estimate is 25 minutes.
 - The customer is third in line and Steve's estimate is 15 minutes.
 - The customer is fifth in line and Steve's estimate is 38 minutes.
 - The customer is first in line and Steve's estimate is 20 minutes.
 - How are these intervals different from confidence intervals?

Worked-Out Answers to Self-Check Exercise

SC 7-5 (a) $\bar{x} \pm 0.74\sigma_{\bar{x}}$. (b) $\bar{x} \pm 1.15\sigma_{\bar{x}}$. (c) $\bar{x} \pm 1.88\sigma_{\bar{x}}$. (d) $\bar{x} \pm 2.33\sigma_{\bar{x}}$.

7.5 CALCULATING INTERVAL ESTIMATES OF THE MEAN FROM LARGE SAMPLES

A large automotive-parts wholesaler needs an estimate of the mean life it can expect from windshield wiper blades under typical driving conditions. Already, management has determined that the standard deviation of the population life is 6 months. Suppose we select a simple random sample of 100 wiper blades, collect data on their useful lives, and obtain these results:

Finding a 95 percent confidence interval

$$\begin{aligned} n &= 100 \leftarrow \text{Sample size} \\ \bar{x} &= 21 \text{ months} \leftarrow \text{Sample mean} \\ \sigma &= 6 \text{ months} \leftarrow \text{Population standard deviation} \end{aligned}$$

Because the wholesaler uses tens of thousands of these wiper blades annually, it requests that we find an interval estimate with a confidence level of 95 percent. The sample size is greater than 30, so the central limit theorem allows us to use the normal distribution as our sampling distribution even if the population isn't normal. We calculate the standard error of the mean by using Equation 6-1:

Population standard deviation is known

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} & [6-1] \\ &= \frac{6 \text{ months}}{\sqrt{100}} \\ &= \frac{6}{10} \\ &= 0.6 \text{ month} \leftarrow \text{Standard error of the mean for an infinite population} \end{aligned}$$

Next, we consider the confidence level with which we are working. Because a 95 percent confidence level will include 47.5 percent of the area on either side of the mean of the sampling distribution, we can search in the body of Appendix Table 1 for the 0.475 value. We discover that 0.475 of the area under the normal curve is contained between the mean and a point 1.96 standard errors to the right of the mean. Therefore, we know that $(2)(0.475) = 0.95$ of the area is located between plus and minus 1.96 standard errors from the mean and that our confidence limits are

$$\bar{x} + 1.96 \sigma_{\bar{x}} \leftarrow \text{Upper confidence limit}$$

$$\bar{x} - 1.96 \sigma_{\bar{x}} \leftarrow \text{Lower confidence limit}$$

Then we substitute numerical values into these two expressions::

$$\begin{aligned}\bar{x} + 1.96 \sigma_{\bar{x}} &= 21 \text{ months} + 1.96(0.6 \text{ month}) \\ &= 21 + 1.18 \text{ months} \\ &= 22.18 \text{ months} \leftarrow \text{Upper confidence limit}\end{aligned}$$

$$\begin{aligned}\bar{x} - 1.96 \sigma_{\bar{x}} &= 21 \text{ months} - 1.96(0.6 \text{ month}) \\ &= 21 - 1.18 \text{ months} \\ &= 19.82 \text{ months} \leftarrow \text{Lower confidence limit}\end{aligned}$$

We can now report that we estimate the mean life of the population of wiper blades to be between 19.82 and 22.18 months with 95 percent confidence.

Our conclusion

When the Population Standard Deviation Is Unknown

A more complex interval estimate problem comes from a social-service agency in a local government. It is interested in estimating the mean annual income of 700 families living in a four-square-block section of a community. We take a simple random sample and find these results:

Finding a 90 percent confidence interval

$$\begin{aligned}n &= 50 \leftarrow \text{Sample size} \\ \bar{x} &= \$11,800 \leftarrow \text{Sample mean} \\ s &= \$950 \leftarrow \text{Sample standard deviation}\end{aligned}$$

The agency asks us to calculate an interval estimate of the mean annual income of all 700 families so that it can be 90 percent confident that the population mean falls within that interval. The sample size is over 30, so once again the central limit theorem enables us to use the normal distribution as the sampling distribution.

Notice that one part of this problem differs from our previous examples: we do *not* know the population standard deviation, and so we will use the sample standard deviation to estimate the *population standard deviation*:

Estimating the population standard deviation

Estimate of the Population Standard Deviation

Estimate of the population standard deviation

$$\hat{\sigma} = s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

[7-1]

The value \$950.00 is our estimate of the standard deviation of the population. We can also symbolize this *estimated* value by $\hat{\sigma}$, which is called *sigma hat*.

Now we can estimate the standard error of the mean. Because we have a finite population size of 700, and because our sample is more than 5 percent of the population, we will use the formula for deriving the standard error of the mean of finite populations:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad [6-3]$$

But because we are calculating the standard error of the mean using an *estimate* of the standard deviation of the population, we must rewrite this equation so that it is correct symbolically:

Estimating the standard error of the mean

Estimated Standard Error of the Mean of a Finite Population

Symbol that indicates an estimated value

Estimate of the population standard deviation

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad [7-2]$$

Continuing our example, we find $\hat{\sigma}_{\bar{x}} = \frac{\$950.00}{\sqrt{50}} \times \sqrt{\frac{700-50}{700-1}}$

$$= \frac{\$950.00}{7.07} \sqrt{\frac{650}{699}}$$

$$= (\$134.37)(0.9643)$$

$$= \$129.57 \leftarrow \text{Estimate of the standard error of the mean of a finite population (derived from an } \hat{\sigma} \text{ of the population standard deviation)}$$

Next we consider the 90 percent confidence level, which would include 45 percent of the area on either side of the mean of the sampling distribution. Looking in the body of Appendix Table 1 for the 0.45 value, we find that about 0.45 of the area under the normal curve is located between the mean and a point 1.64 standard errors away from the mean. Therefore, 90 percent of the area is located between plus and minus 1.64 standard errors away from the mean, and our confidence limits are

$$\begin{aligned} \bar{x} + 1.64 \hat{\sigma}_{\bar{x}} &= \$11,800 + 1.64(\$129.57) \\ &= \$11,800 + \$212.50 \\ &= \$12,012.50 \leftarrow \text{Upper confidence limit} \\ \bar{x} - 1.64 \hat{\sigma}_{\bar{x}} &= \$11,800 - 1.64(\$129.57) \\ &= \$11,800 - \$212.50 \\ &= \$11,587.50 \leftarrow \text{Lower confidence limit} \end{aligned}$$

Our report to the social-service agency would be: With 90 percent confidence, we estimate that the average annual income of all 700 families living in this four-square-block section falls between \$11,587.50 and \$12,012.50.

Our conclusion

HINTS & ASSUMPTIONS

Hint: It's easy to understand how to approach these exercises if you'll go back to Figure 7-2 on page 339 for a minute. When someone states a confidence level, they are referring to the shaded area in the figure, which is defined by how many $\sigma_{\bar{x}}$ (standard errors or standard deviations of the distribution of sample means) there are on either side of the mean. Appendix Table 1 quickly converts any desired confidence level into standard errors. Because we have the information necessary to calculate *one* standard error, we can calculate the endpoints of the shaded area. These are the limits of our confidence interval. Warning: When you don't know the dispersion in the population (the population standard deviation) remember to use Equation 7-1 to estimate it.

EXERCISES 7.5

Self-Check Exercises

- SC 7-6** From a population of 540, a sample of 60 individuals is taken. From this sample, the mean is found to be 6.2 and the standard deviation 1.368.
- Find the estimated standard error of the mean.
 - Construct a 96 percent confidence interval for the mean.
- SC 7-7** In an automotive safety test conducted by the North Carolina Highway Safety Research Center, the average tire pressure in a sample of 62 tires was found to be 24 pounds per square inch, and the standard deviation was 2.1 pounds per square inch.
- What is the estimated population standard deviation for this population? (There are about a million cars registered in North Carolina.)
 - Calculate the estimated standard error of the mean.
 - Construct a 95 percent confidence interval for the population mean.

Basic Concepts

- 7-27** The manager of Cardinal Electric's lightbulb division must estimate the average number of hours that a lightbulb made by each lightbulb machine will last. A sample of 40 lightbulbs was selected from machine A and the average burning time was 1,416 hours. The standard deviation of burning time is known to be 30 hours.
- Compute the standard error of the mean.
 - Construct a 90 percent confidence interval for the true population mean.
- 7-28** Upon collecting a sample of 250 from a population with known standard deviation of 13.7, the mean is found to be 112.4.
- Find a 95 percent confidence interval for the mean.
 - Find a 99 percent confidence interval for the mean.

Applications

- 7-29** The Westview High School nurse is interested in knowing the average height of seniors at this school, but she does not have enough time to examine the records of all 430 seniors. She randomly selects 48 students. She finds the sample mean to be 64.5 inches and the standard deviation to be 2.3 inches.
- (a) Find the estimated standard error of the mean.
 - (b) Construct a 90 percent confidence interval for the mean.
- 7-30** Jon Jackobsen, an overzealous graduate student, has just completed a first draft of his 700-page dissertation. Jon has typed his paper himself and is interested in knowing the average number of typographical errors per page, but does not want to read the whole paper. Knowing a little bit about business statistics, Jon selected 40 pages at random to read and found that the average number of typos per page was 4.3 and the sample standard deviation was 1.2 typos per page.
- (a) Calculate the estimated standard error of the mean.
 - (b) Construct for Jon a 90 percent confidence interval for the true average number of typos per page in his paper.
- 7-31** The Nebraska Cable Television authority conducted a test to determine the amount of time people spend watching television per week. The NCTA surveyed 84 subscribers and found the average number of hours watched per week to be 11.6 hours and the standard deviation to be 1.8 hours.
- (a) What is the estimated population standard deviation for this population? (There are about 95,000 people with cable television in Nebraska.)
 - (b) Calculate the estimated standard error of the mean.
 - (c) Construct a 98 percent confidence interval for the population mean.
- 7-32** Joel Friedlander is a broker on the New York Stock Exchange who is curious about the amount of time between the placement and execution of a market order. Joel sampled 45 orders and found that the mean time to execution was 24.3 minutes and the standard deviation was 3.2 minutes. Help Joel by constructing a 95 percent confidence interval for the mean time to execution.
- 7-33** Oscar T. Grady is the production manager for Citrus Groves Inc., located just north of Ocala, Florida. Oscar is concerned that the last 3 years' late freezes have damaged the 2,500 orange trees that Citrus Groves owns. In order to determine the extent of damage to the trees, Oscar has sampled the number of oranges produced per tree for 42 trees and found that the average production was 525 oranges per tree and the standard deviation was 30 oranges per tree.
- (a) Estimate the population standard deviation from the sample standard deviation.
 - (b) Estimate the standard error of the mean for this finite population.
 - (c) Construct a 98 percent confidence interval for the mean per-tree output of all 2,500 trees.
 - (d) If the mean orange output per tree was 600 oranges 5 years ago, what can Oscar say about the possible existence of damage now?
- 7-34** Chief of Police Kathy Ackert has recently instituted a crackdown on drug dealers in her city. Since the crackdown began, 750 of the 12,368 drug dealers in the city have been caught. The mean dollar value of drugs found on these 750 dealers is \$250,000. The standard deviation of the dollar value of drugs for these 750 dealers is \$41,000. Construct for Chief Ackert a 90 percent confidence interval for the mean dollar value of drugs possessed by the city's drug dealers.

Worked-Out Answers to Self-Check Exercises

SC 7-6 $\hat{\sigma} = 1.368 \quad N = 540 \quad n = 60 \quad \bar{x} = 6.2$

$$(a) \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{1.368}{\sqrt{60}} \times \sqrt{\frac{540-60}{540-1}} = 0.167$$

$$(b) \bar{x} \pm 2.05 \hat{\sigma}_{\bar{x}} = 6.2 \pm 2.05(0.167) = 6.2 \pm 0.342 = (5.86, 6.54)$$

SC 7-7 $s = 2.1 \quad n = 62 \quad \bar{x} = 24$

$$(a) \hat{\sigma} = s = 2.1 \text{ psi}$$

$$(b) \hat{\sigma}_{\bar{x}} = \hat{\sigma}/\sqrt{n} = 2.1/\sqrt{62} = 0.267 \text{ psi}$$

$$(c) \bar{x} \pm 1.96 \hat{\sigma}_{\bar{x}} = 24 \pm 1.96(0.267) = 24 \pm 0.523 = (23.48, 24.52) \text{ psi}$$

7.6 CALCULATING INTERVAL ESTIMATES OF THE PROPORTION FROM LARGE SAMPLES

Statisticians often use a sample to estimate a *proportion* of occurrences in a population. For example, the government estimates by a sampling procedure the unemployment rate, or the proportion of unemployed people, in the U.S. workforce.

Review of the binomial distribution

In Chapter 5, we introduced the binomial distribution, a distribution of discrete, not continuous, data. Also, we presented the two formulas for deriving the mean and the standard deviation of the binomial distribution:

$$\mu = np \quad [5-2]$$

$$\sigma = \sqrt{npq} \quad [5-3]$$

where

- n = number of trials
- p = probability of success
- $q = 1 - p$ = probability of a failure

Theoretically, the binomial distribution is the correct distribution to use in constructing confidence intervals to estimate a population proportion.

Shortcomings of the binomial distribution

Because the computation of binomial probabilities is so tedious (recall that the probability of r successes in n trials is $[n!/(r!(n-r)!)][p^r q^{n-r}]$), using the binomial distribution to form interval estimates of a population proportion is a complex proposition. Fortunately, as the sample size increases, the binomial can be approximated by an appropriate normal distribution, which we can use to approximate the sampling distribution. Statisticians recommend that in estimation, n be large enough for both np and nq to be at least 5 when you use the normal distribution as a substitute for the binomial.

Finding the mean of the sample proportion

Symbolically, let's express the proportion of successes in a sample by \bar{p} (pronounced *p bar*). Then modify Equation 5-2, so

that we can use it to derive the *mean of the sampling distribution of the proportion of successes*. In words, $\mu = np$ shows that the mean of the binomial distribution is equal to the product of the number of trials, n , and the probability of success, p ; that is, np equals the mean number of successes. To change this *number* of successes to the *proportion* of successes, we divide np by n and get p alone. The mean in the left-hand side of the equation becomes $\mu_{\bar{p}}$, or the mean of the sampling distribution of the proportion of successes.

Mean of the Sampling Distribution of the Proportion

$$\mu_{\bar{p}} = p$$

[7-3]

Similarly, we can modify the formula for the standard deviation of the binomial distribution, \sqrt{npq} , which measures the standard deviation in the number of successes. To change the number of successes to the proportion of successes, we divide \sqrt{npq} by n and get $\sqrt{pq/n}$. In statistical terms, the standard deviation for the proportion of successes in a sample is symbolized and is called the *standard error of the proportion*.

Finding the standard deviation of the sample proportion

Standard Error of the Proportion

$$\text{Standard error of the proportion} \longrightarrow \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} \quad [7-4]$$

We can illustrate how to use these formulas if we estimate for a very large organization what proportion of the employees prefer to provide their own retirement benefits in lieu of a company-sponsored plan. First, we conduct a simple random sample of 75 employees and find that 0.4 of them are interested in providing their own retirement plans. Our results are

$$n = 75 \leftarrow \text{Sample size}$$

$$\bar{p} = 0.4 \leftarrow \text{Sample proportion in favor}$$

$$\bar{q} = 0.6 \leftarrow \text{Sample proportion not in favor}$$

Next, management requests that we use this sample to find an interval about which they can be 99 percent confident that it contains the true population proportion.

Estimating a population proportion

But what are p and q for the *population*? We can estimate the population parameters by substituting the corresponding sample statistics \bar{p} and \bar{q} (*p bar* and *q bar*) in the formula for the standard error of the proportion.* Doing this, we get:

*Notice that we do not use the finite population multiplier, because our population is so large compared with the sample size.

Estimated Standard Error of the Proportion

Symbol indicating that the standard error of the proportion estimated

Sample statistics

$$\begin{aligned}\hat{\sigma}_{\bar{p}} &= \sqrt{\frac{\bar{p}\bar{q}}{n}} \\ &= \sqrt{\frac{(0.4)(0.6)}{75}} \\ &= \sqrt{0.0032} \\ &= 0.057 \leftarrow \text{Estimated standard error of the proportion}\end{aligned}$$

[7-5]

Now we can provide the estimate management needs by using the same procedure we have used previously. A 99 percent confidence level would include 49.5 percent of the area on either side of the mean in the sampling distribution. The body of Appendix Table 1 tells us that 0.495 of the area under the normal curve is located between the mean and a point 2.58 standard errors from the mean. Thus, 99 percent of the area is contained between plus *and* minus 2.58 standard errors from the mean. Our confidence limits then become

$$\begin{aligned}\bar{p} + 2.58 \hat{\sigma}_{\bar{p}} &= 0.4 + 2.58(0.057) \\ &= 0.4 + 0.147 \\ &= 0.547 \leftarrow \text{Upper confidence limit} \\ \bar{p} - 2.58 \hat{\sigma}_{\bar{p}} &= 0.4 - 2.58(0.057) \\ &= 0.4 - 0.147 \\ &= 0.253 \leftarrow \text{Lower confidence limit}\end{aligned}$$

Computing the confidence limits

Thus, we estimate from our sample of 75 employees that with 99 percent confidence we believe that the proportion of the total population of employees who wish to establish their own retirement plans lies between 0.253 and 0.547.

Our conclusion

HINTS & ASSUMPTIONS

The same assumptions, hints, and warnings we stated on page 346 apply here as well. The only difference is that now, since we're dealing with a proportion, the binomial distribution is the correct sampling distribution to use. Hint: Remember from Chapter 5 that as long as n is large enough to make both np and nq at least 5, we can use the normal distribution to approximate the binomial. If that is the case, we proceed exactly as we did with interval estimates of the mean. Warning: Since the exact standard error of the proportion depends on the unknown population proportion (p), remember to estimate p by \bar{p} and use \bar{p} in Equation 7-5 to estimate the standard error of the proportion.

EXERCISES 7.6

Self-Check Exercises

- SC 7-8** When a sample of 70 retail executives was surveyed regarding the poor November performance of the retail industry, 66 percent believed that decreased sales were due to unseasonably warm temperatures, resulting in consumers' delaying purchase of cold-weather items.
- Estimate the standard error of the proportion of retail executives who blame warm weather for low sales.
 - Find the upper and lower confidence limits for this proportion, given a 95 percent confidence level.
- SC 7-9** Dr. Benjamin Shockley, a noted social psychologist, surveyed 150 top executives and found that 42 percent of them were unable to add fractions correctly.
- Estimate the standard error of the proportion.
 - Construct a 99 percent confidence interval for the true proportion of top executives who cannot correctly add fractions.

Applications

- 7-35** Pascal, Inc., a computer store that buys wholesale, untested computer chips, is considering switching to another supplier who would provide tested and guaranteed chips for a higher price. In order to determine whether this is a cost-effective plan, Pascal must determine the proportion of faulty chips that the current supplier provides. A sample of 200 chips was tested and of these, 5 percent were found to be defective.
- Estimate the standard error of the proportion of defective chips.
 - Construct a 98 percent confidence interval for the proportion of defective chips supplied.
- 7-36** General Cinema sampled 55 people who viewed *GhostHunter 8* and asked them whether they planned to see it again. Only 10 of them believed the film was worthy of a second look.
- Estimate the standard error of the proportion of moviegoers who will view the film a second time.
 - Construct a 90 percent confidence interval for this proportion.
- 7-37** The product manager for the new lemon-lime Clear 'n Light dessert topping was worried about both the product's poor performance and her future with Clear 'n Light. Concerned that her marketing strategy had not properly identified the attributes of the product, she sampled 1,500 consumers and learned that 956 thought that the product was a floor wax.
- Estimate the standard error of the proportion of people holding this severe misconception about the dessert topping.
 - Construct a 96 percent confidence interval for the true population proportion.
- 7-38** Michael Gordon, a professional basketball player, shot 200 foul shots and made 174 of them.
- Estimate the standard error of the proportion of all foul shots Michael makes.
 - Construct a 98 percent confidence interval for the proportion of all foul shots Michael makes.
- 7-39** SnackMore recently surveyed 95 shoppers and found 80 percent of them purchase SnackMore fat-free brownies monthly.
- Estimate the standard error of the proportion.
 - Construct a 95 percent confidence interval for the true proportion of people who purchase the brownies monthly.

- 7-40** The owner of the Home Loan Company randomly surveyed 150 of the company's 3,000 accounts and determined that 60 percent were in excellent standing.
- Find a 95 percent confidence interval for the proportion in excellent standing.
 - Based on part (a), what kind of interval estimate might you give for the absolute number of accounts that meet the requirement of excellence, keeping the same 95 percent confidence level?
- 7-41** For a year and a half, sales have been falling consistently in all 1,500 franchises of a fast-food chain. A consulting firm has determined that 31 percent of a sample of 95 indicate clear signs of mismanagement. Construct a 98 percent confidence interval for this proportion.
- 7-42** Student government at the local university sampled 45 textbooks at the University Student Store and determined that of these 45 textbooks, 60 percent had been marked up in price more than 50 percent over wholesale cost. Give a 96 percent confidence interval for the proportion of books marked up more than 50 percent by the University Student Store.
- 7-43** Barry Turnbull, the noted Wall Street analyst, is interested in knowing the proportion of individual stockholders who plan to sell at least one-quarter of all their stock in the next month. Barry has conducted a random survey of 800 individuals who hold stock and has learned that 25 percent of his sample plan to sell at least one-quarter of all their stock in the next month. Barry is about to issue his much-anticipated monthly report, "The Wall Street Pulse—the Tape's Ticker," and would like to be able to report a confidence interval to his subscribers. He is more worried about being correct than he is about the width of the interval. Construct a 90 percent confidence interval for the true proportion of individual stockholders who plan to sell at least one-quarter of their stock during the next month.

Worked-Out Answers to Self-Check Exercises

SC 7-8 $n = 70 \quad \bar{p} = 0.66$

$$(a) \hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}\bar{q}}{n}} = \sqrt{\frac{0.66(0.34)}{70}} = 0.0566$$

$$(b) \bar{p} \pm 1.96 \hat{\sigma}_{\bar{x}} = 0.66 \pm 1.96(0.0566) = 0.66 \pm 0.111 = (0.549, 0.771)$$

SC 7-9 $n = 150 \quad \bar{p} = 0.42$

$$(a) \hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}\bar{q}}{n}} = \sqrt{\frac{0.42(0.58)}{150}} = 0.0403$$

$$(b) \bar{p} \pm 2.58 \hat{\sigma}_{\bar{p}} = 0.42 \pm 2.58(0.0403) = 0.42 \pm 0.104 = (0.316, 0.524)$$

7.7 INTERVAL ESTIMATES USING THE t DISTRIBUTION

In our three examples so far, the sample sizes were all larger than 30. We sampled 100 wind-shield wiper blades, 50 families living in a four-square-block section of a community, and 75 employees of a very large organization. Each time, the normal distribution was the appropriate sampling distribution to use to determine confidence intervals.

However, this is not always the case. How can we handle estimates where the normal distribution is *not* the appropriate sampling distribution, that is, when we are estimating the population standard deviation and the sample size is 30 or less? For example, in our chapter-opening problem of coal usage, we had data from only 10 weeks. Fortunately, another distribution exists that is appropriate in these cases. It is called the *t distribution*.

Early theoretical work on *t* distributions was done by a man named W. S. Gosset in the early 1900s. Gosset was employed by the Guinness Brewery in Dublin, Ireland, which did not permit employees to publish research findings under their own names. So Gosset adopted the pen name *Student* and published under that name. Consequently, the *t* distribution is commonly called *Student's t distribution*, or simply *Student's distribution*.

Because it is used when the sample size is 30 or less, statisticians often associate the *t* distribution with small sample statistics. This is misleading, because the size of the sample is only *one* of the conditions that lead us to use the *t* distribution. The second condition is that the population standard deviation must be unknown. **Use of the *t* distribution for estimating is required whenever the sample size is 30 or less and the population standard deviation is not known.** Furthermore, in using the *t* distribution, we assume that the population is normal or approximately normal.

Characteristics of the *t* Distribution

Without deriving the *t* distribution mathematically, we can gain an intuitive understanding of the relationship between the *t* distribution and the *normal distribution*. Both are symmetrical. In general, the *t* distribution is flatter than the normal distribution, for every possible sample size. Even so, as the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t*.

Figure 7-3 compares one normal distribution with two *t* distributions of different sample sizes. This figure shows two characteristics of *t* distributions. **A *t* distribution is lower at the mean and higher at**

Sometimes the normal distribution is not appropriate

Background of the *t* distribution

Conditions for using the *t* distribution

***t* distribution compared to normal distribution**

and there is a different *t* distribution for every sample size, and there is a different *t* distribution for every sample size. Even so, as the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t*.

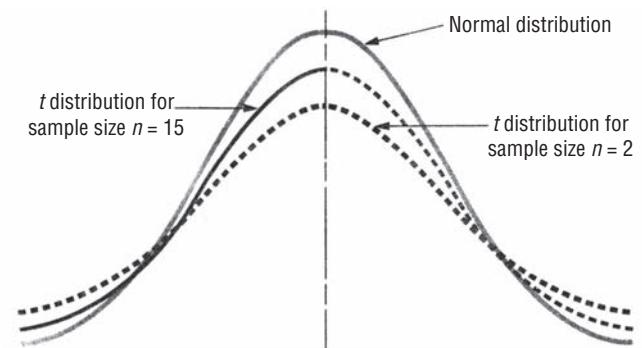


FIGURE 7-3 NORMAL DISTRIBUTION, *t* DISTRIBUTION FOR SAMPLE SIZE $n = 15$, AND *t* DISTRIBUTION FOR SAMPLE SIZE $n = 2$

the tails than a normal distribution. The figure also demonstrates how the t distribution has proportionally more of its area in its tails than the normal does. This is the reason why it will be necessary to go farther out from the mean of a t distribution to include the same area under the curve. Interval widths from t distributions are, therefore, wider than those based on the normal distribution.

Degrees of Freedom

We said earlier that there is a separate t distribution for each sample size. In proper statistical language, we would say, “There is a different t distribution for each of the possible degrees of freedom.” What are degrees of freedom? We can define them as the number of values we can choose freely.

Degrees of freedom defined

Assume that we are dealing with two sample values, a and b , and we know that they have a mean of 18. Symbolically, the situation is

$$\frac{a+b}{2} = 18$$

How can we find what values a and b can take on in this situation? The answer is that a and b can be any two values whose sum is 36, because $36 \div 2 = 18$.

Suppose we learn that a has a value of 10. Now b is no longer free to take on any value but *must* have the value of 26, because

$$\begin{aligned} &\text{if} & a &= 10 \\ &\text{then} & \frac{10+b}{2} &= 18 \\ &\text{so} & 10+b &= 36 \\ &\text{therefore} & b &= 26 \end{aligned}$$

This example shows that when there are two elements in a sample and we know the sample mean of these two elements, we are free to specify only one of the elements because the other element will be determined by the fact that the two elements sum to twice the sample mean. Statisticians say, “We have one degree of freedom.”

Look at another example. There are seven elements in our sample, and we learn that the mean of these elements is 16. Symbolically, we have this situation:

$$\frac{a+b+c+d+e+f+g}{7} = 16$$

Another example

In this case, the degrees of freedom, or the number of variables we can specify freely, are $7 - 1 = 6$. We are free to give values to six variables, and then we are no longer free to specify the seventh variable. It is determined automatically.

With two sample values, we had one degree of freedom ($2 - 1 = 1$), and with seven sample values, we had six degrees of freedom ($7 - 1 = 6$). In each of these two examples, then, we had $n - 1$ degrees of freedom, assuming n is the sample size. Similarly, a sample of 23 would give us 22 degrees of freedom.

We will use degrees of freedom when we select a t distribution to estimate a population mean, and we will use $n - 1$ degrees of

Function of degrees of freedom

freedom, where n is the sample size. For example, if we use a sample of 20 to estimate a population mean, we will use 19 degrees of freedom in order to select the appropriate t distribution.

Using the t Distribution Table

The table of t distribution values (Appendix Table 2) differs in construction from the z table we have used previously. **The t table is more compact and shows areas and t values for only a few percentages (10, 5, 2, and 1 percent).** Because there is a different t distribution for each number of degrees of freedom, a more complete table would be quite lengthy. Although we can conceive of the need for a more complete table, in fact Appendix Table 2 contains all the commonly used values of the t distribution.

***t table compared to z table:
three differences***

A second difference in the t table is that it does *not* focus on the chance that the population parameter being estimated will fall *within* our confidence interval. Instead, it measures the chance that the population parameter we are estimating will *not* be within our confidence interval (that is, that it will lie *outside* it). If we are making an estimate at the 90 percent confidence level, we would look in the t table under the 0.10 column (100 percent – 90 percent = 10 percent). This 0.10 chance of error is symbolized by α , which is the Greek letter *alpha*. We would find the appropriate t values for confidence intervals of 95 percent, 98 percent, and 99 percent under the α columns headed 0.05, 0.02, and 0.01, respectively.

A third difference in using the t table is that we must specify the degrees of freedom with which we are dealing. Suppose we make an estimate at the 90 percent confidence level with a sample size of 14, which is 13 degrees of freedom. Look in Appendix Table 2 under the 0.10 column until you encounter the row labeled 13. Like a z value, the t value there of 1.771 shows that if we mark off plus and minus 1.771 $\hat{\sigma}_{\bar{x}}$'s (estimated standard errors of \bar{x}) on either side of the mean, the area under the curve between these two limits will be 90 percent, and the area outside these limits (the chance of error) will be 10 percent (see Figure 7-4).

Recall that in our chapter-opening problem, the generating plant manager wanted to estimate the coal needed for this year, and he took a sample by measuring coal usage for 10 weeks. The sample data are

$$\begin{aligned} n &= 10 \text{ weeks} \leftarrow \text{Sample size} \\ df &= 9 \leftarrow \text{Degrees of freedom} \end{aligned}$$

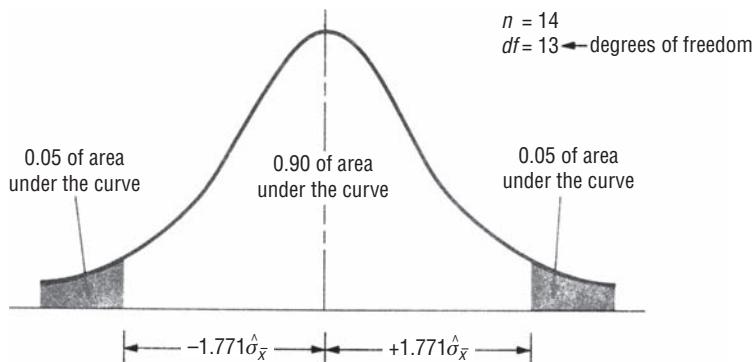


FIGURE 7-4 A t DISTRIBUTION FOR 13 DEGREES OF FREEDOM, SHOWING A 90 PERCENT CONFIDENCE INTERVAL

$$\bar{x} = 11,400 \text{ tons} \leftarrow \text{Sample mean}$$

$$s = 700 \text{ tons} \text{ Sample standard deviation}$$

The plant manager wants an interval estimate of the mean coal consumption, and he wants to be 95 percent confident that the mean consumption falls within that interval. **This problem requires the use of a *t* distribution because the sample size is less than 30, the population standard deviation is unknown, and the manager believes that the population is approximately normal.**

Using the *t* table to compute confidence limits

As a first step in solving this problem, recall that we *estimate* the population standard deviation with the sample standard deviation; thus

$$\hat{\sigma} = s$$

$$= 700 \text{ tons} \quad [7-1]$$

Using this estimate of the population standard deviation, we can estimate the standard error of the mean by modifying Equation 7-2 to omit the finite population multiplier (because the sample size of 10 weeks is less than 5 percent of the 5 years (260 weeks) for which data are available):

Estimated Standard Error of the Mean of an Infinite Population

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \quad [7-6]$$

Continuing our example, we find $\hat{\sigma}_{\bar{x}} = \frac{700}{\sqrt{10}}$

$$= \frac{700}{3.162}$$

= 221.38 tons ← Estimated standard error of the mean of an infinite population

Now we look in Appendix Table 2 down the 0.05 column (100 percent – 95 percent = 5 percent) until we encounter the row for 9 degrees of freedom ($10 - 1 = 9$). There we see the *t* value 2.262 and can set our confidence limits accordingly:

$$\bar{x} + 2.262 \hat{\sigma}_{\bar{x}} = 11,400 \text{ tons} + 2.262(221.38 \text{ tons})$$

$$= 11,400 + 500.76$$

$$= 11,901 \text{ tons} \leftarrow \text{Upper confidence limit}$$

$$\bar{x} - 2.262 \hat{\sigma}_{\bar{x}} = 11,400 \text{ tons} - 2.262(221.38 \text{ tons})$$

$$= 11,400 - 500.76$$

$$= 10,899 \text{ tons} \leftarrow \text{Lower confidence limit}$$

Our confidence interval is illustrated in Figure 7-5. Now we can report to the plant manager with 95 percent confidence that the

Our conclusion

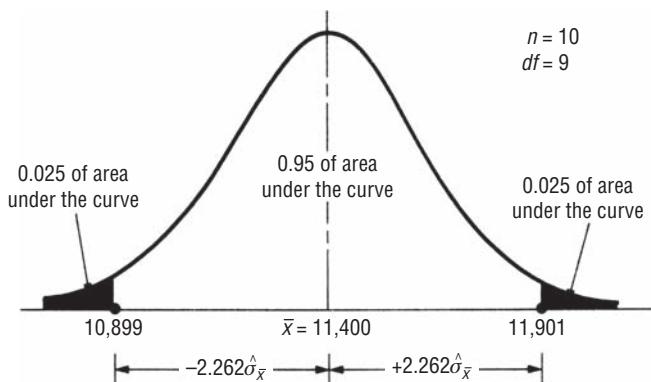


FIGURE 7-5 COAL PROBLEM: A t DISTRIBUTION WITH 9 DEGREES OF FREEDOM AND A 95 PERCENT CONFIDENCE INTERVAL

mean weekly usage of coal lies between 10,899 and 11,901 tons, and he can use the 11,901-ton figure to estimate how much coal to order.

The only difference between the process we used to make this coal-usage estimate and the previous estimating problems is the use of the t distribution as the appropriate distribution. **Remember that in any estimation problem in which the sample size is 30 or less and the standard deviation of the population is unknown and the underlying population can be assumed to be normal or approximately normal, we use the t distribution.**

Summary of Confidence Limits under Various Conditions

Table 7-5 summarizes the various approaches to estimation introduced in this chapter and the confidence limits appropriate for each.

TABLE 7-5 SUMMARY OF FORMULAS FOR CONFIDENCE LIMITS ESTIMATING MEAN AND PROPORTION

	When the Population Is Finite (and $n/N > 0.05$)	When the Population Is Infinite (or $n/N < 0.05$)
Estimating μ (the population mean): When σ (the population standard deviation) is known	$\left\{ \begin{array}{l} \text{Upper limit } \bar{x} + z \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \\ \text{Lower limit } \bar{x} - z \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \end{array} \right.$	$\bar{x} + z \frac{\sigma}{\sqrt{n}}$ $\bar{x} - z \frac{\sigma}{\sqrt{n}}$
When σ (the population standard deviation) is not known ($\hat{\sigma} = s$) When n (the sample size) is larger than 30	$\left\{ \begin{array}{l} \text{Upper limit } \bar{x} + z \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \\ \text{Lower limit } \bar{x} - z \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \end{array} \right.$	$\bar{x} + z \frac{\hat{\sigma}}{\sqrt{n}}$ $\bar{x} - z \frac{\hat{\sigma}}{\sqrt{n}}$

TABLE 7-5 SUMMARY OF FORMULAS FOR CONFIDENCE LIMITS ESTIMATING MEAN AND PROPORTION (Contd.)

When n (the sample size) is 30 or less and the population is normal or approximately normal*

This case is beyond the scope of the text; consult a professional statistician.

$$\bar{x} + t \frac{\hat{\sigma}}{\sqrt{n}}$$

$$\bar{x} - t \frac{\hat{\sigma}}{\sqrt{n}}$$

Estimating p (the population proportion):

When n (the sample size) is larger than 30

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}\bar{q}}{n}}$$

This case is beyond the scope of the text; consult a professional statistician.

$$\bar{p} + z \hat{\sigma}_{\bar{p}}$$

$$\bar{p} - z \hat{\sigma}_{\bar{p}}$$

*Remember that the appropriate t distribution to use is the one with $n - 1$ degrees of freedom.

Interval Estimates using MS Excel

MS-Excel can be used to construct confidence interval for mean, when sample elements are given. For this purpose, go to **Data > Data Analysis > Descriptive Statistics**

Shop No.	2010_Sales (Thousands of Rs.)	2011_Sales (Thousands of Rs.)	Region	Type of Shop
1	59	64	3	1
2	61	66	3	2
3	68	85	2	2
4	69	77	2	1
5	77	86	2	2
6	62	66	3	2
7	62	72	2	2
8	73	74	1	1
9	54	66	1	1
10	74	63	1	1
11	75	69	2	2
12	68	85	2	2
13	67	62	1	1
14	58	80	2	1
15	63	70	2	1
16	72	69	2	2
17	66	63	2	1
18	63	57	1	1
19	54	73	4	2
20	66	65	2	2
21	58	60	1	1
22	78	73	1	1
23	65	62	3	2
24	67	72	4	2
25	75	64	1	2
26	56	80	3	2
27	62	71	1	1
28	63	52	3	1
29	60	71	4	1
30				

When **Descriptive Statistics** dialogue box opens, enter sample-data range in **Input: Input Range**, check **Label in first row**, **Summary Statistics** and **Confidence Interval for Mean** buttons. Level of confidence can be changed from **95%** if situation demands. Press **OK**.

The screenshot shows a Microsoft Excel window with the title "Case_Normal-Data.xls [Compatibility Mode] - Microsoft Excel". The ribbon tabs include Home, Insert, Page Layout, Formulas, Data, Review, View, Add-Ins, and Acrobat. The Data tab is selected, and the ribbon bar has a "Data Analysis" icon. A "Descriptive Statistics" dialog box is open, overlaid on the main worksheet. The dialog box has the following settings:

- Input Range:** \$C\$1:\$G\$30
- Grouped By:** Columns (radio button selected)
- Labels in first row:** checked
- Output options:**
 - Output Range:** \$H\$1 (\$H\$1 is highlighted)
 - New Worksheet (By):** (radio button)
 - New Workbook:** (radio button)
- Summary statistics:** checked
- Confidence Level for Mean:** 95 %
- Kth Largest:** 1
- Kth Smallest:** 1

The main worksheet contains data for 30 rows, columns C through G. Column C is labeled "Shop No.", column D is "2010_Sales (Thousands of Rs.)", column E is "Region", and column F is "Type of Shop". Below this, two small tables are shown: one for "Region" (Northern, Eastern, Western, Southern) and one for "Type" (Urban Shops, Rural Shops).

The result table will be displayed as under

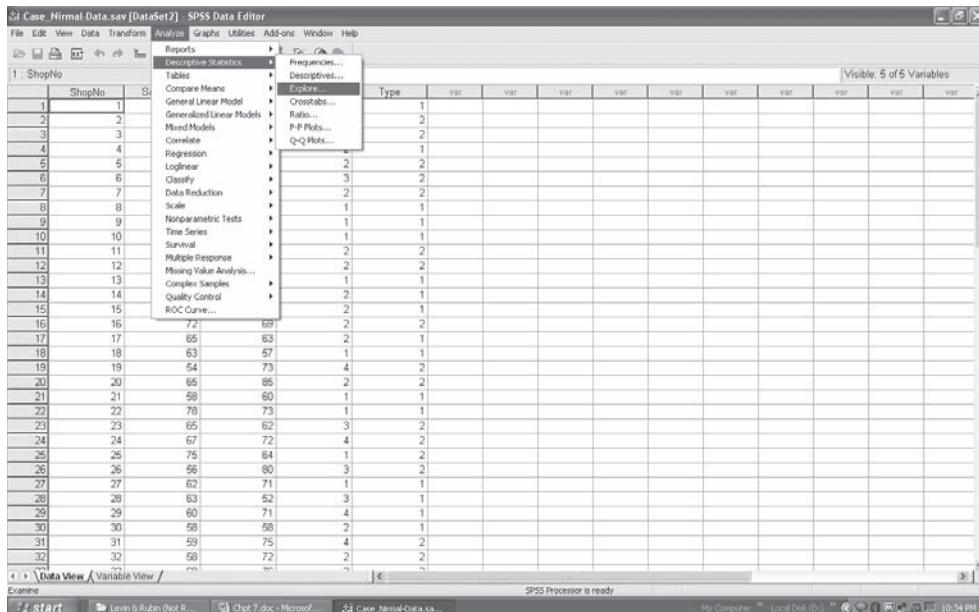
The screenshot shows a Microsoft Excel window with the title "Case_Normal-Data.xls [Compatibility Mode] - Microsoft Excel". The ribbon tabs are identical to the previous screenshot. A new sheet titled "Sheet4" is active, displaying the results of the descriptive statistics analysis. The data is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	2010_Sales (Thousands of Rs.)																
2	Mean	64.08333333															
3	Standard Error	0.872781017															
4	Median	63															
5	Mode	62															
6	Standard Deviation	6.780532691															
7	Sample Variance	46.10480276															
8	Kurtosis	0.041232057															
9	Skewness	0.527208634															
10	Range	32															
11	Minimum	51															
12	Maximum	83															
13	Sum	3845															
14	Count	60															
15	Confidence Level(95.0%)	1.746430767															
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	
30																	
31																	
32																	

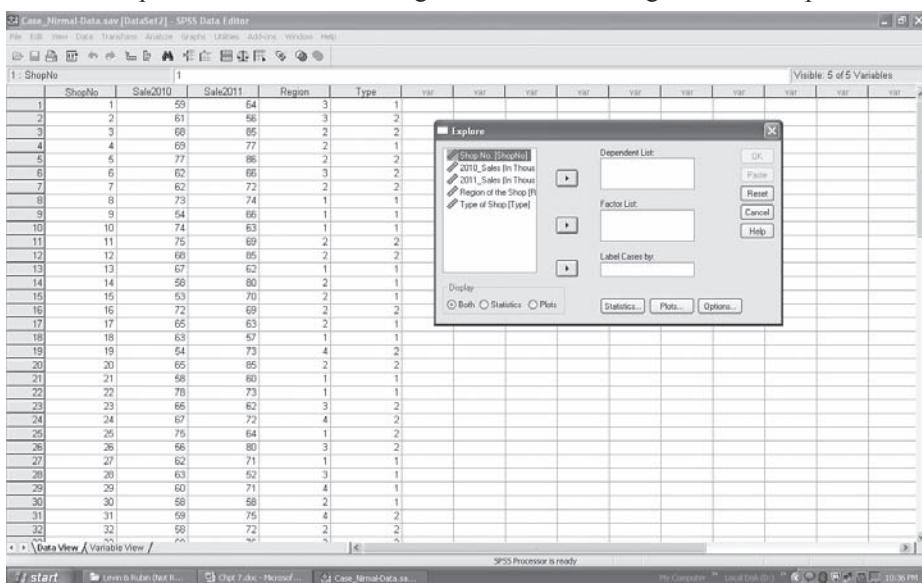
The status bar at the bottom shows "start" and "Levin & Rubin" along with other standard status information.

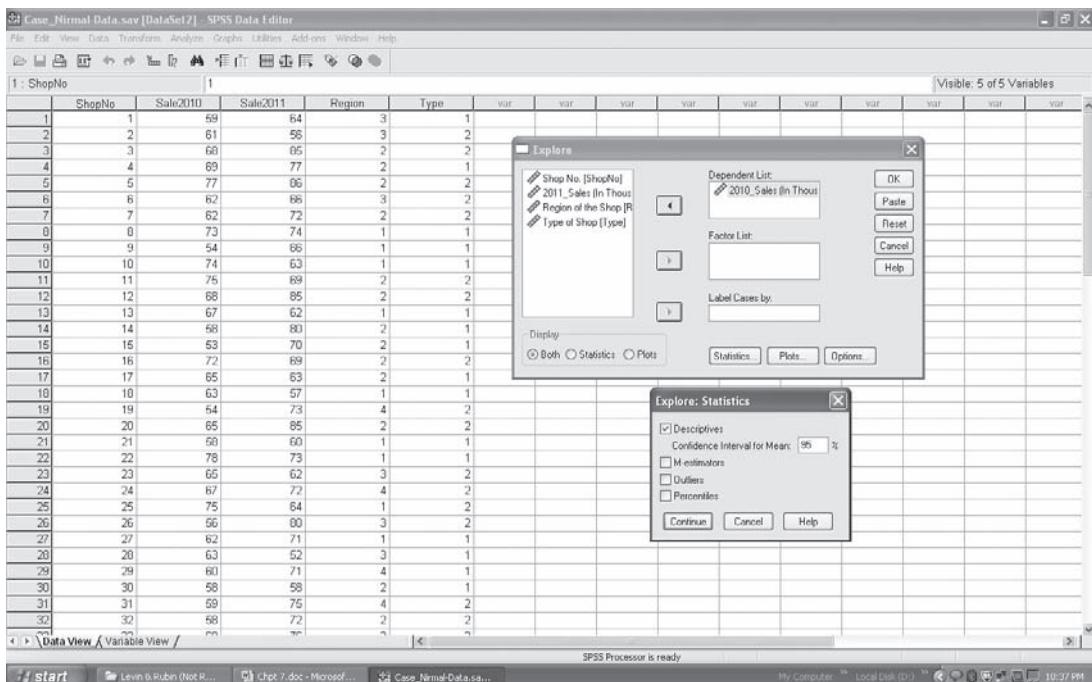
Interval Estimates using SPSS

For obtaining confidence interval for mean from a sample-data series, go to **Analyze > Descriptive Statistics > Explore.**



When **Explore** dialogue box opens, enter variable containing sample-data series in **Dependent List** drop box. Then press **Statistics** tab, the **Explore: Statistics** sub-dialogue box will be opened. Check **Descriptive Statistics** box. The confidence level can be changed from its default value of **95%**, if situation demands Then press **Continue** button to go back to main dialogue box. Then press **OK**.





HINTS & ASSUMPTIONS

The concept of *degrees of freedom* is often difficult to grasp at first. Hint: Think of it as the number of choices you have. If you have peanut butter and cheese in your refrigerator, you can choose either a peanut butter or a cheese sandwich (unless you like peanut butter and cheese sandwiches). If you open the door and the cheese is all gone, Mr. Gosset would probably say, "You now have zero degrees of freedom." That is, if you want lunch, you have no choices left; it's peanut butter or starve. Warning: Although the *t* distribution is associated with small-sample statistics, remember that a sample size of less than 30 is only *one* of the conditions for its use. The others are that the population standard deviation is not known and the population is normally or approximately normally distributed.

EXERCISES 7.7

Self-Check Exercises

- SC 7-10** For the following sample sizes and confidence levels, find the appropriate *t* values for constructing confidence intervals:
- $n = 28$; 95 percent.
 - $n = 8$; 98 percent.
 - $n = 13$; 90 percent.

- (d) $n = 10$; 95 percent.
- (e) $n = 25$; 99 percent.
- (f) $n = 10$; 99 percent.

SC 7-11 Seven homemakers were randomly sampled, and it was determined that the distances they walked in their housework had an average of 39.2 miles per week and a sample standard deviation of 3.2 miles per week. Construct a 95 percent confidence interval for the population mean.

Basic Concepts

- 7-44** For the following sample sizes and confidence levels, find the appropriate t values for constructing confidence intervals:
- (a) $n = 15$; 90 percent.
 - (b) $n = 6$; 95 percent.
 - (c) $n = 19$; 99 percent.
 - (d) $n = 25$; 98 percent.
 - (e) $n = 10$; 99 percent.
 - (f) $n = 41$; 90 percent.
- 7-45** Given the following sample sizes and t values used to construct confidence intervals, find the corresponding confidence levels:
- (a) $n = 27$; $t = \pm 2.056$.
 - (b) $n = 5$; $t = \pm 2.132$.
 - (c) $n = 18$; $t = \pm 2.898$.
- 7-46** A sample of 12 had a mean of 62 and a standard deviation of 10. Construct a 95 percent confidence interval for the population mean.
- 7-47** The following sample of eight observations is from an infinite population with a normal distribution:
- | | | | | | | | |
|------|------|------|------|------|------|------|------|
| 75.3 | 76.4 | 83.2 | 91.0 | 80.1 | 77.5 | 84.8 | 81.0 |
|------|------|------|------|------|------|------|------|
- (a) Find the sample mean.
 - (b) Estimate the population standard deviation.
 - (c) Construct a 98 percent confidence interval for the population mean.

Applications

- 7-48** Northern Orange County has found, much to the dismay of the county commissioners, that the population has a severe problem with dental plaque. Every year the local dental board examines a sample of patients and rates each patient's plaque buildup on a scale from 1 to 100, with 1 representing no plaque and 100 representing a great deal of plaque. This year, the board examined 21 patients and found that they had an average Plaque Rating Score (PRS) of 72 and a standard deviation of 6.2. Construct for Orange County a 98 percent confidence interval for the mean PRS for Northern Orange County.
- 7-49** Twelve bank tellers were randomly sampled and it was determined they made an average of 3.6 errors per day with a sample standard deviation of 0.42 error. Construct a 90 percent confidence interval for the population mean of errors per day. What assumption is implied about the number of errors bank tellers make?

- 7-50** State Senator Hanna Rowe has ordered an investigation of the large number of boating accidents that have occurred in the state in recent summers. Acting on her instructions, her aide, Geoff Spencer, has randomly selected 9 summer months within the last few years and has compiled data on the number of boating accidents that occurred during each of these months. The mean number of boating accidents to occur in these 9 months was 31, and the standard deviation in this sample was 9 boating accidents per month. Geoff was told to construct a 90 percent confidence interval for the true mean number of boating accidents per month, but he was in such an accident himself recently, so you will have to do this for him.

Worked-Out Answers to Self-Check Exercises

SC 7-10 (a) 2.052.

(b) 2.998.

(c) 1.782.

(d) 2.262.

(e) 2.797.

(f) 3.250.

$$\text{SC 7-11 } s = 3.2 \quad n = 7 \quad \bar{x} = 39.2 \quad \hat{\sigma}_{\bar{x}} = s/\sqrt{n} = 3.2/\sqrt{7} = 1.2095$$

$$\bar{x} \pm t\hat{\sigma}_{\bar{x}} = 39.2 \pm 2.447(1.2095) = 39.2 \pm 2.9596$$

$$= (36.240, 42.160) \text{ miles}$$

7.8 DETERMINING THE SAMPLE SIZE IN ESTIMATION

In all our discussions so far, we have used for sample size the symbol n instead of a specific number. Now we need to know how to determine what number to use. How large should the sample be? If it is too small, we may fail to achieve the objective of our analysis. But if it is too large, we waste resources when we gather the sample.

Some sampling error will arise because we have not studied the whole population. Whenever we sample, we always miss *some* helpful information about the population. If we want a high level of precision (that is, if we want to be quite sure of our estimate), we have to sample enough of the population to provide the required information. Sampling error is controlled by selecting a sample that is adequate in size. In general, the more precision you want, the larger the sample you will need to take. Let us examine some methods that are useful in determining what sample size is necessary for any specified level of precision.

What sample size is adequate?

Sample Size for Estimating a Mean

Suppose a university is performing a survey of the annual earnings of last year's graduates from its business school. It knows from past experience that the standard deviation of the annual earnings of the entire population (1,000) of these graduates is about \$1,500. How large a sample size should the university take in order to estimate the mean annual earnings of last year's class within \$500 and at a 95 percent confidence level?

Exactly what is this problem asking? The university is going to take a sample of *some* size, determine the mean of the sample, \bar{x} , and use it as a point estimate of the population mean. It wants

Two ways to express a confidence limit

TABLE 7-6 COMPARISON OF TWO WAYS OF EXPRESSING THE SAME CONFIDENCE LIMITS

Lower Confidence Limit	Upper Confidence Limit
a. $\bar{x} - \$500$	a. $\bar{x} + \$500$
b. $\bar{x} - \sigma_x^-$	b. $\bar{x} + \sigma_x^-$

to be 95 percent certain that the true mean annual earnings of last year's class is not more than \$500 above or below the point estimate. Row *a* in Table 7-6 summarizes in symbolic terms how the university is defining its confidence limits for us. Row *b* shows symbolically how we normally express confidence limits for an infinite population. When we compare these two sets of confidence limits, we can see that

$$z\sigma_x^- = \$500$$

Thus, the university is actually saying that it wants $z\sigma_x^-$ to be equal to \$500. If we look in Appendix Table 1, we find that the necessary *z* value for a 95 percent confidence level is 1.96. Step by step:

$$\begin{aligned} \text{If } z\sigma_x^- &= \$500 \\ \text{and } z &= 1.96 \\ \text{then } 1.96\sigma_x^- &= \$500 \\ \text{and } \sigma_x^- &= \frac{\$500}{1.96} \\ &= \$255 \leftarrow \text{Standard error of the mean} \end{aligned}$$

Remember that the formula for the standard error is Equation 6-1:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \leftarrow \text{Population standard deviation} \quad [6-1]$$

Using Equation 6-1, we can substitute our known population standard deviation value of \$1,500 and our calculated standard error value of \$255 and solve for *n*:

Finding an adequate sample size

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} && [6-1] \\ \$255 &= \frac{\$1,500}{\sqrt{n}} \\ (\sqrt{n})(\$255) &= \$1,500 \\ \sqrt{n} &= \frac{\$1,500}{\$255} \\ \sqrt{n} &= 5.882; \text{ now square both sides} \\ n &= 34.6 \leftarrow \text{Sample size for precision specified} \end{aligned}$$

Therefore, because *n* must be greater than or equal to 34.6, the university should take a sample of 35 business-school graduates to get the precision it wants in estimating the class's mean annual earnings.

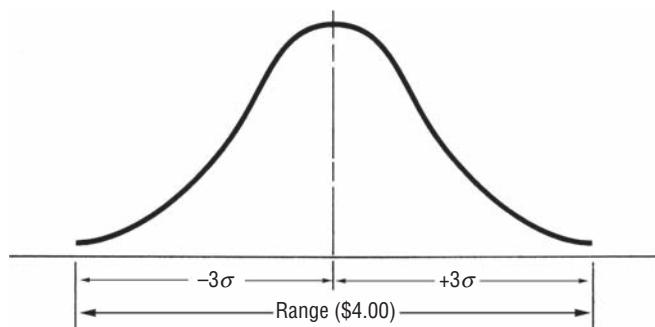


FIGURE 7-6 APPROXIMATE RELATIONSHIP BETWEEN THE RANGE AND THE POPULATION STANDARD DEVIATION

In this example, we knew the standard deviation of the population, but in many cases, the standard deviation of the population is not available. Remember, too, that we have not yet taken the sample, and we are trying to decide how large to make it. We cannot estimate the population standard deviation using methods from the first part of this chapter. If we have a notion about the range of the population, we can use that to get a crude but workable estimate.

Estimating the standard deviation from the range

Suppose we are estimating hourly manufacturing wage rates in a city and are fairly confident that there is a \$4.00 difference between the highest and lowest wage rates. We know that plus and minus 3 standard deviations include 99.7 percent of all the area under the normal curve, that is, plus 3 standard deviations and minus 3 standard deviations include almost all of the distribution. To symbolize this relationship, we have constructed Figure 7-6, in which \$4.00 (the range) equals 6 standard deviations (plus 3 and minus 3). Thus, a rough estimate of the population standard deviation would be

$$6 \hat{\sigma} = \$4.00$$

$$\hat{\sigma} = \frac{\$4.00}{6}$$

Estimate of the population standard deviation $\rightarrow \hat{\sigma} = \0.667

Our estimate of the population standard deviation using this rough method is not precise, but it may mean the difference between getting a working idea of the required sample size and knowing nothing about that sample size.

Sample Size for Estimating a Proportion

The procedures for determining sample sizes for estimating a population proportion are similar to those for estimating a population mean. Suppose we wish to poll students at a large state university. We want to determine what proportion of them is in favor of a new grading system. We would like a sample size that will enable us to be 90 percent certain of estimating the true proportion of the population of 40,000 students that is in favor of the new system within plus and minus 0.02.

We begin to solve this problem by looking in Appendix Table 1 to find the z value for a 90 percent confidence level. That value is ± 1.64 standard errors from the mean. We want our estimate to be within 0.02, so we can symbolize the step-by-step process like this:

If $z\sigma_{\bar{p}} = 0.02$
 and $z = 1.64$
 then $1.64\sigma_{\bar{p}} = 0.02$

If we now substitute the right side of Equation 7-4 for $\sigma_{\bar{p}}$, we get

$$1.64 \sqrt{\frac{pq}{n}} = 0.02$$

$$\sqrt{\frac{pq}{n}} = 0.0122; \text{ now square both sides}$$

$$\frac{pq}{n} = 0.00014884; \text{ now multiply both sides by } n$$

$$pq = 0.00014884n$$

$$n = \frac{pq}{0.00014884}$$

To find n , we still need an estimate of the population parameters p and q . If we have strong feelings about the actual proportion in favor of the new system, we can use that as our best guess to calculate n . But if we have no idea what p is, then our best strategy is to guess at p in such a way that we choose n in a conservative manner (that is, so that the sample size is large enough to supply at least the precision we require no matter what p actually is). At this point in our problem, n is equal to the product of p and q divided by 0.00014884. The way to get the largest n is to generate the largest possible numerator of that expression, which happens if we pick $p = 0.5$ and $q = 0.5$. Then n becomes:

$$n = \frac{pq}{0.00014884}$$

$$= \frac{(0.5)(0.5)}{0.00014884}$$

$$= \frac{0.25}{0.00014884}$$

$$= 1,680 \leftarrow \text{Sample size for precision specified}$$

As a result, to be 90 percent certain of estimating the true proportion within 0.02, we should pick a simple random sample of 1,680 students to interview.

In the problem we have just solved, we picked a value for p that represented the most conservative strategy. The value 0.5 generated the largest possible sample. We would have used another value of p if we had been able to estimate one or if we had a strong feeling about one. Whenever all these solutions are absent, assume the most conservative possible value for p , namely, $p = 0.5$.

To illustrate that 0.5 yields the largest possible sample, Table 7-7 solves the grading-system problem using several different values of p . You can see from the sample sizes associated with these different values that for the range of p 's from 0.3 to 0.7, the change in the appropriate sample size is relatively small. Therefore, even if you knew that the true population proportion was 0.3 and you used a value

Picking the most conservative proportion

TABLE 7-7 SAMPLES SIZE n ASSOCIATED WITH DIFFERENT VALUES OF p AND q

Choose This Value for p	Value of q , or $1 - p$	$\left(\frac{pq}{0.00014884} \right)$	Indicated Sample Size n
0.2	0.8	$\frac{(0.2)(0.8)}{(0.00014884)}$	= 1,075
0.3	0.7	$\frac{(0.3)(0.7)}{(0.00014884)}$	= 1,411
0.4	0.6	$\frac{(0.4)(0.6)}{(0.00014884)}$	= 1,613
0.5	0.5	$\frac{(0.5)(0.5)}{(0.00014884)}$	= 1,680 ← Most conservative
0.6	0.4	$\frac{(0.6)(0.4)}{(0.00014884)}$	= 1,613
0.7	0.3	$\frac{(0.7)(0.3)}{(0.00014884)}$	= 1,411
0.8	0.2	$\frac{(0.8)(0.2)}{(0.00014884)}$	= 1,075

of 0.5 for p anyway, you would have sampled only 269 more people ($1,680 - 1,411$) than was actually necessary for the desired degree of precision. Obviously, guessing values of p in cases like this is not so critical as it seemed at first glance.

HINTS & ASSUMPTIONS

From a commonsense perspective, if the standard deviation of the population is very small, the values cluster very tightly around their mean and just about any sample size will capture them and produce accurate information. On the other hand, if the population standard deviation is very large and the values are quite spread out, it will take a very large sample to include them and turn up accurate information. How do we get an idea about the population standard deviation before we start sampling? Companies planning to conduct market research generally conduct preliminary research on the population to estimate the standard deviation. If the product is like another that has been on the market, often it's possible to rely on previous data about the population without further estimates.

EXERCISES 7.8

Self-Check Exercises

- SC 7-12** For a test market, find the sample size needed to estimate the true proportion of consumers satisfied with a certain new product within ± 0.04 at the 90 percent confidence level. Assume you have no strong feeling about what the proportion is.

- SC 7-13** A speed-reading course guarantees a certain reading rate increase within 2 days. The teacher knows a few people will not be able to achieve this increase, so before stating the guaranteed percentage of people who achieve the reading rate increase, he wants to be 98 percent confident that the percentage has been estimated to within ± 5 percent of the true value. What is the most conservative sample size needed for this problem?

Basic Concepts

- 7-51** If the population standard deviation is 78, find the sample size necessary to estimate the true mean within 50 points for a confidence level of 95 percent.
- 7-52** We have strong indications that the proportion is around 0.7. Find the sample size needed to estimate the proportion within ± 0.02 with a confidence level of 90 percent.
- 7-53** Given a population with a standard deviation of 8.6, what size sample is needed to estimate the mean of the population within ± 0.5 with 99 percent confidence?

Applications

- 7-54** An important proposal must be voted on, and a politician wants to find the proportion of people who are in favor of the proposal. Find the sample size needed to estimate the true proportion to within $\pm .05$ at the 95 percent confidence level. Assume you have no strong feelings about what the proportion is. How would your sample size change if you believe about 75 percent of the people favor the proposal? How would it change if only about 25 percent favor the proposal?
- 7-55** The management of Southern Textiles has recently come under fire regarding the supposedly detrimental effects on health caused by its manufacturing process. A social scientist has advanced a theory that the employees who die from natural causes exhibit remarkable consistency in their life-span: The upper and lower limits of their life-spans differ by no more than 550 weeks (about 10½ years). For a confidence level of 98 percent, how large a sample should be examined to find the average life-span of these employees within ± 30 weeks?
- 7-56** Food Tiger, a local grocery store, sells generic garbage bags and has received quite a few complaints about the strength of these bags. It seems that the generic bags are weaker than the name-brand competitor's bags and, therefore, break more often. John C. Tiger, VP in charge of purchasing, is interested in determining the average maximum weight that can be put into one of the generic bags without its breaking. If the standard deviation of garbage breaking weight is 1.2 lb, determine the number of bags that must be tested in order for Mr. Tiger to be 95 percent confident that the sample average breaking weight is within 0.5 lb of the true average.
- 7-57** The university is considering raising tuition to improve school facilities, and they want to determine what percentage of students favor the increase. The university needs to be 90 percent confident the percentage has been estimated to within 2 percent of the true value. How large a sample is needed to guarantee this accuracy regardless of the true percentage?
- 7-58** A local store that specializes in candles and clocks, Wicks and Ticks, is interested in obtaining an interval estimate for the mean number of customers that enter the store daily. The owners are reasonably sure that the actual standard deviation of the daily number of customers is 15 customers. Help Wicks and Ticks out of a fix by determining the sample size it should use in order to develop a 96 percent confidence interval for the true mean that will have a width of only eight customers.

Worked-Out Answers to Self-Check Exercises

SC 7-12 Assume $p = q = 0.5$.

$$0.04 = 1.64 \sqrt{\frac{pq}{n}} = 1.64 \sqrt{\frac{0.5(0.5)}{n}} \text{ so } n = \left(\frac{1.64(0.5)}{0.04} \right)^2 = 420.25 \text{ i.e. } n \geq 421.$$

SC 7-13 Assume $p = q = 0.5$.

$$0.05 = 2.33 \sqrt{\frac{pq}{n}} = 2.33 \sqrt{\frac{0.5(0.5)}{n}} \text{ so } n = \left(\frac{2.33(0.5)}{0.05} \right)^2 = 542.89 \text{ i.e. } n \geq 543.$$

So take a sample of at least 543 records of prior students.

STATISTICS AT WORK

Loveland Computers

Case 7: Estimation Although Lee Azko had felt nervous about the first job out of college, assignments in production and purchasing had already shown how “book learning” could be applied. The next assignment introduced Lee to another of Loveland Computers’ departments and the no-nonsense approach of its head, Margot Derby.

“Let me tell you the situation,” began Margot, the head of marketing, without bothering with introductions or small talk. “You know that we primarily consider ourselves distributors of hardware—the actual PCs that people use in their homes and businesses. When we started out, we left it up to the customers to seek out software. Sometimes, they bought directly from the companies that wrote the programs, or from national distributors with toll-free numbers. Now there are also retail outlets—almost every suburban mall has at least one store that sells computer programs.

“The reason we stayed clear of software was that there were just too many programs out there—we didn’t want to guess which one would be the ‘hit’ product and end up with a lot of useless inventory on our hands. But the game changed. After some shakeout in software, two or three clear leaders emerged in each field—spreadsheets and word processors, for example. To match the competition we began to bundle some software with the computers for certain promotions.

“Last year, we also started loading the programs onto the hard drive for some customers. We can give them a very competitive price for the software, and preloading turns out to be an important product feature that many people are shopping for. So I’m taking another look at software, to see if we shouldn’t change our strategy and do more in that line. To get some idea of the market, I had a summer intern call up 500 customers who’d owned Loveland machines for about a year. And we asked them how much they’d spent in total on software in the first year.

“I’ve got all the data here; it didn’t take 2 minutes to come up with the mean and standard deviation from our spreadsheet program. Those investment bankers from New York took a look at a draft of my marketing plan for software; when they were down here last week, they asked me how sure I could be that the results of that telephone survey were accurate.

“Every time I pick up the newspaper, I see some opinion poll where they say: ‘This is based on a survey of 1,200 adults and the margin of error is 3 percent.’ How do they know that—do they keep track

of all the surveys and when they're right and wrong? I only have this one set of results. I don't see how I can answer their question."

"It shouldn't be too difficult," said Lee, checking a briefcase to make sure that a calculator and a set of statistical tables were close at hand. "Why don't you show me those numbers and we can figure it out right now."

Study Questions: What distribution will Lee assume for the telephone poll results, and which statistical table will be most useful? How will Lee define *margin of error* for Margot? Is Lee likely to recommend a larger sample?

CHAPTER REVIEW

Terms Introduced in Chapter 7

Confidence Interval A range of values that has some designated probability of including the true population parameter value.

Confidence Level The probability that statisticians associate with an interval estimate of a population parameter, indicating how confident they are that the interval estimate will include the population parameter.

Confidence Limits The upper and lower boundaries of a confidence interval.

Consistent Estimator An estimator that yields values more closely approaching the population parameter as the sample size increases.

Degrees of Freedom The number of values in a sample we can specify freely once we know something about that sample.

Efficient Estimator An estimator with a smaller standard error than some other estimator of the population parameter; that is, the smaller the standard error of an estimator, the more efficient that estimator is.

Estimate A specific observed value of an estimator.

Estimator A sample statistic used to estimate a population parameter.

Interval Estimate A range of values to estimate an unknown population parameter.

Point Estimate A single number used to estimate an unknown population parameter.

Student's *t* Distribution A family of probability distributions distinguished by their individual degrees of freedom, similar in form to the normal distribution, and used when the population standard deviation is unknown and the sample size is relatively small ($n \leq 30$).

Sufficient Estimator An estimator that uses all the information available in the data concerning a parameter.

Unbiased Estimator An estimator of a population parameter that, on the average, assumes values above the population parameter as often, and to the same extent, as it tends to assume values below the population parameter.

Equations Introduced in Chapter 7

7-1

$$\hat{\sigma} = s \times \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

p. 345

This formula indicates that the sample standard deviation can be used to estimate the population standard deviation.

7-2

$$\hat{\sigma} = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

p. 346

This formula enables us to derive an *estimated* standard error of the mean of a *finite* population from an *estimate* of the population standard deviation. The symbol $\hat{\cdot}$, called a hat, indicates that the value is estimated. Equation 7-6 is the corresponding formula for an infinite population.

7-3

$$\mu_{\bar{p}} = p$$

p. 350

Use this formula to derive the *mean* of the sampling distribution of the *proportion* of successes. The right-hand side, p , is equal to $(n \times p)/n$, where the numerator is the expected number of successes in n trials and the denominator is the number of trials. Symbolically, the proportion of successes *in a sample* is written \bar{p} and is pronounced *p bar*.

7-4

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}$$

p. 350

To get the *standard error of the proportion*, take the square root of the product of the probabilities of success and failure divided by the number of trials.

7-5

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}\bar{q}}{n}}$$

p. 351

This is the formula to use to derive an estimated standard error of the proportion when the population proportion is unknown and you are forced to use \bar{p} and \bar{q} , the sample proportions of successes and failures.

7-6

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

p. 357

This formula enables us to derive an *estimated* standard error of the mean of an *infinite* population from an *estimate* of the population standard deviation. It is exactly like Equation 7-2 except that it lacks the finite population multiplier.

Review and Application Exercises

- 7-59** From a sample of 42 gasoline stations statewide, the average price of a gallon of unleaded gas was found to be \$1.12 and the standard deviation was \$0.04 per gallon. Within what interval can we be 99.74 percent confident that the true statewide mean per-gallon price of unleaded gasoline will fall?
- 7-60** What are the advantages of using an interval estimate over a point estimate?
- 7-61** Why is the size of a statistic's standard error important in its use as an estimator? To which characteristic of estimator does this relate?
- 7-62** Suzanne Jones, head registrar for the university system, needs to know what proportion of students have grade-point averages below 2.0. How many students' grades should be looked at in order to determine this proportion to within ± 0.01 with 95 percent confidence?
- 7-63** A 95 percent confidence interval for the population mean is given by (94, 126) and a 75 percent confidence interval is given by (100.96, 119.04). What are the advantages and disadvantages of each of these interval estimates?
- 7-64** The posted speed limit on the Cross-Bronx Expressway is 55 mph. Congestion results in much slower actual speeds. A random sample of 57 vehicles clocked speeds with an average of 23.2 mph and a standard deviation of 0.3 mph.

- (a) Estimate the standard deviation of the population.
 (b) Estimate the standard error of the mean for this population.
 (c) What are the upper and lower limits of the confidence interval for the mean speed given a desired confidence level of 0.95?
- 7-65** Given a sample mean of 8, a population standard deviation of 2.6, and a sample size of 32, find the confidence level associated with each of the following intervals:
 (a) (7.6136, 8.3864).
 (b) (6.85, 9.15).
 (c) (7.195, 8.805).
- 7-66** Based on knowledge about the desirable qualities of estimators, for what reasons might \bar{x} be considered the “best” estimator of the true population mean?
- 7-67** The president of Offshore Oil has been concerned about the number of fights on his rigs and has been considering various courses of action. In an effort to understand the catalysts of offshore fighting, he randomly sampled 41 days on which a crew had returned from mainland leave. For this sample, the average proportion of workers involved in fisticuffs each day is 0.032 and the associated standard deviation is 0.0130.
 (a) Give a point estimate for the average proportion of workers involved in fights on any given day that a crew has returned from the mainland.
 (b) Estimate the population standard deviation associated with this fighting rate.
 (c) Find a 90 percent confidence interval for the average proportion of returning workers who get involved in fights.
- 7-68** Given the following expressions for the limits of a confidence interval, find the confidence level associated with the interval:
 (a) $\bar{x} - 1.25\sigma_{\bar{x}}$ to $\bar{x} + 1.25\sigma_{\bar{x}}$.
 (b) $\bar{x} - 2.4\sigma_{\bar{x}}$ to $\bar{x} + 2.4\sigma_{\bar{x}}$.
 (c) $x - 1.68\sigma_{\bar{x}}$ to $\bar{x} + 1.68\sigma_{\bar{x}}$.
- 7-69** Harris Polls, Inc., is in the business of surveying households. From previous surveys, it is known that the standard deviation of the number of hours of television watched in a week by a household is 1.1 hours. Harris Polls would like to determine the average number of hours of television watched per week per household in the United States. Accuracy is important, so Harris Polls would like to be 98 percent certain that the sample average number of hours falls within ± 0.3 hour of the national average. Conservatively, what sample size should Harris Polls use?
- 7-70** John Bull has just purchased a computer program that claims to pick stocks that will increase in price in the next week with an 85 percent accuracy rate. On how many stocks should John test this program in order to be 98 percent certain that the percentage of stocks that do in fact go up in the next week will be within ± 0.05 of the sample proportion?
- 7-71** Gotchya runs a laser-tag entertainment center where adults and teenagers rent equipment and engage in mock combat. The facility is always used to capacity on weekends. The three owners want to assess the effectiveness of a new advertising campaign aimed at increasing weeknight usage. The number of paying patrons on twenty-seven randomly selected weeknights is given in the following table. Find a 95 percent confidence interval for the mean number of patrons on a weeknight.

61	57	53	60	64	57	54	58	63
59	50	60	60	57	58	62	63	60
61	54	50	54	61	51	53	62	57

- 7-72 Their accountants have told the owners of Gotchya, the laser-tag entertainment center discussed in Exercise 7-71, that they need to have at least fifty-five patrons in order to break even on a weeknight. The partners are willing to continue to operate on weeknights if they can be at least 95 percent certain that they will break even at least half the time. Using the data in Exercise 7-71, find a 95 percent confidence interval for the proportion of weeknights on which Gotchya will break even. Should Gotchya continue to stay open on weeknights? Explain.
- 7-73 In evaluating the effectiveness of a federal rehabilitation program, a survey of 52 of a prison's 900 inmates found that 35 percent were repeat offenders.
- Estimate the standard error of the proportion of repeat offenders.
 - Construct a 90 percent confidence interval for the proportion of repeat offenders among the inmates of this prison.
- 7-74 From a random sample of 60 buses, Montreal's mass-transit office has calculated the mean number of passengers per kilometer to be 4.1. From previous studies, the population standard deviation is known to be 1.2 passengers per kilometer.
- Find the standard error of the mean. (Assume that the bus fleet is very large.)
 - Construct a 95 percent confidence interval for the mean number of passengers per kilometer for the population.
- 7-75 The Internal Revenue Service sampled 200 tax returns recently and found that the sample average income tax refund amounted to \$425.39 and the sample standard deviation was \$107.10.
- Estimate the population mean tax refund and standard deviation.
 - Using the estimates of part (a), construct an interval in which the population mean is 95 percent certain to fall.
- 7-76 The Physicians Care Group operates a number of walk-in clinics. Patient charts indicate the time that a patient arrived at the clinic and the time that the patient was actually seen by a physician. Administrator Val Likmer has just received a stinging phone call from a patient complaining of an excessive wait at the Rockridge clinic. Val pulls 49 charts at random from last week's workload and calculates an average wait time of 15.2 minutes. A previous large-scale study of waiting time over several clinics had a standard deviation of 2.5 minutes. Construct a confidence interval for the average wait time with confidence level
- 90 percent.
 - 99 percent.
- 7-77 Bill Wenslaff, an engineer on the staff of a water purification plant, measures the chlorine content in 200 different samples daily. Over a period of years, he has established the population standard deviation to be 1.4 milligrams of chlorine per liter. The latest samples averaged 4.6 milligrams of chlorine per liter.
- Find the standard error of the mean.
 - Establish the interval around 5.2, the population mean, that will include the sample mean with a probability of 68.3 percent.
- 7-78 Ellen Harris, an industrial engineer, was accumulating normal times for various tasks on a labor-intensive assembly process. This process included 300 separate job stations, each performing the same assembly tasks. She sampled seven stations and obtained the following assembly times for each station: 1.9, 2.5, 2.9, 1.3, 2.6, 2.8, and 3.0 minutes.
- Calculate the mean assembly time and the corresponding standard deviation for the sample.

TABLE RW7-1 FINANCIAL DATA FOR A SAMPLE OF 35 MUTUAL FUNDS

Fund Name	NAV	OP	ΔNAV	%YTD
AHA Balanced	12.54	12.54	-0.01	3.9%
Ambassador Index Stock	11.36	11.36	0.01	1.9
American Capital Global Equity (A)	10.44	11.08	0.01	8.2
American Capital Municipal Bond	10.33	10.85	-0.01	5.1
Atlas Growth & Income	13.69	14.04	-0.05	2.2
Babson Enterprise	16.13	16.13	0.08	6.0
Blanchard Flexible Income	5.11	5.11	0.00	5.9
Colonial Growth	14.08	14.94	-0.05	0.1
Columbia Common Stock	14.54	14.54	-0.02	3.8
Evergreen Total Return	19.96	19.96	-0.07	5.9
Fidelity Equity-Income	31.24	31.88	-0.14	8.6
Fidelity Spartan Municipal Income	11.02	11.02	0.00	5.9
First Union Value (B)	17.30	18.02	-0.04	1.8
Flag Investors Value	10.89	11.40	-0.05	2.9
Fortis Capital	17.48	18.35	0.03	-5.3
GT Global Europe	9.11	9.56	0.03	7.1
Helmsman Equity Index	11.68	11.68	0.02	1.8
Homestead Value	13.48	13.48	-0.01	7.9
IAI Emerging Growth	13.64	13.64	0.09	-2.8
John Hancock Tax Exempt	11.32	11.85	0.00	5.1
Kemper Blue Chip	13.30	14.11	0.02	-0.2
Keystone International	6.50	6.50	0.01	8.0
Marshall Stock	9.90	9.90	0.03	-1.9
MAS Equity	54.37	54.37	-0.11	-1.9
MFS Research	12.86	13.64	0.01	4.6
MIM Bond Income	9.24	9.24	0.02	-0.5
PFAMCo MidCap Growth	12.51	12.51	-0.03	2.8
Pilgrim GNMA	14.02	14.45	-0.01	3.2
PIMCO Short Term	10.03	10.03	0.01	1.8
Prudential Municipal Maryland	11.35	11.35	0.00	4.8
Putnam Global Growth	8.18	8.68	-0.01	10.1
Rightime Blue Chip	31.07	32.62	0.02	1.2
Schwab 1000	12.11	12.11	-0.01	1.3
Shearson Appreciation (A)	10.72	11.28	-0.03	0.6
Weiss Peck Greer Tudor	24.90	24.90	0.19	0.2

NAV net asset value, the price (in \$) at which an investor can redeem shares of the fund

OP offering price, the price (in \$) which an investor pays to purchase shares of the fund

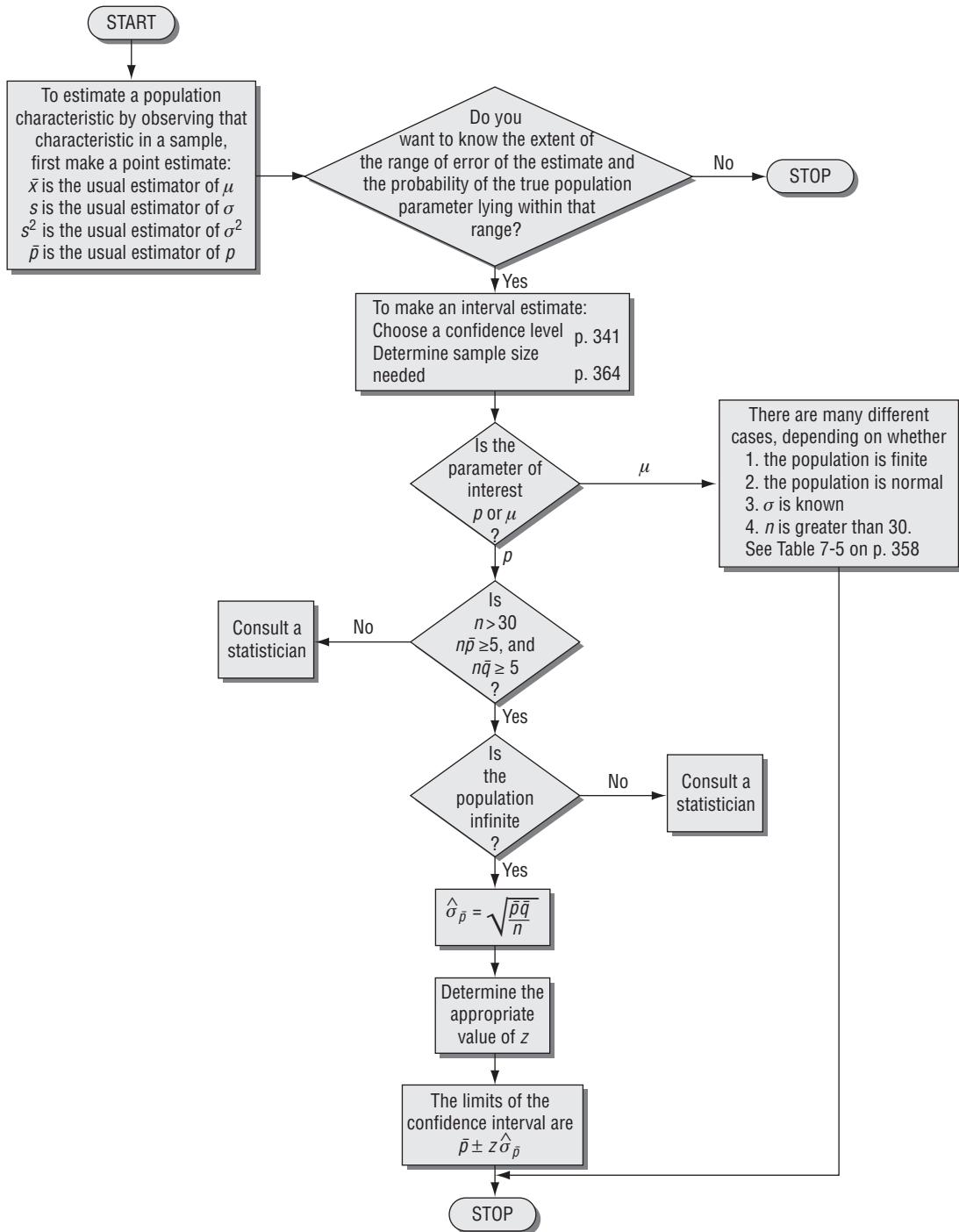
ΔNAV the change in NAV from the previous day

%YTD the year-to-date percentage change in the value of an investment in the fund, assuming all dividends are reinvested

Source: The Wall Street Journal (17 May 1993): C16–C19.

- (b) Estimate the population standard deviation.
(c) Construct a 98 percent confidence interval for the mean assembly time.
- 7-79** Larry Culler, the federal grain inspector at a seaport, found spoilage in 40 of 120 randomly selected lots of wheat shipped from the port. Construct a 95 percent confidence interval for him for the actual proportion of lots with spoilage in shipments from that port.
- 7-80** High Fashion Marketing is considering reintroducing paisley ties. In order to avoid a fashion flop, High Fashion interviewed 90 young executives (their primary market) and found that of the 90 interviewed, 79 believed that paisley ties were fashionable and were interested in purchasing one. Using a confidence level of 98 percent, construct a confidence interval for the proportion of all young executives who find paisley ties fashionable.
- 7-81** The Department of Transportation has mandated that the average speed of cars on interstate highways must be no more than 67 miles per hour in order for state highway departments to retain their federal funding. North Carolina troopers, in unmarked cars, clocked a sample of 186 cars and found that the average speed was 66.3 miles per hour and the standard deviation was 0.6 mph.
(a) Find the standard error of the mean.
(b) What is the interval around the sample mean that would contain the population mean 95.5 percent of the time?
(c) Can North Carolina truthfully report that the true mean speed on its highways is 67 mph or less with 95.5 percent confidence?
- 7-82** Mark Semmes, owner of the Aurora Restaurant, is considering purchasing new furniture. To help him decide on the amount he can afford to invest in tables and chairs, he wishes to determine the average revenue per customer. The checks for 9 randomly sampled customers had an average of \$18.30 and a standard deviation of \$3.60. Construct a 95 percent confidence interval for the size of the average check per customer.
- 7-83** John Deer, a horticulturist at Northern Carrboro State University, knows that a certain strain of corn will always produce between 80 and 140 bushels per acre. For a confidence level of 90 percent, how many 1-acre samples must be taken in order to estimate the average production per acre to within ± 5 bushels per acre?
- 7-84** Nirmal Pvt. Limited is a FMCG company, selling a range of products. It has 1150 sales outlets. A sample of 60 sales outlets was chosen, using random sampling for the purpose of sales analysis. The sample consists of sales outlets from rural and urban areas belonging to the four regions of the country—Northern, Eastern, Western, Southern. The information related to annual sales has been collected from them in the month of December 2010. This process has been repeated in December 2011. In the meanwhile, in 2010 a comprehensive sales-promotion program was launched to augment the sales. The information is presented in the data sheet provided in the DVD (Nirmal Pvt. Ltd). Analyze the data and give answer to the following questions.
(a) Construct 95% confidence interval around the “mean 2010 and 2011 sales” separately. Compare the results and comment.
(b) Construct 99% confidence interval around the “mean 2010 sale of Urban Shops and Rural Shops” separately. Compare the results and comment.
(c) Construct 95% confidence interval for the “proportion of urban shops.”
(d) Construct 99% confidence interval of the “proportion of Northern shops.”
(e) Point estimate of the standard deviation of 2011 sale.
(f) Compare the sample means of “2010 sale” with respect to the four regions and comment on it.

Flow Chart: Estimation



This page is intentionally left blank.

8 Testing Hypotheses: One-sample Tests

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn how to use samples to decide whether a population possesses a particular characteristic
 - To determine how unlikely it is that an observed sample could have come from a hypothesized population
 - To understand the two types of errors possible when testing hypotheses
 - To learn when to use one-tailed tests and when to use two-tailed tests
 - To learn the five-step process for testing hypotheses
 - To understand how and when to use the normal and t distributions for testing hypotheses about population means and proportions
-

CHAPTER CONTENTS

- | | |
|---|---|
| 8.1 Introduction 380 | 8.6 Hypothesis Testing of Proportions:
Large Samples 405 |
| 8.2 Concepts Basic to the Hypothesis-Testing
Procedure 381 | 8.7 Hypothesis Testing of Means When the
Population Standard Deviation is not
Known 411 |
| 8.3 Testing Hypotheses 385 | ■ Statistics at Work 418 |
| 8.4 Hypothesis Testing of Means When
the Population Standard Deviation is
Known 393 | ■ Terms Introduced in Chapter 8 418 |
| 8.5 Measuring the Power of a Hypothesis
Test 402 | ■ Review and Application Exercises 419 |
| | ■ Flow Chart: One-Sample Tests
of Hypotheses 424 |

The roofing contract for a new sports complex in San Francisco has been awarded to Parkhill Associates, a large building contractor. Building specifications call for a movable roof covered by approximately 10,000 sheets of 0.04-inch-thick aluminum. The aluminum sheets cannot be appreciably thicker than 0.04 inch because the structure could not support the additional weight. Nor can the sheets be appreciably thinner than 0.04 inch because the strength of the roof would be inadequate. Because of this restriction on thickness, Parkhill carefully checks the aluminum sheets from its supplier. Of course, Parkhill does not want to measure each sheet, so it randomly samples 100. The sheets in the sample have a mean thickness of 0.0408 inch. From past experience with this supplier, Parkhill believes that these sheets come from a thickness population with a standard deviation of 0.004 inch. On the basis of these data, Parkhill must decide whether the 10,000 sheets meet specifications. In Chapter 7, we used sample statistics to estimate population parameters. Now, to solve problems like Parkhill's, we shall learn how to use characteristics of samples to test an assumption we have about the population from which that sample came. Our test for Parkhill, later in the chapter, may lead Parkhill to accept the shipment or it may indicate that Parkhill should reject the aluminum sheets sent by the supplier because they do not meet the architectural specifications. ■

8.1 INTRODUCTION

Hypothesis testing begins with an assumption, called a *hypothesis*, that we make about a population parameter. Then we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct. Say that we assume a certain value for a population mean. To test the validity of our assumption, we gather sample data and determine the difference between the hypothesized value and the actual value of the sample mean. Then we judge whether the difference is significant. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference, the smaller the likelihood.

Unfortunately, the difference between the hypothesized population parameter and the actual statistic is more often neither so large that we automatically reject our hypothesis nor so small that we just as quickly accept it. So in hypothesis testing, as in most significant real-life decisions, clear-cut solutions are the exception, not the rule.

Suppose a manager of a large shopping mall tells us that the average work efficiency of her employees is at least 90 percent. How can we test the validity of her hypothesis? Using the sampling methods we learned in Chapter 6, we could calculate the efficiency of a *sample* of her employees. If we did this and the sample statistic came out to be 95 percent, we would readily accept the manager's statement. However, if the sample statistic were 46 percent, we would reject her assumption as untrue. We can interpret both these outcomes, 95 percent and 46 percent, using our common sense.

Now suppose that our sample statistic reveals an efficiency of 88 percent. This value is relatively close to 90 percent, but is it close enough for us to accept the manager's hypothesis? Whether we accept or reject the manager's hypothesis, we cannot be absolutely certain that our decision is correct; therefore, we will have to learn to deal with uncertainty in our decision making. **We cannot accept or reject a hypothesis about a population parameter simply by intuition. Instead, we need to learn how to decide objectively, on the basis of sample information, whether to accept or reject a hunch.**

Function of hypothesis testing

When to accept or reject the hypothesis

The basic problem is dealing with uncertainty

Making Big Jumps

College students often see ads for learning aids. One very popular such aid is a combination outline, study guide, and question set for various courses. Advertisements about such items often claim better examination scores with less studying time. Suppose a study guide for a basic statistics course is available through an organization that produces such guides for 50 different courses. If this study guide for basic statistics has been tested (and let us assume properly), the firm may advertise that “our study guides have been statistically proven to raise grades and lower study time.” Of course, this assertion is quite true, but only as it applies to the basic statistics experience. There may be no evidence of statistical significance that establishes the same kind of results for the other 49 guides.

Projecting too far

Another product may be advertised as being beneficial in removing crabgrass from your lawn and may assert that the product has been “thoroughly tested” on real lawns. Even if we assume that the proper statistical procedures were, in fact, used during the tests, such claims still involve big jumps. Suppose that the test plot was in Florida and your lawn problems are in Utah. Differences in rainfall, soil fertility, airborne pollutants, temperature, dormancy hours and germination conditions may vary widely between these two locations. Claiming results for a statistically valid test under a completely different set of test conditions is invalid. One such test cannot measure effectiveness under a wide variety of environmental conditions.

Different test conditions

EXERCISES 8.1

- 8-1 Why must we be required to deal with uncertainty in our decisions, even when using statistical techniques?
- 8-2 Theoretically speaking, how might one go about testing the hypothesis that a coin is fair? That a die is fair?
- 8-3 Is it possible that a false hypothesis will be accepted? How would you explain this?
- 8-4 Describe the hypothesis-testing process.
- 8-5 How would you explain a large difference between a hypothesized population parameter and a sample statistic if, in fact, the hypothesis is true?

8.2 CONCEPTS BASIC TO THE HYPOTHESIS-TESTING PROCEDURE

Before we introduce the formal statistical terms and procedures, we’ll work our chapter-opening sports-complex problem all the way through. Recall that the aluminum roofing sheets have a claimed average thickness of 0.04 inch and that they will be unsatisfactory if they are too thick *or* too thin. The contractor takes a sample of 100 sheets and determines that the sample mean thickness is 0.0408 inch. On the basis of past experience, he knows that the population standard deviation is 0.004 inch. Does this sample evidence indicate that the batch of 10,000 sheets of aluminum is suitable for constructing the roof of the new sports complex?

Sports-complex problem

If we assume that the true mean thickness is 0.04 inch, and we know that the population standard deviation is 0.004 inch, how likely is it that we would get a sample mean of 0.0408 or more from that population? In other words, *if*

Formulating the hypothesis

the true mean is 0.04 inch and the standard deviation is 0.004 inch, what are the chances of getting a sample mean that differs from 0.04 inch by 0.0008 (= 0.0408 – 0.04) inch or more?

These questions show that to determine whether the population mean is actually 0.04 inch, we must calculate the probability that a random sample with a mean of 0.0408 inch will be selected from a population with a μ of 0.04 inch and a σ of 0.004 inch. This probability will indicate whether it is *reasonable* to observe a sample like this if the population mean is actually 0.04 inch. If this probability is far too low, we must conclude that the aluminum company's statement is false and that the mean thickness of the aluminum sheets is not 0.04 inch.

Let's answer the question illustrated in Figure 8-1: If the hypothesized population mean is 0.04 inch and the population standard deviation is 0.004 inch, what are the chances of getting a sample mean (0.0408 inch) that differs from 0.04 inch by 0.0008 inch? First, we calculate the standard error of the mean from the population standard deviation:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.004 \text{ in.}}{\sqrt{100}} \\ &= \frac{0.004 \text{ in.}}{10} \\ &= 0.0004 \text{ in.}\end{aligned}\quad [6-1]$$

Calculating the standard error of the mean

Next we use Equation 6-2 to discover that the mean of our sample (0.0408 inch) lies 2 standard errors to the right of the hypothesized population mean:

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \\ &= \frac{0.0408 - 0.04}{0.0004} \\ &= 2 \leftarrow \text{Standard errors of the mean}\end{aligned}\quad [6-2]$$

Interpreting the probability associated with this difference

Using Appendix Table 1, we learn that 4.5 percent is the total chance of our sample mean differing from the population mean by 2 or more standard errors; that is, the chance that the sample mean would be 0.0408 inch or larger or 0.0392 inch or smaller is only 4.5 percent ($P(z \geq 2 \text{ or } z \leq -2) = 2(0.5 - 0.4772) = 0.0456$, or about 4.5 percent). **With this low a chance, Parkhill could conclude that a population with a true mean of 0.04 inch would not be likely to produce a sample like this.** The project supervisor would reject the aluminum company's statement about the mean thickness of the sheets.

In this case, the difference between the sample mean and the hypothesized population mean is too large, and the chance that the population would produce such a random sample is far too low. Why this probability of 4.5 percent is too low, or wrong, is a judgment for decision makers to make. Certain situations demand that decision makers be very sure about the characteristics of the items being

The decision maker's role in formulating hypotheses

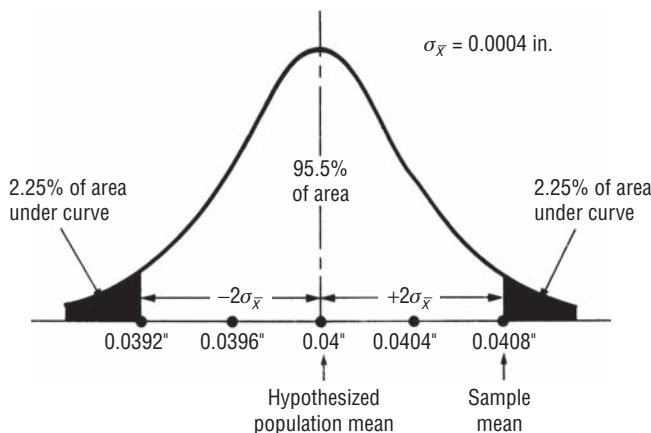


FIGURE 8-1 PROBABILITY THAT \bar{X} WILL DIFFER FROM HYPOTHESIZED μ BY 2

tested, and then even 2 percent is too high to be attributable to chance. Other processes allow for a wider latitude or variation, and a decision maker might accept a hypothesis with a 4.5 percent probability of chance variation. In each situation, we must try to determine the costs resulting from an incorrect decision and the precise level of risk we are willing to assume.

In our example, we rejected the aluminum company's contention that the population mean is 0.04 inch. But suppose for a moment that the population mean is *actually* 0.04 inch. If we then stuck to our rejection rule of 2 standard errors or more (the 4.5 percent probability or less in the tails of Figure 8-1), we would reject a perfectly good lot of aluminum sheets 4.5 percent of the time. Therefore, **our minimum standard for an acceptable probability, 4.5 percent, is also the risk we take of rejecting a hypothesis that is true. In this or any decision making, there can be no risk-free trade-off.**

HINTS & ASSUMPTIONS

Although *hypothesis testing* sounds like some formal statistical term completely unrelated to business decision making, in fact managers propose and test hypotheses all the time. “If we drop the price of this car model by \$1,500, we’ll sell 50,000 cars this year” is a hypothesis. To test this hypothesis, we have to wait until the end of the year and count sales. Managerial hypotheses are based on intuition; the marketplace decides whether the manager’s intuitions were correct. Hint: Hypothesis testing is about making inferences about a population from only a small sample. The bottom line in hypothesis testing is when we ask ourselves (and then decide) whether a population like we *think* this one is would be likely to produce a sample like the one we are looking at.

EXERCISES 8.2

Self-Check Exercises

- SC 8-1** How many standard errors around the hypothesized value should we use to be 99.44 percent certain that we accept the hypothesis when it is true?

- SC 8-2** An automobile manufacturer claims that a particular model gets 28 miles to the gallon. The Environmental Protection Agency, using a sample of 49 automobiles of this model, finds the sample mean to be 26.8 miles per gallon. From previous studies, the population standard deviation is known to be 5 miles per gallon. Could we reasonably expect (within 2 standard errors) that we could select such a sample if indeed the population mean is actually 28 miles per gallon?

Basic Concepts

- 8-6** What do we mean when we reject a hypothesis on the basis of a sample?
- 8-7** Explain why there is no single standard level of probability used to reject or accept in hypothesis testing.
- 8-8** If we reject a hypothesized value because it differs from a sample statistic by more than 1.75 standard errors, what is the probability that we have rejected a hypothesis that is in fact true?
- 8-9** How many standard errors around the hypothesized value should we use to be 98 percent certain that we accept the hypothesis when it is true?

Applications

- 8-10** Sports and media magnate Ned Sterner is interested in purchasing the Atlanta Stalwarts if he can be reasonably certain that operating the team will not be too costly. He figures that average attendance would have to be about 28,500 fans per game to make the purchase attractive to him. Ned randomly chooses 64 home games over the past 4 years and finds from figures reported in *Sporting Reviews* that average attendance at these games was 26,100. A study he commissioned the last time he purchased a team showed that the population standard deviation for attendance at similar events had been quite stable for the past 10 years at about 6,000 fans. Using 2 standard errors as the decision criterion, should Ned purchase the Stalwarts? Can you think of any reason(s) why your conclusion might not be valid?
- 8-11** *Computing World* has asserted that the amount of time owners of personal computers spend on their machines averages 23.9 hours per week and has a standard deviation of 12.6 hours per week. A random sampling of 81 of its subscribers revealed a sample mean usage of 27.2 hours per week. On the basis of this sample, is it reasonable to conclude (using 2 standard errors as the decision criterion) that *Computing World*'s subscribers are different from average personal computer owners?
- 8-12** A grocery store has specially packaged oranges and has claimed a bag of oranges will yield 2.5 quarts of juice. After randomly selecting 42 bags, a stacker found the average juice production per bag to be 2.2 quarts. Historically, we know the population standard deviation is 0.2 quart. Using this sample and a decision criterion of 2.5 standard errors, could we conclude the store's claims are correct?

Worked-Out Answers to Self-Check Exercises

- SC 8-1** To leave a probability of $1 - 0.9944 = 0.0056$ in the tails, the absolute value of z must be greater than or equal to 2.77, so the interval should be ± 2.77 standard errors about the hypothesized value.

SC 8-2 $\sigma = 5 \quad n = 49 \quad \bar{x} = 26.8 \quad \mu = 28$

$$\mu \pm 2\sigma_{\bar{x}} = \mu \pm 2\sigma/\sqrt{n} = 28 \pm 2(5)/\sqrt{49} = 28 \pm 1.429 = (26.571, 29.429)$$

Because $\bar{x} = 26.8 > 26.57$, it is not unreasonable to see such sample results if μ really is 28 mpg.

8.3 TESTING HYPOTHESES

In hypothesis testing, we must state the assumed or hypothesized value of the population parameter *before* we begin sampling. The assumption we wish to test is called the *null hypothesis* and is symbolized H_0 or “H sub-zero.”

Making a formal statement of the null hypothesis

Suppose we want to test the hypothesis that the population mean is equal to 500. We would symbolize it as follows and read it, “The null hypothesis is that the population mean is equal to 500”:

$$H_0: \mu = 500$$

The term *null hypothesis* arises from earlier agricultural and medical applications of statistics. In order to test the effectiveness of a new fertilizer or drug, the tested hypothesis (the null hypothesis) was that it had *no effect*, that is, there was no difference between treated and untreated samples.

Why is it called the null hypothesis?

If we use a hypothesized value of a population mean in a problem, we would represent it symbolically as

$$\mu_{H_0}$$

This is read, “The hypothesized value of the population mean.”

If our sample results fail to support the null hypothesis, we must conclude that something else is true. **Whenever we reject the hypothesis, the conclusion we do accept is called the *alternative hypothesis* and is symbolized H_1 (“H sub-one”).** For the null hypothesis

$$H_0: \mu = 200 \text{ (Read: "The null hypothesis is that the population mean is equal to 200.")}$$

we will consider three possible alternative hypotheses:

- $H_1: \mu \neq 200 \leftarrow \text{"The alternative hypothesis is that the population mean is } \text{not} \text{ equal to 200"}$
- $H_1: \mu > 200 \leftarrow \text{"The alternative hypothesis is that the population mean is } \text{greater than 200"}$
- $H_1: \mu < 200 \leftarrow \text{"The alternative hypothesis is that the population mean is } \text{less than 200"}$

Making a formal statement of the alternate hypothesis

Interpreting the Significance Level

The purpose of hypothesis testing is not to question the computed value of the sample statistic but to make a judgment about the *difference* between that sample statistic and a hypothesized population parameter. The next step after stating the null and alternative hypotheses, then, is to decide what criterion to use for deciding whether to accept or reject the null hypothesis.

Goal of hypothesis testing

In our sports-complex example, we decided that a difference observed between the sample mean \bar{x} and the hypothesized population mean μ_{H_0} had only a 4.5 percent, or 0.045, chance of occurring. Therefore, we rejected the null hypothesis that the population mean was 0.04 inch ($H_0: \mu = 0.04$ inch). In statistical terms, the value 0.045 is called the *significance level*.

Function of the significance level

What if we test a hypothesis at the 5 percent level of significance? This means that we will reject the null hypothesis if the difference between the sample statistic and the hypothesized population

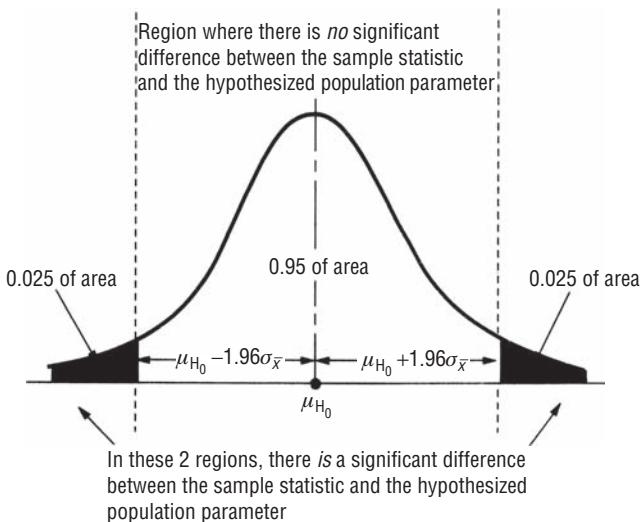


FIGURE 8-2 REGIONS OF SIGNIFICANT DIFFERENCE AND OF NO SIGNIFICANT DIFFERENCE AT A 5 PERCENT LEVEL OF SIGNIFICANCE

parameter is so large that it or a larger difference would occur, on the average, only five or fewer times in every 100 samples when the hypothesized population parameter is correct. **If we assume the hypothesis is correct, then the significance level will indicate the percentage of sample means that is outside certain limits.** (In estimation, you remember, the confidence level indicated the percentage of sample means that fell *within* the defined confidence limits.)

Figure 8-2 illustrates how to interpret a 5 percent level of significance. Notice that 2.5 percent of the area under the curve is located in each tail. From Appendix Table 1, we can determine that 95 percent of all the area under the curve is included in an interval extending $1.96\sigma_{\bar{x}}$ on either side of the hypothesized mean. In 95 percent of the area, then, there is no significant difference between the observed value of the sample statistic and the hypothesized value of the population parameter. In the remaining 5 percent (the colored regions in Figure 8-2), a significant difference does exist.

Figure 8-3 examines this same example in a different way. Here, the 0.95 of the area under the curve is where we would accept the null hypothesis. The two colored parts under the curve, representing a total of 5 percent of the area, are the regions where we would reject the null hypothesis.

A word of caution is appropriate here. Even if our sample statistic in Figure 8-3 does fall in the nonshaded region (the region that makes up 95 percent of the area under the curve), **this does not prove that our null hypothesis (H_0) is true; it simply does not provide statistical evidence to reject it.** Why? Because the only way in which the hypothesis can be accepted with certainty is for us to know the population parameter; unfortunately, this is not possible. Therefore, whenever we say that we accept the null hypothesis, we actually mean that there is not sufficient statistical evidence to reject it. **Use of the term *accept*, instead of *do not reject*, has become standard. It means simply that when sample data do not cause us to reject a null hypothesis, we behave as if that hypothesis is true.**

Area where no significant difference exists

Also called the area where we accept the null hypothesis

Hypotheses are accepted, not proved

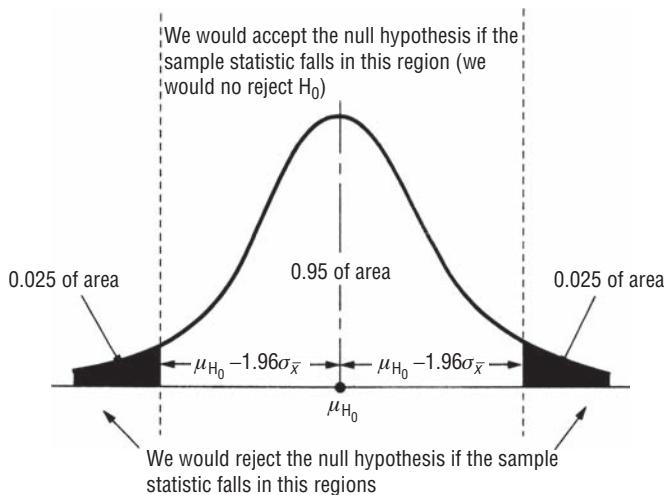


FIGURE 8-3 A 5 PERCENT LEVEL OF SIGNIFICANCE, WITH ACCEPTANCE AND REJECTION REGIONS DESIGNATED

Selecting a Significance Level

There is no single standard or universal level of significance for testing hypotheses. In some instances, a 5 percent level of significance is used. Published research results often test hypotheses at the 1 percent level of significance. It is possible to test a hypothesis at *any* level of significance. But remember that our choice of the minimum standard for an acceptable probability, or the significance level, is also the risk we assume of rejecting a null hypothesis when it is true. **The higher the significance level we use for testing a hypothesis, the higher the probability of rejecting a null hypothesis when it is true.**

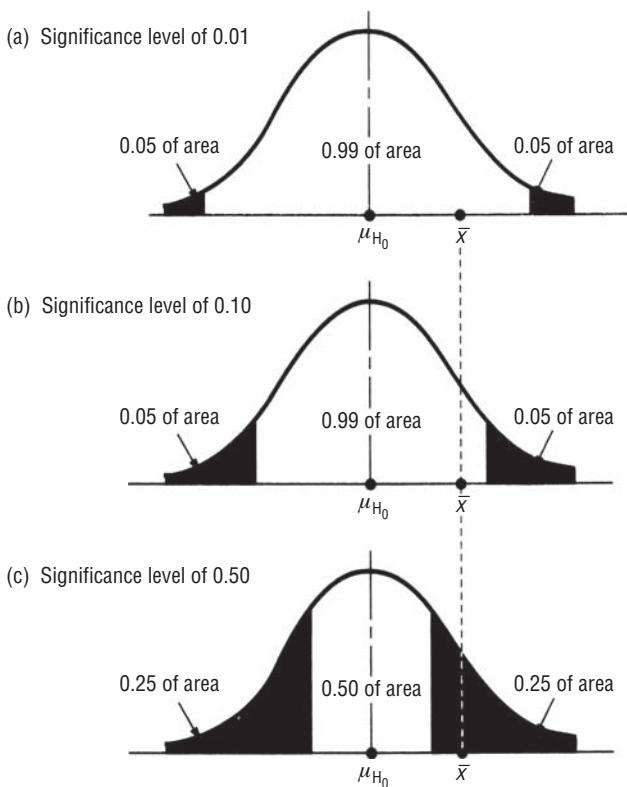
Trade-offs when choosing a significance level

Examining this concept, we refer to Figure 8-4. Here we have illustrated a hypothesis test at three different significance levels: 0.01, 0.10, and 0.50. Also, we have indicated the location of the same sample mean \bar{x} on each distribution. In parts *a* and *b*, we would accept the null hypothesis that the population mean is equal to the hypothesized value. But notice that in part *c*, we would reject this same null hypothesis. Why? Our significance level there of 0.50 is so high that we would rarely accept the null hypothesis when it is *not* true but, at the same time, often reject it when it *is* true.

Type I and Type II Errors

Statisticians use specific definitions and symbols for the concept illustrated in Figure 8-4. Rejecting a null hypothesis when it is true is called a *Type I error*, and its probability (which, as we have seen, is also the significance level of the test) is symbolized α (*alpha*). Alternatively, accepting a null hypothesis when it is false is called a *Type II error*, and its probability is symbolized β (*beta*). There is a trade-off between these two errors: The probability of making one type of error can be reduced only if we are willing to increase the probability of making the other type of error. Notice in part *c*, Figure 8-4, that our acceptance region is quite small (0.50 of the area under the curve). With an acceptance region this small, we will rarely accept a null hypothesis when it is not true, but as a cost of being this sure, we will

Type I and Type II errors defined

**FIGURE 8-4 THREE DIFFERENT LEVELS OF SIGNIFICANCE**

often reject a null hypothesis when it is true. Put another way, in order to get a low β , we will have to put up with a high α . To deal with this trade-off in personal and professional situations, decision makers decide the appropriate level of significance by examining the costs or penalties attached to both types of errors.

Suppose that making a Type I error (rejecting a null hypothesis when it is true) involves the time and trouble of reworking a batch of chemicals that should have been accepted. At the same time, making a Type II error (accepting a null hypothesis when it is false) means risking a chance that an entire group of users of this chemical compound will be poisoned. Obviously, the management of this company will prefer a Type I error to a Type II error and, as a result, will set very high levels of significance in its testing to get low β s.

Suppose, on the other hand, that making a Type I error involves disassembling an entire engine at the factory, but making a Type II error involves relatively inexpensive warranty repairs by the dealers. Then the manufacturer is more likely to prefer a Type II error and will set lower significance levels in its testing.

Deciding Which Distribution to Use in Hypothesis Testing

After deciding what level of significance to use, our next task in hypothesis testing is to determine the appropriate probability distribution. We have a choice between the normal distribution,

Preference for a Type I error

Preference for a Type II error

Selecting the correct distribution before the test

TABLE 8-1 CONDITIONS FOR USING THE NORMAL AND t DISTRIBUTIONS IN TESTING HYPOTHESES ABOUT MEANS

	When the Population Standard Deviation is Known	When the Population Standard Deviation is Not Known
Sample size n is larger than 30	Normal distribution, z table	Normal distribution, z table
Sample size n is 30 or less and we assume the population is normal or approximately so	Normal distribution, z table	t distribution, t table

Appendix Table 1, and the t distribution, Appendix Table 2. The rules for choosing the appropriate distribution are similar to those we encountered in Chapter 7 on estimation. Table 8-1 summarizes when to use the normal and t distributions in making tests of means. Later in this chapter, we shall examine the distributions appropriate for testing hypotheses about proportions.

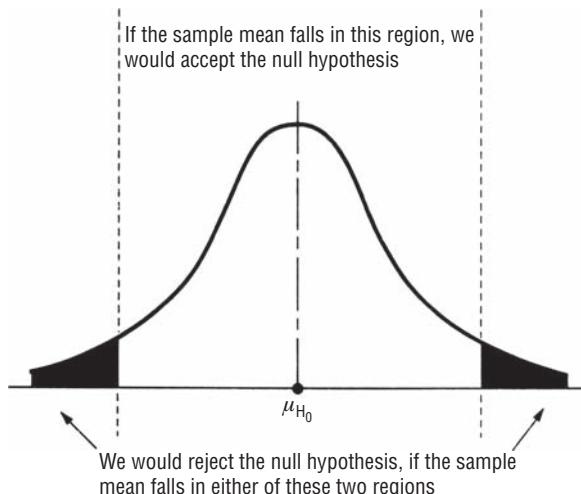
Remember one more rule when testing the hypothesized value of a mean. As in estimation, use the *finite population multiplier* whenever the population is finite in size, sampling is done without replacement, and the sample is more than 5 percent of the population.

Use of the finite population multiplier

Two-Tailed and One-Tailed Tests of Hypotheses

In the tests of hypothesized population means that follow, we shall illustrate two-tailed tests and one-tailed tests. These new terms need a word of explanation. A *two-tailed test* of a hypothesis will reject the null hypothesis if the sample mean is significantly higher than *or* lower than the hypothesized population mean. Thus, in a two-tailed test, there are *two* rejection regions. This is illustrated in Figure 8-5.

Description of a two-tailed hypothesis test

**FIGURE 8-5** TWO-TAILED TEST OF A HYPOTHESIS, SHOWING THE TWO REJECTION REGIONS

A two-tailed test is appropriate when the null hypothesis is $\mu = \mu_{H_0}$ (μ_{H_0} being some specified value) and the alternative hypothesis is $\mu \neq \mu_{H_0}$. Assume that a manufacturer of lightbulbs wants to produce bulbs with a mean life of $\mu = \mu_{H_0} = 1,000$ hours. If the lifetime is shorter, he will lose customers to his competition; if the lifetime is longer, he will have a very high production cost because the filaments will be excessively thick. In order to see whether his production process is working properly, he takes a sample of the output to test the hypothesis $H_0: \mu = 1,000$. Because he does not want to deviate significantly from 1,000 hours *in either direction*, the appropriate alternative hypothesis is $H_1: \mu \neq 1,000$, and he uses a two-tailed test. That is, he rejects the null hypothesis if the mean life of bulbs in the sample is *either too far above 1,000 hours or too far below 1,000 hours*.

However, there are situations in which a two-tailed test is not appropriate, and we must use a one-tailed test. Consider the case of a wholesaler that buys lightbulbs from the manufacturer discussed earlier. The wholesaler buys bulbs in large lots and does not want to accept a lot of bulbs unless their mean life is at least 1,000 hours. As each shipment arrives, the wholesaler tests a sample to decide whether it should accept the shipment. The company will reject the shipment only if it feels that the mean life is below 1,000 hours. If it feels that the bulbs are *better* than expected (with a mean life above 1,000 hours), it certainly will not reject the shipment because the longer life comes at no extra cost. So the wholesaler's hypotheses are $H_0: \mu = 1,000$ hours and $H_1: \mu < 1,000$ hours. It rejects H_0 only if the mean life of the sampled bulbs is significantly *below* 1,000 hours. This situation is illustrated in Figure 8-6. From this figure, we can see why this test is called a *left-tailed test* (or a *lower-tailed test*).

Sometimes a one-tailed test is appropriate

In general, a left-tailed (lower-tailed) test is used if the hypotheses are $H_0: \mu = \mu_{H_0}$ and $H_1: \mu < \mu_{H_0}$. In such a situation, it is sample evidence with the sample mean significantly below the hypothesized population mean that leads us to reject the null hypothesis in favor of the alternative hypothesis. Stated differently, the rejection region is in the lower tail (left tail) of the distribution of the sample mean, and that is why we call this a lower-tailed test.

Left-tailed tests

A left-tailed test is one of the two kinds of one-tailed tests. As you have probably guessed by now, the other kind of one-tailed test

Right-tailed tests

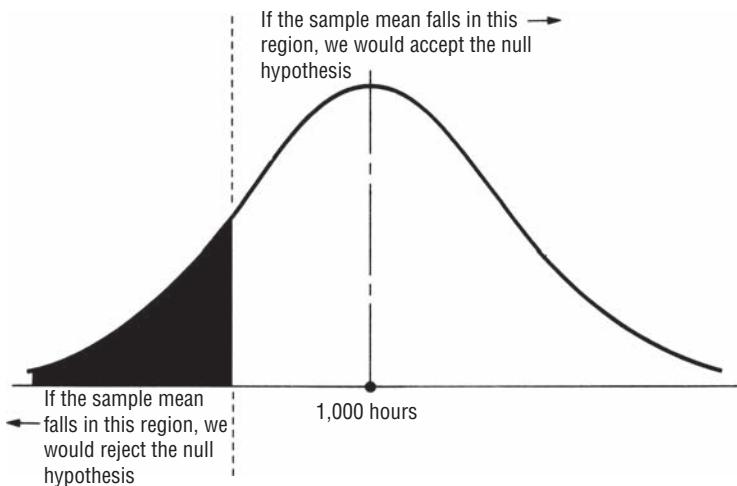
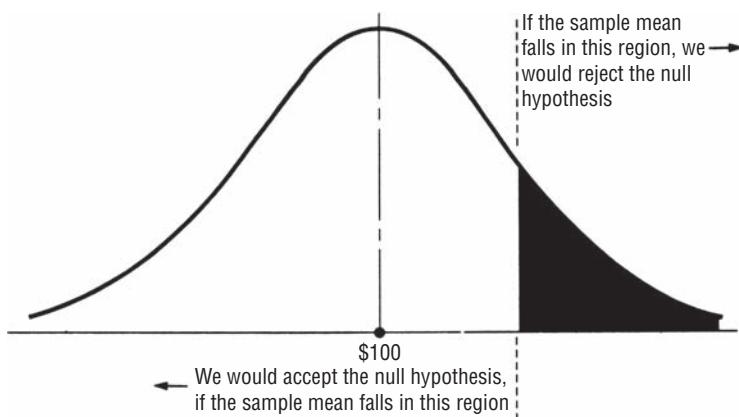


FIGURE 8-6 LEFT-TAILED TEST (A LOWER-TAILED TEST) WITH THE REJECTION REGION ON THE LEFT SIDE (LOWER SIDE)

**FIGURE 8-7 RIGHT-TAILED (UPPER-TAILED) TEST**

is a *right-tailed test* (or an *upper-tailed test*). An upper upper-tailed test is used when the hypotheses are $H_0: \mu = \mu_{H_0}$ and $H_1: \mu > \mu_{H_0}$. Only values of the sample mean that are *significantly above* the hypothesized population mean will cause us to reject the null hypothesis in favor of the alternative hypothesis. This is called an upper-tailed test because the rejection region is in the upper tail of the distribution of the sample mean.

The following situation is illustrated in Figure 8-7; it calls for the use of an upper-tailed test. A sales manager has asked her salespeople to observe a limit on traveling expenses. The manager hopes to keep expenses to an average of \$100 per salesperson per day. One month after the limit is imposed, a sample of submitted daily expenses is taken to see whether the limit is being observed. The null hypothesis is $H_0: \mu = \$100.00$, but the manager is concerned only with excessively high expenses. Thus, the appropriate alternative hypothesis here is $H_1: \mu > \$100.00$, and an upper-tailed test is used. The null hypothesis is rejected (and corrective measures taken) only if the sample mean is significantly higher than \$100.00.

Finally, we should remind you again that in each example of hypothesis testing, when we accept a null hypothesis on the basis of sample information, we are really saying that there is no statistical evidence to reject it. We are not saying that the null hypothesis is true. The only way to *prove* a null hypothesis is to know the population parameter, and that is not possible with sampling. Thus, we accept the null hypothesis and behave as if it is true simply because we can find no evidence to reject it.

Accepting H_0 doesn't guarantee that H_0 is true

HINTS & ASSUMPTIONS

Warning: Don't use sample results to decide whether to use a two-tailed, upper-tailed, or lower-tailed test. *Before* any data are collected, the form of the test is determined by what the decision maker believes or wants to detect. Hint: If marketing researchers suspect that people who purchase Sugar Frosted Flakes also buy *more* sugar than folks who purchase unsweetened cereals, they try to verify their belief by subjecting the data to an *upper-tailed* test. Should the sample mean (surprisingly) turn out smaller than the hypothesized value, that doesn't turn it around into a *lower-tailed* test—the data just don't support their original belief.

EXERCISES 8.3

Self-Check Exercises

- SC 8-3** For the following cases, specify which probability distribution to use in a hypothesis test:
- $H_0: \mu = 27, H_1: \mu \neq 27, \bar{x} = 33, \hat{\sigma} = 4, n = 25.$
 - $H_0: \mu = 98.6, H_1: \mu > 98.6, \bar{x} = 99.1, \sigma = 1.5, n = 50.$
 - $H_0: \mu = 3.5, H_1: \mu < 3.5, \bar{x} = 2.8, \hat{\sigma} = 0.6, n = 18.$
 - $H_0: \mu = 382, H_1: \mu \neq 382, \bar{x} = 363, \sigma = 68, n = 12.$
 - $H_0: \mu = 57, H_1: \mu > 57, \bar{x} = 65, \hat{\sigma} = 12, n = 42.$
- SC 8-4** Martha Inman, a highway safety engineer, decides to test the load-bearing capacity of a bridge that is 20 years old. Considerable data are available from similar tests on the same type of bridge. Which is appropriate, a one-tailed or a two-tailed test? If the minimum load-bearing capacity of this bridge must be 10 tons, what are the null and alternative hypotheses?

Basic Concepts

- 8-13** Formulate null and alternative hypotheses to test whether the mean annual snowfall in Buffalo, New York, exceeds 45 inches.
- 8-14** Describe what the null and alternative hypotheses typically represent in the hypothesis-testing process.
- 8-15** Define the term *significance level*.
- 8-16** Define Type I and Type II errors.
- 8-17** In a trial, the null hypothesis is that an individual is innocent of a certain crime. Would the legal system prefer to commit a Type I or a Type II error with this hypothesis?
- 8-18** What is the relationship between the significance level of a test and Type I error?
- 8-19** If our goal is to accept a null hypothesis that $\mu = 36.5$ with 96 percent certainty when it's true, and our sample size is 50, diagram the acceptance and rejection regions for the following alternative hypotheses:
- $\mu \neq 36.5.$
 - $\mu > 36.5.$
 - $\mu < 36.5.$
- 8-20** For the following cases, specify which probability distribution, to use in a hypothesis test:
- $H_0: \mu = 15, H_1: \mu \neq 15, \bar{x} = 14.8, \hat{\sigma} = 3.0, n = 35.$
 - $H_0: \mu = 9.9, H_1: \mu \neq 9.9, \bar{x} = 10.6, \sigma = 2.3, n = 16.$
 - $H_0: \mu = 42, H_1: \mu > 42, \bar{x} = 44, \sigma = 4.0, n = 10.$
 - $H_0: \mu = 148, H_1: \mu > 148, \bar{x} = 152, \hat{\sigma} = 16.4, n = 29.$
 - $H_0: \mu = 8.6, H_1: \mu < 8.6, \bar{x} = 8.5, \hat{\sigma} = 0.15, n = 24.$
- 8-21** Your null hypothesis is that the battery for a heart pacemaker has an average life of 300 days, with the alternative hypothesis being that the battery life is more than 300 days. You are the quality control engineer for the battery manufacturer.
- Would you rather make a Type I or a Type II error?
 - Based on your answer to part (a), should you use a high or a low significance level?
- 8-22** Under what conditions is it appropriate to use a one-tailed test? A two-tailed test?
- 8-23** If you have decided that a one-tailed test is the appropriate test to use, how do you decide whether it should be a lower-tailed test or an uppertailed test?

Applications

- 8-24** The statistics department installed energy-efficient lights, heaters, and air conditioners last year. Now they want to determine whether the average monthly energy usage has decreased. Should they perform a one- or two-tailed test? If their previous average monthly energy usage was 3,124 kilowatt hours, what are the null and alternative hypotheses?
- 8-25** Dr. Ross Darrow believes that nicotine in cigarettes causes cigarette smokers to have higher daytime heart rates on average than do nonsmokers. He also believes that smokers crave the nicotine in cigarettes rather than just smoking for the physical satisfaction of the act and, accordingly, that the average smoker will smoke more cigarettes per day if he or she switches from a brand with a high nicotine content to one with a low level of nicotine.
- Suppose Ross knows that nonsmokers have an average daytime heart rate of 78 beats per minute. What are the appropriate null and alternative hypotheses for testing his first belief?
 - For the past 3 months, he has been observing a sample of 48 individuals who smoke an average of 15 high-nicotine cigarettes per day. He has just switched them to a brand with a low nicotine content. State null and alternative hypotheses for testing his second belief.

Worked-Out Answers to Self-Cheek Exercises

- SC 8-3** (a) t with 24 df. (b) Normal. (c) t with 17 df.
 (d) Normal. (e) t with 41 df (so we use the normal table).
- SC 8-4** The engineer would be interested in whether a bridge of this age could withstand minimum load-bearing capacities necessary for safety purposes. She therefore wants its capacity to be *above* a certain minimum level, so a one-tailed test (specifically an upper-tailed or right-tailed test) would be used. The hypotheses are

$$H_0: \mu = 10 \text{ tons}$$

$$H_1: \mu > 10 \text{ tons}$$

8.4 HYPOTHESIS TESTING OF MEANS WHEN THE POPULATION STANDARD DEVIATION IS KNOWN

Two-Tailed Tests of Means: Testing in the Scale of the Original Variable

A manufacturer supplies the rear axles for U.S. Postal Service mail trucks. These axles must be able to withstand 80,000 pounds per square inch in stress tests, but an excessively strong axle raises production costs significantly. Long experience indicates that the standard deviation of the strength of its axles is 4,000 pounds per square inch. The manufacturer selects a sample of 100 axles from production, tests them, and finds that the mean stress capacity of the sample is 79,600 pounds per square inch. Written symbolically, the data in this case are

$$\mu_{H_0} = 80,000 \leftarrow \text{Hypothesized value of the population mean}$$

$$\sigma = 4,000 \leftarrow \text{Population standard deviation}$$

$$n = 100 \leftarrow \text{Sample size}$$

$$\bar{x} = 79,600 \leftarrow \text{Sample mean}$$

*Setting up the problem
symbolically*

If the axle manufacturer uses a significance level (α) of 0.05 in testing, will the axles meet his stress requirements? Symbolically, we can state the problem:

$$H_0: \mu = 80,000 \leftarrow \text{Null hypothesis: The true mean is 80,000 pounds per square inch.}$$

$$H_1: \mu \neq 80,000 \leftarrow \text{Alternative hypothesis: The true mean is not 80,000 pounds per square inch.}$$

$$\alpha = 0.05 \leftarrow \text{Level of significance for testing this hypothesis}$$

Because we know the population standard deviation, and because the size of the population is large enough to be treated as infinite, we can use the normal distribution in our testing. First, we calculate the standard error of the mean using Equation 6-1:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} & [6-1] \\ &= \frac{4,000}{\sqrt{100}} \\ &= \frac{4,000}{10} \\ &= 400 \text{ pounds per square inch} \leftarrow \text{Standard error of the mean} \end{aligned}$$

Figure 8-8 illustrates this problem, showing the significance level of 0.05 as the two shaded regions that each contain 0.025 of the area. The 0.95 acceptance region contains two equal areas of 0.475 each. From the normal distribution table (Appendix Table 1), we can see that the appropriate z value for 0.475 of the area under the curve is 1.96. Now we can determine the limits of the acceptance region:

$$\begin{aligned} \mu_{H_0} + 1.96\sigma_{\bar{x}} &= 80,000 + 1.96(400) \\ &= 80,000 + 784 \\ &= 80,784 \text{ pounds per square inch} \leftarrow \text{Upper limit} \end{aligned}$$

Calculating the standard error of the mean

Illustrating the problem

Determining the limits of the acceptance region

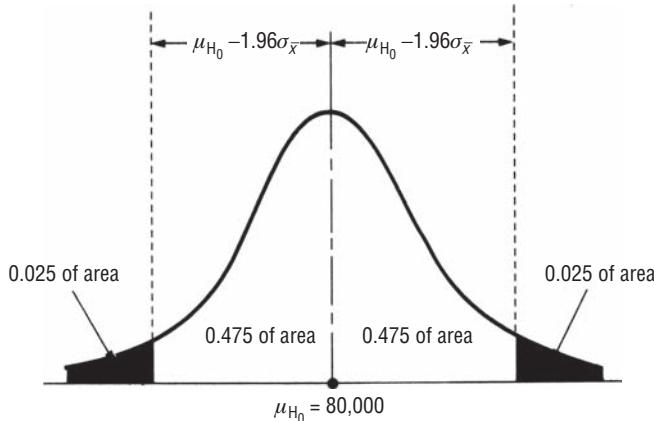


FIGURE 8-8 TWO-TAILED HYPOTHESIS TEST AT THE 0.05 SIGNIFICANCE LEVEL

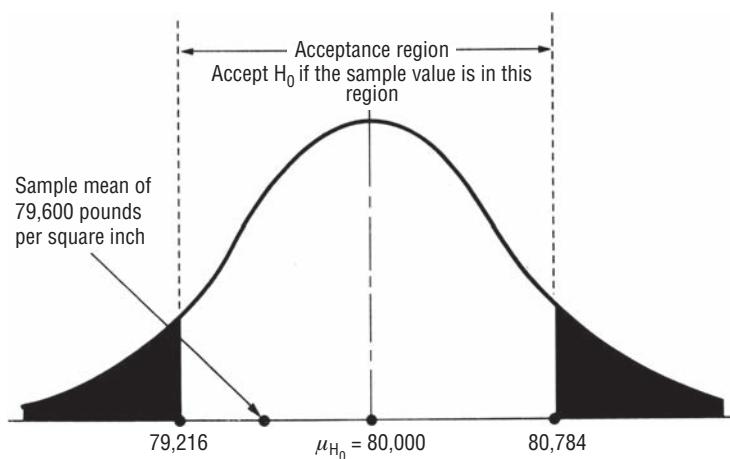


FIGURE 8-9 TWO-TAILED HYPOTHESIS TEST AT THE 0.05 SIGNIFICANCE LEVEL, SHOWING THE ACCEPTANCE REGION AND THE SAMPLE MEAN

and

$$\begin{aligned}\mu_{H_0} - 1.96\sigma_{\bar{x}} &= 80,000 - 1.96(400) \\ &= 80,000 - 784 \\ &= 79,216 \text{ pounds per square inch} \leftarrow \text{Lower limit}\end{aligned}$$

Note that we have defined the limits of the acceptance region (Interpreting the results) (80,784 and 79,216) and the sample mean (79,600), and illustrated them in Figure 8-9 in the scale of the original variable (pounds per square inch). In a moment, we'll show you another way to define the limits of the acceptance region and the value of the sample mean. Obviously, the sample mean lies within the acceptance region; the manufacturer should accept the null hypothesis because there is no significant difference between the hypothesized mean of 80,000 and the observed mean of the sample axles. On the basis of this sample, the manufacturer should accept the production run as meeting the stress requirements.

Hypothesis Testing Using the Standardized Scale

In the hypothesis test we just completed, two numbers were needed to make our decision: an *observed* value computed from the sample, and a *critical value* defining the boundary between the acceptance and rejection regions. Let's look carefully at how we obtained that critical value: After establishing our significance level of $\alpha = 0.05$, we looked in Appendix Table 1—the standard normal probability distribution—to find that ± 1.96 were the z values that left 0.025 of probability in each tail of the distribution.

Recall our discussion of standardizing normal variables in Chapter 5 (pp. 262–267): Instead of measuring the variable in its original units, the standardized variable z tells how many standard deviations above ($z > 0$) or below ($z < 0$) the mean our observation falls. So there are two different scales of measurement we are using, the original scale, or *raw scale*, and the *standardized scale*. Figure 8-10 repeats Figure 8-9, but includes both scales. Notice that our sample mean of 79,600 pounds is given on the raw scale, but that the critical z values of ± 1.96 are given on the standardized scale. **Because these two**

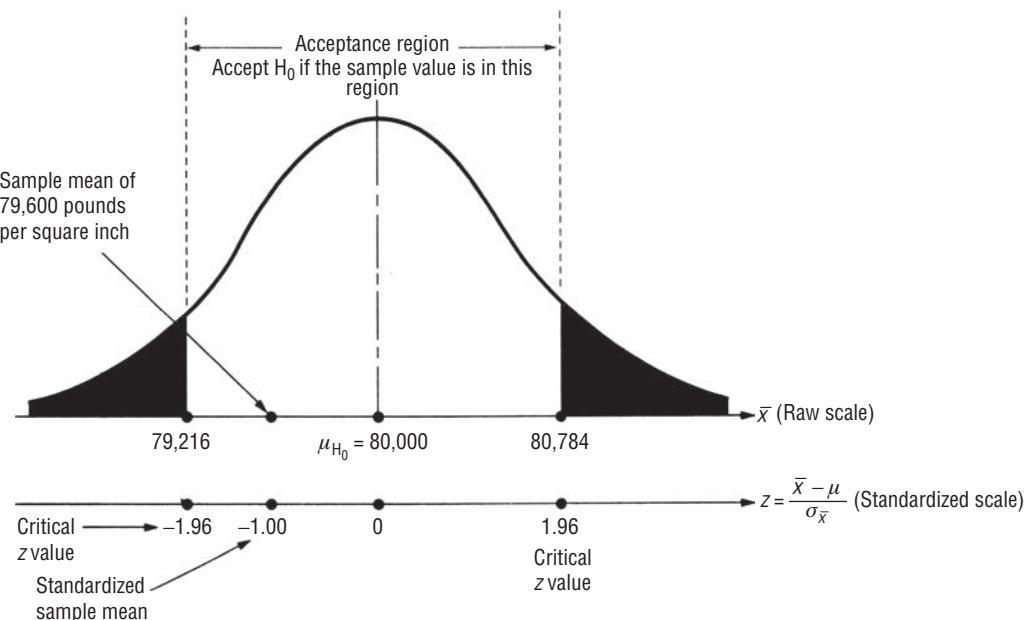


FIGURE 8-10 TWO-TAILED HYPOTHESIS TEST AT THE 0.05 SIGNIFICANCE LEVEL, SHOWING THE ACCEPTANCE REGION AND THE SAMPLE MEAN ON BOTH RAW AND STANDARDIZED SCALES

numbers are given on two different scales, we cannot compare them directly when we test our hypotheses. We must convert one of them to the scale of the other.

We did our hypothesis testing on the original scale by converting the critical z values of ± 1.96 to critical values of \bar{x} on the original scale. Then because the observed value of \bar{x} (79,600) fell between the lower and upper limits of the acceptance region (79,216 and 80,784), we accepted the null hypothesis. Instead of converting the critical z values to the original scale to get numbers directly comparable to the observed value of \bar{x} , we could have converted our observed value of \bar{x} to the standardized scale, using Equation 6-2, to get an observed z value, a number directly comparable to the critical z values:

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_{\bar{x}}}$$

$= \frac{79,600 - 80,000}{400}$

$= -1.00$

The standard error of the mean from Equation 6-1 The sample mean is one standard error below the population mean

Converting the observed value to the standardized scale

In Figure 8-10, we have also illustrated this observed value on the standardized scale. Notice that it falls between the ± 1.96 lower and upper limits of the acceptance region on this scale. Once again, we conclude that H_0 should be accepted: The manufacturer should accept the production run as meeting the stress requirements.

What is the difference between the two methods we have just used to test our hypothesis? Only that we define the units (or scale of measurement) differently in each method. **However, the two methods will always lead to the same conclusions.** Some people are more comfortable using the scale of the original variable; others prefer to use the standardized scale we just explained. The output from most computer statistical packages uses the standardized scale. For the remainder of this chapter and in Chapter 9, we'll test hypotheses using the standardized scale. Our suggestion: Use the method that's more comfortable for you.

How do the two methods differ?

The Five-Step Process for Hypothesis Testing Using the Standardized Scale

Table 8-2 summarizes the five-step process that we will use in the remainder of this chapter and throughout Chapter 9 to test hypotheses.

One-Tailed Test of Means

For a one-tailed test of a mean, suppose a hospital uses large quantities of packaged doses of a particular drug. The individual dose of this drug is 100 cubic centimeters (100 cc). The action of the drug is such that the body will harmlessly pass off excessive doses. On the other hand, insufficient doses do not produce the desired medical effect, and they interfere with patient treatment. The hospital has purchased this drug from the same manufacturer for a number of years and knows that the population standard deviation is 2 cc. The hospital inspects 50 doses of this drug at random from a very large shipment and finds the mean of these doses to be 99.75 cc.

$$\begin{aligned}\mu_{H_0} &= 100 && \leftarrow \text{Hypothesized value of the population mean} \\ \sigma &= 2 && \leftarrow \text{Population standard deviation} \\ n &= 50 && \leftarrow \text{Sample size} \\ \bar{x} &= 99.75 && \leftarrow \text{Sample mean}\end{aligned}$$

TABLE 8-2 SUMMARY OF THE FIVE-STEP PROCESS

Step	Action
1	Decide whether this is a two-tailed or a one-tailed test. State your hypotheses. Select a level of significance appropriate for this decision.
2	Decide which distribution (t or z) is appropriate (see Table 8-1) and find the <i>critical value(s)</i> for the chosen level of significance from the appropriate table.
3	Calculate the <i>standard error of the sample statistic</i> . Use the standard error to convert the observed value of the sample statistic to a standardized value.
4	Sketch the distribution and mark the position of the standardized sample value and the critical value(s) for the test.
5	Compare the value of the standardized sample statistic with the critical value(s) for this test and interpret the result.

If the hospital sets a 0.10 significance level and asks us whether the dosages in this shipment are too small, how can we find the answer?

To begin, we can state the problem symbolically:

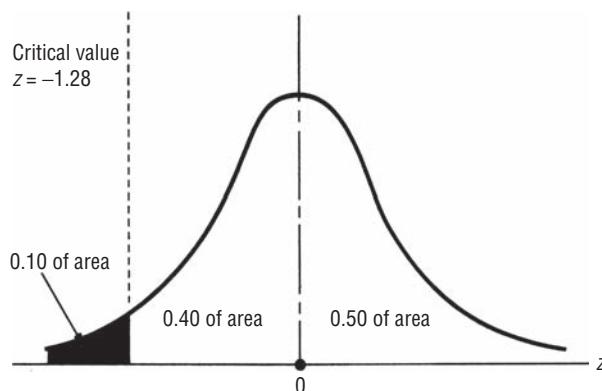
$$\begin{aligned} H_0: \mu = 100 &\leftarrow \text{Null hypothesis: The mean of the shipment's dosages is } 100 \text{ cc} \\ H_1: \mu < 100 &\leftarrow \text{Alternative hypothesis: The mean is less than } 100 \text{ cc} \\ \alpha = 0.10 &\leftarrow \text{Level of significance for testing this hypothesis} \end{aligned}$$

Because we know the population standard deviation, and n is larger than 30, we can use the normal distribution. From Appendix Table 1, we can determine that the value of z for 40 percent of the area under the curve is 1.28, so the critical value for our *lower-tailed* test is -1.28.

The hospital wishes to know whether the actual dosages are 100 cc or whether, in fact, the dosages are too small. The hospital must determine that the dosages are *more* than a certain amount, or it must reject the shipment. This is a *left-tailed* test, which we have shown graphically in Figure 8-11. Notice that the colored region corresponds to the 0.10 significance level. Also notice that the acceptance region consists of 40 percent on the left side of the distribution *plus* the entire right side (50 percent), for a total area of 90 percent.

Now we can calculate the standard error of the mean, using the known population standard deviation and Equation 6-1 (because the population size is large enough to be considered infinite):

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{2}{\sqrt{50}} \\ &= \frac{2}{7.07} \\ &= 0.2829 \text{ cc} \leftarrow \text{Standard error of the mean} \end{aligned} \tag{[6-1]}$$



Step 1: State your hypotheses, type of test, and significance level

Step 2: Choose the appropriate distribution and find the critical value

Step 3: Compute the standard error and standardize the sample statistic

FIGURE 8-11 LEFT-TAILED HYPOTHESIS TEST AT THE 0.10 SIGNIFICANCE LEVEL

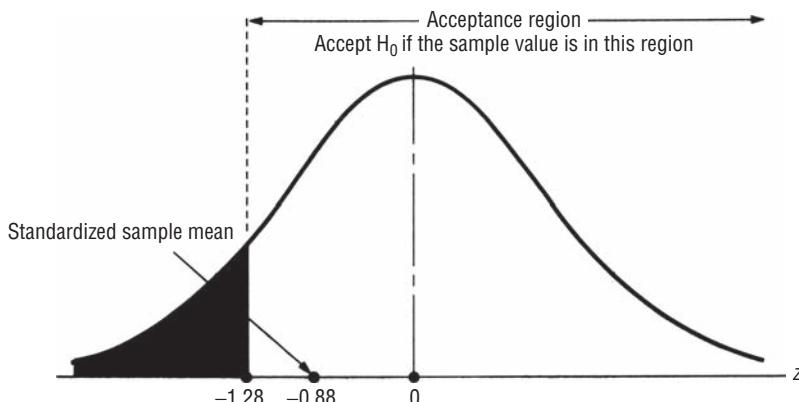


FIGURE 8-12 LEFT-TAILED HYPOTHESIS TEST AT THE 0.10 SIGNIFICANCE LEVEL, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE MEAN

Next we use Equation 6-2 to *standardize* the sample mean, \bar{x} , by subtracting μ_{H_0} , the hypothesized mean, and dividing by $\sigma_{\bar{x}}$, the standard error of the mean.

$$\begin{aligned} z &= \frac{\bar{x} - \mu_{H_0}}{\sigma_{\bar{x}}} \\ &= \frac{99.75 - 100}{0.2829} \\ &= -0.88 \end{aligned} \quad [6-2]$$

Placing the standardized value on the z scale shows that this sample mean falls well within the acceptance region, as shown in Figure 8-12.

Therefore, the hospital should accept the null hypothesis because the observed mean of the sample is not significantly lower than our hypothesized mean of 100 cc. On the basis of this sample of 50 doses, the hospital should conclude that the doses in the shipment are sufficient.

Step 4: Sketch the distribution and mark the sample value and the critical value

Step 5: Interpret the result

HINTS & ASSUMPTIONS

There are a lot of managerial situations that call for a one-tailed test. For example, a concert promoter is interested in attracting enough fans to *break even or more*. If he fills up the coliseum and has to turn away customers, that adds to the prestige of the event but costs him nothing. But failing to attract enough customers can lead to financial problems. He would set up a one-tailed test worded as “greater than or equal to 10,000 fans” (if 10,000 is his break-even point). A water district that is designing pressure limits in its supply system has quite another perspective. If the pressure is too low, customers are inconvenienced and some cannot get an adequate water supply. If the pressure is too high, pipes and hoses can burst. The water engineer is interested in keeping the water pressure *close to a certain value* and would use a two-tailed test. Hint: If the question to be answered is worded as *less than, more than, less than or equal to, or more than or equal to*, a one-tailed test is appropriate. If the question concerns *different from* or *changed from*, use a two-tailed test.

EXERCISES 8.4

Self-Check Exercises

- SC 8-5** Hinton Press hypothesizes that the average life of its largest web press is 14,500 hours. They know that the standard deviation of press life is 2,100 hours. From a sample of 25 presses, the company finds a sample mean of 13,000 hours. At a 0.01 significance level, should the company conclude that the average life of the presses is less than the hypothesized 14,500 hours?
- SC 8-6** American Theaters knows that a certain hit movie ran an average of 84 days in each city, and the corresponding standard deviation was 10 days. The manager of the southeastern district was interested in comparing the movie's popularity in his region with that in all of American's other theaters. He randomly chose 75 theaters in his region and found that they ran the movie an average of 81.5 days.
- (a) State appropriate hypotheses for testing whether there was a significant difference in the length of the picture's run between theaters in the southeastern district and all of American's other theaters.
 - (b) At a 1 percent significance level, test these hypotheses.

Applications

- 8-26** Atlas Sporting Goods has implemented a special trade promotion for its propane stove and feels that the promotion should result in a price change for the consumer. Atlas knows that before the promotion began, the average retail price of the stove was \$44.95, and the standard deviation was \$5.75. Atlas samples 25 of its retailers after the promotion begins and finds the mean price for the stoves is now \$42.95. At a 0.02 significance level, does Atlas have reason to believe that the average retail price to the consumer has decreased?
- 8-27** From 1980 until 1985, the mean price/earnings (P/E) ratio of the approximately 1,800 stocks listed on the New York Stock Exchange was 14.35 and the standard deviation was 9.73. In a sample of 30 randomly chosen NYSE stocks, the mean P/E ratio in 1986 was 11.77. Does this sample present sufficient evidence to conclude (at the 0.05 level of significance) that in 1986 the mean P/E ratio for NYSE stocks had changed from its earlier value?
- 8-28** Generally Electric has developed a new bulb whose design specifications call for a light output of 960 lumens compared to an earlier model that produced only 750 lumens. The company's data indicate that the standard deviation of light output for this type of bulb is 18.4 lumens. From a sample of 20 new bulbs, the testing committee found an average light output of 954 lumens per bulb. At a 0.05 significance level, can Generally Electric conclude that its new bulb is producing the specified 960 lumen output?
- 8-29** Maxwell's Hot Chocolate is concerned about the effect of the recent year-long coffee advertising campaign on hot chocolate sales. The average weekly hot chocolate sales two years ago was 984.7 pounds and the standard deviation was 72.6 pounds. Maxwell's has randomly selected 30 weeks from the past year and found average sales of 912.1 pounds.
- (a) State appropriate hypotheses for testing whether hot chocolate sales have decreased.
 - (b) At the 2 percent significance level, test these hypotheses.
- 8-30** The average commission charged by full-service brokerage firms on a sale of common stock is \$144, and the standard deviation is \$52. Joel Frelander has taken a random sample of 121 trades by his clients and determined that they paid an average commission of \$151. At a 0.10 significance level, can Joel conclude that his clients' commissions are higher than the industry average?

- 8-31** Each day, the United States Customs Service has historically intercepted about \$28 million in contraband goods being smuggled into the country, with a standard deviation of \$16 million per day. On 64 randomly chosen days in 1992, the U.S. Customs Service intercepted an average of \$30.3 million in contraband goods. Does this sample indicate (at a 5 percent level of significance) that the Customs Commissioner should be concerned that smuggling has increased above its historic level?
- 8-32** Before the 1973 oil embargo and subsequent increases in the price of crude oil, gasoline usage in the United States had grown at a seasonally adjusted rate of 0.57 percent per month, with a standard deviation of 0.10 percent per month. In 15 randomly chosen months between 1975 and 1985, gasoline usage grew at an average rate of only 0.33 percent per month. At a 0.01 level of significance, can you conclude that the growth in the use of gasoline had decreased as a result of the embargo and its consequences?
- 8-33** The Bay City Bigleaguers, a semiprofessional baseball team, have the player who led the league in batting average for many years. For the past several years, Joe Carver's batting average has had a mean of .343, and a standard deviation of .018. This year, however, Joe's average was only .306. Joe is renegotiating his contract for next year, and the salary he will be able to obtain is highly dependent on his ability to convince the team's owner that his batting average this year was not significantly worse than in previous years. If the owner is willing to use a 0.02 significance level, will Joe's salary be cut next year?

Worked-Out Answers to Self-Check Exercises

SC 8-5 $\sigma = 2,100$ $n = 25$ $\bar{x} = 13,000$

$$H_0: \mu = 14,500 \quad H_1: \mu < 14,500 \quad \alpha = 0.01$$

The lower limit of the acceptance region is $z = -2.33$, or

$$\bar{x} = \mu_{H_0} - z\sigma/\sqrt{n} = 14,500 - \frac{2.33(2,100)}{\sqrt{25}} = 13,521.4 \text{ hours}$$

$$\text{Because the observed } z \text{ value} = \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} = \frac{13,000 - 14,500}{2,100/\sqrt{25}} = -3.57 < -2.33$$

(or $\bar{x} < 13,521.4$), we should reject H_0 . The average life is significantly less than the hypothesized value.

SC 8-6 $\sigma = 10$ $n = 75$ $\bar{x} = 81.5$

$$H_0: \mu = 84 \quad H_1: \mu \neq 84 \quad \alpha = 0.01$$

The limits of the acceptance region are $z = \pm 2.58$, or

$$\bar{x} = \mu_{H_0} \pm z\sigma/\sqrt{n} = 84 \pm \frac{2.58(10)}{\sqrt{75}} = (81.02, 86.98) \text{ days}$$

$$\text{Because the observed } z \text{ value} = \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} = \frac{81.5 - 84}{10/\sqrt{75}} = -2.17, \text{ it and } \bar{x} \text{ are in the acceptance}$$

region, so we do not reject H_0 . The length of run in the southeast is not significantly different from the length of run in other regions.

8.5 MEASURING THE POWER OF A HYPOTHESIS TEST

Now that we have considered two examples of hypothesis testing, a step back is appropriate to discuss what a good hypothesis test *should do*. Ideally, α and β (the probabilities of Type I and Type II errors) should both be small. Recall that a Type I error occurs when we reject a null hypothesis that is true, and that α (the significance level of the test) is the probability of making a Type I error. In other words, once we decide on the significance level, there is nothing else we can do about α . A Type II error occurs when we accept a null hypothesis that is false; the probability of a Type II error is β . What can we say about β ?

What should a good hypothesis test do?

Suppose the null hypothesis *is* false. Then managers would like the hypothesis test to reject it all the time. Unfortunately, hypothesis tests cannot be foolproof; sometimes when the null hypothesis is false, a test does not reject it, and thus a Type II error is made. When the null hypothesis is false, μ (the *true* population mean) does not equal μ_{H_0} (the *hypothesized* population mean); instead, μ equals some other value. For each possible value of μ for which the alternative hypothesis is true, there is a different probability (β) of incorrectly accepting the null hypothesis. Of course, we would like this β (the probability of accepting a null hypothesis when it is false) to be as small as possible, or, equivalently, we would like $1 - \beta$ (the probability of rejecting a null hypothesis when it is false) to be as large as possible.

Meaning of β and $1 - \beta$

Because rejecting a null hypothesis when it is false is exactly what a good test should do, a high value of $1 - \beta$ (something near 1.0) means the test is working quite well (it is rejecting the null hypothesis when it is false); a low value of $1 - \beta$ (something near 0.0) means that the test is working very poorly (it's not rejecting the null hypothesis when it is false). Because the value of $1 - \beta$ is the measure of how well the test is working, it is known as the power of the test. If we plot the values of $1 - \beta$ for each value of μ for which the alternative hypothesis is true, the resulting curve is known as a *power curve*.

Interpreting the values of $1 - \beta$

In part *a* of Figure 8-13, we reproduce the left-tailed test from Figure 8-11, but now we are looking at the raw scale. In Figure 8-13(b), we show the power curve associated with this test. Computing the values of $1 - \beta$ to plot the power curve is not difficult; three such points are shown in Figure 8-13(b). Recall that with this test we were deciding whether to accept a drug shipment. Our test dictated that we should reject the null hypothesis if the standardized sample mean is less than -1.28 , that is, if sample mean dosage is less than $100.00 - 1.28$ (0.2829), or 99.64 cc.

Computing the values of $1 - \beta$

Consider point *C* on the power curve in Figure 8-13(b). The population mean dosage is 99.42 cc. Given that the population mean is 99.42 cc, we must compute the probability that the mean of a random sample of 50 doses from this population will be less than 99.64 cc (the point below which we decided to reject the null hypothesis). Now look at Figure 8-13(c). Earlier we computed the standard error of the mean to be 0.2829 cc, so 99.64 cc is $(99.64 - 99.42)/0.2829$, or 0.78 standard error above 99.42 cc. Using Appendix Table 1, we can see that the probability of observing a sample mean less than 99.64 cc and thus rejecting the null hypothesis is 0.7823, the colored area in Figure 8-13(c). Thus, the power of the test ($1 - \beta$) at $\mu = 99.42$ is 0.7823. This simply means that if $\mu = 99.42$, the probability that this test will reject the null hypothesis when it is false is 0.7823.

Interpreting a point on the power curve

Now look at point *D* in Figure 8-13(b). For this population mean dosage of 99.61 cc, what is the probability that the mean of a random sample of 50 doses from this population will be less than 99.64 cc and thus cause the test to reject the null hypothesis? Look at Figure 8-13(d). Here we see that 99.64 is

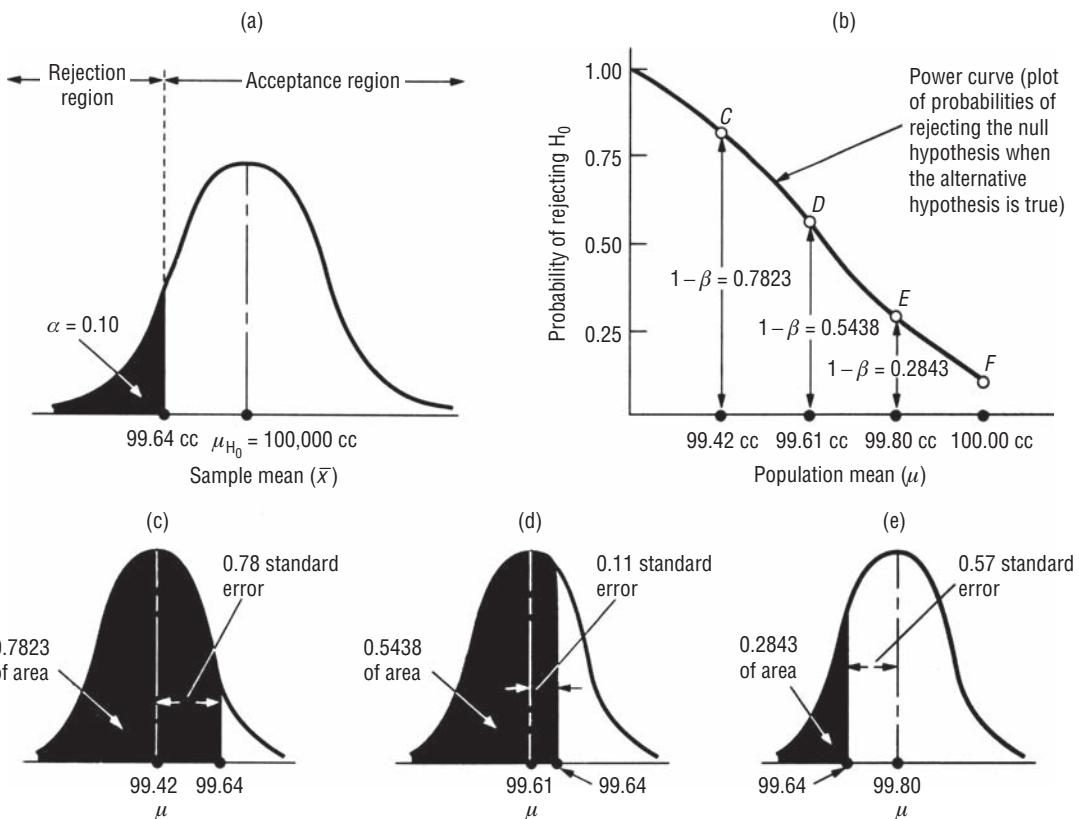


FIGURE 8-13 LEFT-TAILED HYPOTHESIS TEST, ASSOCIATED POWER CURVE, AND THREE VALUES OF μ .

$(99.64 - 99.61)/0.2829$, or 0.11 standard error above 99.61 cc. Using Appendix Table 1 again, we can see that the probability of observing a sample mean less than 99.64 cc and thus rejecting the null hypothesis is 0.5438, the colored area in Figure 8-13(d). Thus, the power of the test ($1 - \beta$) at $\mu = 99.61$ cc is 0.5438.

Using the same procedure at point E, we find the power of the test at $\mu = 99.80$ cc is 0.2843; this is illustrated as the colored area in Figure 8-13(e). The values of $1 - \beta$ continue to decrease to the right of point E. How low do they get? As the population mean gets closer and closer to 100.00 cc, the power of the test ($1 - \beta$) must get closer and closer to the probability of rejecting the null hypothesis when the population mean is exactly 100.00 cc. And we know that probability is nothing but the significance level of the test—in this case, 0.10. Thus, the curve terminates at point F, which lies at a height of 0.10 directly over the population mean.

What does our power curve in Figure 8-13(b) tell us? Just that as the shipment becomes less satisfactory (as the doses in the shipment become smaller), our test is more powerful (it has a greater probability of recognizing that the shipment is unsatisfactory). It also shows us, however, that because of sampling error, when the dosage is only slightly less than 100.00 cc, the power of the test to recognize this situation is quite low. Thus, if having *any* dosage below 100.00 cc is completely unsatisfactory, the test we have been discussing would not be appropriate.

Termination point of the power curve

Interpreting the power curve

HINTS & ASSUMPTIONS

Of course, we'd always like to use the hypothesis test with the greatest power. But we also know that a certain proportion of the time, all hypothesis tests will fail to reject the null hypothesis when it is false or accept it when it's true (that's statistical language that really means that when the test does fail, it will persuade us that things haven't changed when in fact they have, or persuade us that things have changed when they really haven't). That's just the price we pay for using sampling in hypothesis testing. The failure of a test to perform perfectly is due to sampling error. The only way to avoid such error is to examine everything in the population and that is either physically impossible or too expensive.

EXERCISES 8.5**Self-Check Exercises**

- SC 8-7** See Exercise 8-32. Compute the power of the test for $\mu = 0.50$, 0.45 , and 0.40 percent per month.
- SC 8-8** In Exercise 8-32, what happens to the power of the test for $\mu = 0.50$, 0.45 , and 0.40 percent per month if the significance level is changed to 0.04 ?

Applications

- 8-34** See Exercise 8-31. Compute the power of the test for $\mu = \$28$, $\$29$, and $\$30$ million.
- 8-35** See Exercise 8-30. Compute the power of the test for $\mu = \$140$, $\$160$, and $\$175$.
- 8-36** In Exercise 8-31, what happens to the power of the test for $\mu = \$28$, $\$29$, and $\$30$ million if the significance level is changed to 0.02 ?
- 8-37** In Exercise 8-30, what happens to the power of the test for $\mu = \$140$, $\$160$, and $\$175$ if the significance level is changed to 0.05 ?

Worked-Out Answers to Self-Check Exercises

- SC 8-7** From Exercise 8-32, we have $\sigma = 0.10$, $n = 15$, $H_0: \mu = 0.57$, $H_1: \mu < 0.57$. At $\alpha = 0.01$, the lower limit of the acceptance region is

$$\mu_{H_0} - 2.33\sigma/\sqrt{n} = 0.57 - 2.33(0.10)/\sqrt{15} = 0.510$$

- (a) At $\mu = 0.50$, the power of the test is

$$P(\bar{x} < 0.510) = P(z < \frac{0.510 - 0.50}{0.10/\sqrt{15}}) = P(z < 0.39) = 0.5 + 0.1517 = 0.6517$$

- (b) At $\mu = 0.45$, the power of the test is

$$P(\bar{x} < 0.510) = P(z < \frac{0.510 - 0.45}{0.10/\sqrt{15}}) = P(z < 2.32) = 0.5 + 0.4898 = 0.9898$$

- (c) At $\mu = 0.40$, the power of the test is

$$P(\bar{x} < 0.510) = P(z < \frac{0.510 - 0.40}{0.10/\sqrt{15}}) = P(z < 4.26) = 1.0000$$

SC 8-8 At $\alpha = 0.04$, the lower limit of the acceptance region is

$$\mu_{H_0} - 1.75\sigma / \sqrt{n} = 0.57 - 1.75(0.10) / \sqrt{15} = 0.525$$

(a) At $\mu = 0.50$, the power of the test is

$$P(\bar{x} < 0.525) = P(z < \frac{0.525 - 0.50}{0.10 / \sqrt{15}}) = P(z < 0.97) = 0.5 + 0.3340 = 0.8340$$

(b) At $\mu = 0.45$, the power of the test is

$$P(\bar{x} < 0.525) = P(z < \frac{0.525 - 0.45}{0.10 / \sqrt{15}}) = P(z < 2.90) = 0.5 + 0.4981 = 0.9981$$

(c) At $\mu = 0.40$, the power of the test is

$$P(\bar{x} < 0.525) = P(z < \frac{0.525 - 0.40}{0.10 / \sqrt{15}}) = P(z < 4.84) = 1.0000$$

8.6 HYPOTHESIS TESTING OF PROPORTIONS: LARGE SAMPLES

Two-Tailed Tests of Proportions

In this section, we'll apply what we have learned about tests concerning means to tests for *proportions* (that is, the proportion of occurrences in a population). But before we apply it, we'll review the important conclusions we made about proportions in Chapter 7. First, remember that the binomial is the theoretically correct distribution to use in dealing with proportions. As the sample size increases, the binomial distribution approaches the normal in its characteristics, and we can use the normal distribution to approximate the sampling distribution. Specifically, np and nq each need to be at least 5 before we can use the normal distribution as a substitute for the binomial.

Dealing with proportions

Consider, as an example, a company that is evaluating the promotability of its employees, that is, determining the proportion whose ability, training, and supervisory experience qualify them for promotion to the next higher level of management. The human resources director tells the president that roughly 80 percent, or 0.8, of the employees in the company are "promotable." The president assembles a special committee to assess the promotability of all employees. This committee conducts in-depth interviews with 150 employees and finds that in its judgment only 70 percent of the sample are qualified for promotion.

$p_{H_0} = 0.8$ ← Hypothesized value of the population proportion of successes (judged promotable, in this case)

$q_{H_0} = 0.2$ ← Hypothesized value of the population proportion of failures (judged not promotable)

$n = 150$ ← Sample size

$\bar{p} = 0.7$ ← Sample proportion of promotables

$\bar{q} = 0.3$ ← Sample proportion judged not promotable

Step 1: State your hypotheses,

The president wants to test at the 0.05 significance level the hypothesis that 0.8 of the employees are promotable:

type of test, and significance level

$H_0: p = 0.8$ ← Null hypothesis, 80 percent of the employees are promotable

$H_1: p \neq 0.8$ ← Alternative hypothesis. The proportion of promotable employees is not 80 percent

$\alpha = 0.05$ ← Level of significance for testing the hypothesis

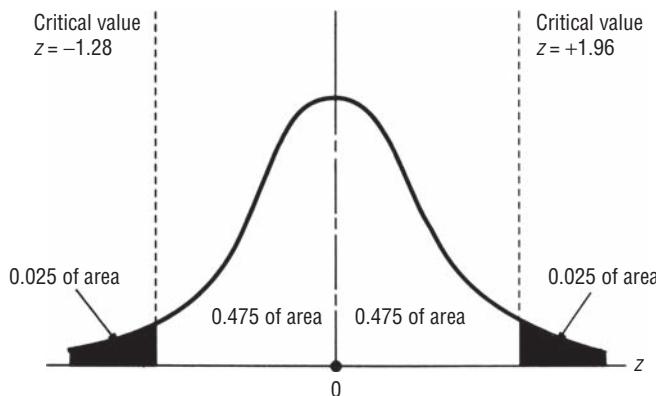


FIGURE 8-14 TWO-TAILED HYPOTHESIS TEST OF A PROPORTION AT THE 0.05 LEVEL OF SIGNIFICANCE

In this instance, the company wants to know whether the true proportion is larger or smaller than the hypothesized proportion. Thus, a two-tailed test of a proportion is appropriate, and we have shown it graphically in Figure 8-14. The significance level corresponds to the two colored regions, each containing 0.025 of the area. The acceptance region of 0.95 is illustrated as two areas of 0.475 each. Because np and nq are each larger than 5, we can use the normal approximation of the binomial distribution. From Appendix Table 1, we can determine that the critical value of z for 0.475 of the area under the curve is 1.96.

We can calculate the standard error of the proportion, using the hypothesized values of P_{H_0} and Q_{H_0} in Equation 7-4:

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{P_{H_0}Q_{H_0}}{n}} \\ &= \sqrt{\frac{(0.8)(0.2)}{150}} \\ &= \sqrt{0.0010666} \\ &= 0.0327 \quad \leftarrow \text{Standard error of the proportion}\end{aligned}$$

Step 2: Choose the appropriate distribution and find the critical value

Step 3: Compute the standard error and standardize the sample statistic

Next we standardize the sample proportion by dividing the difference between the observed sample proportion, \bar{p} , and the hypothesized proportion, P_{H_0} , by the standard error of the proportion.

$$\begin{aligned}z &= \frac{\bar{p} - P_{H_0}}{\sigma_{\bar{p}}} \\ &= \frac{0.7 - 0.8}{0.0327} \\ &= -3.06\end{aligned}$$

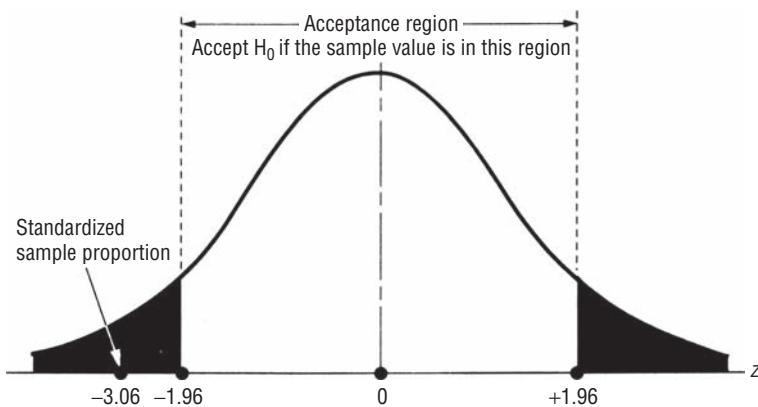


FIGURE 8-15 TWO-TAILED HYPOTHESIS TEST OF A PROPORTION AT THE 0.05 SIGNIFICANCE LEVEL, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE PROPORTION

By marking the calculated standardized sample proportion, -3.06 , on a sketch of the sampling distribution, it is clear that this sample falls outside the region of acceptance, as shown in Figure 8-15.

Therefore, in this case, the president should reject the null hypothesis and conclude that there *is* a significant difference between the director of human resources' hypothesized proportion of promotable employees (0.8) and the observed proportion of promotable employees in the sample. From this, he should infer that the true proportion of promotable employees in the entire company is not 80 percent.

Step 4: Sketch the distribution and mark the sample value and the critical values

Step 5: Interpret the result

One-Tailed Tests of Proportions

A one-tailed test of a proportion is conceptually equivalent to a one-tailed test of a mean, as can be illustrated with this example. A member of a public interest group concerned with environmental pollution asserts at a public hearing that “fewer than 60 percent of the industrial plants in this area are complying with air-pollution standards.” Attending this meeting is an official of the Environmental Protection Agency who believes that 60 percent of the plants *are* complying with the standards; she decides to test that hypothesis at the 0.02 significance level.

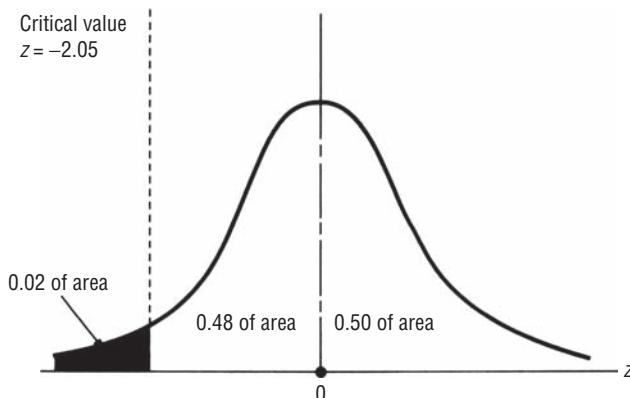
$H_0: p = 0.6$ ← Null hypothesis: The proportion of plants complying with the air-pollution standards is 0.6

$H_1: p < 0.6$ ← Alternative hypothesis: The proportion complying with the standards is less than 0.6

$\alpha = 0.02$ ← Level of significance for testing the hypothesis

Step 1: State your hypotheses, type of test, and significance level

The official makes a thorough search of the records in her office. She samples 60 plants from a population of over 10,000 plants and finds that 33 are complying with air-pollution standards. Is the assertion by the member of the public interest group a valid one?

**FIGURE 8-16 ONE-TAILED HYPOTHESIS TEST AT THE 0.02 LEVEL OF SIGNIFICANCE**

We begin by summarizing the case symbolically:

$P_{H_0} = 0.6 \leftarrow$ Hypothesized value of the population proportion that is complying with air-pollution standards

$q_{H_0} = 0.4 \leftarrow$ Hypothesized value of the population proportion that is not complying and thus polluting

$n = 60 \leftarrow$ Sample size

$\bar{p} = 33/60, \text{ or } 0.55 \leftarrow$ Sample proportion complying

$\bar{q} = 27/60, \text{ or } 0.45 \leftarrow$ Sample proportion polluting

This is a one-tailed test: The EPA official wonders only whether the actual proportion is less than 0.6. Specifically, this is a left-tailed test. In order to reject the null hypothesis that the true proportion of plants in compliance is 60 percent, the EPA representative must accept the alternative hypothesis that fewer than 0.6 have complied. In Figure 8-16, we have shown this hypothesis test graphically.

Because np and nq are each over 5, we can use the normal approximation of the binomial distribution. The critical value of z from Appendix Table 1 for 0.48 of the area under the curve is 2.05.

Next, we can calculate the standard error of the proportion using the hypothesized population proportion as follows:

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{P_{H_0} q_{H_0}}{n}} \\ &= \sqrt{\frac{(0.6)(0.4)}{60}} \\ &= \sqrt{0.004} \\ &= 0.0632 \leftarrow \text{Standard error of the proportion}\end{aligned}$$

Step 2: Choose the appropriate distribution and find the critical value

Step 3: Compute the standard error and standardize the sample statistic

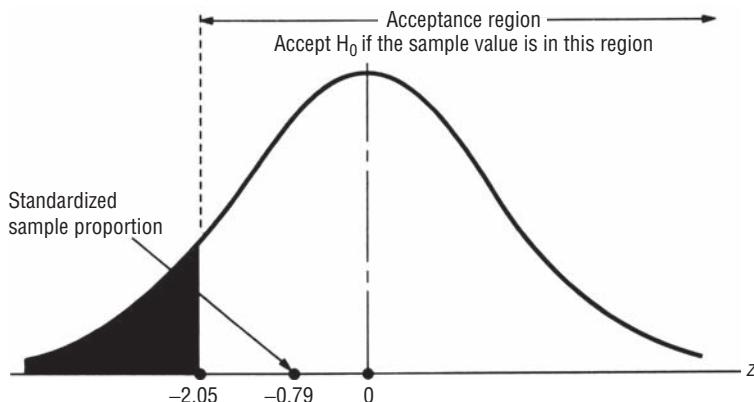


FIGURE 8-17 ONE-TAILED (LEFT-TAILED) HYPOTHESIS TEST AT THE 0.02 SIGNIFICANCE LEVEL, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE PROPORTION

And we standardize the sample proportion by dividing the difference between the observed sample proportion, \bar{p} , and the hypothesized proportion, P_{H_0} , by the standard error of the proportion.

$$\begin{aligned} z &= \frac{\bar{p} - P_{H_0}}{\sigma_{\bar{p}}} \\ &= \frac{0.55 - 0.6}{0.0632} \\ &= -0.79 \end{aligned}$$

Figure 8-17 illustrates where the sample proportion lies in relation to the critical value, -2.05 . Looking at this figure, we can see that the sample proportion lies within the acceptance region. Therefore, the EPA official should accept the null hypothesis that the true proportion of complying plants is 0.6 . **Although the observed sample proportion is below 0.6 , it is not significantly below 0.6 ;** that is, it is not far enough below 0.6 to make us accept the assertion by the member of the public interest group.

Step 4: Sketch the distribution and mark the sample value and the critical value

Step 5: Interpret the result

HINTS & ASSUMPTIONS

Warning: When we're doing hypothesis tests involving proportions, we use the binomial distribution as the sampling distribution, unless np and nq are *both* at least 5. In that case, we can use the normal distribution as an approximation of the binomial without worry. Fortunately, in practice, hypothesis tests about proportions almost always involve sufficiently large samples so that this condition is met. Even when they aren't, the arithmetic of the binomial distribution and the binomial table is not that difficult to use.

EXERCISES 8.6

Self-Check Exercises

- SC 8-9** A ketchup manufacturer is in the process of deciding whether to produce a new extra-spicy brand. The company's marketing-research department used a national telephone survey of 6,000 households and found that the extra-spicy ketchup would be purchased by 335 of them. A much more extensive study made 2 years ago showed that 5 percent of the households would purchase the brand then. At a 2 percent significance level, should the company conclude that there is an increased interest in the extra-spicy flavor?
- SC 8-10** Steve Cutter sells Big Blade lawn mowers in his hardware store, and he is interested in comparing the reliability of the mowers he sells with the reliability of Big Blade mowers sold nationwide. Steve knows that only 15 percent of all Big Blade mowers sold nationwide require repairs during the first year of ownership. A sample of 120 of Steve's customers revealed that exactly 22 of them required mower repairs in the first year of ownership. At the 0.02 level of significance, is there evidence that Steve's Big Blade mowers differ in reliability from those sold nationwide?

Applications

- 8-38** Grant, Inc., a manufacturer of women's dress blouses, knows that its brand is carried in 19 percent of the women's clothing stores east of the Mississippi River. Grant recently sampled 85 women's clothing stores on the West Coast and found that 14.12 percent of the stores carried the brand. At the 0.04 level of significance, is there evidence that Grant has poorer distribution on the West Coast than it does east of the Mississippi?
- 8-39** From a total of 10,200 loans made by a state employees' credit union in the most recent 5-year period, 350 were sampled to determine what proportion was made to women. This sample showed that 39 percent of the loans were made to women employees. A complete census of loans 5 years ago showed that 41 percent of the borrowers then were women. At a significance level of 0.02, can you conclude that the proportion of loans made to women has changed significantly in the past 5 years?
- 8-40** Feronetics specializes in the use of gene-splicing techniques to produce new pharmaceutical compounds. It has recently developed a nasal spray containing interferon, which it believes will limit the transmission of the common cold within families. In the general population, 15.1 percent of all individuals will catch a rhinovirus-caused cold once another family member contracts such a cold. The interferon spray was tested on 180 people, one of whose family members subsequently contracted a rhinovirus-caused cold. Only 17 of the test subjects developed similar colds.
- At a significance level of 0.05, should Feronetics conclude that the new spray effectively reduces transmission of colds?
 - What should it conclude at $\alpha = 0.02$?
 - On the basis of these results, do you think Feronetics should be allowed to market the new spray? Explain.
- 8-41** Some financial theoreticians believe that the stock market's daily prices constitute a "random walk with positive drift." If this is accurate, then the Dow Jones Industrial Average should show a gain on more than 50 percent of all trading days. If the average increased on 101 of 175 randomly chosen days, what do you think about the suggested theory? Use a 0.01 level of significance.

- 8-42** MacroSwift estimated last year that 35 percent of potential software buyers were planning to wait to purchase the new operating system, Window Panes, until an upgrade had been released. After an advertising campaign to reassure the public, MacroSwift surveyed 3,000 people and found 950 who were still skeptical. At the 5 percent significance level, can the company conclude the proportion of skeptical people has decreased?
- 8-43** Rick Douglas, the new manager of Food Barn, is interested in the percentage of customers who are totally satisfied with the store. The previous manager had 86 percent of the customers totally satisfied, and Rick claims the same is true today. Rick sampled 187 customers and found 157 were totally satisfied. At the 1 percent significance level, is there evidence that Rick's claim is valid?

Worked-Out Answers to Self-Check Exercises

SC 8-9 $n = 6,000 \quad \bar{p} = 335/6,000 = 0.05583$

$$H_0: p = 0.05 \quad H_1: p > 0.05 \quad \alpha = 0.02$$

The upper limit of the acceptance region is $z = 2.05$, or

$$\bar{p} = p_{H_0} + z\sqrt{\frac{p_{H_0}q_{H_0}}{n}} = 0.05 + 2.05\sqrt{\frac{0.05(0.95)}{6,000}} = 0.05577$$

$$\text{Because the observed } z \text{ value} = \frac{\bar{p} - p_{H_0}}{\sqrt{p_{H_0}q_{H_0}/n}} = \frac{0.05583 - 0.05}{\sqrt{0.05(0.95)/6,000}} = 2.07$$

> 2.05 (or $\bar{p} > 0.05577$), we should reject H_0 (but just barely). The current interest is significantly greater than the interest of 2 years ago.

SC 8-10 $n = 120 \quad \bar{p} = 22/120 = 0.1833$

$$H_0: p = 0.15 \quad H_1: p \neq 0.15 \quad \alpha = 0.02$$

The limits of the acceptance region are $z = \pm 2.33$, or

$$\bar{p} = p_{H_0} \pm z\sqrt{\frac{p_{H_0}q_{H_0}}{n}} = 0.15 \pm 2.33\sqrt{\frac{0.15(0.85)}{120}} = (0.0741, 0.2259)$$

$$\text{Because the observed } z \text{ value} = \frac{\bar{p} - p_{H_0}}{\sqrt{p_{H_0}q_{H_0}/n}} = \frac{0.1833 - 0.15}{\sqrt{0.15(0.85)/120}} = 1.02$$

< 2.33 (or $\bar{p} = 0.1833$, which is between 0.0741 and 0.2259), we do not reject H_0 . Steve's mowers are not significantly different in reliability from those sold nationwide.

8.7 HYPOTHESIS TESTING OF MEANS WHEN THE POPULATION STANDARD DEVIATION IS NOT KNOWN

When we estimated confidence intervals in Chapter 7, we learned that the difference in size between large and small samples is important when the population standard deviation σ is unknown and must be estimated from the sample standard deviation. If the sample size n is 30 or less and σ is not known, we should use the t distribution. The appropriate t distribution has $n - 1$ degrees of freedom. These rules apply to hypothesis testing, too.

When to use the t distribution

Two-Tailed Tests of Means Using the *t* Distribution

A personnel specialist of a major corporation is recruiting a large number of employees for an overseas assignment. During the testing process, management asks how things are going, and she replies, "Fine. I think the average score on the aptitude test will be around 90." When management reviews 20 of the test results compiled, it finds that the mean score is 84, and the standard deviation of this score is 11.

$$\begin{aligned}\mu_{H_0} &= 90 && \leftarrow \text{Hypothesized value of the population mean} \\ n &= 20 && \leftarrow \text{Sample size} \\ \bar{x} &= 84 && \leftarrow \text{Sample mean} \\ s &= 11 && \leftarrow \text{Sample standard deviation}\end{aligned}$$

If management wants to test her hypothesis at the 0.10 level of significance, what is the procedure?

$$\begin{aligned}H_0: \mu &= 90 \leftarrow \text{Null hypothesis, the true population mean score is 90} \\ H_1: \mu &\neq 90 \leftarrow \text{Alternative hypothesis, the mean score is not 90} \\ \alpha &= 0.10 \leftarrow \text{Level of significance for testing this hypothesis}\end{aligned}$$

Step 1: State your hypotheses, type of test, and significance level

Figure 8-18 illustrates this problem graphically. Because management is interested in knowing whether the true mean score is *larger or smaller* than the hypothesized score, a *two-tailed test* is the appropriate one to use. The significance level of 0.10 is shown in Figure 8-18 as the two colored areas, each containing 0.05 of the area under the *t* distribution. Because the sample size is 20, the appropriate number of degrees of freedom is 19, that is, $20 - 1$. Therefore, we look in the *t* distribution table, Appendix Table 2, under the 0.10 column until we reach the 19 degrees of freedom row. There we find the critical value of *t*, 1.729.

Step 2: Choose the appropriate distribution and find the critical value

Because the population standard deviation is not known, we must estimate it using the sample standard deviation and Equation 7-1:

$$\hat{\sigma} = s \\ = 11 \quad [7-1]$$

Now we can compute the standard error of the mean. Because we are using $\hat{\sigma}$, an estimate of the population standard deviation, the standard error of the mean will also be an estimate. We can use Equation 7-6, as follows:

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \frac{\hat{\sigma}}{\sqrt{n}} \\ &= \frac{11}{\sqrt{20}} \\ &= \frac{11}{4.47} \\ &= 2.46 \leftarrow \text{Estimated standard error of the mean}\end{aligned} \quad [7-6]$$

Step 3: Compute the standard error and standardize the sample statistic

Next we standardize the sample mean, \bar{x} , by subtracting μ_{H_0} , the hypothesized mean, and dividing by $\hat{\sigma}_{\bar{x}}$, the estimated standard error of the mean. Because our test of hypotheses is based on the *t* distribution,

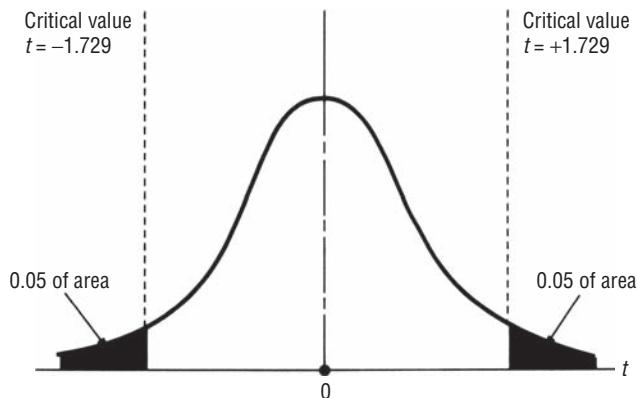


FIGURE 8-18 TWO-TAILED TEST OF HYPOTHESIS AT THE 0.10 LEVEL OF SIGNIFICANCE USING THE t DISTRIBUTION

we use t to denote the standardized statistic.

$$\begin{aligned} t &= \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_{\bar{x}}} \\ &= \frac{84 - 90}{2.46} \\ &= -2.44 \end{aligned}$$

Drawing this result on a sketch of the sampling distribution, we see that the sample mean falls outside the acceptance region, as shown in Figure 8-19.

Therefore, management should reject the null hypothesis (the personnel specialist's assertion that the true mean score of the employees being tested is 90).

Step 4 Sketch the distribution and mark the sample value and the critical values

Step 5: Interpret the result

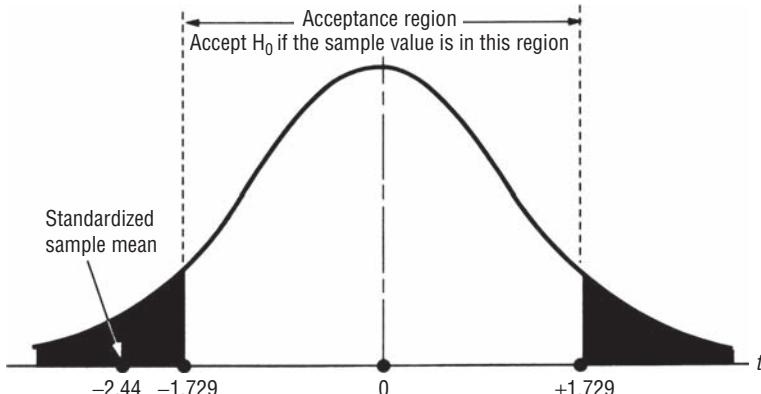


FIGURE 8-19 TWO-TAILED HYPOTHESIS TEST AT THE 0.10 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE MEAN

One-Tailed Tests of Means Using the *t* Distribution

The procedure for a one-tailed hypothesis test using the *t* distribution is the same conceptually as for a one-tailed test using the normal distribution and the *z* table. Performing such one-tailed tests may cause some difficulty, however. Notice that the column headings in Appendix Table 2 represent the *area in both tails combined*. Thus, they are appropriate to use in a two-tailed test with two rejection regions.

One difference from the *z* tables

If we use the *t* distribution for a one-tailed test, we need to determine the area located in only one tail. So to find the appropriate *t* value for a one-tailed test at a significance level of 0.05 with 12 degrees of freedom, we would look in Appendix Table 2 under the 0.10 column opposite the 12 degrees of freedom row. The answer in this case is 1.782. **This is true because the 0.10 column represents 0.10 of the area under the curve contained in both tails combined, so it also represents 0.05 of the area under the curve contained in each of the tails separately.**

Using the *t* table for one-tailed tests

In the next chapter, we continue our work on hypothesis testing by looking at situations where decisions must be made on the basis of two samples that may or may not come from the same underlying population.

Looking ahead

HINTS & ASSUMPTIONS

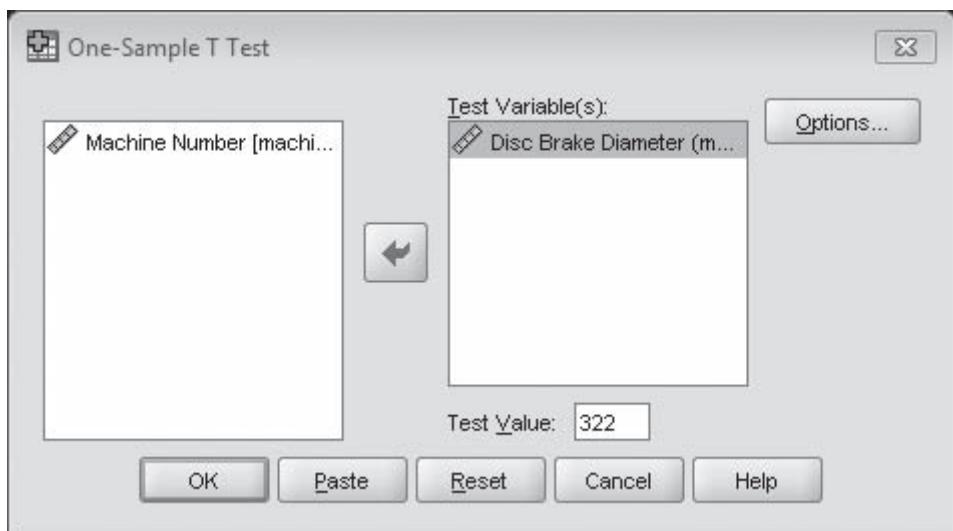
Doing hypothesis tests with the *t* distribution is no different from doing them with the normal distribution except that we use a different table and we have to supply the number of degrees of freedom. Hint: The number of degrees of freedom in a single-sample test is always one fewer than the sample size. Warning: Use the *t* distribution whenever the sample size is less than 30, the population standard deviation is not known, and the population is normal or approximately normal.

One Sample Test Using SPSS

1 machine		1	brake	var											
1		1	322.0003												
2		1	322.0048												
3		1	322.0215												
4		1	321.9907												
5		1	322.0109												
6		1	321.9954												
7		1	322.0059												
8		1	321.9759												
9		1	321.9981												
10		1	321.9957												
11		1	321.9041												
12		1	321.9836												
13		1	322.0037												
14		1	322.0000												
15		1	322.0033												
16		1	322.0023												
17		2	322.0069												
18		2	322.0306												
19		2	322.0114												
20		2	322.0269												
21		2	322.0087												
22		2	322.0267												
23		2	322.0176												
24		2	322.0070												
25		2	922.0182												

Above data are from a manufacturer of high-performance automobiles produces disc-brakes that must measure 322 millimeters in diameter. Quality control randomly draws 16 discs made by each of eight production machines and measures their diameters.

For one sample t-test go to **Analyze>compare means>One sample t-test>Select test variable>select test value>Ok**



The SPSS Viewer window displays the output for a T-Test. The syntax at the top is:

```

GET
FILE='D:\spss\Samples\brakes.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
T-TEST
/TESTVAL=322
/MISSING=ANALYSIS
/VARIABLES=brake
/CRITERIA=CI(.9500).

```

The 'T-Test' section shows the command used:

```

[T-Test]
[DataSet1] D:\spss\Samples\brakes.sav

```

The 'One-Sample Statistics' table is as follows:

	N	Mean	Std. Deviation	Std. Error Mean
Disc Brake Diameter (mm)	128	3.2200E2	.0108224	.0009566

The 'One-Sample Test' table is as follows:

Test Value = 322						
			95% Confidence Interval of the Difference			
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Disc Brake Diameter (mm)	2.101	127	.030	.0020094	.000116	.003902

EXERCISES 8.7

Self-check Exercises

- SC 8-11** Given a sample mean of 83, a sample standard deviation of 12.5, and a samplesize of 22, test the hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use the 0.025 significance level.
- SC 8-12** Picosoft, Ltd., a supplier of operating system software for personal computers, was planning the initial public offering of its stock in order to raise sufficient working capital to finance the development of a radically new, seventh-generation integrated system. With current earnings of \$1.61 a share, Picosoft and its underwriters were contemplating an offering price of \$21, or about 13 times earnings. In order to check the appropriateness of this price, they randomly chose seven publicly traded software firms and found that their average price/earnings ratio was 11.6, and the sample standard deviation was 1.3. At $\alpha = 0.02$, can Picosoft conclude that the stocks of publicly traded software firms have an average P/E ratio that is significantly different from 13?

Basic Concepts

- 8-44** Given a sample mean of 94.3, a sample standard deviation of 8.4, and a sample size of 6, test the hypothesis that the value of the population mean is 100 against the alternative hypothesis that it is less than 100. Use the 0.05 significance level.
- 8-45** If a sample of 25 observations reveals a sample mean of 52 and a sample variance of 4.2, test the hypothesis that the population mean is 65 against the alternative hypothesis that it is some other value. Use the 0.01 significance level.

Application

- 8-46** Realtor Elaine Snyderman took a random sample of 12 homes in a prestigious suburb of Chicago and found the average appraised market value to be \$780,000, and the standard deviation was \$49,000. Test the hypothesis that for all homes in the area, the mean appraised value is \$825,000 against the alternative that it is less than \$825,000. Use the 0.05 level of significance.
- 8-47** For a sample of 60 women taken from a population of over 5,000 enrolled in a weight-reducing program at a nationwide chain of health spas, the sample mean diastolic blood pressure is 101 and the sample standard deviation is 42. At a significance level of 0.02, on average, did the women enrolled in the program have diastolic blood pressure that exceeds the value of 75?
- 8-48** The data-processing department at a large life insurance company has installed new color video display terminals to replace the monochrome units it previously used. The 95 operators trained to use the new machines averaged 7.2 hours before achieving a satisfactory level of performance. Their sample variance was 16.2 squared hours. Long experience with operators on the old monochrome terminals showed that they averaged 8.1 hours on the machines before their performances were satisfactory. At the 0.01 significance level, should the supervisor of the department conclude that the new terminals are easier to learn to operate?
- 8-49** As the bottom fell out of the oil market in early 1986, educators in Texas worried about how the resulting loss of state revenues (estimated to be about \$100 million for each \$1 decrease

in the price of a barrel of oil) would affect their budgets. The state board of education felt the situation would not be critical as long as they could be reasonably certain that the price would stay above \$18 per barrel. They surveyed 13 randomly chosen oil economists and asked them to predict how low the price would go before it bottomed out. The 13 predictions average \$21.60, and the sample standard deviation was \$4.65. At $\alpha = 0.01$, is the average prediction significantly higher than \$18.00? Should the board conclude that a budget crisis is unlikely? Explain.

- 8-50** A television documentary on overeating claimed that Americans are about 10 pounds overweight on average. To test this claim, eighteen randomly selected individuals were examined; their average excess weight was found to be 12.4 pounds, and the sample standard deviation was 2.7 pounds. At a significance level of 0.01, is there any reason to doubt the validity of the claimed 10-pound value?

- 8-51** XCO, a multinational manufacturer, uses a batch process to produce widgets. Each batch of widgets takes 8 hours to produce and has material and labor costs of \$8,476. Because of variations in machine efficiency and raw material purity, the number of widgets per batch is random. All widgets made can be sold for \$2.50 each, and widget production is profitable so long as the batches sell for more than \$12,500 on average. XCO sampled 16 batches and found 5,040 widgets per batch on average, with a standard deviation of 41.3 widgets. At $\alpha = 0.025$, can XCO conclude that its widget operation is profitable?

Worked-Out Answers to Self-Check Exercises

SC 8-11 $s = 12.5$ $n = 22$ $\bar{x} = 83$

$$H_0: \mu = 70 \quad H_1: \mu > 70 \quad \alpha = 0.025$$

The upper limit of the acceptance region is $t = 2.080$, or

$$\bar{x} = \mu_{H_0} + ts/\sqrt{n} = 70 + 2.080(12.5)/\sqrt{22} = 75.54$$

$$\text{Because the observed } t \text{ value} = \frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}} = \frac{83 - 70}{12.5/\sqrt{22}} = 4.878 > 2.080$$

(or $\bar{x} > 75.54$), we reject H_0 .

SC 8-12 $s = 1.3$ $n = 7$ $\bar{x} = 11.6$

$$H_0: \mu = 13 \quad H_1: \mu \neq 13 \quad \alpha = 0.02$$

The limits of the acceptance region are $t = \pm 3.143$, or

$$\bar{x} = \mu_{H_0} \pm ts/\sqrt{n} = 13 \pm 3.143(1.3)/\sqrt{7} = (11.46, 14.54)$$

$$\text{Because the observed } t \text{ value} = \frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}} = \frac{11.6 - 13}{1.3/\sqrt{7}} = -2.849 > -3.143$$

(or $\bar{x} = 11.6$, which is between 11.46 and 14.54), we do not reject H_0 . The average P/E ratio of publicly traded software firms is not significantly different from 13.

STATISTICS AT WORK

Loveland Computers

Case 8: One-Sample Tests of Hypotheses “Here’s the other thing that has me thinking more about adding a software division,” said Margot Derby, the head of Marketing at Loveland Computers, as she pulled a *Wall Street Journal* column from her desk drawer. “As you know, prices on PCs have been dropping. But, to everyone’s surprise, PC buyers seem to be spending the same in total—they are making up for the discount price by buying more bells and whistles—and more software.

“The article quotes a figure for the average amount spent on software by people in the first year they own the machine. That’s the same figure that we asked when we did our telephone survey, but our number came in much lower than the amount they quoted in the article. The trouble is I’m not sure which figure to use to make our business plan for a software division.”

“Well, why would your number be different?” asked Lee Azko.

“We don’t intend to appeal to everyone,” Margot replied. “We probably have more of a ‘techie’ image, so our customers may be different from the ‘average’ customer they talk about in that article. Maybe they use custom programs they write themselves.”

“Or maybe the difference doesn’t mean anything at all and it’s just the result of sampling error,” Lee suggested.

“But I don’t know how we could decide for sure. We’ve calculated the mean and standard deviation for our telephone sample, but the *Journal* article only gives us the mean. And I remember enough from my one stat course in college to know that we can’t run a test if we don’t know the population standard deviation.”

Study Questions: Assume that the mean software expenditure figure quoted in the newspaper is a reliable *population mean*. Is Margot right that Lee also needs to know the population standard deviation in order to perform a test? What idea is Margot exploring here? How would the idea be stated in hypothesis-testing terms?

CHAPTER REVIEW

Terms Introduced in Chapter 8

Alpha (α) The probability of a Type I error.

Alternative Hypothesis The conclusion we accept when the data fail to support the null hypothesis.

Beta (β) The probability of a Type II error.

Critical Value The value of the standard statistic (z or t) beyond which we reject the null hypothesis; the boundary between the acceptance and rejection regions.

Hypothesis An assumption or speculation we make about a population parameter.

Lower-Tailed Test A one-tailed hypothesis test in which a sample value significantly below the hypothesized population value will lead us to reject the null hypothesis.

Null Hypothesis The hypothesis, or assumption, about a population parameter we wish to test, usually an assumption of the status quo.

One-Tailed Test A hypothesis test in which there is only one rejection region; that is where we are concerned only with whether the observed value deviates from the hypothesized value in one direction.

Power Curve A graph of the values of the power of a test for each value of μ , or other population parameter, for which the alternative hypothesis is true.

Power of the Hypothesis Test The probability of rejecting the null hypothesis when it is false, that is, a measure of how well the hypothesis test is working.

Raw Scale Measurement in the variable's original units.

Significance Level A value indicating the percent-age of sample values that is outside certain limits, assuming the null hypothesis is correct, that is, the probability of rejecting the null hypothesis when it is true.

Standardized Scale Measurement in standard deviations from the variable's mean.

Two-Tailed Test A hypothesis test in which the null hypothesis is rejected if the sample value is significantly higher or lower than the hypothesized value of the population parameter; a test involving two rejection regions.

Type I Error Rejecting a null hypothesis when it is true.

Type II Error Accepting a null hypothesis when it is false.

Upper-Tailed Test A one-tailed hypothesis test in which a sample value significantly above the hypothesized population value will lead us to reject the null hypothesis.

Review and Application Exercises

- 8-52** For the following situations, state appropriate null and alternative hypotheses.
- The Census Bureau wants to determine whether the percentage of homeless people in New York City is the same as the national average.
 - A local hardware store owner wants to determine whether sales of garden supplies are better than usual after a spring promotion.
 - The Weather Channel wants to know whether average annual snowfall in the 1980s was significantly different from the 8-inch average recorded over the past 100 years.
 - A consumer-products investigative magazine wonders whether the fuel efficiency of a new subcompact car is significantly less than the 34 miles per gallon stated on the window sticker.
- 8-53** Health Electronics, Inc., a manufacturer of pacemaker batteries, specifies that the life of each battery is greater than or equal to 28 months. If scheduling for replacement surgery for the batteries is to be based on this claim, explain to the management of this company the consequences of Type I and Type II errors.
- 8-54** A manufacturer of petite women's sportswear has hypothesized that the average weight of the women buying its clothing is 110 pounds. The company takes two samples of its customers and finds one sample's estimate of the population mean is 98 pounds, and the other sample produces a mean weight of 122 pounds. In the test of the company's hypothesis that the population mean is 110 pounds versus the hypothesis that the mean does not equal 110 pounds, is one of these sample values more likely to lead us to accept the null hypothesis? Why or why not?

- 8-55** Many cities have adopted High Occupancy Vehicle (HOV) lanes to speed commuter traffic to downtown business districts. Planning for Metro Transportation District has depended on a well-established average of 3.4 passengers per HOV. But a summer intern notes that because many firms are sponsoring van pools, the average number of passengers per car is probably higher. The intern takes a sample of 23 vehicles going through the HOV lane of a toll plaza and reports a sample mean of 4.3 passengers, and a standard deviation of 1.5 passengers. At the 0.01 level of significance, does the sample suggest that the mean number of passengers has increased?
- 8-56** In Exercise SC 8-5, what would be the power of the test for $\mu = 14,000, 13,500$, and 13,000 if the significance level were changed to 0.10?
- 8-57** On an average day, about 5 percent of the stocks on the New York Stock Exchange set a new high for the year. On Friday, September 18, 1992, the Dow Jones Industrial Average closed at 3,282 on a robust volume of over 136 million shares traded. A random sample of 120 stocks showed that sixteen had set new annual highs that day. Using a significance level of 0.01, should we conclude that more stocks than usual set new highs on that day?
- 8-58** In response to criticism concerning lost mail, the U.S. Postal Service initiated new procedures to alleviate this problem. The postmaster general had been assured that this change would reduce losses to below the historic loss rate of 0.3 percent. After the new procedures had been in effect for 2 months, the USPS sponsored an investigation in which a total of 8,000 pieces of mail were mailed from various parts of the country. Eighteen of the test pieces failed to reach their destinations. At a significance level of 0.10, can the postmaster general conclude that the new procedures achieved their goal?
- 8-59** What is the probability that we are rejecting a true null hypothesis when we reject the hypothesized value because
- The sample statistic differs from it by more than 2.15 standard errors in either direction?
 - The value of the sample statistic is more than 1.6 standard errors above it?
 - The value of the sample statistic is more than 2.33 standard errors below it?
- 8-60** If we wish to accept the null hypothesis 85 percent of the time when it is correct, within how many standard errors around the hypothesized mean should the sample mean fall, in order to be in the acceptance region? What if we want to be 98 percent certain of accepting the null hypothesis when it is true?
- 8-61** Federal environmental statutes applying to a particular nuclear power plant specify that recycled water must, on average, be no warmer than 84°F (28.9°C) before it can be released into the river beside the plant. From 70 samples, the average temperature of recycled water was found to be 86.3°F (30.2°C). If the population standard deviation is 13.5 Fahrenheit (7.5 Celsius) degrees, should the plant be cited for exceeding the limitations of the statute? State and test appropriate hypotheses at $\alpha = 0.05$.
- 8-62** State inspectors, investigating charges that a Louisiana soft-drink bottling company underfills its product, have sampled 200 bottles and found the average contents to be 31.7 fluid ounces. The bottles are advertised to contain 32 fluid ounces. The population standard deviation is known to be 1.5 fluid ounces. Should the inspectors conclude, at the 2 percent significance level, that the bottles are being underfilled?
- 8-63** In 1995, the average 2-week-advance-purchase airfare between Raleigh-Durham, North Carolina, and New York City was \$235. The population standard deviation was \$68. A 1996 survey of 90 randomly chosen travelers between these two cities found that they had paid \$218.77, on average, for their tickets. Did the average airfare on this route change significantly

- between 1995 and 1996? What is the largest α at which you would conclude that the observed average fare is not significantly different from \$235?
- 8-64** Audio Sounds runs a chain of stores selling stereo systems and components. It has been very successful in many university towns, but it has had some failures. Analysis of its failures has led it to adopt a policy of not opening a store unless it can be reasonably certain that more than 15 percent of the students in town own stereo systems costing \$1,100 or more. A survey of 300 of the 2,400 students at a small, liberal arts college in the Midwest has discovered that 57 of them own stereo systems costing at least \$1,100. If Audio Sounds is willing to run a 5 percent risk of failure, should it open a store in this town?
- 8-65** The City of Oakley collects a 1.5 percent transfer tax on closed real estate transactions. In an average week, there are usually 32 closed transactions, with a standard deviation of 2.4. At the 0.10 level of significance, would you agree with the tax collector's conclusion that "sales are off this year" if a sample of 16 weeks had a mean of 28.25 closed sales?
- 8-66** In 1996, it was estimated that about 72 percent of all U.S. households were cable TV subscribers. *Newstime* magazine's editors were sure that their readers subscribed to cable TV at a higher rate than the general population and wanted to use this fact to sell advertising space for premium cable channels. To verify this, they sampled 250 of *Newstime*'s subscribers and found that 194 subscribed to cable TV. At a significance level of 2 percent, do the survey data support the editors' belief?
- 8-67** A company, recently criticized for not paying women as much as men working in the same positions, claims that its average salary paid to all employees is \$23,500. From a random sample of 29 women in the company, the average salary was calculated to be \$23,000. If the population standard deviation is known to be \$1,250 for these jobs, determine whether we could reasonably (within 2 standard errors) expect to find \$23,000 as the sample mean if, in fact, the company's claim is true.
- 8-68** Drive-a-Lemon rents cars that are mechanically sound, but older than those rented by the large national rent-a-car chains. As a result, it advertises that its rates are considerably lower than rates of its larger competitors. An industry survey has established that the average total charge per rental at one of the major firms is \$77.38. A random sample of 18 completed transactions at Drive-a-Lemon showed an average total charge of \$87.61 and a sample standard deviation of \$19.48. Verify that at $\alpha = 0.025$, Drive-a-Lemon's average total charge is significantly *higher* than that of the major firms. Does this result indicate that Drive-a-Lemon's rates, in fact, are not lower than the rates charged by the major national chains? Explain.
- 8-69** Refer to Exercise 8-26. Compute the power of the test for $\mu = \$41.95$, $\$42.95$, and $\$43.95$.
- 8-70** A personnel manager believed that 18 percent of the company's employees work overtime every week. If the observed proportion this week is 13 percent in a sample of 250 of the 2,500 employees, can we accept her belief as reasonable or must we conclude that some other value is more appropriate? Use $\alpha = 0.05$.
- 8-71** Refer to Exercise SC 8-5. Compute the power of the test for $\mu = 14,000$, $13,500$, and $13,000$.
- 8-72** A stockbroker claims that she can predict with 85 percent accuracy whether a stock's market value will rise or fall during the coming month. As a test, she predicts the outcome of 60 stocks and is correct in 45 of the predictions.

TABLE RW 8-1 PERSONAL DATA FOR A SAMPLE OF 20 CEOS

Company Name	Age	Status	Kids
Parkdale Mills Inc.	68	M	3
SAS Institute Inc.	50	M	3
Cogentrix Inc.	65	M	3
House of Raeford Farms Inc.	66	M	3
Harriet & Henderson Yarns Inc.	52	M	1
Harvey Enterprises and Affiliates	44	M	4
Radiator Specialty Co.	77	M	3
Parrish Tire Co.	43	M	2
Spectrum Dyed Yarns Inc.	59	M	2
Southeastern Hospital Supply Corp.	45	M	4
Miller Building Corp.	55	M	3
Pneumafil Corp.	55	S	0
Kroehler Furniture Industries Inc.	50	M	3
Carolina Petroleum Distributors Inc.	42	D	2
Tanner Cos.	64	M	4
Raycom Inc.	43	M	2
Cummins Atlantic Inc.	57	M	4
W. R. Bonsal Co.	62	M	3
Maola Milk & Ice Cream Co.	67	M	2
Waste Industries Inc.	56	M	2

Status = marital status (Single, Married, or Divorced)

Kids = number of children

Source: "Getting a Grip on Closely Held Companies," Business North Carolina 13(2), (June 1993):28–63.

Do these data present conclusive evidence (at $\alpha=0.04$) that her prediction accuracy is significantly less than the asserted 85 percent?

8-73 In Exercise 8-26, what would be the power of the test for $\mu = \$41.95$, $\$42.95$, and $\$43.95$ if the significance level were changed to 0.05?

8-74 A manufacturer of a vitamin supplement for newborns inserts a coupon for a free sample of its product in a package that is distributed at hospitals to new parents. Historically, about 18 percent of the coupons have been redeemed. Given current trends for having fewer children and starting families later, the firm suspects that today's new parents are better educated, on average, and, as a result, more likely to use a vitamin supplement for their infants. A sample of 1,500 new parents redeemed 295 coupons. Does this support, at a significance level of 2 percent, the firm's belief about today's new parents?

8-75 An innovator in the motor-drive industry felt that its new electric motor drive would capture 48 percent of the regional market within 1 year, because of the product's low price and superior performance. There are 5,000 users of motor drives in the region. After sampling

10 percent of these users a year later, the company found that 43 percent of them were using the new drives. At $\alpha = 0.01$, should we conclude that the company failed to reach its market-share goal?

- 8-76** According to machine specifications, the one-armed bandits in gambling casinos should pay off once in 11.6 turns, with a standard deviation of 2.7 turns. A lawyer believes that the machines at Casino World have been tampered with and observes a payoff once in 12.4 turns, over 36 machines. At $\alpha = 0.01$, is the lawyer right in concluding that the machines have a lower payoff frequency?

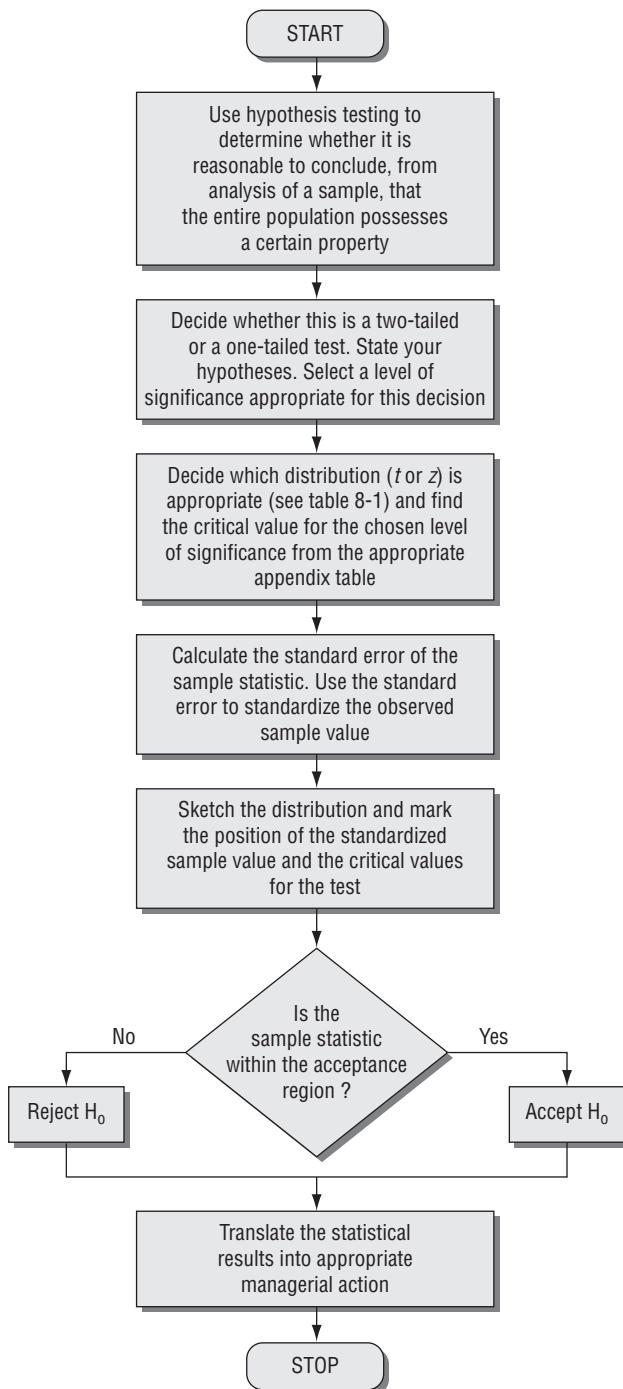


Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Test the hypothesis that the bank customers on an average is satisfied with the e-banking services offered by their banks. (Question 9)
2. Do the customers on an average agree that the e-banking facilities offered by private sector banks are better than public sector banks? (Question 13b)
3. Do the bank customers in general believe that the information provided by them for using the e-banking services, are misused. (Question 13d)



Flow Chart: One-Sample Tests of Hypotheses



0 Testing Hypotheses: Two-sample Tests

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn how to use samples from two populations to test hypotheses about how the populations are related
 - To learn how hypothesis tests for differences between population means take different forms, depending on whether the samples are large or small
 - To distinguish between independent and dependent samples when comparing two means
 - To learn how to reduce a hypothesis test for the difference of means from dependent samples to a test about a single mean
 - To learn how to test hypotheses that compare the proportions of two populations having some attribute of interest
 - To understand how prob values can be used in testing hypotheses
 - To get a feel for the kinds of outputs computer statistical packages produce for testing hypotheses
-

CHAPTER CONTENTS

- 9.1 Hypothesis Testing for Differences between Means and Proportions 426
- 9.2 Tests for Differences between Means: Large Sample Sizes 428
- 9.3 Tests for Differences between Means: Small Sample Sizes 434
- 9.4 Testing Differences between Means with Dependent Samples 445
- 9.5 Tests for Differences between Proportions: Large Sample Sizes 455
- 9.6 Prob Values: Another Way to Look at Testing Hypotheses 464
- Statistics at Work 469
- Terms Introduced in Chapter 9 470
- Equations Introduced in Chapter 9 470
- Review and Application Exercises 471
- Flow Chart: Two-Sample Tests of Hypotheses 477

A manufacturer of personal computers has a large number of employees from the local Spanish-speaking community. To improve the productivity of its workforce, it wants to increase the sensitivity of its managers to the needs of this ethnic group. It started by scheduling several informal question-and-answer sessions with leaders of the Spanish-speaking community. Later, it designed a program involving formal classroom contact between its managers and professional psychologists and sociologists. The newer program is much more expensive, and the company president wants to know whether this expenditure has resulted in greater sensitivity. In this chapter, we'll show you how to test whether these two methods have had essentially the same effects on the managers' sensitivity or if the expense of the newer program is justified by its improved results. ■

9.1 HYPOTHESIS TESTING FOR DIFFERENCES BETWEEN MEANS AND PROPORTIONS

In many decision-making situations, people need to determine whether the parameters of two populations are alike or different.

Comparing two populations

A company may want to test, for example, whether its female employees receive lower salaries than its male employees for the same work. A training director may wish to determine whether the proportion of promotable employees at one government installation is different from that at another. A drug manufacturer may need to know whether a new drug causes one reaction in one group of experimental animals but a different reaction in another group.

In each of these examples, decision makers are concerned with the parameters of two populations. In these situations, they are not as interested in the actual value of the parameters as they are in the *relation between* the values of the two parameters—that is, how these parameters differ. *Do* female employees earn less than male employees for the same work? *Is* the proportion of promotable employees at one installation different from that at another? *Did* one group of experimental animals react differently from the other? In this chapter, we shall introduce methods by which these questions can be answered, using hypothesis-testing procedures.

Sampling Distribution for the difference Between Two Population Parameters: Basic Concepts

In Chapter 6, we introduced the concept of the sampling distribution of the mean as the foundation for the work we would do in estimation and hypothesis testing. For a quick review of the sampling distribution of the mean, you may refer to Figure 6-2.

Because we now wish to study two populations, not just one, the sampling distribution of interest is the *sampling distribution of the difference between sample means*. Figure 9-1 may help us conceptualize this particular sampling distribution. At the top of this figure, we have drawn two populations, identified as Population 1 and Population 2. These two have means of μ_1 and μ_2 and standard deviations of σ_1 and σ_2 , respectively. Beneath each population, we show the sampling distribution of the sample mean for that population. At the bottom of the figure is the sampling distribution of the difference between the sample means.

Deriving the sampling distribution of the difference between sample means

The two theoretical sampling distributions of the mean in Figure 9-1 are each made up from all possible samples of a given size that can be drawn from the corresponding population distribution. Now,

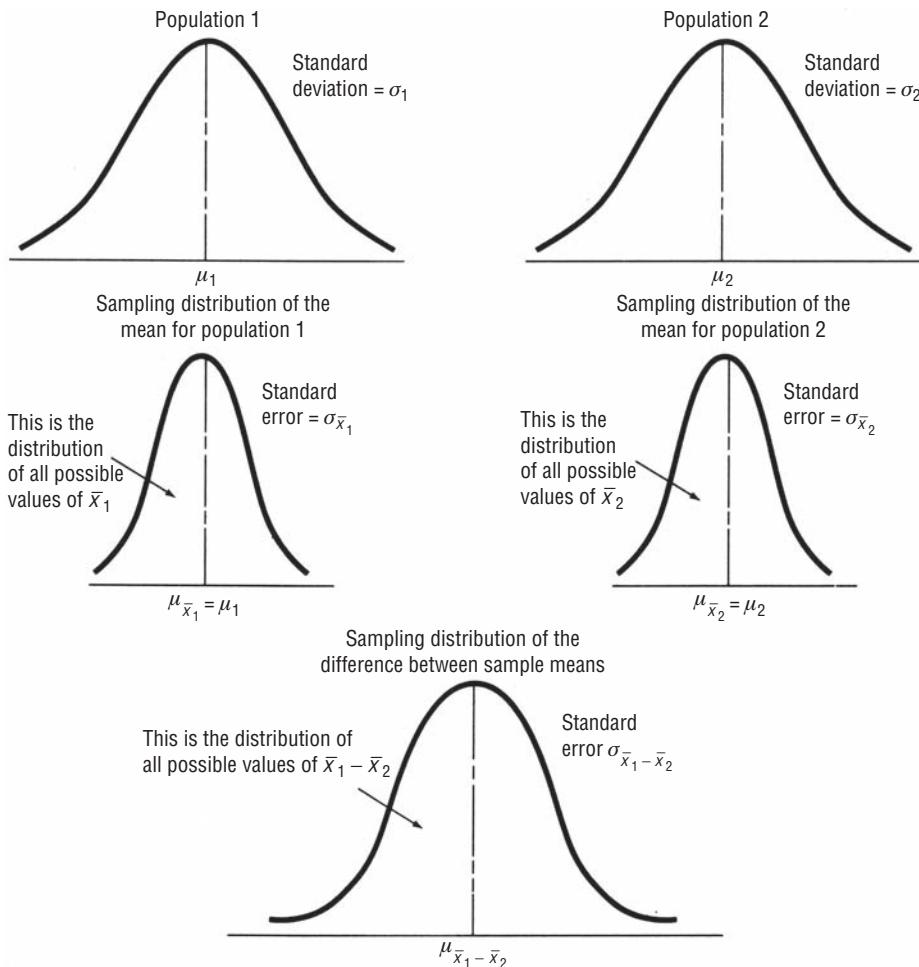


FIGURE 9-1 BASIC CONCEPTS OF POPULATION DISTRIBUTIONS, SAMPLING DISTRIBUTION OF THE MEAN, AND THE SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN SAMPLE MEANS

suppose we take a random sample from the distribution of Population 1 and another random sample from the distribution of Population 2. If we then subtract the two sample means, we get

$$\bar{x}_1 - \bar{x}_2 \leftarrow \text{Difference between sample means}$$

This difference will be positive if \bar{x}_1 is larger than \bar{x}_2 and negative if \bar{x}_2 is greater than \bar{x}_1 . By constructing a distribution of *all* possible sample differences of $\bar{x}_1 - \bar{x}_2$, we end up with the sampling distribution of the difference between sample means, which is shown at the bottom of Figure 9-1.

The *mean of the sampling distribution of the difference between sample means* is symbolized $\mu_{\bar{x}_1 - \bar{x}_2}$ and is equal to $\mu_{\bar{x}_1} - \mu_{\bar{x}_2}$, which as we saw in Chapter 6, is the same as $\mu_1 - \mu_2$. If $\mu_1 = \mu_2$, then $\mu_{\bar{x}_1} - \mu_{\bar{x}_2} = 0$.

Parameters of this sampling distribution

The standard deviation of the distribution of the difference between the sample means is called the *standard error of the difference between two means* and is calculated using this formula:

Standard Error of the Difference between Two Means

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad [9-1]$$

If the two population standard deviations are *not* known, we can *estimate* the standard error of the difference between two means. We can use the same method of estimating the standard error that we have used before by letting sample standard deviations estimate the population standard deviations as follows:

How to estimate the standard error of this sampling distribution

$$\hat{\sigma} = s \leftarrow \text{Sample standard deviation} \quad [7-1]$$

Therefore, the formula for the estimated standard error of the difference between two means becomes

Estimated Standard Error of the Difference between Two Means

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad [9-2]$$

As the following sections show, depending on the sample sizes, we shall use different estimates for $\hat{\sigma}_1$ and $\hat{\sigma}_2$ in Equation 9-2.

9.2 TESTS FOR DIFFERENCES BETWEEN MEANS: LARGE SAMPLE SIZES

When both sample sizes are greater than 30, this example illustrates how to do a two-tailed test of a hypothesis about the difference between two means. A manpower-development statistician is asked to determine whether the hourly wages of semiskilled

Step 1: State your hypotheses, type of test, and significance level

TABLE 9-1 DATA FROM A SAMPLE SURVEY OF HOURLY WAGES

City	Mean Hourly Earnings from Sample	Standard Deviation of Sample	Size of Sample
Apex	\$8.95	\$.40	200
Eden	9.10	.60	175

workers are the same in two cities. The results of this survey are presented in Table 9-1. Suppose the company wants to test the hypothesis at the 0.05 level that there is no difference between hourly wages for semiskilled workers in the two cities:

$$H_0: \mu_1 = \mu_2 \leftarrow \text{Null hypothesis: there is no difference}$$

$$H_1: \mu_1 \neq \mu_2 \leftarrow \text{Alternative hypothesis: a difference exists}$$

$$\alpha = 0.05 \leftarrow \text{Level of significance for testing this hypothesis}$$

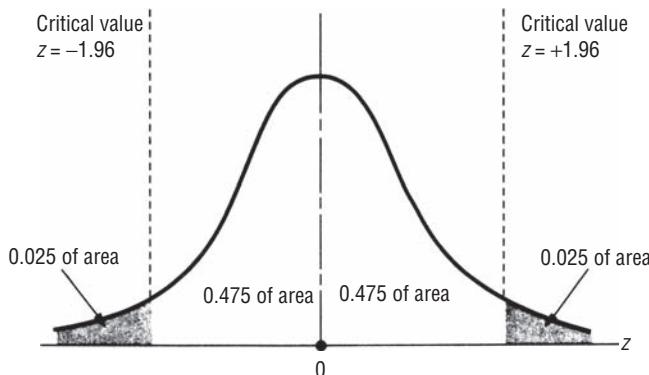
Because the company is interested only in whether the means are *or* are not equal, this is a two-tailed test.

We can illustrate this hypothesis test graphically. In Figure 9-2, the significance level of 0.05 corresponds to the two colored areas, each of which contains 0.025 of the area. The acceptance region contains two equal areas of 0.475 each. Because both samples are large, we can use the normal distribution. From Appendix Table 1, we can determine the critical value of z for 0.475 of the area under the curve to be 1.96.

Step 2: Choose the appropriate distribution and find the critical value

The standard deviations of the two populations are not known. Therefore, our first step is to estimate them, as follows:

$$\begin{aligned} \hat{\sigma}_1 &= s_1 & \hat{\sigma}_2 &= s_2 \\ &= \$0.40 & &= \$0.60 \end{aligned} \quad [7-1]$$

**FIGURE 9-2** TWO-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO MEANS AT THE 0.05 LEVEL OF SIGNIFICANCE

Now the estimated standard error of the difference between the two means can be determined by

Step 3: Compute the standard error and standardize the sample statistic

$$\begin{aligned}\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \\ &= \sqrt{\frac{(0.40)^2}{200} + \frac{(0.60)^2}{175}} \\ &= \sqrt{0.00286} \\ &= \$0.053 \leftarrow \text{Estimated standard error}\end{aligned}\quad [9-2]$$

Next we standardize the difference of sample means, $\bar{x}_1 - \bar{x}_2$. First, we subtract $(\mu_1 - \mu_2)_{H_0}$, the hypothesized difference of the population means. Then we divide by $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$, the estimated standard error of the difference between the sample means.

$$\begin{aligned}z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \\ &= \frac{(8.95) - (9.10) - 0}{0.053} \\ &= -2.83\end{aligned}$$

We mark the standardized difference on a sketch of the sampling distribution and compare with the critical value, as shown in Figure 9-3. It demonstrates that the standardized difference between the two sample means lies outside the acceptance

Step 4: Sketch the distribution and mark the sample value and critical values

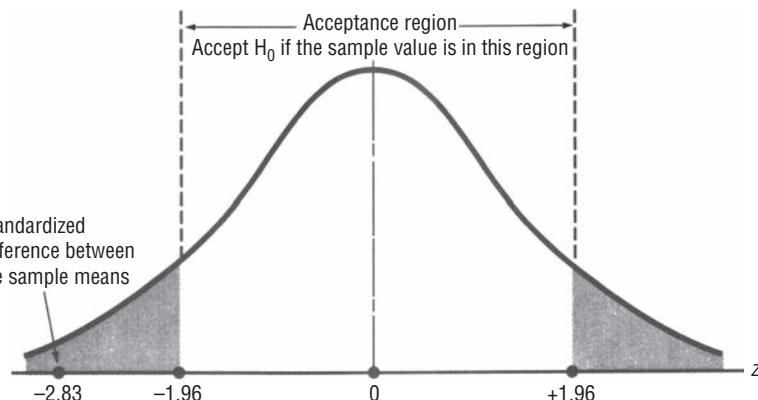


FIGURE 9-3 TWO-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO MEANS AT THE 0.05 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED DIFFERENCE BETWEEN SAMPLE MEANS

region. Thus, we reject the null hypothesis of no difference and conclude that the population means (the average semiskilled wages in these two cities) differ.

In this example, and in most of the examples we will encounter, we will be testing whether two populations have the same means. Because of this, $(\mu_1 - \mu_2)_{H_0}$, the hypothesized difference between the two means, was zero. However, we could also have investigated whether the average wages were *about* ten cents per hour *lower* in Apex than in Eden. In that case our hypotheses would have been:

$$H_0: \mu_1 = \mu_2 - 0.10 \leftarrow \text{null hypothesis wages are \$0.10 lower in Apex than in Eden}$$

$$H_1: \mu_1 \neq \mu_2 - 0.10 \leftarrow \text{Alternative hypothesis, wages are not \$0.10 lower in Apex than in Eden}$$

In this case, the hypothesized difference between the two means would be $(\mu_1 - \mu_2)_{H_0} = -0.10$, and the standardized difference between the sample means would be

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \\ &= \frac{(8.95 - 9.10) - (-0.10)}{0.053} \\ &= -0.94 \end{aligned}$$

In this case, we would not reject the null hypothesis.

Although our example was a two-tailed test, we can also perform one-tailed tests of the differences between two population means. Those one-tailed tests are conceptually similar to the one-tailed tests of a single mean that we discussed in Chapter 8. For example, if we had wanted to test whether wages in Apex were *lower than* wages in Eden (or equivalently if wages in Eden were *higher than* wages in Apex), our hypotheses would have been

$$H_0: \mu_1 = \mu_2 \leftarrow \text{Null hypothesis: wages are the same in Eden and Apex}$$

$$H_1: \mu_1 < \mu_2 \leftarrow \text{Alternative hypothesis wages are lower in Apex than in Eden}$$

This would be a one-tailed test with $(\mu_1 - \mu_2)_{H_0} = 0$.

Finally, if we had wanted to test whether wages in Apex were *more than ten cents per hour lower than* wages in Eden, then our hypotheses would have been

$$H_0: \mu_1 = \mu_2 - 0.10 \leftarrow \text{Null hypothesis: wages are \$0.10 lower in Apex than in Eden}$$

$$H_1: \mu_1 < \mu_2 - 0.10 \leftarrow \text{Alternative hypothesis: wages are more than \$0.10 lower in Apex than in Eden}$$

This would be a one-tailed test with $(\mu_1 - \mu_2)_{H_0} = -0.10$.

Step 5: Interpret the result

Testing the difference between means when $\mu_1 - \mu_2 \neq 0$

One-tailed tests of the difference between means

HINTS & ASSUMPTIONS

Hint: In testing for differences between two means, you must choose whether to use a one-tailed hypothesis test or a two-tailed test. If the test concerns whether two means *are or are not equal*, use a two-tailed test that will measure whether one mean is different from the other (higher or lower). If the test concerns whether *one mean is significantly higher or significantly lower than the other*, a one-tailed test is appropriate.

EXERCISES 9.1

Self-Check Exercises

- SC 9-1** Two independent samples of observations were collected. For the first sample of 60 elements, the mean was 86 and the standard deviation 6. The second sample of 75 elements had a mean of 82 and a standard deviation of 9.
- Compute the estimated standard error of the difference between the two means.
 - Using $\alpha = 0.01$, test whether the two samples can reasonably be considered to have come from populations with the same mean.
- SC 9-2** In 1993, the Financial Accounting Standards Board (FASB) was considering a proposal to require companies to report the potential effect of employees' stock options on earnings per share (EPS). A random sample of 41 high-technology firms revealed that the new proposal would reduce EPS by an average of 13.8 percent, with a standard deviation of 18.9 percent. A random sample of 35 producers of consumer goods showed that the proposal would reduce EPS by 9.1 percent on average, with a standard deviation of 8.7 percent. On the basis of these samples, is it reasonable to conclude (at $\alpha = 0.10$) that the FASB proposal will cause a greater reduction in EPS for high-technology firms than for producers of consumer goods?

Basic Concepts

- 9-1** Two independent samples were collected. For the first sample of 42 items, the mean was 32.3 and the variance 9. The second sample of 57 items had a mean of 34 and a variance of 16.
- Compute the estimated standard error of the difference between the two means.
 - Using $\alpha = 0.05$, test whether there is sufficient evidence to show the second population has a larger mean.

Applications

- 9-2** Block Enterprises, a manufacturer of chips for computers, is in the process of deciding whether to replace its current semiautomated assembly line with a fully automated assembly line. Block has gathered some preliminary test data about hourly chip production, which is summarized in the following table, and it would like to know whether it should upgrade its assembly line. State (and test at $\alpha = 0.02$) appropriate hypotheses to help Block decide.

	\bar{x}	s	n
Semiautomatic line	198	32	150
Automatic line	206	29	200

- **9-3** Two research laboratories have independently produced drugs that provide relief to arthritis sufferers. The first drug was tested on a group of 90 arthritis sufferers and produced an average of 8.5 hours of relief, and a sample standard deviation of 1.8 hours. The second drug was tested on 80 arthritis sufferers, producing an average of 7.9 hours of relief, and a sample standard deviation of 2.1 hours. At the 0.05 level of significance, does the second drug provide a significantly shorter period of relief?
- **9-4** A sample of 32 money-market mutual funds was chosen on January 1, 1996, and the average annual rate of return over the past 30 days was found to be 3.23 percent, and the sample

standard deviation was 0.51 percent. A year earlier, a sample of 38 money-market funds showed an average rate of return of 4.36 percent, and the sample standard deviation was 0.84 percent. Is it reasonable to conclude (at $\alpha = 0.05$) that money-market interest rates declined during 1995?

- 9-5** In September 1995, the Automobile Confederation of the Carolinas surveyed 75 randomly chosen service stations in North and South Carolina and determined that the average price for regular unleaded gasoline at self-service pumps was \$1.059, and the sample standard deviation was 3.9¢. Three months later, another survey of 50 service stations found an average price of \$1.089, and the sample standard deviation was 6.8¢. At $\alpha = 0.02$, had the Carolinas' average price of self-service regular unleaded gasoline changed significantly in this 3-month period?
- 9-6** Notwithstanding the Equal Pay Act of 1963, in 1993 it still appeared that men earned more than women in similar jobs. A random sample of 38 male machine-tool operators found a mean hourly wage of \$11.38, and the sample standard deviation was \$1.84. A random sample of 45 female machine-tool operators found their mean wage to be \$8.42, and the sample standard deviation was \$1.31. On the basis of these samples, is it reasonable to conclude (at $\alpha = 0.01$) that the male operators are earning over \$2.00 more per hour than the female operators?
- 9-7** BullsEye Discount store has always prided itself on customer service. The store hopes that all BullsEye stores are providing the same level of service from coast to coast, so they have surveyed some customers. In the Southeast region, a random sample of 97 customers yielded an average overall satisfaction rating of 8.8 out of 10 and the sample standard deviation was 0.7. In the Northeast region, a random sample of 84 customers resulted in an average rating of 9.0 and the sample standard deviation was 0.6. Can BullsEye conclude, at $\alpha = 0.05$, that the levels of customer satisfaction in the two markets are significantly different?

Worked-Out Answers to Self-Check Exercises

$$\text{SC 9-1} \quad s_1 = 6 \quad n_1 = 60 \quad \bar{x}_1 = 86 \quad s_2 = 9 \quad n_2 = 75 \quad \bar{x}_2 = 82$$

$$(a) \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{36}{60} + \frac{81}{75}} = 1.296$$

$$(b) H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2 \quad \alpha = 0.01$$

The limits of the acceptance region are $z = \pm 2.58$, or

$$\bar{x}_1 - \bar{x}_2 = 0 \pm z\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \pm 2.58(1.296) = \pm 3.344$$

$$\text{Because the observed } z \text{ value} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$$= \frac{(86 - 82) - 0}{1.296} = 3.09 > 2.58 \text{ (or } \bar{x}_1 - \bar{x}_2 = 86 - 82 = 4 > 3.344)$$

we reject H_0 . It is reasonable to conclude that the two samples come from different populations.

- SC 9-2** Sample 1 (HT firms): $s_1 = 18.9$ $n_1 = 41$ $\bar{x}_1 = 13.8$
 Sample 2 (CG producers): $s_2 = 8.7$ $n_2 = 35$ $\bar{x}_2 = 9.1$

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 > \mu_2 \quad \alpha = 0.10$$

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(18.9)^2}{41} + \frac{(8.7)^2}{35}} = 3.298 \text{ percent}$$

The upper limit of the acceptance region is $z = 1.28$, or

$$\bar{x}_1 - \bar{x}_2 = 0 + z\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = 1.28(3.298) = 4.221 \text{ percent}$$

Because the observed z value

$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} = \frac{(13.8 - 9.1) - 0}{3.298} = 1.43 > 1.28 \text{ (or } \bar{x}_1 - \bar{x}_2 = 4.7 > 4.221\text{)}$$

we reject H_0 and conclude that the FASB proposal will cause a significantly greater reduction in EPS for high-tech firms.

9.3 TESTS FOR DIFFERENCES BETWEEN MEANS: SMALL SAMPLE SIZES

When the sample sizes are small, there are two technical changes in our procedure for testing the differences between means. The first involves the way we compute the estimated standard error of the difference between the two sample means. The second will remind you of what we did in Chapter 8 with small-sample tests of a single mean. Once again we will base our small-sample tests on the t distribution, rather than the normal distribution. To explore the details of these changes, let's return to our chapter-opening illustration concerning the sensitivity of the managers at a personal-computer manufacturer to the needs of their Spanish-speaking employees.

Recall that the company has been investigating two education programs for increasing the sensitivity of its managers. The original program consisted of several informal question-and-answer sessions with leaders of the Spanish-speaking community. Over the past few years, a program involving formal classroom contact with professional psychologists and sociologists has been developed. The new program is considerably more expensive, and the president wants to know at the 0.05 level of significance whether this expenditure has resulted in greater sensitivity. Let's test the following:

$H_0: \mu_1 = \mu_2 \leftarrow$ Null hypothesis: There is no difference in sensitivity levels achieved by the two programs

$H_1: \mu_1 > \mu_2 \leftarrow$ Alternative hypothesis: The new program results in higher sensitivity levels

$\alpha = 0.05 \leftarrow$ Level of significance for testing this hypothesis

**Step 1: State your hypotheses,
type of test, and significance
level**

Table 9-2 contains the data resulting from a sample of the managers trained in both programs. Because only limited data are available for the two programs, the population standard deviations are estimated from the data. The sensitivity level is measured as a percentage on a standard psychometric scale.

The company wishes to test whether the sensitivity achieved by the new program is *significantly higher* than that achieved under the older, more informal program. To reject the null hypothesis (a result that the company desires), the observed difference of sample means would need to fall sufficiently high in the *right* tail of the distribution. Then we would accept the alternative hypothesis that the new program leads to higher sensitivity levels and that the extra expenditures on this program are justified.

TABLE 9-2 DATA FROM SAMPLE OF TWO SENSITIVITY PROGRAMS

Program Sampled	Mean Sensitivity after This Program	Number of Managers Observed	Estimated Standard Deviation of Sensitivity after This Program
Formal	92%	12	15%
Informal	84	15	19

The second step in our five-step process for hypothesis testing now requires us to choose the appropriate distribution and find the critical value. Recall from the opening paragraph in this section that the test will be based on a t distribution, but we don't yet know which t distribution to use. *How many degrees of freedom are there?* The answer to this question will be more apparent after we see how to compute the estimated standard error.

Our first task in performing the test is to calculate the standard error of the difference between the two means. Because the population standard deviations are not known, we must use Equation 9-2.

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad [9-2]$$

In the previous example, where the sample sizes were large (both greater than 30), we used Equation 7-1 and estimated $\hat{\sigma}_1^2$ by s_1^2 , and $\hat{\sigma}_2^2$ by s_2^2 . Now, with small sample sizes, that procedure is not appropriate. If we can assume that the unknown population variances are equal (and this assumption can be tested using a method discussed in Section 6 of Chapter 11), we can continue. If we cannot assume that $\sigma_1^2 = \sigma_2^2$, then the problem is beyond the scope of this text.

Assuming for the moment that $\sigma_1^2 = \sigma_2^2$, how can we estimate the common variance σ^2 ? If we use either s_1^2 or s_2^2 , we get an unbiased estimator of σ^2 , but we don't use all the information available to us because we ignore one of the samples. Instead we use a weighted average of s_1^2 and s_2^2 , and the weights are the numbers of degrees of freedom in each sample. This weighted average is called a "pooled estimate" of σ^2 . It is given by:

Pooled Estimate of σ^2

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad [9-3]$$

Because we have to use the sample variances to estimate the unknown σ^2 , the test will be based on the t distribution. This is just like the test of a single mean from a sample of size n when we did not know σ^2 . There we used a t distribution with $n - 1$ degrees of freedom, because once we knew the sample mean, only $n - 1$ of the sample observations could be freely specified. (You may review the discussion of degrees of freedom on pages 355–356.) Here we

Postponing Step 2 until we know how many degrees of freedom to use

Estimating σ^2 with small sample sizes

For this test, we have $n_1 + n_2 - 2$ degrees of freedom

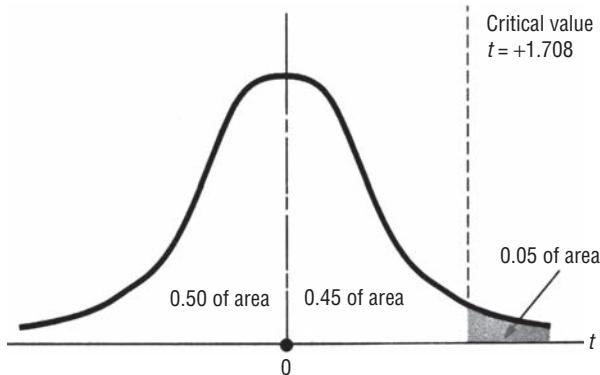


FIGURE 9-4 RIGHT-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO MEANS AT THE 0.05 LEVEL OF SIGNIFICANCE

have $n_1 - 1$ degrees of freedom in the first sample and $n_2 - 1$ degrees of freedom in the second sample, so when we pool them to estimate σ^2 , we wind up with $n_1 + n_2 - 2$ degrees of freedom. Hence the appropriate sampling distribution for our test of the two sensitivity programs is the t distribution with $12 + 15 - 2 = 25$ degrees of freedom. Because we are doing an upper-tailed test at a 0.05 significance level, the critical value of t is 1.708, according to Appendix Table 2.

Now that we have the critical value for our hypothesis test, we can illustrate it graphically in Figure 9-4. The colored region at the right of the distribution represents the 0.05 significance level of our test.

Continuing on to Step 3, we plug the formula for s_p^2 from Equation 9-3 into Equation 9-2 and simplify to get an equation for the estimated standard error of $\bar{x}_1 - \bar{x}_2$:

Returning to Step 2: Choose the appropriate distribution and find the critical value

Beginning Step 3: Compute the standard error

Estimated Standard Error of the Difference between Two Sample Means with Small Samples and Equal Population Variances

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad [9-4]$$

Applying these results to our sensitivity example:

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(12 - 1)(15)^2 + (15 - 1)(19)^2}{12 + 15 - 2} \\ &= \frac{11(225) + 14(361)}{25} \\ &= 301.160 \end{aligned} \quad [9-3]$$

Taking square roots on both sides, we get $s_p = \sqrt{301.160}$, or 17.354, so :

$$\begin{aligned}\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} &= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 17.354 \sqrt{\frac{1}{12} + \frac{1}{15}} \\ &= 17.354(0.387) \\ &= 6.721 \leftarrow \text{Estimated standard error of the difference}\end{aligned}\quad [9-4]$$

Next we standardize the difference of sample means, $\bar{x}_1 - \bar{x}_2$. First, we subtract $(\mu_1 - \mu_2)_{H_0}$, the hypothesized difference of the population means. Then we divide by $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$, the estimated standard error of the difference between the sample means.

Concluding Step 3: Standardize the sample statistic

$$\begin{aligned}t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \\ &= \frac{(92 - 84) - 0}{6.721} \\ &= 1.19\end{aligned}$$

Because our test of hypotheses is based on the t distribution, we use t to denote the standardized statistic.

Then we mark the standardized difference on a sketch of the sampling distribution and compare it with the critical value of $t = 1.708$, as shown in Figure 9-5. We can see in Figure 9-5 that the standardized difference between the two sample means lies within the acceptance region. Thus, we accept the null hypothesis

Step 4: Sketch the distribution and mark the sample value and critical value

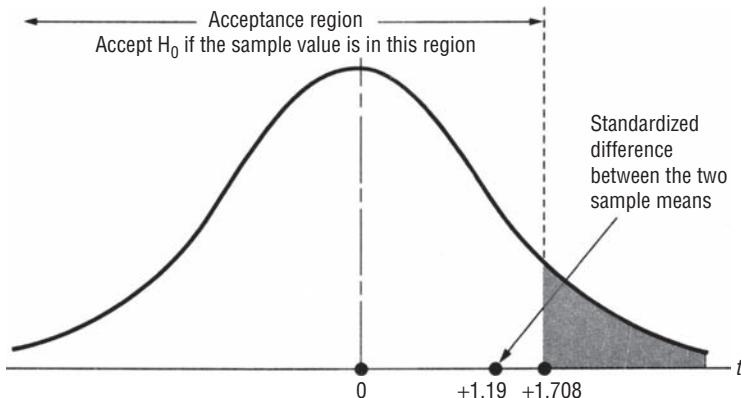


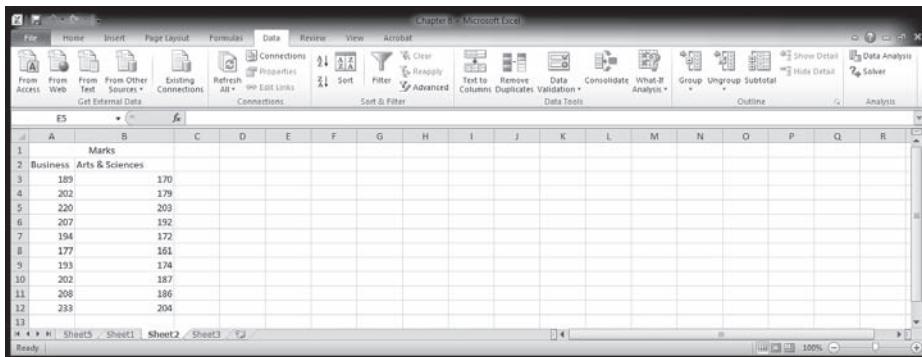
FIGURE 9-5 ONE-TAILED TEST OF THE DIFFERENCE BETWEEN TWO MEANS AT THE 0.05 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED DIFFERENCE BETWEEN THE SAMPLE MEANS

that there is no difference between the sensitivities achieved by the two programs. The company's expenditures on the formal instructional program have not produced significantly higher sensitivities among its managers.

Step 5: Interpret the result**HINTS & ASSUMPTIONS**

Hint: Because our sample sizes here are small (less than 30) and we do not know the standard deviations of the populations; the t distribution is appropriate. Like the one-sample t test we've already studied, here too, we need to determine the appropriate degrees of freedom. In a one-sample test, degrees of freedom were the sample size minus one. Here, because we are using two samples, the correct degrees of freedom are the first sample size minus one plus the second sample size less one, or symbolically, $n_1 + n_2 - 2$. Assumption: We are assuming that the variances of the two populations are equal. If this is not the case, we can not do this test using the methods we have covered. Warning: To use the method explained in this section, the two samples (one from each population) must have been chosen independently of each other.

To Test the Differences between Two Means Using MS Excel: (Equal Variances Assumed)

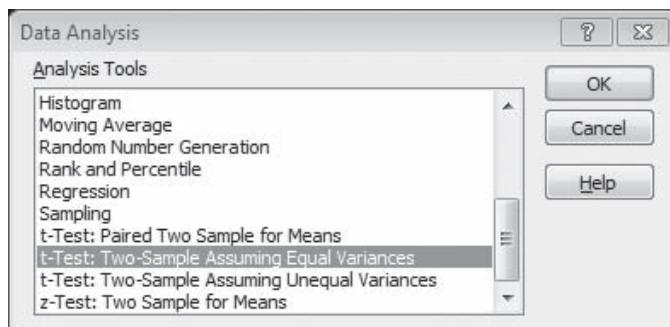


The screenshot shows a Microsoft Excel spreadsheet titled "Chapter 8 in Microsoft Excel". The data is organized into columns A through R. Column A contains row numbers from 1 to 13. Column B contains the stream names: "Business" and "Arts & Sciences". Columns C and D contain the marks for each student. For example, student 1 has marks 189 and 170 respectively. The data is as follows:

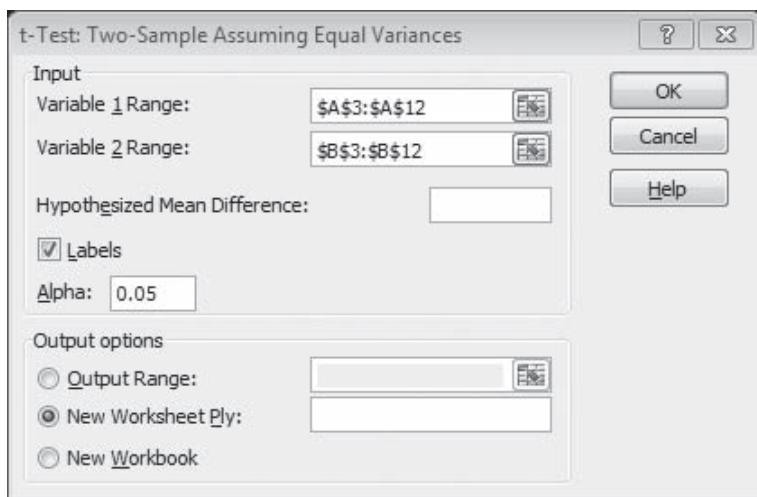
	Marks	
1	Business	Arts & Sciences
2	189	170
3	202	179
4	220	203
5	207	192
6	194	172
7	177	161
8	193	174
9	202	187
10	208	186
11	233	204
12		
13		

Above data are the marks of students of two different streams.

For performing two sample t-test assuming equal variance, go to **DATA>DATA ANALYSIS >t-Test: Two sample Assuming Equal Variances>Select marks of both groups**



Above data are the marks of students of two different streams.

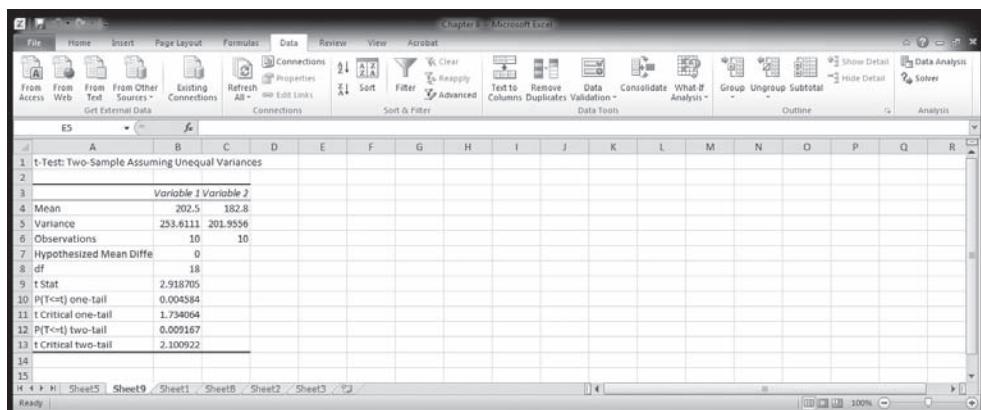
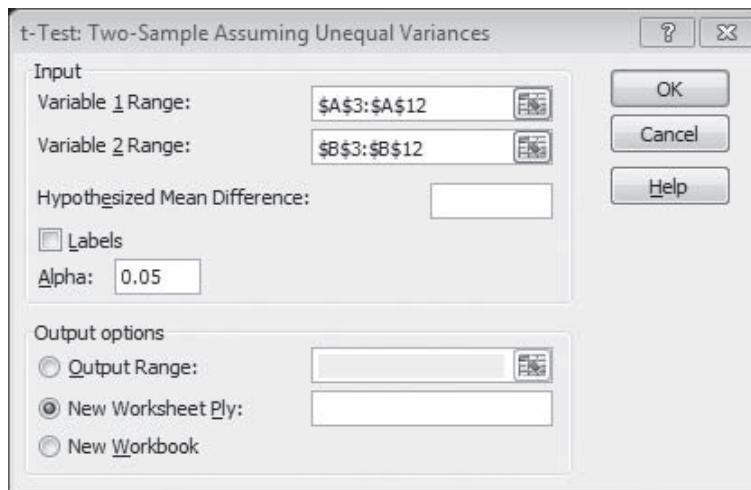
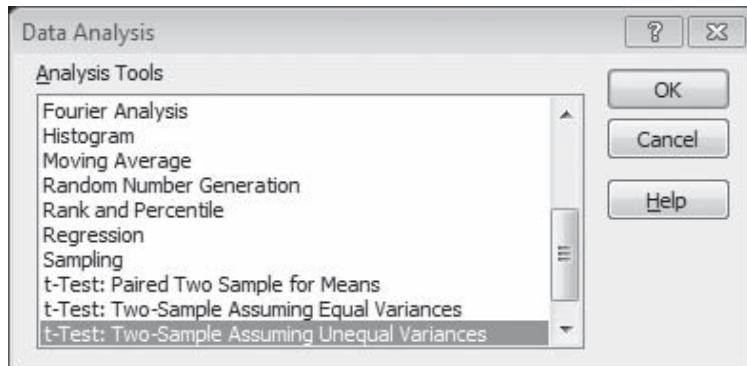


t-Test: Two-Sample Assuming Equal Variances	
<u>Variable 1</u> <u>Variable 2</u>	
Mean	202.5 182.8
Variance	253.6111 201.9556
Observations	10 10
Pooled Variance	227.7833
Hypothesized Mean Difference	0
df	18
t Stat	2.918705
P(T<=t) one-tail	0.004584
t Critical one-tail	1.734064
P(T >t) two-tail	0.009167
t Critical two-tail	2.309922

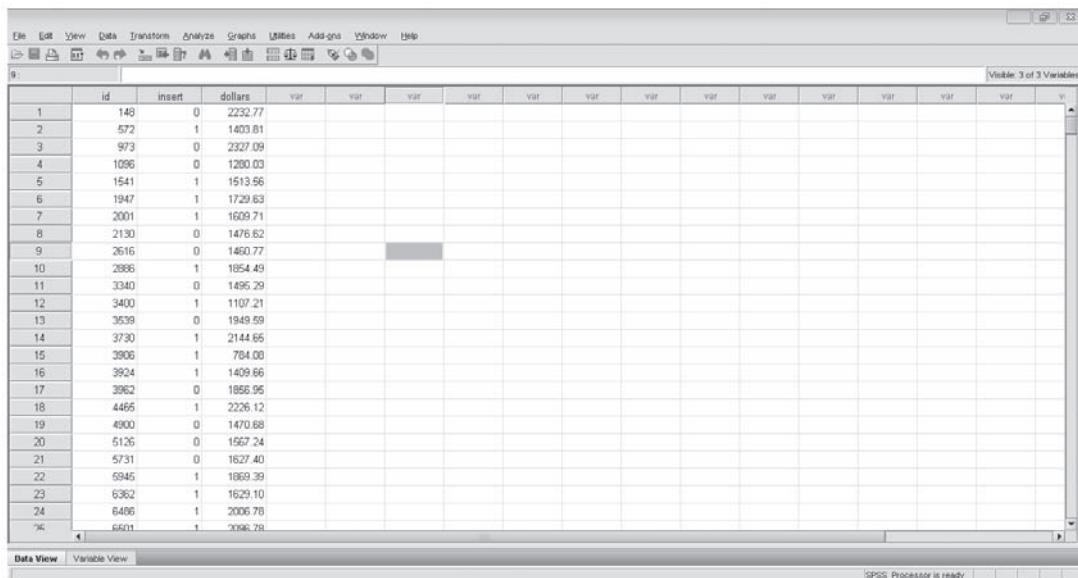
To Test the Differences between Two Means Using MS Excel: (Unequal Variances Assumed)

J18	
1	Weight (pounds)
2	Before After
3	189 170
4	202 179
5	220 203
6	207 192
7	194 172
8	177 161
9	193 174
10	202 187
11	208 186
12	233 204
13	

For two sample t-test assuming unequal variance go to **DATA>DATA ANALYSIS >t-Test: Two sample Assuming Unequal Variances>Select marks of both groups**



To Test the Differences between Two Means Using SPSS

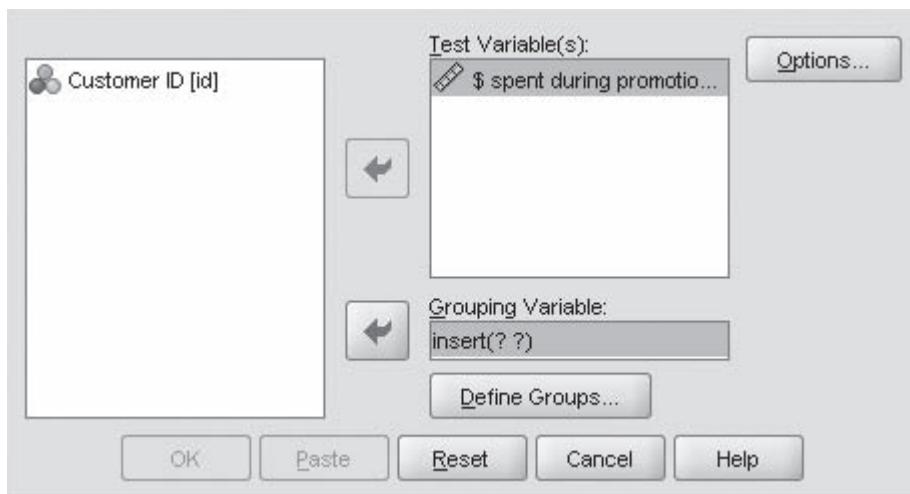


The screenshot shows the SPSS Data View window. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. The toolbar has icons for opening files, saving, printing, and other functions. The main area displays a data table with 50 rows and 5 columns. The columns are labeled id, insert, dollars, var, and var. The 'insert' column contains binary values (0 or 1) representing different promotional groups. The 'dollars' column contains numerical values representing spending amounts. The 'var' columns are empty. A status bar at the bottom indicates 'Visible: 3 of 3 Variables'.

	id	insert	dollars	var													
1	149	0	223.77														
2	572	1	1403.81														
3	973	0	2327.09														
4	1096	0	1200.03														
5	1541	1	1513.56														
6	1947	1	1729.63														
7	2001	1	1609.71														
8	2130	0	1476.62														
9	2616	0	1460.77														
10	2886	1	1854.49														
11	3340	0	1496.29														
12	3400	1	1107.21														
13	3539	0	1949.59														
14	3730	1	2144.66														
15	3906	1	784.00														
16	3924	1	1409.66														
17	3962	0	1856.96														
18	4466	1	2226.12														
19	4900	0	1470.68														
20	5126	0	1567.24														
21	5731	0	1627.40														
22	5945	1	1869.39														
23	6362	1	1629.10														
24	6496	1	2006.78														
25	6501	1	2006.78														

In above data, an analyst at a department store wants to evaluate a recent credit card promotion program. To this end, 500 cardholders were randomly selected. Half of them received an advertisement promising a reduced interest rate on purchases made over the next three months, and the other half received the standard seasonal advertisement.

For Independent sample t test, go to **Analyze>Compare means>independent sample t test>Insert test variable>Insert grouping variable>Define groups>Ok**



T-TEST GROUPS=insert(0 1)
/MISSING=ANALYSIS
/VARIABLES=dollars
/CRITERIA=CI(.9500).

T-Test

[Dataset2] D:\spss\Samples\creditpromo.sav

Group Statistics				
	Type of mail mail received	N	Mean	Std. Deviation
\$ spent during promotional period	Standard	250	1.5664E3	346.67305
	New Promotion	250	1.6375E3	356.70317
				21.92553
				22.55988

Independent Samples Test								
	Levene's Test for Equality of Variances			Test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
\$ spent during promotional period	1.190	.276	-2.260	498	.024	-71.11095	31.45914	.-132.91995 .-9.30196
	Equal variances assumed		-2.260	497.595	.024	-71.11095	31.45914	.-132.92007 .-9.30183
	Equal variances not assumed							

EXERCISES 9.2

Self-Check Exercises

- SC 9-3** A consumer-research organization routinely selects several car models each year and evaluates their fuel efficiency. In this year's study of two similar subcompact models from two different automakers, the average gas mileage for 12 cars of brand A was 27.2 miles per gallon, and the standard deviation was 3.8 mpg. The nine brand B cars that were tested averaged 32.4 mpg, and the standard deviation was 4.3 mpg. At $\alpha = 0.01$, should it conclude that brand A cars have lower average gas mileage than do brand B cars?
- SC 9-4** Connie Rodrigues, the Dean of Students at Midstate College, is wondering about grade distributions at the school. She has heard grumblings that the GPAs in the Business School are about 0.25 lower than those in the College of Arts and Sciences. A quick random sampling produced the following GPAs.

Business: 2.86 2.77 3.18 2.80 3.14 2.87 3.19 3.24 2.91 3.00 2.83
Arts & Sciences: 3.35 3.32 3.36 3.63 3.41 3.37 3.45 3.43 3.44 3.17 3.26 3.18 3.41

Do these data indicate that there is a factual basis for the grumblings? State and test appropriate hypotheses at $\alpha = 0.02$.

Applications

- 9-8** A credit-insurance organization has developed a new high-tech method of training new sales personnel. The company sampled 16 employees who were trained the original way and found average daily sales to be \$688 and the sample standard deviation was \$32.63. They also sampled 11 employees who were trained using the new method and found average daily sales to be \$706 and the sample standard deviation was \$24.84. At $\alpha = 0.05$, can the company conclude that average daily sales have increased under the new plan?

- 9-9** A large stock-brokerage firm wants to determine how successful its new account executives have been at recruiting clients. After completing their training, new account execs spend several weeks calling prospective clients, trying to get the prospects to open accounts with the firm. The following data give the numbers of new accounts opened in their first 2 weeks by 10 randomly chosen female account execs and by 8 randomly chosen male account execs. At $\alpha = 0.05$, does it appear that the women are more effective at generating new accounts than the men are?

	Number of New Accounts									
Female account execs	12	11	14	13	13	14	13	12	14	12
Male account execs	13	10	11	12	13	12	10	12		

- 9-10** To celebrate their first anniversary, Randy Nelson decided to buy a pair of diamond earrings for his wife Debbie. He was shown nine pairs with marquise gems weighing approximately 2 carats per pair. Because of differences in the colors and qualities of the stones, the prices varied from set to set. The average price was \$2,990, and the sample standard deviation was \$370. He also looked at six pairs with pear-shaped stones of the same 2-carat approximate weight. These earrings had an average price of \$3,065, and the standard deviation was \$805. On the basis of this evidence, can Randy conclude (at a significance level of 0.05) that pear-shaped diamonds cost more, on average, than marquise diamonds?
- 9-11** A sample of 30-year conventional mortgage rates at 11 randomly chosen banks in California yielded a mean rate of 7.61 percent and a standard deviation of 0.39 percent. A similar sample taken at 8 randomly chosen banks in Pennsylvania had a mean rate of 7.43 percent, and a standard deviation of 0.56 percent. Do these samples provide evidence to conclude (at $\alpha = 0.10$) that conventional mortgage rates in California and Pennsylvania come from populations with different means?
- 9-12** Because refunds are paid more quickly on tax returns that are filed electronically, the Commissioner of the Internal Revenue Service was wondering whether refunds due on returns filed by mail were smaller than those due on returns filed electronically. Looking only at returns claiming refunds, a sample of 17 filed by mail had an average refund of \$563, and a standard deviation of \$378. The average refund on a sample of 13 electronically filed returns was \$958, and the sample standard deviation was \$619. At $\alpha = 0.01$, do these data support the commissioner's speculation?
- 9-13** Greatyear tires currently produces tires at their Wilmington, North Carolina plant during two 12-hour shifts. The night-shift employees are planning to ask for a raise because they believe they are producing more tires per shift than the day shift. "Because Greatyear is making more money during the night shift, those employees should also make more money" according to the night-shift spokesman, I. M. Checking, the Greatyear production supervisor, randomly selected some daily production runs from each shift with the results given below (in 1,000s of tires produced).

Shift	Production (in 1,000s)									
Day	107.5	118.6	124.6	101.6	113.6	119.6	120.6	109.6	105.9	
Night	115.6	109.4	121.6	128.7	136.6	125.4	121.3	108.6	117.5	

Do these data indicate, at $\alpha = 0.01$, that the night shift is producing more tires per shift?

Worked-Out Answers to Self-Check Exercises

SC 9-3 $s_A = 3.8$ $n_A = 12$ $\bar{x}_A = 27.2$ $s_B = 4.3$ $n_B = 9$ $\bar{x}_B = 32.1$

$$H_0: \mu_A = \mu_B \quad H_1: \mu_A < \mu_B \quad \alpha = 0.01$$

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = \sqrt{\frac{11(3.8)^2 + 8(4.3)^2}{19}} = 4.0181 \text{ mpg}$$

The lower limit of the acceptance region is $t = -2.539$, or

$$\begin{aligned} \bar{x}_A - \bar{x}_B &= 0 - ts_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = -2.539(4.0181) \sqrt{\frac{1}{12} + \frac{1}{9}} \\ &= -4.499 \text{ mpg} \end{aligned}$$

Because the observed t value $= \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)_{H_0}}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{(27.2 - 32.1) - 0}{4.0181 \sqrt{\frac{1}{12} + \frac{1}{9}}}$

$= -2.766 < -2.539$ (or $\bar{x}_A - \bar{x}_B = -4.9 < -4.499$), we reject H_0 . Brand B delivers significantly higher mileage than does brand A.

SC 9-4 Sample 1 (Business): $s_B = 0.176$ $n_B = 11$ $\bar{x}_B = 2.98$

Sample 2 (Arts & Sciences): $s_A = 0.121$ $n_A = 13$ $\bar{x}_A = 3.368$

$$H_0: \mu_B - \mu_A = -0.25 \quad H_1: \mu_B - \mu_A \neq -0.25 \quad \alpha = 0.02$$

$$s_p = \sqrt{\frac{(n_B - 1)s_B^2 + (n_A - 1)s_A^2}{n_B + n_A - 2}} = \sqrt{\frac{10(0.176)^2 + 12(0.121)^2}{22}} = 0.1485$$

The limits of the acceptance region are $t = \pm 2.508$, or

$$\bar{x}_B - \bar{x}_A = (\mu_B - \mu_A)_{H_0} \pm ts_p \sqrt{\frac{1}{n_B} + \frac{1}{n_A}} = -0.25$$

$$\pm 2.508(0.1485) \sqrt{\frac{1}{11} + \frac{1}{13}} = (-0.4026, -0.0974)$$

Because the observed t value $= \frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)_{H_0}}{s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$

$$= \frac{(2.980 - 3.368) - 0.25}{0.1485 \sqrt{\frac{1}{11} + \frac{1}{13}}}$$

$= -2.268 > -2.508$ (or $\bar{x}_B - \bar{x}_A = -0.388 > -0.403$), we do not reject H_0 . The Business School GPAs are about .25 below those in the College of Arts & Sciences.

9.4 TESTING DIFFERENCES BETWEEN MEANS WITH DEPENDENT SAMPLES

In the examples in Sections 9.2 and 9.3, our samples were chosen *independent* of each other. In the wage example, the samples were taken from two different cities. In the sensitivity example, samples were taken of those managers who had gone through two different training programs. Sometimes, however, it makes sense to take samples that are not independent of each other. Often the use of such *dependent* (or *paired*) *samples* enables us to perform a more precise analysis, because these allow us to control for extraneous factors. With dependent samples, we still follow the same basic procedure that we have followed in all our hypothesis testing. The only differences are that we will use a different formula for the estimated standard error of the sample differences and that we will require that both samples be of the same size.

A health spa has advertised a weight-reducing program and has claimed that the average participant in the program loses more than 17 pounds. A somewhat overweight executive is interested in the program but is skeptical about the claims and asks for some hard evidence. The spa allows him to select randomly the records of 10 participants and record their weights before and after the program. These data are recorded in Table 9-3. Here we have two samples (a *before* sample and an *after* sample) that are clearly dependent on each other, because the same 10 people have been observed twice.

The overweight executive wants to test at the 5 percent significance level the claimed average weight loss of more than 17 pounds. Formally, we may state this problem:

Conditions under which paired samples aid analysis

Step 1: State your hypotheses, type of test, and significance level

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 17 && \leftarrow \text{Null hypothesis: average weight loss is only 17 pounds} \\ H_1: \mu_1 - \mu_2 &> 17 && \leftarrow \text{Alternative hypothesis: average weight loss exceeds 17 pounds} \\ \alpha &= 0.05 && \leftarrow \text{Level of significance} \end{aligned}$$

What we are really interested in is not the weights before and after, but their *differences*. **Conceptually, what we have is not two samples of before and after weights, but rather one sample of weight losses.** If the population of weight losses has a mean μ_l , we can restate our hypotheses as

$$\begin{aligned} H_0: \mu_l &= 17 \\ H_1: \mu_l &> 17 \end{aligned}$$

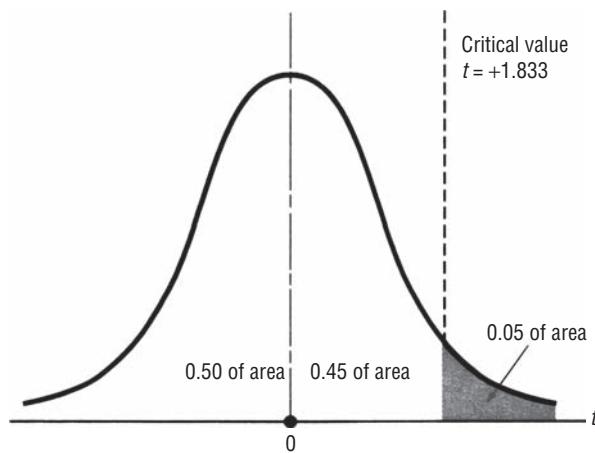
Conceptual understanding of differences

Step 2: Choose the appropriate distribution and find the critical value

Figure 9-6 illustrates this problem graphically. Because we want to know whether the mean weight loss *exceeds* 17 pounds, an upper-tailed test is appropriate. The 0.05 significance level is shown in Figure 9-6 as the colored area under the *t* distribution. We use the *t* distribution because the sample size is only 10;

TABLE 9-3 WEIGHTS BEFORE AND AFTER A REDUCING PROGRAM

Before	189	202	220	207	194	177	193	202	208	233
After	170	179	203	192	172	161	174	187	186	204

**FIGURE 9-6 ONE-TAILED HYPOTHESIS TEST AT THE 0.05 LEVEL OF SIGNIFICANCE**

the appropriate number of degrees of freedom is 9, $(10 - 1)$. Computing the paired differences Appendix Table 2 gives the critical value of t , 1.833.

We begin by computing the individual losses, their mean, and standard deviation, and proceed exactly as we did when testing hypotheses about a single mean. The computations are done in Table 9-4.

TABLE 9-4 FINDING THE MEAN WEIGHT LOSS AND ITS STANDARD DEVIATION

Before	After	Loss x	Loss Squared x^2
189	170	19	361
202	179	23	529
220	203	17	289
207	192	15	225
194	172	22	484
177	161	16	256
193	174	19	361
202	187	15	225
208	186	22	484
233	204	29	841
		$\sum x = 197$	$\sum x^2 = 4,055$
		$\bar{x} = \frac{\sum x}{n} \quad [3-2]$ $= \frac{197}{10}$ $= 19.7$	
		$s = \sqrt{\frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1}} \quad [3-18]$ $= \sqrt{\frac{4,055}{9} - \frac{10(19.7)^2}{9}}$ $= \sqrt{19.34}$ $= 4.40$	

Next, we use Equation 7-1 to estimate the unknown population standard deviation:

$$\begin{aligned}\hat{\sigma} &= s \\ &= 4.40\end{aligned}\quad [7-1]$$

Step 3: Compute the standard error and standardize the sample statistic

and now we can estimate the standard error of the mean:

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \frac{\hat{\sigma}}{\sqrt{n}} \\ &= \frac{4.40}{\sqrt{10}} \\ &= \frac{4.40}{3.16} \\ &= 1.39 \leftarrow \text{Estimated standard error of the mean}\end{aligned}\quad [7-6]$$

Next we standardize the observed average weight loss, $\bar{x} = 19.7$ pounds, by subtracting μ_{H_0} , the hypothesized average loss, and dividing by $\hat{\sigma}_{\bar{x}}$, the estimated standard error of the mean.

$$\begin{aligned}t &= \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_{\bar{x}}} \\ &= \frac{19.7 - 17}{1.39} \\ &= 1.94\end{aligned}$$

Because our test of hypotheses is based on the t distribution, we use t to denote the standardized statistic.

Figure 9-7 illustrates the location of the sample mean weight loss on the standardized scale. We see that the sample mean lies outside the acceptance region, so the executive can

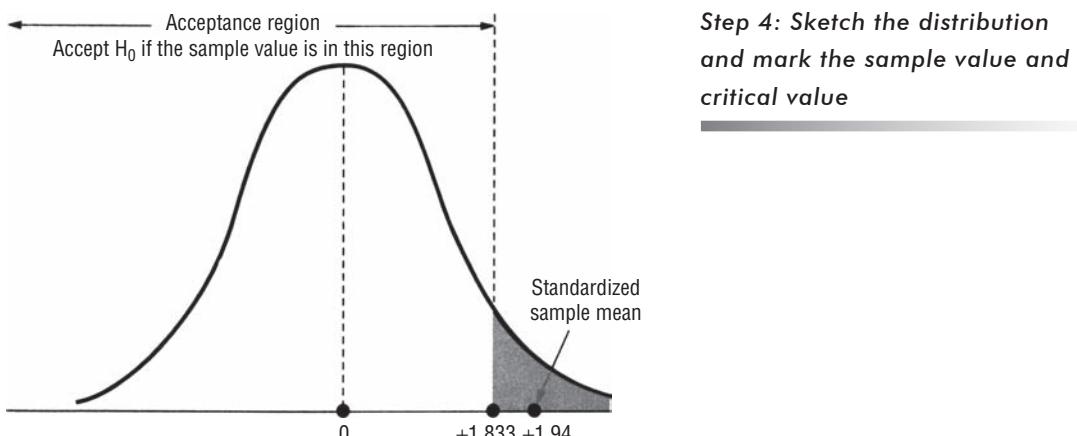


FIGURE 9-7 ONE-TAILED HYPOTHESIS TEST AT THE 0.05 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND STANDARDIZED SAMPLE MEAN

reject the null hypothesis and conclude that the claimed weight loss in the program is legitimate.

Let's see how this *paired difference test* differs from a test of the difference of means of *two independent* samples. Suppose that the data in Table 9-4 represent two independent samples of 10 individuals *entering* the program and *another* 10 randomly selected individuals *leaving* the program. The means and variances of the two samples are given in Table 9-5.

Because the sample sizes are small, we use Equation 9-3 to get a pooled estimate of σ^2 and Equation 9-4 to estimate $\sigma_{\bar{x}_1 - \bar{x}_2}$:

Step 5: Interpret the result

How does the paired difference test differ?

TABLE 9-5 BEFORE AND AFTER MEANS AND VARIANCES

Sample	Size	Mean	Variance
Before	10	202.5	253.61
After	10	182.8	201.96

A pooled estimate of σ^2

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} & [9-3] \\ &= \frac{(10 - 1)(253.61) + (10 - 1)(201.96)}{10 + 10 - 2} \\ &= \frac{2282.49 + 1817.64}{18} \\ &= 227.79 \leftarrow \text{Estimate of common population variance} \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} &= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} & [9-4] \\ &= \sqrt{227.79} \sqrt{\frac{1}{10} + \frac{1}{10}} \\ &= 15.09(0.45) \\ &= 6.79 \leftarrow \text{Estimate of } \sigma_{\bar{x}_1 - \bar{x}_2} \end{aligned}$$

The appropriate test is now based on the *t* distribution with 18 degrees of freedom ($10 + 10 - 2$). With a significance level of 0.05, the critical value of *t* from Appendix Table 2 is 1.734. The observed difference of the sample means is

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 &= 202.5 - 182.8 \\ &= 19.7 \text{ pounds} \end{aligned}$$

Now when we standardize the difference of the sample means for this independent-samples test, we get

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \\ &= \frac{(202.5 - 182.8) - 17}{6.79} \\ &= 0.40 \end{aligned}$$

Once again, because our test of hypotheses is based on the t distribution, we use t to denote the standardized statistic. Comparing the standardized difference of the sample means (0.40) with the critical value of t , (1.734), we see that the standardized sample statistic no longer falls outside the acceptance region, so this test will *not* reject H_0 .

Why did these two tests give such different results? In the paired sample test, the sample standard deviation of the individual differences was relatively small, so 19.7 pounds was significantly larger than the hypothesized weight loss of 17 pounds.

With independent samples, however, the estimated standard deviation of the difference between the means depended on the standard deviations of the before weights and the after weights. Because both of these were relatively large, $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$, was also large, and thus 19.7 was not significantly larger than 17. The paired sample test controlled this initial and final variability in weights by looking only at the individual changes in weights. Because of this, it was better able to detect the significance of the weight loss.

We conclude this section with two examples showing when to treat two samples of equal size as dependent or independent:

1. An agricultural extension service wishes to determine whether a new hybrid seed corn has a greater yield than an old standard variety. If the service asks 10 farmers to record the yield of an acre planted with the new variety and asks another 10 farmers to record the yield of an acre planted with the old variety, the two samples are independent. However, if it asks 10 farmers to plant one acre with each variety and record the results, then the samples are dependent, and the paired difference test is appropriate. In the latter case, differences due to fertilizer, insecticide, rainfall, and so on, are controlled, because each farmer treats his two acres identically. Thus, any differences in yield can be attributed solely to the variety planted.
2. The director of the secretarial pool at a large legal office wants to determine whether typing speed depends on the word-processing software used by a secretary. If she tests seven secretaries using PicosoftWrite and seven using WritePerfect, she should treat her samples as independent. If she tests the same seven secretaries twice (once on each word processor), then the two samples are dependent. In the paired difference test, differences among the secretaries are eliminated as a contributing factor, and the differences in typing speeds can be attributed to the different word processors.

HINTS & ASSUMPTIONS

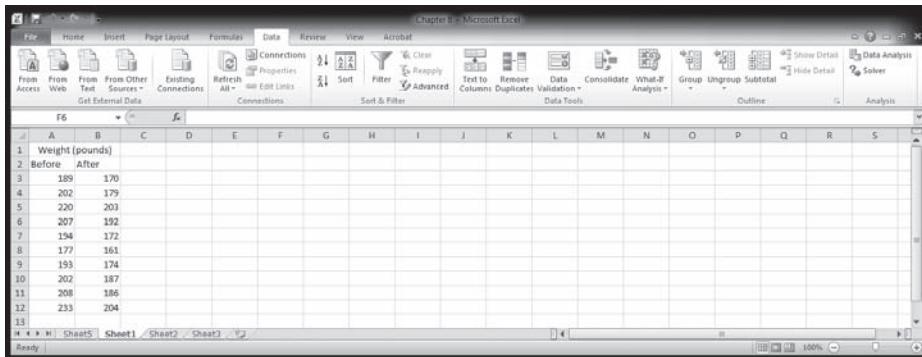
Often in testing differences between means, it makes sense to take samples that are *not* independent of each other. For example, if you are measuring the effect of a rust inhibitor on the buildup of rust on metal pipe, you would normally sample the rust on the same pipes before and after applying the inhibitor. Doing that controls for the effects of different locations, heat, and moisture. Because the same pipe is involved twice, the samples are not independent. Hint: If we measure the rust on each pipe before and six months after the application, we have a single sample of the grams of rust that have appeared since the application.

*With independent samples,
 H_0 cannot be rejected*

Explaining differing results

*Should we treat samples as
dependent or independent?*

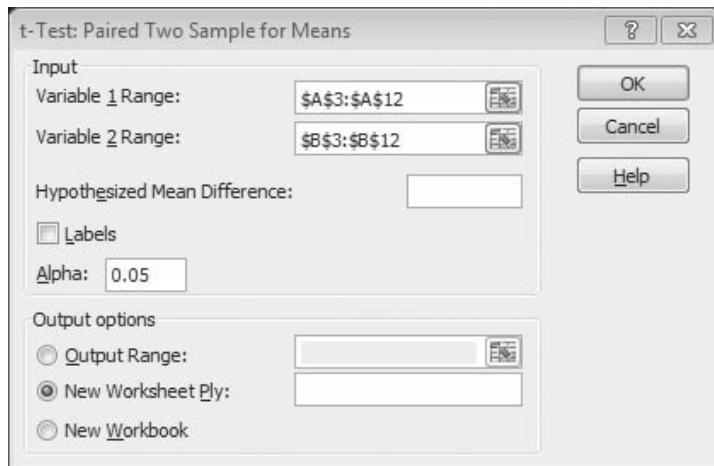
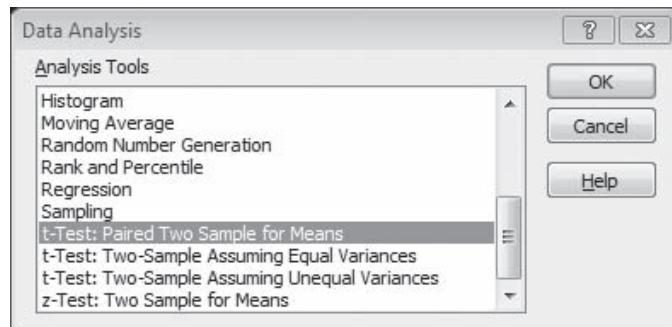
Testing Differences between Means with Dependent Samples (Paired t-test) Using MS Excel

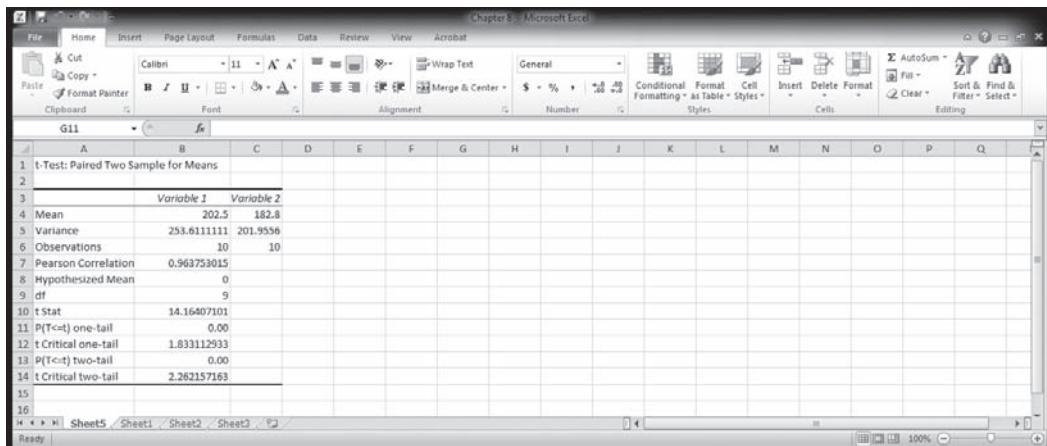


The screenshot shows a Microsoft Excel spreadsheet titled "Chapter 8 - Microsoft Excel". The data is located in the range A1:D13. Column A is labeled "Weight (pounds)" and contains two rows: "1. Before" and "2. After". Columns B and C contain 12 data points each, representing individual weights. The data is as follows:

	B	C
1. Before	189	170
2. After	202	179
	220	203
	207	192
	194	172
	177	161
	193	174
	202	187
	208	186
	233	204

The above data is the weight of a group of people before and after joining a weight reduction program. For paired t-test, go to **DATA>DATA ANALYSIS >t-Test: Paired Two Sample for Means>Select pre and post intervention data**

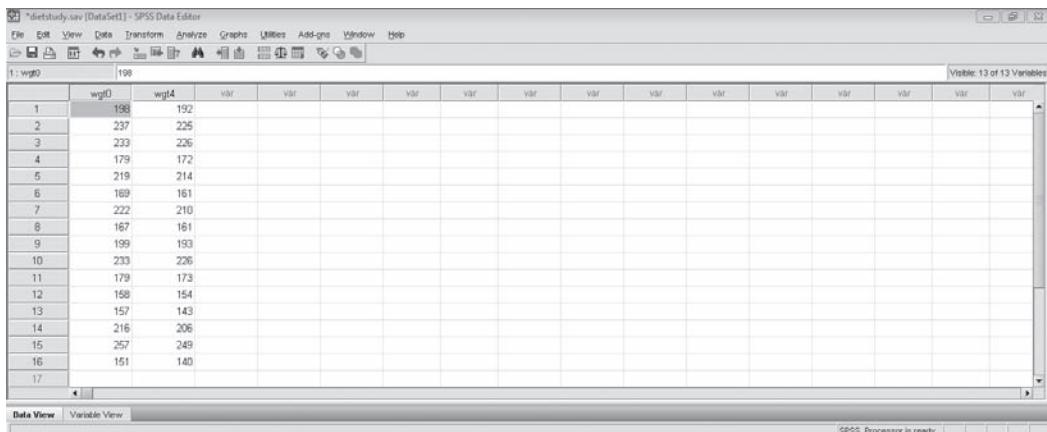




A screenshot of Microsoft Excel showing a t-test output for paired sample means. The data is organized into two columns: Variable 1 and Variable 2. The output includes the following statistics:

	Variable 1	Variable 2
Mean	202.5	182.8
Variance	253.611111	201.9556
Observations	10	10
Pearson Correlation	0.963753015	
Hypothesized Mean	0	
df	9	
t Stat	14.16407101	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.833112933	
P(T >t) two-tail	0.00	
t Critical two-tail	2.262157163	

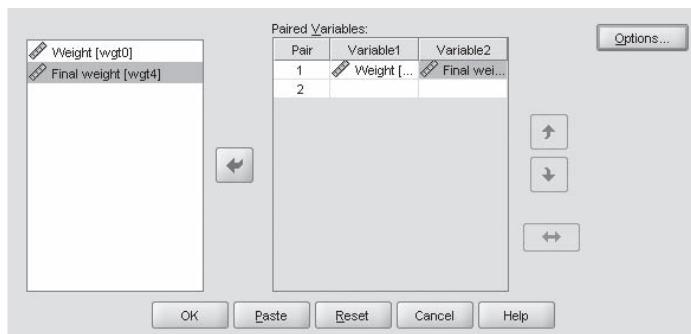
Testing Differences between Means with Dependent Samples (Paired t-test) Using SPSS



A screenshot of the SPSS Data View window. The data consists of two variables: wgt0 (before weight) and wgt4 (after weight). There are 17 rows of data, each containing a value for both variables.

	wgt0	wgt4
1	198	192
2	237	225
3	233	226
4	179	172
5	219	214
6	169	161
7	222	210
8	167	161
9	199	193
10	233	226
11	179	173
12	158	154
13	157	143
14	216	206
15	257	249
16	151	140
17		

Above data is the weight of a group of people before and after joining a weight reduction program. For paired t-test, go to **Analyze>compare means>paired sample t-test>Select paired variables>Ok**



The screenshot shows the SPSS Viewer window with the following details:

- Output1 [Document1] - SPSS Viewer**
- File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help**
- Log** (selected)
 - T-Test
 - Title
 - Notes
 - Active Dataset
 - Paired Samples Statistics
 - Paired Samples Correlations
 - Paired Samples Test
- GET**

```
FILE='D:\spss\Samples\diestudy.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
T-TEST PAIRS=wgt0 WITH wgt4 (PAIRED)
 /CRITERIA=CI(.9500)
 /MISSING=ANALYSIS.
```
- T-Test**
- [DataSet1] D:\spss\Samples\diestudy.sav**
- Paired Samples Statistics**

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Weight	198.38	16	33.472	8.368
	Final weight	190.31	33.508	8.377

- Paired Samples Correlations**

	N	Correlation	Sig.
Pair 1 Weight & Final weight	16	.996	.000

- Paired Samples Test**

	Paired Differences			95% Confidence Interval of the Difference			1	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper				
				-.602	2.886	.722			
Pair 1 Weight - Final weight	-.602	2.886	.722	6.525	9.600	11.175	15	.000	

EXERCISES 9.3

Self-Check Exercises

- SC 9-5** Sherri Welch is a quality control engineer with the windshield wiper manufacturing division of Emsco, Inc. Emsco is currently considering two new synthetic rubbers for its wiper blades, and Sherri was charged with seeing whether blades made with the two new compounds wear equally well. She equipped 12 cars belonging to other Emsco employees with one blade made of each of the two compounds. On cars 1 to 6, the right blade was made of compound A and the left blade was made of compound B; on cars 7 to 12, compound A was used for the left blade. The cars were driven under normal operating conditions until the blades no longer did a satisfactory job of clearing the windshield of rain. The data below give the usable life (in days) of the blades. At $\alpha = 0.05$, do the two compounds wear equally well?

Car	1	2	3	4	5	6	7	8	9	10	11	12
Left Blade	162	323	220	274	165	271	233	156	238	211	241	154
Right blade	183	347	247	269	189	257	224	178	263	199	263	148

- SC 9-6** Nine computer-components dealers in major metropolitan areas were asked for their prices on two similar color inkjet printers. The results of this survey are given below. At $\alpha = 0.05$, is it reasonable to assert that, on average, the Apson printer is less expensive than the Okaydata printer?

Dealer	1	2	3	4	5	6	7	8	9
Apson price	\$250	319	285	260	305	295	289	309	275
Okaydata price	\$270	325	269	275	289	285	295	325	300

Applications

- 9-14** The data below are a random sample of 9 firms chosen from the “Digest of Earnings Reports” in *The Wall Street Journal* on February 6, 1992:
- Find the mean change in earnings per share between 1991 and 1992.
 - Find the standard deviation of the change and the standard error of the mean.
 - Were average earnings per share different in 1991 and 1992? Test at $\alpha = 0.02$.

Firm	1	2	3	4	5	6	7	8	9
1991 earnings	1.38	1.26	3.64	3.50	2.47	3.21	1.05	1.98	2.72
1992 earnings	2.48	1.50	4.59	3.06	2.11	2.80	1.59	0.92	0.47

- 9-15** Jeff Richardson, the receiving clerk for a chemical-products distributor, is faced with the continuing problem of broken glassware, including test-tubes, petri dishes, and flasks. Jeff has determined some additional shipping precautions that can be undertaken to prevent breakage, and he has asked the Purchasing Director to inform the suppliers of the new measures. Data for 8 suppliers are given below in terms of average number of broken items per shipment. Do the data indicate, at $\alpha = 0.05$, that the new measures have lowered the average number of broken items?

Supplier	1	2	3	4	5	6	7	8
Before	16	12	18	7	14	19	6	17
After	14	13	12	6	9	15	8	15

- 9-16** Additives-R-Us has developed an additive to improve fuel efficiency for trucks that pull very heavy loads. They tested the additive by randomly selecting 18 trucks and dividing them into 9 pairs. In each pair, both trucks hauled the same type of load over the same roadway, but only one truck used fuel with the new additive. Different pairs followed different routes and carried different loads. The resulting fuel efficiencies (in miles per gallon) are given below. Do the data indicate, at $\alpha = 0.01$, that trucks using fuel with the additive achieved significantly better fuel efficiency than trucks using regular fuel?

Pair	1	2	3	4	5	6	7	8	9
Regular	5.7	6.1	5.9	6.2	6.4	5.1	5.9	6.0	5.5
Additive	6.0	6.2	5.8	6.6	6.7	5.3	5.7	6.1	5.9

- 9-17** Aquarius Health Club has been advertising a rigorous program for body conditioning. The club claims that after 1 month in the program, the average participant should be able to do eight more push-ups in 2 minutes than he or she could do at the start. Does the random sample of 10 program participants given below support the club’s claim? Use the 0.025 level of significance.

Participant	1	2	3	4	5	6	7	8	9	10
Before	38	11	34	25	17	38	12	27	32	29
After	45	24	41	39	30	44	30	39	40	41

- 9-18** Donna Rose is a production supervisor on the disk-drive assembly line at Winchester Technologies. Winchester recently subscribed to an easy listening music service at its factory, hoping that this would relax the workers and lead to greater productivity. Donna is skeptical about this hypothesis and fears the music will be distracting, leading to lower productivity. She sampled weekly production for the same six workers before the music was installed and after it was installed. Her data are given below. At $\alpha = 0.02$, has the average production changed at all?

Employee	1	2	3	4	5	6
Week without music	219	205	226	198	209	216
Week with music	235	186	240	203	221	205

- 9-19** Modems transmit information across telephone lines from one computer to another. Their speed is rated in baud, the number of bits per second that they can transmit. Because of several technical factors, actual transmission rate varies from file to file. Anne Evans was shopping for a new 28,800 baud modem. In testing two modems to decide which to purchase, she transmitted 7 randomly chosen files with both modems and recorded the following rates (in thousands of baud).

File	1	2	3	4	-5	6	7
Haynes Ultima 28.8	9.52	10.17	10.33	10.02	10.72	9.62	9.17
Extel PerFAXtion 28.8	10.92	11.46	11.18	12.21	10.42	11.36	10.47

A review in *PC Reports* said that the magazine's tests had found the Extel PerFAXtion to be significantly faster than the Haynes Ultima. At $\alpha = 0.01$, do Anne's results confirm that conclusion?

Worked-Out Answers to Self-Check Exercises

SC 9-5

Car	1	2	3	4	5	6	7	8	9	10	11	12
Blade A	183	347	247	269	189	257	233	156	238	211	241	154
Blade B	162	323	220	274	165	271	224	178	263	199	263	148
Difference	21	24	27	-5	24	-14	9	-22	-25	12	-22	6

$$\bar{x} = \frac{\sum x}{n} = \frac{35}{12} = 2.9167 \text{ days}$$

$$s^2 = \frac{1}{n-1} (\sum x^2 - n\bar{x}^2) = \frac{1}{11} [4397 - 12(2.9167)^2] = 390.45, s = \sqrt{s^2}$$

$$= 19.76 \text{ days}$$

$$\hat{\sigma}_{\bar{x}} = s/\sqrt{n} = 19.76/\sqrt{12} = 5.7042 \text{ days}$$

$$H_0: \mu_A = \mu_B \quad H_1: \mu_A \neq \mu_B \quad \alpha = 0.05$$

The limits of the acceptance region are $t = \pm 2.201$, or

$$\bar{x} = 0 \pm t\hat{\sigma}_{\bar{x}} = \pm 2.201(5.7042) = \pm 12.55 \text{ days}$$

Because the observed t value $= \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_{\bar{x}}} = \frac{2.9167 - 0}{5.7042} = 0.511 < 2.201$ (or $\bar{x} = 2.9167 < 12.55$),

we do not reject H_0 . The two compounds are not significantly different with respect to usable life.

SC 9-6

Dealer	1	2	3	4	5	6	7	8	9
Apson price	250	319	285	260	305	295	289	309	275
Okaydata price	270	325	269	275	289	285	295	325	300
Difference	20	6	-16	15	-16	-10	6	16	25

$$\bar{x} = \frac{\sum x}{n} = \frac{46}{9} = \$5.1111$$

$$s^2 = \frac{1}{n-1}(\sum x^2 - n\bar{x}^2) = \frac{1}{8}(2,190 - 9(5.1111)^2) = 244.36, s = \sqrt{s^2} = \$15.63$$

$$\hat{\sigma}_{\bar{x}} = s/\sqrt{n} = 15.63/\sqrt{9} = \$5.21$$

$$H_0: \mu_0 = \mu_A \quad H_1: \mu_0 > \mu_A \quad \alpha = 0.05$$

The upper limit of the acceptance region is $t = 1.860$, or

$$\bar{x} = 0 + t\hat{\sigma}_{\bar{x}} = 1.860(5.21) = \$9.69$$

Because the observed t value $= \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_{\bar{x}}} = \frac{5.1111 - 0}{5.21} = 0.981 < 1.860$ (or $\bar{x} = \$5.11 < \9.69)

we do not reject H_0 . On average, the Apson inkjet printer is not significantly less expensive than the Okaydata inkjet printer.

9.5 TESTS FOR DIFFERENCES BETWEEN PROPORTIONS: LARGE SAMPLE SIZES

Suppose you are interested in finding out whether the Republican party is stronger in New York than in California. Or perhaps you would like to know whether women are as likely as men to purchase sports cars. To reach a conclusion in situations like these, you can take samples from each of the two groups in question (voters in New York and California or women and men) and use the sample proportions to test the difference between the two populations.

The big picture here is very similar to what we did in Section 9.2, when we compared two means using independent samples: We standardize the difference between the two sample proportions and base our tests on the normal distribution. The only major difference will be in the way we find an estimate for the standard error of the difference between the two sample proportions. Let's look at some examples.

Two-Tailed Tests for Differences between Proportions

Consider the case of a pharmaceutical manufacturing company testing two new compounds intended to reduce blood-pressure levels. The compounds are administered to two different sets of laboratory animals. In group one, 71 of 100 animals tested respond to drug 1 with lower blood-pressure levels. In group two, 58 of 90 animals tested respond to drug 2 with lower blood-pressure levels. The company wants to test at the 0.05 level whether there is a difference between the efficacies of these two drugs. How should we proceed with this problem?

$$\begin{aligned}
 \bar{p}_1 &= 0.71 && \leftarrow \text{Sample proportion of successes with drug 1} \\
 \bar{q}_1 &= 0.29 && \leftarrow \text{Sample proportion of failures with drug 1} \\
 n_1 &= 100 && \leftarrow \text{Sample size for testing drug 1} \\
 \bar{p}_2 &= 0.644 && \leftarrow \text{Sample proportion of successes with drug 2} \\
 \bar{q}_2 &= 0.356 && \leftarrow \text{Sample proportion of failures with drug 2} \\
 n_2 &= 90 && \leftarrow \text{Sample size for testing drug 2} \\
 H_0: p_1 &= p_2 && \leftarrow \text{Null hypothesis: There is no difference between these two drugs} \\
 H_1: p_1 &\neq p_2 && \leftarrow \text{Alternative hypothesis: There is a difference between them} \\
 \alpha &= 0.05 && \leftarrow \text{Level of significance for testing this hypothesis}
 \end{aligned}$$

Step 1: State your hypotheses, type of test, and significance level

Step 2: Choose the appropriate distribution and find the critical value

Figure 9-8 illustrates this hypothesis test graphically. Because the management of the pharmaceutical company wants to know whether there is a difference between the two compounds, this is a two-tailed test. The significance level of 0.05 corresponds to the colored regions in the figure. Both samples are large enough to justify using the normal distribution to approximate the binomial. From Appendix Table 1, we can determine that the critical value of z for 0.475 of the area under the curve is 1.96.

As in our previous examples, we can begin by calculating the standard deviation of the sampling distribution we are using in our hypothesis test. In this example, the binomial distribution is the correct sampling distribution.

We want to find the *standard error of the difference between two proportions*; therefore, we should recall the formula for the *standard error of the proportion*:

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}$$

[7-4]

Step 3: Compute the standard error and standardize the sample statistic

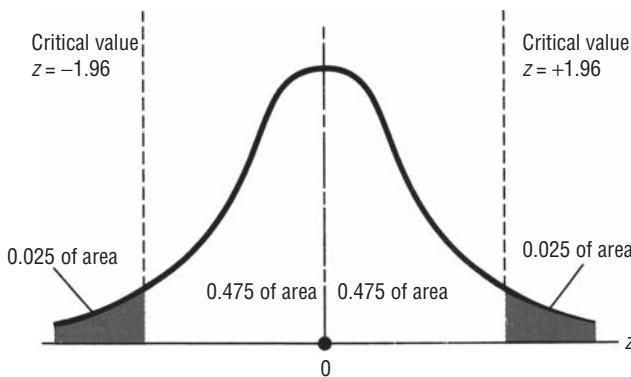


FIGURE 9-8 TWO-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO PROPORTIONS AT THE 0.05 LEVEL OF SIGNIFICANCE

Using this formula and the same form we previously used in Equation 9-1 for the standard error of the difference between two *means*, we get

Standard Error of the Difference between Two Proportions

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \quad [9-5]$$

To test the two compounds, we do not know the population parameters p_1, p_2, q_1 and q_2 , and thus we need to estimate them from the sample statistics $\bar{p}_1, \bar{p}_2, \bar{q}_1$, and \bar{q}_2 . In this case, we might suppose that the practical formula to use would be

How to estimate this standard error

Estimated Standard Error of the Difference between Two Proportions

Sample proportions
for sample 1 Sample proportions
for sample 2

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}} \quad [9-6]$$

But think about this a bit more. After all, if we hypothesize that there is *no difference* between the two population proportions, then our best estimate of the overall population proportion of successes is probably the *combined proportion of successes* in both samples, that is:

Best estimate of the overall proportion of successes in the population if the two proportions are hypothesized to be equal

$$= \frac{\text{number of successes in sample 1} + \text{number of successes in sample 2}}{\text{total size of both samples}}$$

And in the case of the two compounds, we use this equation with symbols rather than words:

Estimated Overall Proportion of Successes in Two Populations

$$\begin{aligned}
 \hat{p} &= \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} & [9-7] \\
 &= \frac{(100)(0.71) + (90)(0.644)}{100 + 90} \\
 &= \frac{71 + 58}{190} \\
 &= 0.6789 \leftarrow \text{Estimate of the overall proportion of success in the combined populations using combined proportions from both samples} \\
 &\quad (\hat{q} \text{ would be } 1 - 0.6789 = 0.3211)
 \end{aligned}$$

Now we can appropriately modify Equation 9-6 using the values \hat{p} and \hat{q} from Equation 9-7:

Estimated Standard Error of the Difference between Two Proportions Using Combined Estimates from Both Samples

Estimate of the population proportions using combined proportions from both samples

$$\begin{aligned}
 \hat{\sigma}_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} & [9-8] \\
 &= \sqrt{\frac{(0.6789)(0.3211)}{100} + \frac{(0.6789)(0.3211)}{90}} \\
 &= \sqrt{\frac{0.2180}{100} + \frac{0.2180}{90}} \\
 &= \sqrt{0.004602} \\
 &= 0.6789 \leftarrow \text{Estimated standard error of the difference between two proportions}
 \end{aligned}$$

We standardize the difference between the two observed sample proportions, $\bar{p}_1 - \bar{p}_2$, by dividing by the estimated standard error of the difference between two proportions:

$$\begin{aligned}
 z &= \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - \bar{p}_2)_{H_0}}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} \\
 &= \frac{(0.71 - 0.644) - 0}{0.0678} \\
 &= 0.973
 \end{aligned}$$

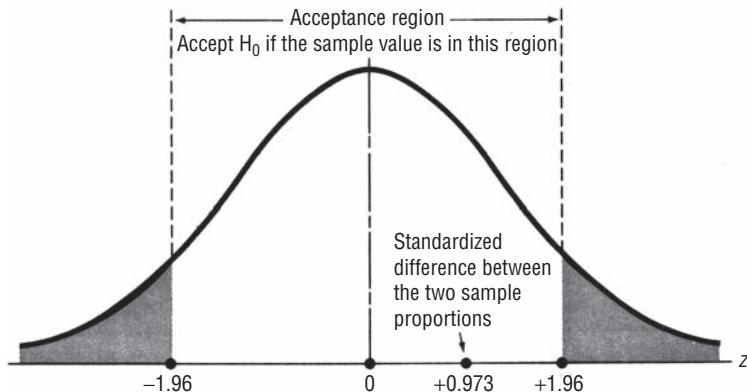


FIGURE 9-9 TWO-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO PROPORTIONS AT THE 0.05 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED DIFFERENCE BETWEEN THE SAMPLE PROPORTIONS

Next we plot the standardized value on a sketch of the sampling distribution in Figure 9-9.

We can see in Figure 9-9 that the standardized difference between the two sample proportions lies within the acceptance region. Thus, we accept the null hypothesis and conclude that these two new compounds produce effects on blood pressure that are *not* significantly different.

Step 4: Sketch the distribution and mark the sample value and critical values

Step 5: Interpret the result

One-Tailed Tests for Differences between Proportions

Conceptually, the one-tailed test for the difference between two population proportions is similar to a one-tailed test for the difference between two means. Suppose that for tax purposes, a city government has been using two methods of listing property. The first requires the property owner to appear in person before a tax lyster, but the second permits the property owner to mail in a tax form. The city manager thinks the personal-appearance method produces far fewer mistakes than the mail-in method. She authorizes an examination of 50 personal-appearance listings and 75 mail-in listings. Ten percent of the personal-appearance forms contain errors; 13.3 percent of the mail-in forms contain them. The results of her sample can be summarized:

$$\bar{p}_1 = 0.10 \leftarrow \text{Proportion of personal-appearance forms with errors}$$

$$\bar{q}_1 = 0.90 \leftarrow \text{proportion of personal-appearance forms without errors}$$

$$n_1 = 50 \leftarrow \text{Sample size of personal-appearance forms}$$

$$\bar{p}_2 = 0.133 \leftarrow \text{Proportion of mail-in forms with errors}$$

$$\bar{q}_2 = 0.867 \leftarrow \text{Proportion of mail-in forms without errors}$$

$$n_2 = 75 \leftarrow \text{Sample size of mail-in forms}$$

The city manager wants to test at the 0.15 level of significance the hypothesis that the personal-appearance method produces a lower proportion of errors. What should she do?

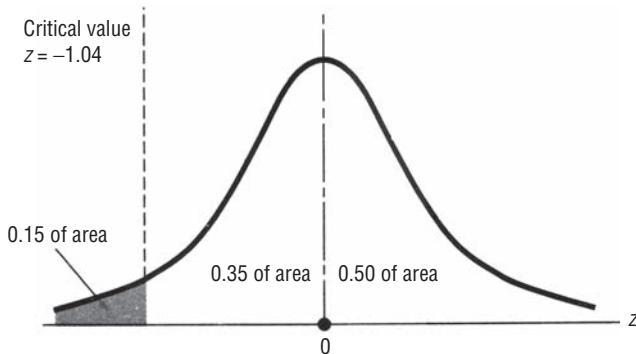


FIGURE 9-10 ONE-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO PROPORTIONS AT THE 0.15 LEVEL OF SIGNIFICANCE

$H_0: p_1 = p_2 \leftarrow$ Null hypothesis: There is no difference between the two methods

$H_1: p_1 < p_2 \leftarrow$ Alternative hypothesis: The personal-appearance method has a lower proportion of errors than the mail-in method

$\alpha = 0.15 \leftarrow$ Level of significance for testing the hypothesis

Step 1: State your hypotheses, type of test, and significance level

With samples of this size, we can use the standard normal distribution and Appendix Table 1 to determine the critical value of z for 0.35 of the area under the curve ($0.50 - 0.15$). We can use this value, 1.04, as the boundary of the acceptance region.

Step 2: Choose the appropriate distribution and find the critical value

Figure 9-10 illustrates this hypothesis test. Because the city manager wishes to test whether the personal-appearance listing is better than the mailed-in listing, the appropriate test is a one-tailed test. Specifically, it is a *left-tailed* test, because to reject the null hypothesis, the test result must fall in the colored portion of the left tail, indicating that *significantly fewer errors* exist in the personal-appearance forms. This colored region in Figure 9-10 corresponds to the 0.15 significance level.

To estimate the *standard error of the difference between two proportions*, we first use the combined proportions from both samples to estimate the overall proportion of successes:

Step 3: Compute the standard error and standardize the sample statistic

$$\hat{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad [9-7]$$

$$= \frac{(50)(0.10) + (75)(0.133)}{50 + 75}$$

$$= \frac{5 + 10}{125}$$

$$= 0.12 \leftarrow \text{Estimate of the overall proportion of successes in the population using combined proportions from both samples}$$

Now this answer can be used to calculate the estimated standard error of the difference between the two proportions, using Equation 9-8:

$$\begin{aligned}
 \hat{\sigma}_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} & [9-8] \\
 &= \sqrt{\frac{(0.12)(0.88)}{50} + \frac{(0.12)(0.88)}{75}} \\
 &= \sqrt{\frac{0.10560}{50} + \frac{0.10560}{75}} \\
 &= \sqrt{0.00352} \\
 &= 0.0593 \leftarrow \text{Estimated standard error of the difference between two proportions using combined estimates}
 \end{aligned}$$

We use the estimated standard error of the difference, $\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}$, to convert the observed difference between the two sample proportions, $\bar{p}_1 - \bar{p}_2$, to a standardized value:

$$\begin{aligned}
 z &= \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - \bar{p}_2)_{H_0}}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} \\
 &= \frac{(0.10 - 0.133) - 0}{0.0593} \\
 &= -0.556
 \end{aligned}$$

Step 4: Sketch the distribution and mark the sample value and critical value

Figure 9-11 shows where this standardized difference lies in comparison to the critical value.

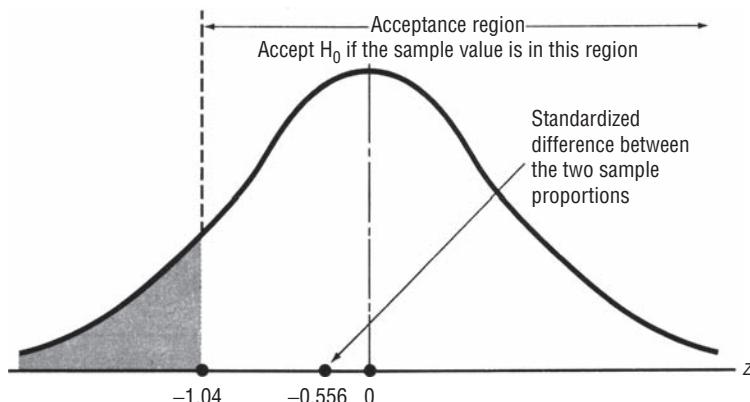


FIGURE 9-11 ONE-TAILED HYPOTHESIS TEST OF THE DIFFERENCE BETWEEN TWO PROPORTIONS AT THE 0.15 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED DIFFERENCE BETWEEN THE SAMPLE PROPORTIONS

This figure shows us that the standardized difference between the two sample proportions lies well within the acceptance region, and the city manager should accept the null hypothesis that there is no difference between the two methods of tax listing. Therefore, if mailed-in listing is considerably less expensive to the city, the city manager should consider increasing the use of this method.

HINTS & ASSUMPTIONS

The procedure here is almost like the one we used earlier in comparing differences between two means using independent samples. The only difference here is that we first use the combined proportions from both samples to estimate the overall proportion, then we use that answer to estimate the standard error of the difference between the two proportions. Hint: If the test is concerned with whether one proportion is significantly *different* from the other, use a two-tailed test. If the test asks whether one proportion is significantly *higher* or significantly *lower* than the other, then a one-tailed test is appropriate.

EXERCISES 9.4

Self-Check Exercises

- SC 9-7** A large hotel chain is trying to decide whether to convert more of its rooms to nonsmoking rooms. In a random sample of 400 guests last year, 166 had requested nonsmoking rooms. This year, 205 guests in a sample of 380 preferred the nonsmoking rooms. Would you recommend that the hotel chain convert more rooms to nonsmoking? Support your recommendation by testing the appropriate hypotheses at a 0.01 level of significance.
- SC 9-8** Two different areas of a large eastern city are being considered as sites for day-care centers. Of 200 households surveyed in one section, the proportion in which the mother worked full-time was 0.52. In another section, 40 percent of the 150 households surveyed had mothers working at full-time jobs. At the 0.04 level of significance, is there a significant difference in the proportions of working mothers in the two areas of the city?

Applications

- 9-20** On Friday, 11 stocks in a random sample of 40 of the roughly 2,500 stocks traded on the New York Stock Exchange advanced; that is, their price of their shares increased. In a sample of 60 NYSE stocks taken on Thursday, 24 advanced. At $\alpha = 0.10$, can you conclude that a smaller proportion of NYSE stocks advanced on Friday than did on Thursday?
- 9-21** MacroSwift has recently released a new word-processing product, and they are interested in determining whether people in the 30–39 age group rate the program any differently than members of the 40–49 age group. MacroSwift randomly sampled 175 people in the 30–39 age group who purchased the product and found 87 people who rated the program as excellent, with 52 people who would purchase an upgrade. They also sampled 220 people in the 40–49 age group and found 94 people who gave an excellent rating, with 37 people who plan to purchase an upgrade. Is there any significant difference in the proportions of people in the two age groups who rate the program as excellent at the $\alpha = 0.05$ level? Is the same result true for proportions of people who plan to purchase an upgrade?

- 9-22** A coal-fired power plant is considering two different systems for pollution abatement. The first system has reduced the emission of pollutants to acceptable levels 68 percent of the time, as determined from 200 air samples. The second, more expensive system has reduced the emission of pollutants to acceptable levels 76 percent of the time, as determined from 250 air samples. If the expensive system is significantly more effective than the inexpensive system in reducing pollutants to acceptable levels, then the management of the power plant will install the expensive system. Which system will be installed if management uses a significance level of 0.02 in making its decision?
- 9-23** A group of clinical physicians is performing tests on patients to determine the effectiveness of a new antihypertensive drug. Patients with high blood pressure were randomly chosen and then randomly assigned to either the control group (which received a well-established anti-hypertensive) or the treatment group (which received the new drug). The doctors noted the percentage of patients whose blood pressure was reduced to a normal level within 1 year. At the 0.01 level of significance, test appropriate hypotheses to determine whether the new drug is significantly more effective than the older drug in reducing high blood pressure.

Group	Proportion That Improved	Number of Patients
Treatment	0.45	120
Control	0.36	150

- 9-24** The University Bookstore is facing significant competition from off-campus bookstores, and they are considering targeting a specific class in order to retain student business. The bookstore randomly sampled 150 freshmen and 175 sophomores. They found that 46 percent of the freshmen and 40 percent of the sophomores purchase all of their textbooks at the University Bookstore. At $\alpha = 0.10$, is there a significant difference in the proportions of freshman and sophomores who purchase entirely at the University Bookstore?
- 9-25** In preparation for contract-renewal negotiations, the United Manufacturing Workers surveyed its members to see whether they preferred a large increase in retirement benefits or a smaller increase in salary. In a group of 1,000 male members who were polled, 743 were in favor of increased retirement benefits. Of 500 female members surveyed, 405 favored the increase in retirement benefits.
- Calculate \hat{p} .
 - Compute the standard error of the difference between the two proportions.
 - Test the hypothesis that equal proportions of men and women are in favor of increased retirement benefits. Use the 0.05 level of significance.

Worked-Out Answers to Self-Check Exercises

SC 9-7 $n_1 = 400 \quad \bar{p}_1 = 0.415 \quad n_2 = 380 \quad \bar{p}_2 = 0.5395$

$$H_0: p_1 = p_2 \quad H_1: p_1 < p_2 \quad \alpha = 0.01$$

$$\hat{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{400(0.415) + 380(0.5395)}{400 + 380} = 0.4757$$

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.4757(0.5243)\left(\frac{1}{400} + \frac{1}{380}\right)} = 0.0358$$

The lower limit of the acceptance region is $z = -2.33$, or

$$\bar{p}_1 - \bar{p}_2 = 0 - z\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = -2.33(0.0358) = -0.0834$$

Because the observed z value = $\frac{\bar{p}_1 - \bar{p}_2}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} = \frac{0.415 - 0.5395}{0.0358} = -3.48$

< -2.33 (or $\bar{p}_1 - \bar{p}_2 = -0.1245 < -0.0834$) we reject H_0 . The hotel chain should convert more rooms to nonsmoking because there was a significant increase in the proportion of guests requesting these rooms over the last year.

SC 9-8 $n_1 = 200$ $\bar{p}_1 = 0.52$ $n_2 = 150$ $\bar{p}_2 = 0.40$

$$H_0: p_1 = p_2 \quad H_1: p_1 \neq p_2 \quad \alpha = 0.40$$

$$\hat{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{200(0.52) + 150(0.40)}{200 + 150} = 0.4686$$

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.4686(0.5314)\left(\frac{1}{200} + \frac{1}{150}\right)} = 0.0539$$

The limits of the acceptance region are $z = \pm 2.05$, or

$$\bar{p}_1 - \bar{p}_2 = 0 \pm z\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \pm 2.05(0.0539) = \pm 0.1105$$

Because the observed z value = $\frac{\bar{p}_1 - \bar{p}_2}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} = \frac{0.52 - 0.40}{0.0539} = 2.23 > 2.05$ (or $p_1 - p_2 = 0.12 > 0.1105$),

we reject H_0 . The proportions of working mothers in the two areas differ significantly.

9.6 PROB VALUES: ANOTHER WAY TO LOOK AT TESTING HYPOTHESES

In all the work we've done so far on hypothesis testing, one of the first things we had to do was choose a level of significance, α , for the test. It has been traditional to choose a significance level of $\alpha = 10$ percent, 5 percent, 2 percent, or 1 percent, and almost all our examples have been done at these levels. But why use only these few values?

When we discussed Type I and Type II errors on page 387, we saw that the choice of the significance level depended on a trade-off between the costs of each of these two kinds of errors. If the cost of a Type I error (incorrectly rejecting H_0) is relatively high, we want to avoid making this kind of error, so we choose a small value of α . On the other hand, if a Type II error (incorrectly accepting H_0) is relatively more expensive, we are more willing to make a Type I error, and we choose a high value of α . **However, understanding the nature of the trade-off still doesn't tell us how to choose a significance level.**

How do we choose a significance level?

When we test the hypotheses:

$$H_0: \mu = \mu_{H_0}$$

$$H_1: \mu \neq \mu_{H_0}$$

$$\alpha = 0.05$$

Deciding before we take a sample

we take a sample, compute \bar{x} and reject H_0 if \bar{x} is so far from μ_{H_0} that the probability of seeing a value of \bar{x} this far (or farther) from μ_{H_0} is less than 0.05. In other words, **before we take the sample**, we specify how unlikely the observed results will have to be in order for us to reject H_0 . There is another way to approach this decision about rejecting or accepting H_0 that doesn't require that we specify the significance level before taking the sample. Let's see how it works.

Suppose we take our sample, compute \bar{x} , and then ask the question, "Supposing H_0 were true, what's the probability of getting a value of \bar{x} this far or farther from μ_{H_0} ?" This probability is called a *prob value* or a *p-value*. **Whereas before we asked, "Is the probability of what we've observed less than α ?" now we are merely asking, "How unlikely is the result we have observed?" Once the prob value for the test is reported, then the decision maker can weigh all the relevant factors and decide whether to accept or reject H_0 , without being bound by a prespecified significance level.**

Another benefit of using prob values is that they provide more information. If you know that I rejected H_0 at $\alpha = 0.05$, you know only that \bar{x} was *at least* 1.96 standard errors away from μ_{H_0} . However, a prob value of 0.05 tells you that \bar{x} was *exactly* 1.96 standard errors away from μ_{H_0} . Let's look at an example.

Prob values

Another advantage

Two-Tailed Prob Values When σ Is Known

A machine is used to cut wheels of Swiss cheese into blocks of specified weight. On the basis of long experience, it has been observed that the weight of the blocks is normally distributed with a standard deviation of 0.3 ounce. The machine is currently set to cut blocks that weigh 12 ounces. A sample of nine blocks is found to have an average weight of 12.25 ounces. Should we conclude that the cutting machine needs to be recalibrated?

Written symbolically, the data in our problem are

$$\mu_{H_0} = 12 \quad \leftarrow \text{Hypothesized value of the population mean}$$

$$\sigma = 0.3 \quad \leftarrow \text{Population standard deviation}$$

$$n = 9 \quad \leftarrow \text{Sample size}$$

$$\bar{x} = 12.25 \quad \leftarrow \text{Sample mean}$$

Setting up the problem symbolically

The hypotheses we wish to test are

$$H_0: \mu = 12 \quad \leftarrow \text{Null hypothesis: The true population mean weight is 12 ounces}$$

$$H_1: \mu \neq 12 \quad \leftarrow \text{Alternative hypothesis: The true population mean weight is not 12 ounces}$$

Because this is a two-tailed test, our prob value is the probability of observing a value of \bar{x} at least as far away (on either side) from 12 as 12.25, if H_0 is true. In other words, the prob value is the probability

of getting $\bar{x} \geq 12.25$ or $\bar{x} \leq 11.75$ if H_0 is true. To find this probability, we first use Equation 6-1 to calculate the standard error of the mean:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} & [6-1] \\ &= \frac{0.3}{\sqrt{9}} \\ &= \frac{0.3}{3} \\ &= 0.1 \text{ ounce} \leftarrow \text{Standard error of the mean}\end{aligned}$$

Calculating the standard error of the mean

Then we use this to convert \bar{x} to a standard z score:

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} & [6-2] \\ &= \frac{12.25 - 12}{0.1} \\ &= \frac{0.25}{0.1} \\ &= 2.5\end{aligned}$$

Finding the z score and the prob value

From Appendix Table 1, we see that the probability that z is greater than 2.5 is $0.5000 - 0.4938 = 0.0062$. Hence, because this is a two-tailed hypothesis test, the prob value is $2(0.0062) = 0.0124$. Our results are illustrated in Figure 9-12. Given this information, our cheese packer can now decide whether to recalibrate the machine (reject H_0) or not (accept H_0).

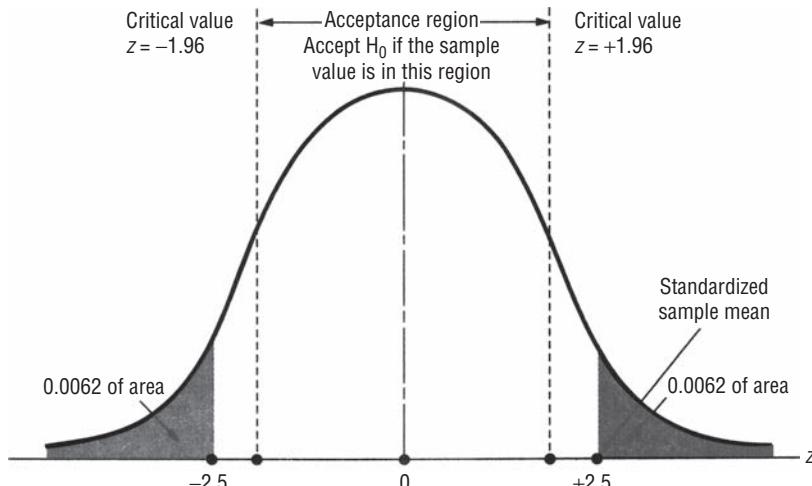


FIGURE 9-12 TWO-TAILED HYPOTHESIS TEST, SHOWING PROB VALUE OF 0.0124 (IN BOTH TAILS COMBINED)

How is this related to what we did before, when we specified a significance level? If a significance level of $\alpha = 0.05$ were adopted, we would reject H_0 . You can easily see this by looking at Figure 9-12. At a significance level of $\alpha = 0.05$, we reject H_0 if \bar{x} is so far from μ_{H_0} that less than 0.05 of the area under the curve is left in the two tails. Because our observed value of $\bar{x} = 12.25$ leaves only 0.0124 of the total area in the tails, we would reject H_0 at a significance level of $\alpha = 0.05$. (You can also verify this result by noting in Appendix Table 1 that the critical z values for $\alpha = 0.05$ are ± 1.96 . Thus, the standardized value of \bar{x} (2.5) is *outside* the acceptance region.)

Relationship between prob values and significance levels

Similarly, we can see that at a significance level of $\alpha = 0.01$, we would accept H_0 , because $\bar{x} = 12.25$ leaves more than 0.01 of the total area in the tails. (In this case, the critical z values for $\alpha = 0.01$ would be ± 2.58 , and now the standardized value of \bar{x} , 2.5, is *inside* the acceptance region.) In fact, at any level of α above 0.0124, we would reject H_0 . **Thus, we see that the prob value is precisely the largest significance level at which we would accept H_0 .**

Prob Values under Other Conditions

In our example, we did a two-tailed hypothesis test using the normal distribution. How would we proceed in other circumstances?

- If σ was known, and we were doing a one-tailed test, we would compute the prob value in exactly the same way except that we would not multiply the probability that we got from Appendix Table 1 by 2, because that table gives one-tailed probabilities directly.
- If σ was not known, we would use the t distribution with $n - 1$ degrees of freedom and Appendix Table 2. This table gives two-tailed probabilities, but only a few of them, so we can not get exact prob values from it. For example, for a two-tailed test, if $\mu_{H_0} = 50$, $\bar{x} = 49.2$, $s = 1.4$, and $n = 16$, we find that

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \frac{\hat{\sigma}}{\sqrt{n}} \\ &= \frac{1.4}{\sqrt{16}} \\ &= 0.35\end{aligned}\quad [7-6]$$

and that \bar{x} is 2.286 estimated standard errors below μ_{H_0} [$(49.2 - 50)/0.35 = -2.286$]. Looking at the 15 degrees of freedom row in Appendix Table 2, we see that 2.286 is between 2.131 ($\alpha = 0.05$) and 2.602 ($\alpha = 0.02$). Our prob value is therefore something between 0.02 and 0.05, but we can't be more specific.

Most computer statistics packages report exact prob values, not only for tests about means based on the normal distribution, but for other tests such as chi-square and analysis of variance (which we will discuss in Chapter 11) and tests in the context of linear regression (which we will discuss in Chapters 12 and 13). The discussion we have provided in this section

One-tailed prob values

Using the t distribution

Prob values in other contexts

will enable you to understand prob values in those contexts too. Although different statistics and distributions will be involved, the ideas are the same.

HINTS & ASSUMPTIONS

Prob values and computers eliminate having to look up values from a z or t distribution table, and take the drudgery out of hypothesis testing. Warning: The *smaller* the prob value, the greater the significance of the findings. Hint: You can avoid confusion here by remembering that a prob value is the chance that the result you have could have occurred by sampling error, thus, smaller prob values mean smaller chances of sampling error and higher significance.

EXERCISES 9.5

Self-Check Exercises

- SC 9-9** The Coffee Institute has claimed that more than 40 percent of American adults regularly have a cup of coffee with breakfast. A random sample of 450 individuals revealed that 200 of them were regular coffee drinkers at breakfast. What is the prob value for a test of hypotheses seeking to show that the Coffee Institute's claim was correct? (*Hint:* Test $H_0: p = 0.4$, versus $H_1: p > 0.4$)
- SC 9-10** Approximately what is the prob value for the test in Self-Check Exercise 9-3 on page 448?

Applications

- 9-26** A car retailer thinks that a 40,000-mile claim for tire life by the manufacturer is too high. She carefully records the mileage obtained from a sample of 64 such tires. The mean turns out to be 38,500 miles. The standard deviation of the life of all tires of this type has previously been calculated by the manufacturer to be 7,600 miles. Assuming that the mileage is normally distributed, determine the largest significance level at which we would accept the manufacturer's mileage claim, that is, at which we would not conclude the mileage is significantly less than 40,000 miles.
- 9-27** The North Carolina Department of Transportation has claimed that at most, 18 percent of passenger cars exceed 70 mph on Interstate 40 between Raleigh and Durham. A random sample of 300 cars found 48 cars exceeding 70 mph. What is the prob value for a test of hypothesis seeking to show the NCDOT's claim is correct?
- 9-28** Kelly's machine shop uses a machine-controlled metal saw to cut sections of tubing used in pressure-measuring devices. The length of the sections is normally distributed with a standard deviation of 0.06". Twenty-five pieces have been cut with the machine set to cut sections 5.00" long. When these pieces were measured, their mean length was found to be 4.97". Use prob values to determine whether the machine should be recalibrated because the mean length is significantly different from 5.00"?
- 9-29** SAT Services advertises that 80 percent of the time, its preparatory course will increase an individual's score on the College Board exams by at least 50 points on the combined verbal and quantitative total score. Lisle Johns, SAT's marketing director, wants to see whether this is a reasonable claim. He has reviewed the records of 125 students who took the course and found that 94 of them did, indeed, increase their scores by at least

50 points. Use prob values to determine whether SAT's ads should be changed because the percentage of students whose scores increase by 50 or more points is significantly different from 80 percent.

- 9-30 What is the prob value for the test in Exercise 9-2?
- 9-31 What is the prob value for the test in Exercise 9-3?
- 9-32 Approximately what is the prob value for the test in Exercise 9-8?
- 9-33 Approximately what is the prob value for the test in Exercise 9-11?
- 9-34 Approximately what is the prob value for the test in Exercise 9-14?
- 9-35 Approximately what is the prob value for the test in Exercise 9-15?
- 9-36 What is the prob value for the test in Exercise 9-22?
- 9-37 What is the prob value for the test in Exercise 9-25?

Worked-Out Answers to Self-Check Exercises

SC 9-9 $n = 450 \quad \bar{p} = 200/450 = 0.4444$

$$H_0: p = 0.4 \quad H_1: p > 0.4$$

The prob value is the probability that $\bar{p} \geq 0.4444$, that is,

$$P\left(z \geq \frac{0.4444 - 0.4}{\sqrt{0.4(0.6)/450}}\right) = P(z \geq 1.92) = 0.5 - 0.4726 = 0.0274$$

SC 9-10 From the solution to exercise SC 9-3 on page 422, we have $t = -2.766$, with $12 + 9 - 2 = 19$ degrees of freedom. From the row in Appendix Table 2 for 19 degrees of freedom, we see that -2.766 is between -2.861 (corresponding to a probability of $.01/2 = .005$ in the lower tail) and -2.539 (corresponding to a probability of $.02/2 = .01$ in the lower tail). Hence the prob value for our test is between $.005$ and $.01$.

STATISTICS AT WORK

Loveland Computers

Case 9: Two-Sample Tests of Hypotheses When Lee Azko looked over the results of the telephone survey conducted by the marketing department of Loveland Computers, something was troubling. “Hmm, you wouldn’t still have the data for the ‘Total spent on software’ on computer, would you, Margot?” Lee asked the head of the department.

“Hey, I keep *everything*,” Margot replied. “It’s in a worksheet file on the computer over there. I had the intern camp out in my office last summer. Why do you need to see the data?”

“Well, give me a minute and I’ll show you,” said Lee, turning on the machine. After a few minutes of muttering over the keyboard, Lee pushed back from the screen. “Thought so! Take a look at that. It looks like there are really two groups of customers here—see how there are two different peaks on this graph?”

“I guess we should have done more than just print out the mean and standard deviation last summer,” said Margot disconsolately. “I guess this means the data are no good.”

“Not necessarily,” said Lee with more optimism. “I’ll bet your ‘big spenders’ are your business customers and the lower peak is the home users. You wouldn’t have any way to know which category the response was in, would you?”

"Well, we capture that automatically," Margot said, leaning over and clearing the graph from the screen. "If you look at the first column, you'll see that it's the customer number. All the business customers have a customer number that begins with a 1 and all the home users have a customer number that begins with a 2."

"Let me copy this file onto a floppy," Lee said, opening the briefcase. "I'll be back this afternoon with the answer."

Study Questions: What graph did Lee plot using the worksheet program? What hypothesis is being tested and what is the appropriate statistical test? Is this a one-tailed or a two-tailed problem?

CHAPTER REVIEW

Terms Introduced in Chapter 9

Combined Proportion of Successes In comparing two population proportions, the total number of successes in both samples divided by the total size of both samples; used to estimate the proportion of successes common to both populations.

Dependent Samples Samples drawn from two populations in such a way that the elements in one sample are matched or paired with the elements in the other sample, in order to allow a more precise analysis by controlling for extraneous factors.

Paired Difference Test A hypothesis test of the difference between two population means based on the means of two dependent samples.

Paired Samples Another name for dependent samples.

Pooled Estimate of σ^2 A weighted average of s_1^2 and s_2^2 used to estimate the common variance, σ^2 , when using small samples to test the difference between two population means.

Prob Value The largest significance level at which we would accept the null hypothesis. It enables us to test hypotheses without first specifying a value for α .

P-value Another name for a prob value.

Two-Sample Tests Hypothesis tests based on samples taken from two populations in order to compare their means or proportions.

Equations Introduced in Chapter 9

$$9-1 \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad p. 428$$

This formula enables us to derive the standard deviation of the distribution of the difference between two sample means, that is, *the standard error of the difference between two means*. To do this, we take the square root of the sum of Population 1's variance divided by its sample size plus Population 2's variance divided by its sample size.

$$9-2 \quad \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad p. 428$$

If the two population standard deviations are unknown, we can use this formula to derive the *estimated* standard error of the difference between two means. We can use this equation after we have used the two sample standard deviations and Equation 7-1 to determine the estimated standard deviations of Population 1 and Population 2, ($\hat{\sigma} = s$).

9-3 $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ p. 435

With this formula, we can get a pooled estimate of σ^2 . It uses a weighted average of s_1^2 and s_2^2 , where the weights are the numbers of degrees of freedom in each sample. Use of this formula assumes that $\sigma_1^2 = \sigma_2^2$ (that the unknown population variances are equal). We use this formula when testing for the differences between means in situations with small sample sizes (less than 30).

9-4 $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ p. 436

Given the pooled estimate of σ^2 obtained from Equation 9-3, we put this value into Equation 9-2 and simplify the expression. This gives us a formula to estimate the standard error of the difference between sample means when we have small samples (less than 30) but equal population variances.

9-5 $\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$ p. 457

This is the formula used to derive the standard error of the difference between two *proportions*. The symbols and p_1 and p_2 represent the proportions of successes in Population 1 and Population 2, respectively, and q_1 and q_2 are the proportions of failures in Populations 1 and 2, respectively.

9-6 $\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}}$ p. 457

If the population parameters p and q are unknown, we can use the sample statistics \bar{p} and \bar{q} and this formula to *estimate* the standard error of the difference between two proportions.

9-7 $\hat{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$ p. 458

Because the null hypothesis assumes that there is *no difference* between the two population proportions, it would be more appropriate to modify Equation 9-6 and to use the combined proportions from both samples to estimate the overall proportion of successes in the combined populations. Equation 9-7 combines the proportions from both samples. Note that the value of \hat{q} is equal to $1 - \hat{p}$.

9-8 $\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$ p. 458

Now we can substitute the results of Equation 9-7, both \hat{p} and \hat{q} , into Equation 9-6 and get a more correct version of Equation 9-6. This new equation, 9-8, gives us the *estimated* standard error of the difference between the two proportions using combined estimates from both samples.

Review and Application Exercises

- 9-38 Clic Pens has tested two types of point-of-purchase displays for its new erasable pen. A shelf display was placed in a random sample of 40 stores in the test market, and a floor display was placed in 40 other stores in the area. The mean number of pens sold per store in one month

with the shelf display was 42, and the sample standard deviation was 8. With the floor display, the mean number of pens sold per store in the same month was 45, and the sample standard deviation was 7. At $\alpha = 0.02$, was there a significant difference between sales with the two types of displays?

- 9-39** In 1992, a survey of 50 municipal hospitals revealed an average occupancy rate of 73.6 percent, and the sample standard deviation was 18.2 percent. Another survey of 75 municipal hospitals in 1995 found an average occupancy rate of 68.9 percent, and the sample standard deviation was 19.7 percent. At $\alpha = 0.10$, can we conclude that the average occupancy rate changed significantly during the 3 years between surveys?

- 9-40** General Cereals has just concluded a new advertising campaign for Fruit Crunch, its all-natural breakfast cereal with nuts, grains, and dried fruits. To test the effectiveness of the campaign, brand manager Alan Neebe surveyed 11 customers before the campaign and another 11 customers after the campaign. Given are the customers' reported weekly consumption (in ounces) of Fruit Crunch:

Before	14	5	18	18	30	10	8	26	13	29	24
After	23	14	13	29	33	11	12	25	21	26	34

- (a) At $\alpha = 0.05$, can Alan conclude that the campaign has succeeded in increasing demand for Fruit Crunch?
- (b) Given Alan's initial survey before the campaign, can you suggest a better sampling procedure for him to follow after the campaign?

- 9-41** Students Against Drunk Driving has targeted seat-belt usage as a positive step to reduce accidents and injuries. Before a major campaign at one high school, 44 percent of 150 drivers entering the school parking lot were using their seat belts. After the seat-belt awareness program, the proportion using seat belts had risen to 52 percent in a sample of 200 vehicles. At a 0.04 significance level, can the students conclude that their campaign was effective?

- 9-42** Allen Distributing Company hypothesizes that a phone call is more effective than a letter in speeding up collection of slow accounts. Two groups of slow accounts were contacted, one by each method, and the length of time between mailing the letter or making the call and the receipt of payment was recorded:

Method Used	Days to Collection						
	10	8	9	11	11	14	10
Letter							
Phone call	7	4	5	4	8	6	9

- (a) At $\alpha = 0.025$, should Allen conclude that slow accounts are collected more quickly with calls than with letters?
- (b) Can Allen conclude that slow accounts respond more quickly to calls?

- 9-43** A buffered aspirin recently lost some of its market share to a new competitor. The competitor advertised that its brand enters the bloodstream faster than the buffered aspirin does and, as a result, it relieves pain sooner. The buffered-aspirin company would like to prove that there is no significant difference between the two products and, hence, that the competitor's claim is false. As a preliminary test, 9 subjects were given buffered aspirin once a day for 3 weeks. For another 3 weeks, the same subjects were given the competitive product. For

each medication, the average number of minutes it took to reach each subject's bloodstream was recorded:

Subject	1	2	3	4	5	6	7	8	9
Buffered aspirin	16.5	25.5	23.0	14.5	28.0	10.0	21.5	18.5	15.5
Competitor	12.0	20.5	25.0	16.5	24.0	11.5	17.0	15.0	13.0

At $\alpha = 0.10$, is there any significant difference in the times the two medications take to reach the bloodstream?

- 9-44 Eros India Pvt. Ltd is a leading manufacturer of washing machines. Its semi-automatic washing machine named INVA and fully-automatic washing machine INTA are market leaders in their respective segments. In the 4th quarter (Jan.–March) of 2010–11, there was a slight decline in the sales of INVA as compared to the previous quarter, but the sales of INTA increased.

	Sales in 3rd Quarter (Oct.–Dec.) 2010–11	Sales in 4th quarter (Jan.–March) 2010–11
INVA	59.2 %	57.9%
INTA	40.8%	42.1%
Total units Sold	88,841	88,057

Is the % change in share of INVA versus INTA significant, at 5% level?

- 9-45 A chemist developing insect repellents wishes to know whether a newly developed formula gives greater protection from insect bites than that given by the leading product on the market. In an experiment, 14 volunteers each had one arm sprayed with the old product and the other sprayed with the new formula. Then each subject placed his arms into two chambers filled with equal numbers of mosquitoes, gnats, and other biting insects. The numbers of bites received on each arm follow. At $\alpha = 0.01$, should the chemist conclude that the new formula is, indeed, more effective than the current market leader?

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Old formula	5	2	5	4	3	6	2	4	2	6	5	7	1	3
New formula	3	1	5	1	1	4	4	2	5	2	3	3	1	2

- 9-46 Long Distance Carrier is trying to see the effect of offering “1 month free” with a monthly fixed fee of \$10.95, versus an offer of a low monthly fee \$8.75 with no free month. To test which might be more attractive to consumers, Long Distance runs a brief market test: 12 phone reps make calls using one approach, and 10 use the other. The following number of customers agreed to switch from their present carrier to LDC:

Offer	Number of Switches											
1 month free	118	115	122	99	106	125	102	100	92	103	113	129
Low monthly fee	115	126	113	110	135	102	124	137	108	128		

Test at a significance level of 10 percent whether there are significant productivity differences with the two offers.

- 9-47 Is the perceived level of responsibility for an action related to the severity of its consequences? That question was the basis of a study of responsibility in which the subjects read a description

of an accident on an interstate high-way. The consequences, in terms of cost and injury, were described as either very minor or serious. A questionnaire was used to rate the degree of responsibility that the subjects believed should be placed on the main figure in the story. Below are the ratings for both the mild-consequences and the severe-consequences groups. High ratings correspond to higher responsibility attributed to the main figure. If a 0.025 significance level was used, did the study conclude that severe consequences lead to a greater attribution of responsibility?

Consequences	Degree of Responsibility							
	Mild	4	5	3	3	4	1	2
Severe	4	5	4	6	7	8	6	5

- 9-48** In October 1992, a survey of 120 macroeconomists found 87 who believed that the recession had already ended. A survey of 150 purchasing agents found 89 who believed the recession had ended. At $\alpha=0.10$, should you conclude that the purchasing agents were more pessimistic about the economy than the macroeconomists were?
- 9-49** The MBA program at Piedmont Business School offers Analytic Skills Workshop (ASW) during the summer to help entering students brush up on their accounting, economics, and mathematics. Program Director Andy Bunch wonders whether ASW has been advantageous to the students enrolled. He has taken random samples of grade-point averages for students enrolled in ASW over the past 5 years and for students who started the MBA program without ASW during the same time span. At $\alpha=0.02$, have the ASW students gotten significantly higher GPAs? Should Andy advertise that ASW helps student achievement in the MBA program?

	\bar{x}	s	n
ASW	3.37	1.13	26
Non-ASW	3.15	1.89	35

- 9-50** Fifty-eight of 2,000 randomly sampled corporations had their 1995 federal income tax returns audited. In another sample of 2,500 corporations, 61 had their 1994 returns audited. Was the fraction of corporate returns audited in 1995 significantly different from the 1994 fraction? Test the appropriate hypotheses at $\alpha=0.01$.
- 9-51** Ellen Singer asserted to one of her colleagues at Triangle Realty that homes in southern Durham County sold for about \$15,000 less than similar homes in Chapel Hill. To test this assertion, her colleague randomly chose 10 recent sales in Chapel Hill and matched them with 10 recent sales in southern Durham County in terms of style, size, age, number of rooms, and size of lot. At $\alpha=0.05$, do the following data (selling prices in thousands of \$) support Ellen's claim?

Chapel Hill	97.3	108.4	135.7	142.3	151.8	158.5	177.4	183.9	195.2	207.6
Durham County	81.5	92.0	115.8	137.8	150.9	149.2	168.2	173.9	175.9	194.4

- 9-52** The following table gives the closing prices stocks of ten FMCGs Companies as on January 31st, 2012. These share-prices are to be compared with share-prices of the same companies as in the last year. Analyze the data at level of significance of 1 percent and comment whether there is a significant decrease as compared to the last year.

Company	Closing Price (January 31st, 2012)	Closing Price (January 31st, 2011)
XML Company	33.89	43.78
IWL Enterprises	26.32	22.56
TYE Company	72.67	90.91
NIR Company	18.53	20.45
SUP Enterprises	61.45	61.45
ANT Company	45.49	54.21
ZNN Enterprises	80.00	79.27
KYE Enterprises	58.12	58.75
PCU Company	41.38	40.86
NTR Company	39.06	40.12

- 9-53** TV network executive Terri Black has just received a proposal and a pilot tape for a new show. *Empty Nest No Longer* is a situation comedy about a middle-aged couple whose two college-graduate offspring cannot find jobs and have returned home. Terri wonders whether the show will appeal to twenty-somethings as well as to an older audience. Figuring that people in her office are reasonably representative of their age group in the population as a whole, she asks them to evaluate the pilot tape on a scale from 0 to 100, and gets the following responses.

Age	Responses							
	20–29	86	74	73	65	82	78	79
≥ 30	63	72	68	75	73	80		

- (a) At a 0.05 significance level, should Terri conclude the show will be equally attractive to the two age groups?
(b) Independent of your answer to (a), do you think Terry should use the results of her office survey to decide how to design an advertising campaign for *Empty Nest No Longer*? Explain.
- 9-54** A manufacturer of pet foods was wondering whether cat owners and dog owners reacted differently to premium pet foods. They commissioned a consumer survey that yielded the following data.

Pet	Owners Surveyed	Number Using Premium Food
Cat	280	152
Dog	190	81

- Is it reasonable to conclude, at $\alpha = 0.02$, that cat owners are more likely than dog owners to feed their pets premium food?
- 9-55** Robin Wendell has been offered a transfer from Pittsburgh to Boston, but is holding out for more money, “because the cost of living there is so much more.” Looking at a grocery receipt and deleting big-ticket items, Robin came up with 36 items under \$2 with a mean of \$0.98, standard deviation \$0.43 in Pittsburgh. The recruiting manager stops by a Boston grocery store, and with the same \$2.00 limit, buys 42 items, with a mean price of \$1.07, standard deviation \$0.38. Is Robin right that the cost of groceries is more in Boston than in Pittsburgh, at a confidence level $\alpha = 0.01$? What could be done to improve the analysis of cost of living in the two cities?

9-56 Nirmal Pvt. Limited is a FMCG company, selling a range of products. It has 1150 sales-outlets. A sample of 60 sales-outlets was chosen, using random sampling for the purpose of sales analysis. The sample consists of sales-outlets from rural and urban areas belonging to the four regions of the country as Northern, Eastern, Western and Southern. The information related to annual sales has been collected from them in the month of December 2010. This process has been repeated in December 2011. In the meanwhile, in 2010 a comprehensive sales-promotion program was launched to augment the sales. The information is presented in the data sheet in the DVD (Nirmal Pvt. Ltd DATA). Analyze the data and give answer to the following questions:

- (a) Can you conclude that there is significant difference between sales of urban outlets and rural outlets in 2010?
- (b) Does the pattern of differentiation between urban and rural outlets remain the same in 2011 also?
- (c) Do the data indicate that there is significant increase in the sales in 2011 as compared to 2010? Comment on the success of the sales-promotion program.

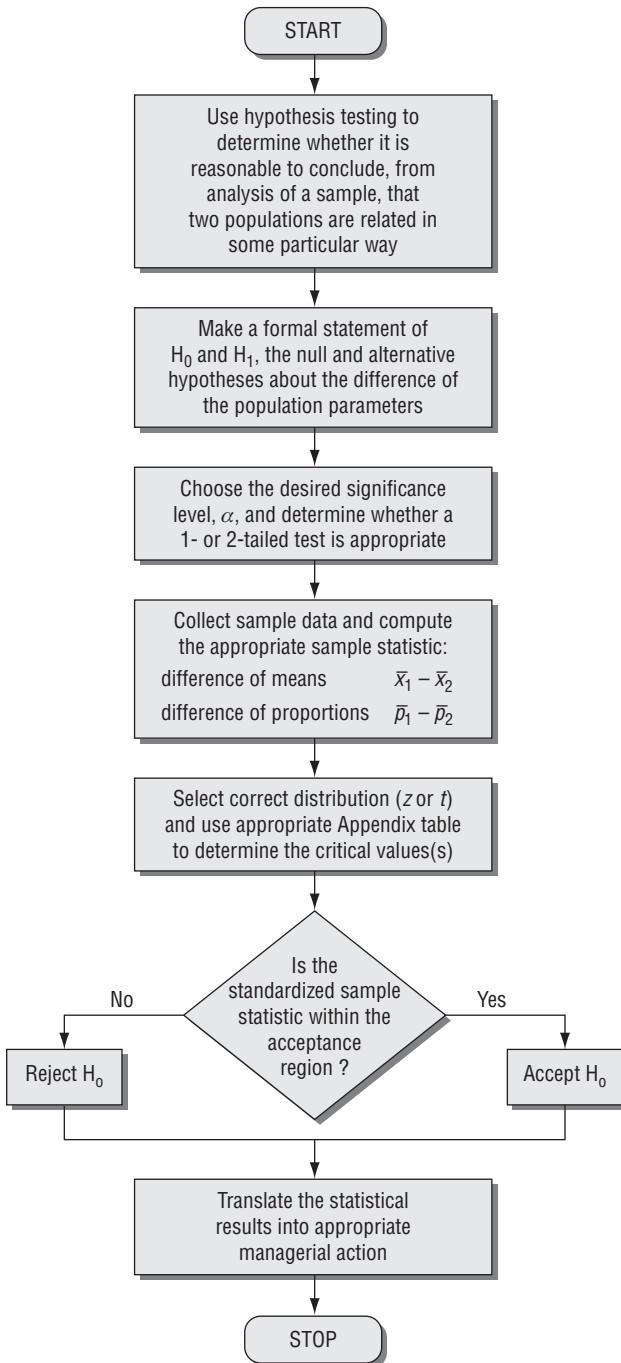


Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across the gender. (Q9 & Q15)
2. Test the hypothesis that both males and females perceive that private sector banks are better than the public sector banks in terms of the e-banking services provided by them. (Q13b & Q15)
3. Test the hypothesis that irrespective of the marital status of the respondents, people felt that e-banking lacks personal touch. (Q13k & Q16).



Flow Chart: Two-Sample Tests of Hypotheses



This page is intentionally left blank.

10

Quality and Quality Control

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To examine why the concept of quality—making sure that a product or service is consistent, reliable, and free from errors and defects—is important in decision making
 - To learn how to use control charts to monitor the output of a process and see whether it is meeting established quality standards
 - To recognize patterns that indicate that a process is out-of-control
 - To understand how to construct \bar{x} , R , and p charts
 - To introduce basic concepts of Total Quality Management
 - To learn how acceptance sampling is used to monitor the input to a process to ensure that it meets established quality standards
-

CHAPTER CONTENTS

- 10.1 Introduction 480
- 10.2 Statistical Process Control 482
- 10.3 \bar{x} Charts: Control Charts for Process Means 484
- 10.4 R Charts: Control Charts for Process Variability 495
- 10.5 p Charts: Control Charts for Attributes 501
- 10.6 Total Quality Management 508
- 10.7 Acceptance Sampling 514

- Statistics at Work 522
- Terms Introduced in Chapter 10 523
- Equations Introduced in Chapter 10 524
- Review and Application Exercises 525
- Flow Chart: Quality and Quality Control 529

The Durham City Executive for TransCarolina Bank has just established an express teller to handle transactions consisting of a single deposit or withdrawal. She hopes ultimately to be able to complete express transactions in an average of less than 60 seconds. For the moment, however, she wants to be sure that the express line works smoothly and consistently. Once the process is in-control, she can devote her attention to reducing its average time to meet her 60-second goal. For the past month, she has randomly sampled six express transactions each business day:

Day	Date	Transaction Time (seconds)						Day	Date	Transaction Times (seconds)					
M	5/03	63	55	56	53	61	64	M	5/17	57	63	56	64	62	59
T	5/04	60	63	60	65	61	66	T	5/18	66	63	65	59	70	61
W	5/05	57	60	61	65	66	62	W	5/19	63	53	69	60	61	58
Th	5/06	58	64	60	61	57	65	Th	5/20	68	67	59	58	65	59
F	5/07	79	68	65	61	74	71	F	5/21	70	62	66	80	71	76
M	5/10	55	66	62	63	56	52	M	5/24	65	59	60	61	62	65
T	5/11	57	61	58	64	55	63	T	5/25	63	69	58	56	66	61
W	5/12	58	51	61	57	66	59	W	5/26	61	56	62	59	57	55
Th	5/13	65	66	62	68	61	67	Th	5/27	65	57	69	62	58	72
F	5/14	73	66	61	70	72	78	F	5/28	70	60	67	79	75	68

Using the techniques discussed in this chapter, the Durham city executive can determine whether the express-teller line is in-control or out-of-control. ■

10.1 INTRODUCTION

We often hear that everyone talks about the weather but no one is prepared to do anything about it. Until recently, the same was true about American business and quality. However, as the relative isolation of national economies has been succeeded by the increasing globalization of commerce, American industry has had to respond to challenges from abroad. One of those challenges was a dedication to quality control and the management of quality in production that came to be epitomized by such Japanese products as automobiles and consumer electronics. In response to this challenge, the philosophy and techniques of quality control and management are becoming widespread in American manufacturing. In addition, the rapidly growing circle of applications of *Total Quality Management* (known by the acronym TQM) has expanded from manufacturing industries to service industries such as health care and legal services.

Responding to global challenge

TQM for services as well as for manufacturing

In this chapter, you'll see how simple applications of some of the ideas we've already discussed about estimation and hypothesis testing can be used for quality control and improvement. We'll look at *control charts* and *acceptance sampling*, two commonly used quality control techniques. Along the way, you'll meet some of the pioneers in the field and learn some of the language of quality control.

What Is Quality?

When you hear a commercial for a “high-quality automobile,” do you conjure up images of such luxurious options as leather seats and fancy sound systems? Most of us do connect *luxury* and *quality*. But expensive leather seats don’t mean very much if the engine won’t start on a cold morning, and the latest noise-reduction technology is hard to appreciate if the tape deck starts chewing up your tapes. These examples show us that it’s important to separate the notion of luxury from our discussion of quality.

Distinguish between luxury and quality

In fact, some of the cheapest items in everyday life can have very high quality. Consider the paper used in a copying machine. For little more than a penny a page, you can buy smooth white paper, less than one hundredth of an inch thick and of uniform size. You have come to expect such high quality in copier paper that you don’t examine individual pages before loading it into a copier. You wouldn’t think of measuring the thickness of each page to make sure that it was thin enough not to jam the copier, but thick enough so that you could print on both sides and not have the two images interfere with each other.

Quality means fitness for use

The copy paper example gives us a clue to a working definition of *quality*. Things that are of high quality are those that work in the way we expect them to. As quality expert Joseph M. Juran has put it, *quality implies fitness for use*. In this sense, *quality means conformance to requirements*. Note that this is not quite the same as conformance to specifications. Copy paper that is cut to size for American copy machines won’t fit European machines that demand the slightly narrower A4 metric format.

Consistency, reliability, and lack of defects

Note that the idea of “things that work in the way we expect them to” points out that quality is defined by customers as well as by producers. As you shall see, meeting the needs of customers is central to TQM. Working definitions of quality vary in different contexts, especially when we contrast goods and services. But in keeping with our notion of conformance to requirements, most working definitions of quality will include the concepts of *consistency*, *reliability*, and *lack of errors and defects*.

Variability Is the Enemy of Quality

When a craftsman makes something by hand, there is a continuous process of checking, measuring, and reworking. If you’d watched Michelangelo completing a sculpture, you wouldn’t have seen a final “quality control” step before he shipped his artwork to his patron. Indeed, quality control is not an issue when you are producing goods and services that are essentially unique. However, when mass production became common during the nineteenth century, it was soon realized that individual pieces could not be identical—a certain amount of variation is inevitable. But this leads to a problem: With too much variation, parts that are supposed to fit together won’t fit! In this sense, you can see why variability is the enemy of quality.

Mass production makes quality an issue

Controlling Variability: Inspection vs. Prevention

How should we deal with variability? Think about a stack of two-by-fours in a lumberyard. Most of them will conform to requirements, but some won’t because of twisting and warping as they dried, splitting when the saw hit a knot, or sundry other causes. One approach to mass production says it’s cheaper to push material through the process and sort out the defects at the end. This leads our lumberyard to

have an inspector examine two-by-fours as they come out of the drying kiln. Defective pieces go to the scrap heap.

In the early days of mass production, sorting out defects became the chief method of quality control. Armies of white-coated inspectors tested goods at the end of a production line and released only some of them to customers. It was widely believed that the cost of a few rejects didn't amount to much because the marginal cost of each unit was small. But by the late 1970s, people were pointing out that the costs of defects were much higher than supposed. The armies of inspectors had to be paid, and if defective products slipped through, there were warranty costs and loss of customers' goodwill.

They argued that it is simply cheaper to do things right the first time. They preached the concept of *zero defects*. If your electrical power was 99 percent reliable, you'd spend a lot of time resetting your clocks. A major airline with a 99.9 percent safety record would have several crashes each week! If we demand near-perfect performance from power companies and airlines, perhaps we should expect no less from the producers of all our goods and services.

When poorly made parts are passed down a production line, all subsequent work is wasted when the final product is rejected by quality control inspectors. But it's expensive to keep inspecting components to make sure they conform to requirements. Imagine how much time would be wasted if you had to examine each piece of paper for defects before loading a copier. This leads to the goal of preventing defects at each stage of manufacturing a product or delivering a service. To accomplish this, the people who make things are given the responsibility to check their work before it is passed on, rather than just letting sloppy work slide by to be caught at a final inspection. This also has the benefit of giving workers a greater investment of pride in the work they are doing—in this sense, they are more like craftsmen.

Early quality control: Sorting out defective finished goods

Zero defects as a goal

Preventing defects and increasing workers' pride

EXERCISES 10.1

Basic Concepts

- 10-1 Give an example of a very expensive product that has very low quality.
- 10-2 Give an example of a very inexpensive product that has very high quality.
- 10-3 What is a reasonable working definition of quality?
- 10-4 What actually makes quality control an issue of concern to management?
- 10-5 What kinds of costs would you gather to perform an “inspection versus prevention” analysis?
- 10-6 Define *zero defects* as a concept.

10.2 STATISTICAL PROCESS CONTROL

The key to managing for quality is to believe that excessive variability is not inevitable. When the output of some process is found to be unreliable, not always conforming to requirements, we must carefully examine the process and see how it can be controlled.

Variability is not inevitable.

In the 1920s, Walter A. Shewhart, a researcher at Bell Labs, created a system for tracking variation and identifying its causes. Shewhart's system of *statistical process control* (or *SPC*) was developed further and championed by his one-time colleague, W. Edwards Deming. For many years, Deming was a prophet without honor in the United States, but when Japan was rebuilding its economy after World

War II, Japanese managers incorporated Deming's ideas into their management philosophy. Many American industries, including automobiles and consumer electronics, encountered severe competitive pressures from the Japanese in the late 1970s and 1980s. As a result, the contributions made to quality control by Deming and others were reconsidered by American managers.

Let's look at some basic ideas of Shewhart's statistical process control. Consider a production line that makes driveshafts for automobiles. Requirements for well-functioning shafts have been established. We would like to monitor and improve the quality of the shafts we produce. The shafts are made in large quantities on an automatic lathe. If we measured the diameter of each shaft after manufacture, we would expect to see some variability (perhaps a normal distribution) of the measurements around a mean value. These observed random variations in the measurements could result from variations in the hardness of the steel used for the shafts, power surges affecting the lathe, or even errors in making the measurements on the finished shafts.

But imagine what happens as the cutting tool begins to dull. The average diameter will gradually increase unless the lathe is recalibrated. And if the bearings on the lathe wear over time, the cutting edge might move around. Then some shafts would be too large, and some too small. Although the mean diameter might well be the same, the variability in the measurements would increase. It would be important to note such nonrandom (or *systematic*) variation, to identify its source, and to correct the problem.

From this discussion, you can see that there are two kinds of variation that are observed in the output from most processes, in general, and from our automatic lathe, in particular:

- Random variation (sometimes called *common*, or *inherent*, variation)
- Systematic variation (sometimes called *assignable*, or *special cause*, variation)

These two kinds of variation call for different managerial responses. Although one of the goals of quality management is *constant improvement* by the reduction of inherent variation, this cannot ordinarily be accomplished without changing the process. And you should not change the process until you are sure that all assignable variation has been identified and brought under control. So the idea is this: **If the process is *out-of-control* because there is still some special cause variation present, identify and correct the cause of that variation. Then, when the process has been brought *in-control*, quality can be improved by redesigning the process to reduce its inherent variability.**

In the next three sections, we shall look at control charts, devices that Shewhart invented for monitoring process outputs to identify when they slip out of control.

Random variation in process output

Nonrandom variation in the output

Managerial responses to inherent and assignable variation

HINTS & ASSUMPTIONS

There are a lot of catch phrases associated with quality control programs today: "Put Quality First," "Variation Is the Enemy of Quality," "Make It Right the First Time," and "Zero Defects" are only a few. As you read these phrases in the popular press it may seem like a paradox that *statistical process control*, the topic of this chapter, focuses on *variation*. Hint: Until we can measure a process and find out the sources of the variation (random variation and systematic variation) we are not able to bring the process into control. Warning: Quality control programs based solely on slogans instead of sound statistical methods do not work.

EXERCISES 10.2

Basic Concepts

- 10-7** What happened in the 1970s and 1980s to cause American managers to pay more attention to Deming's ideas?
- 10-8** Explain why work produced by a robot might have less random variation than work produced by a human.
- 10-9** When the manager of a baseball team decides to change pitchers, is he responding to random or assignable variation? Explain.
- 10-10** What kinds of systematic variation are the managers of supermarkets trying to control when they establish express lanes at some of their cash registers?

10.3 \bar{x} CHARTS: CONTROL CHARTS FOR PROCESS MEANS

The essence of statistical process control is to identify a parameter that is easy to measure and whose value is important for the quality of the process output (the shaft diameter in our example), plot it in such a way that we can recognize nonrandom variations, and decide when to make adjustments to a process. These plots are known genetically as *control charts*. Suppose, for the moment, that we want to produce shafts whose diameters are distributed normally with $\mu = 60$ millimeters and $\sigma = 1$ millimeter. (**An assumption of normality with μ and σ known is unreasonable in most situations, and we will drop it later. However, it facilitates our discussion of the basic ideas of control charts.**)

Plot the data to find nonrandom variations

To monitor the process, we take a random sample of 16 measurements each day and compute their means, \bar{x} . From Chapter 6, we know that the sample means have a sampling distribution with

$$\begin{aligned}\mu_{\bar{x}} &= \mu = 60 & [6-1] \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{1}{\sqrt{16}} \\ &= 0.25\end{aligned}$$

For a period of 2 weeks, let's plot the daily sample means against time. This is called an \bar{x} chart. In Figure 10-1, we have plotted the results from three hypothetical sets of two weeks' worth of daily sample means. In each of these three \bar{x} charts, we have also included

\bar{x} charts

- A center line (CL), with value $\mu_{\bar{x}} = 60$
- An upper control limit (UCL) line, with value $\mu_{\bar{x}} + 3\sigma_{\bar{x}} = 60 + 3(0.25) = 60.75$
- A lower control limit (LCL) line, with value $\mu_{\bar{x}} - 3\sigma_{\bar{x}} = 60 - 3(0.25) = 59.25$

The number 3 in the upper and lower control limits is used by standard convention. Where does it come from? Recall Chebyshev's theorem, which we discussed on page 116. No matter what the underlying distribution, at least 89 percent of all observations fall within ± 3 standard

$\pm 3\sigma$ control limits should contain most of the observations

deviations from the mean. And recall that for normal populations (see Appendix Table 1), over 99.7 percent of all observations fall within that interval.

So, if a process is in-control, essentially all observations should fall within the control limits. Conversely, observations that fall outside those limits suggest that the process is out-of-control, and they warrant further investigation to see if some special cause can be found to explain why they fall outside the limits. With this in mind, let's look at Figure 10-1.

Basic Interpretation of Control Charts

In Figure 10-1(a), all observations fall within the control limits, so the process is in-control. In Figure 10-1(b), the second and eighth observations are *outliers*—they fall outside the control limits. In this instance, the process is out-of-control. The production staff should try to find out whether something out of the ordinary happened on those 2 days. Perhaps the lathe was not recalibrated those mornings, or maybe the regular operator was out sick. An investigation may not turn up anything. After all, even purely random variation will produce outliers 0.3 percent of the time. In such cases, concluding that something has gone awry corresponds to making a Type I error in hypothesis testing. However, because legitimate outliers happen so infrequently, it makes good sense to investigate whenever an outlier is observed.

Outliers should be investigated

What should we conclude about Figure 10-1(c)? Even though all 10 observations fall within the control limits, they exhibit anything but random variation. They show a distinct pattern of increase over time. Whenever you see such lack of randomness, you should assume that something systematic is causing it and seek to determine what that assignable cause is. Even though all the observations fall within the control limits, we still say that the process is out-of-control. In this example, the lathe blade was getting duller each day, and the maintenance department had neglected to sharpen it as scheduled.

Patterns in the data points also indicate out-of-control processes

What sort of patterns should you be looking out for? Among the more commonly noted patterns are

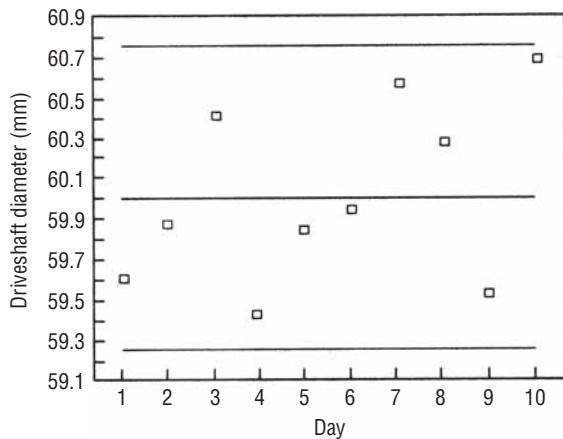
Common out-of-control patterns

- **Individual outliers** (Figure 10-1(b)).
- **Increasing or decreasing trends** (Figure 10-1(c)). These indicate that the process mean may be drifting.
- **Jumps in the level around which the observations vary** (Figure 10-2(a)). These indicate that the process mean may have shifted.
- **Cycles** (Figure 10-2(b)). Such regularly repeating waves above and below the center line could indicate such things as worker fatigue and changeover between work shifts.
- **“Hugging the control limits”** (Figure 10-2(c)). Uniformly large deviations from the mean can indicate that two distinct populations are being observed.
- **“Hugging the center line”** (Figure 10-2(d)). Uniformly small deviations from the mean indicate that variability has been reduced from historic levels; this is generally desirable. If it can be maintained, the control limits should be tightened to make sure that this improved quality continues.

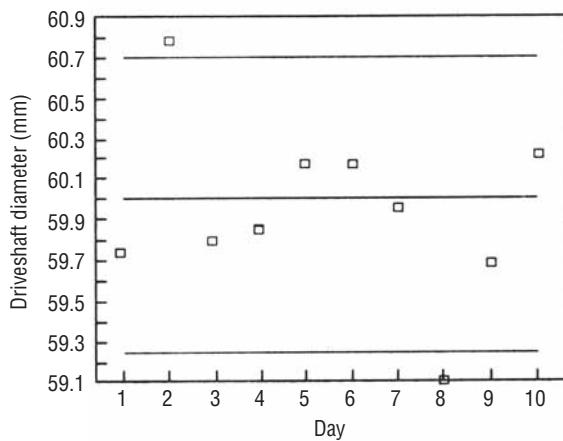
\bar{x} Charts when μ and σ Are Not Known

Now that you understand the basic ideas for interpreting \bar{x} charts, let's see how to construct them when μ and σ aren't known. Recall the express-teller line at TransCarolina Bank, which opened this chapter. Lisa Klein, Durham City Executive for TransCarolina, wants express-teller transactions to be completed

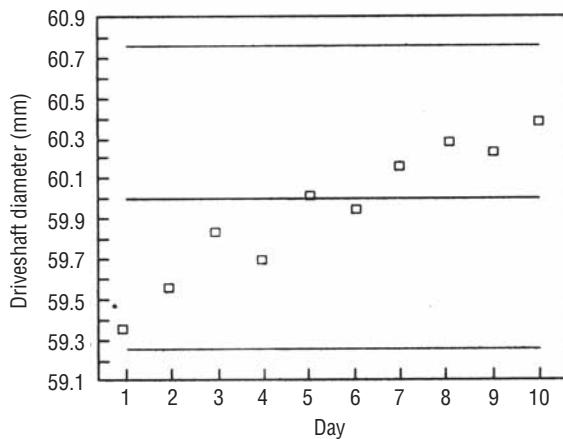
(a) Process is in-control



(b) Process is out-of-control: outliers beyond control limits



(C) Process is out-of-control: increasing trend in observations

**FIGURE 10-1 THREE \bar{x} CHARTS FOR THE DRIVESHAFT PRODUCTION PROCESS**

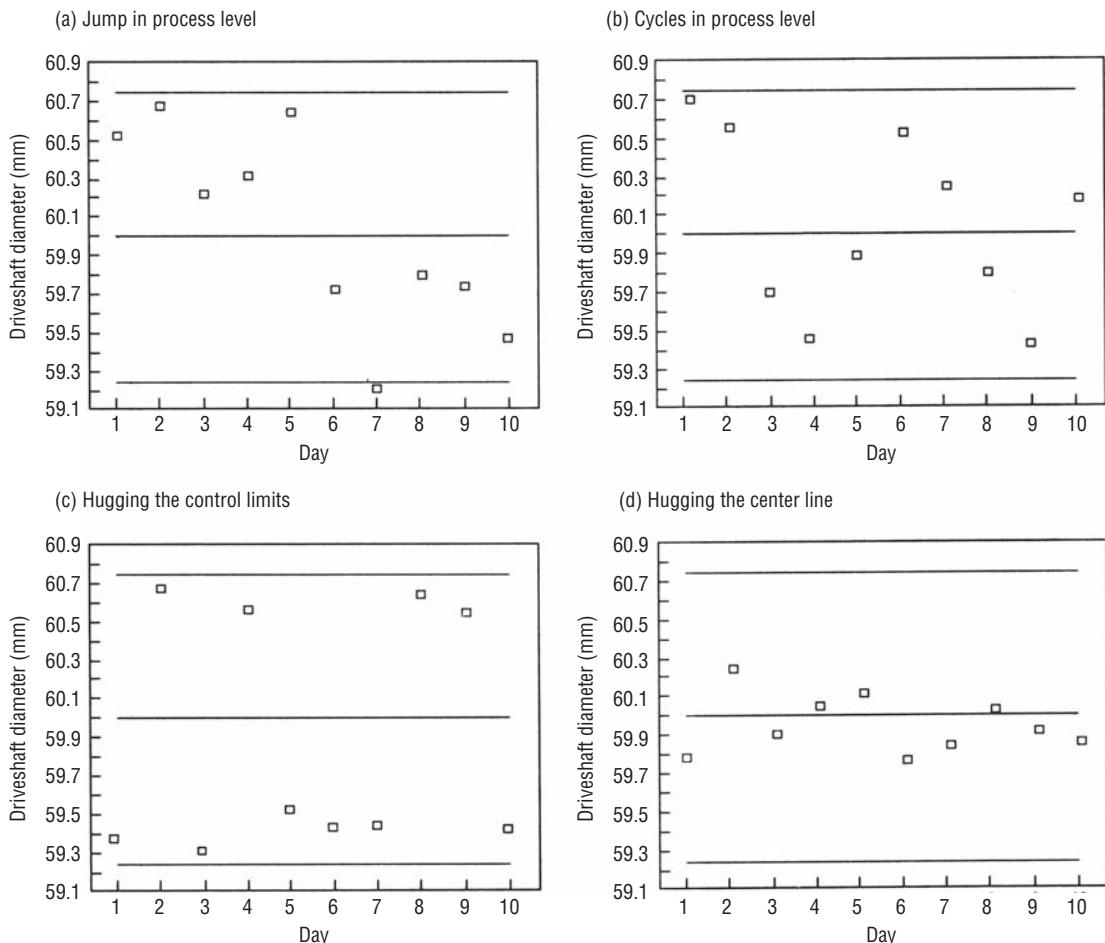


FIGURE 10-2 NONRANDOM PATTERNS IN CONTROL CHARTS

in an average of less than 60 seconds. Her sample data for last month are repeated in Table 10-1, which also includes daily sample means and ranges.

As we saw in Chapters 7–9, a common theme in statistics is the use of sample information to estimate unknown parameters. Because Lisa doesn't know the true process mean, μ , she will use the sample mean, \bar{x} , in its place. But which of the twenty daily \bar{x} 's should she use? None of them! Each of them contains information from only six observations, but she has a total of 120 observations available (six observations for each of 20 days). She captures all this information by using $\bar{\bar{x}}$, the *grand mean*, which can be calculated in two equivalent ways:

Estimate μ by $\bar{\bar{x}}$

Grand Mean from Several Samples of the Same Size

$$\bar{\bar{x}} = \frac{\sum x}{n \times k} = \frac{\sum \bar{x}}{k}$$

[10-1]

TABLE 10-1 RAW DATA AND DAILY SAMPLE MEANS AND RANGES FOR TRANSCAROLINA BANK EXPRESS-TELLER LINE

Day	Date	Transaction Times (seconds)						Mean \bar{x}	Range (R)
M	5/03	63	55	56	53	61	64	58.7	11
T	5/04	60	63	60	65	61	66	62.5	6
W	5/05	57	60	61	65	66	62	61.8	9
TH	5/06	58	64	60	61	57	65	60.8	8
F	5/07	79	68	65	61	74	71	69.7	18
M	5/10	55	66	62	63	56	52	59.0	14
T	5/11	57	61	58	64	55	63	59.7	9
W	5/12	58	51	61	57	66	59	58.7	15
TH	5/13	65	66	62	68	61	67	64.8	7
F	5/14	73	66	61	70	72	78	70.0	17
M	5/17	57	63	56	64	62	59	60.2	8
T	5/18	66	63	65	59	70	61	64.0	11
W	5/19	63	53	69	60	61	58	60.7	16
TH	5/20	68	67	59	58	65	59	62.7	10
F	5/21	70	62	66	80	71	76	70.8	18
M	5/24	65	59	60	61	62	65	62.0	6
T	5/25	63	69	58	56	66	61	62.2	13
W	5/26	61	56	62	59	57	55	58.3	7
TH	5/27	65	57	69	62	58	72	63.8	15
F	5/28	70	60	67	79	75	68	69.8	19
								$\sum \bar{x} = 1,260.2$	$\Sigma R = 237$

where

- \bar{x} = grand mean
- $\sum x$ = sum of all observations
- $\sum \bar{x}$ = sum of the sample means
- n = number of observations in each sample
- k = number of samples taken

In our example, $n = 6$ and $k = 20$, so we find

$$\bar{\bar{x}} = \frac{\sum x}{n \times k} = \frac{7,561}{6(20)} = 63.0 \text{ or} \quad [10-1]$$

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{k} = \frac{1,260.2}{20} = 63.0$$

Once $\bar{\bar{x}}$ has been calculated, its value is used as the center line (CL) in the \bar{x} chart.

How should Lisa estimate σ ? In Chapters 7-9, we used s , the sample standard deviation, to estimate σ . However, in control charts, it has become customary to base an estimate of σ on \bar{R} , the average of the sample ranges. This custom arose because control charts were often plotted on the factory floor, and it was a lot easier for workers to compute sample ranges (the difference between the highest and lowest observations in the sample) than to compute sample standard deviations using Equation 3-18 (see p. 124). The relationship between σ and \bar{R} is captured in a factor called d_2 , which depends on n , the sample size. The values of d_2 are given in Appendix Table 9.

Estimate σ from \bar{R} using d_2

The upper and lower control limits (UCL and LCL) for an \bar{x} chart are computed with the following formulas:

Control Limits for an \bar{x} Chart

$$\begin{aligned} \text{UCL} &= \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} & [10-2] \\ \text{LCL} &= \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} \end{aligned}$$

where

- $\bar{\bar{x}}$ = grand mean
- \bar{R} = average of the sample ranges ($= \Sigma R/k$)
- d_2 = control chart factor from Appendix Table 9
- n = number of observations in each sample

To make life simple on the factory floor, these limits are often calculated as $\bar{\bar{x}} \pm A_2\bar{R}$, where $A_2 = 3/(d_2\sqrt{n})$. Appendix Table 9 also gives the values of A_2 .

Using Equation 10-2, Lisa computes $\bar{R} = \Sigma R/k = 237/20 = 11.85$, looks up d_2 for $n = 6$ in Appendix Table 9 ($d_2 = 2.534$), and then finds the control limits for her \bar{x} chart:

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 63.0 + \frac{3(11.85)}{2.534\sqrt{6}} = 63.0 + 5.7 = 68.7 \quad [10-2]$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 63.0 - \frac{3(11.85)}{2.534\sqrt{6}} = 63.0 - 5.7 = 57.3$$

Lisa now plots the CL, UCL, LCL, and the daily values of \bar{x} , to get the \bar{x} chart in Figure 10-3. A quick glance at the chart shows her that something is awry: Every Friday, the average service time jumps above the UCL. When she investigates more closely, Lisa discovers that the experienced express-line teller is spending Fridays in a professional development course. On those days, a trainee is manning the express-teller line. Lisa decides to provide more supervision to the trainee to help him improve his processing speed.

Investigating the pattern in the \bar{x} chart

Now that she has found out why Fridays are out-of-control, Lisa can see whether the experienced express-line teller is meeting her goal of completing transactions in under 60 seconds on average. To do this, she goes back to the data in Table 10-1, excludes the four Friday outliers, and plots a new control

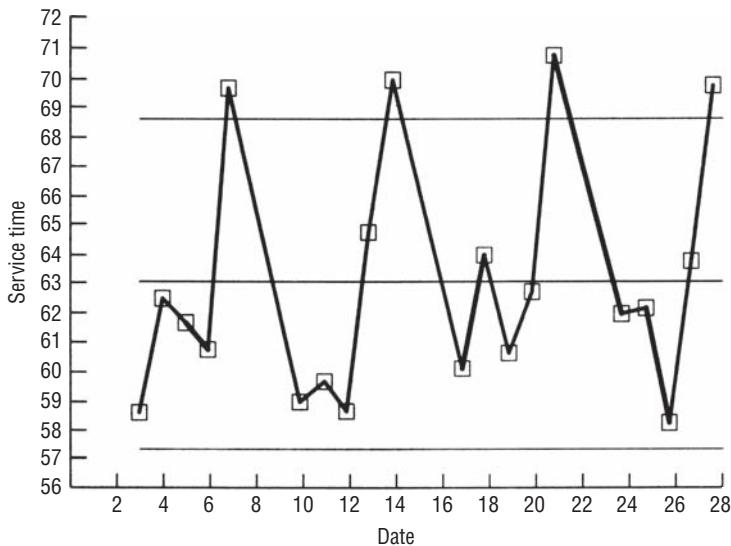
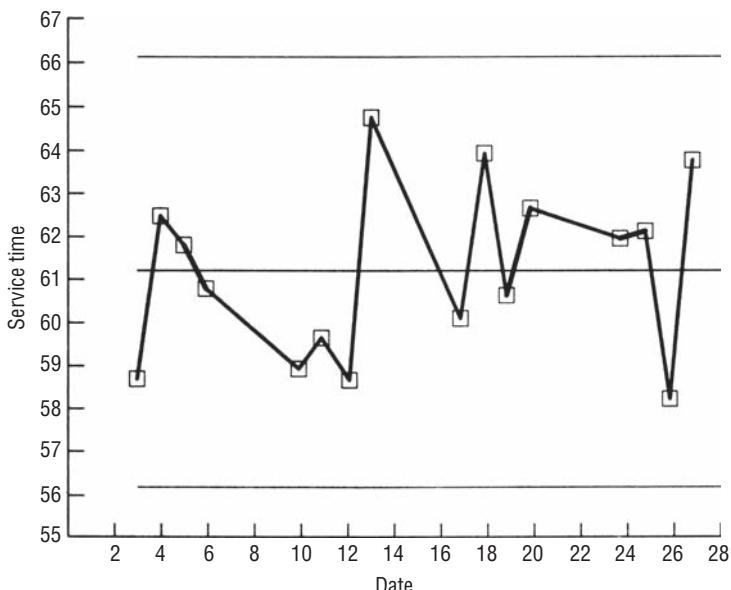
**FIGURE 10-3** \bar{x} CHART FOR THE EXPRESS-TELLER LINE AT THE TRANSCAROLINA BANK

chart from the remaining $k = 16$ daily samples. For that chart, displayed in Figure 10-4, the center line and control limits are

Redo the chart, excluding the outliers

**FIGURE 10-4** \bar{x} CHART FOR EXPRESS-TELLER LINE AT TRANSCAROLINA BANK, WITH FRIDAYS DELETED

given by

$$\bar{\bar{x}} = \frac{\Sigma x}{n \times k} = \frac{5,879}{6(16)} = 61.2 \quad [10-1]$$

$$UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 61.2 + \frac{3(10.3)}{2.534\sqrt{6}} = 61.2 + 5.0 = 66.2$$

$$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 61.2 - \frac{3(10.3)}{2.534\sqrt{6}} = 61.2 - 5.0 = 56.2 \quad [10-2]$$

From Figure 10-4, Lisa sees that the process is in-control. However, with a sample grand mean of 61.2 seconds, even the experienced teller is not yet meeting the under-60-seconds goal. Being in-control does not mean that a process is meeting its goals. In this case, Lisa and the teller will have to work together to analyze the way in which transactions are handled. Perhaps they can redesign procedures to achieve their goal. Or, because the current process is behaving well, they may decide that 61.2 seconds is good enough and not run the risk of spoiling a good system by tinkering with it. This is a managerial decision, not a statistical one. But the statistical analysis has provided Lisa with information she can use in making her managerial decision.

Managerial vs. statistical decisions

HINTS & ASSUMPTIONS

Recognizing patterns in quality control measurements is the key to fixing an out-of-control situation. When, they exist, these patterns focus our attention on something *systematic* that is causing our problem. Hint: The distribution of the variable we measure in quality control does not have to be normal in order for us to use statistical methods to control the process. As we take successive samples, the use of upper and lower control limits is a very practical example of Chebyshev's theorem. You will remember that Chebyshev assured us back in Chapter 3 that even when the underlying distribution is not normally distributed, we can still make useful statements about the population from information contained in samples. Warning: The statistical quality control methods we will illustrate in this chapter *illuminate* problems. From that point on, it takes focused management and effective communication to *correct* the situation.

EXERCISES 10.3

Self-Check Exercise

SC 10-1 For each of the following cases, find the CL, UCL, and LCL for an \bar{x} chart based on the given information.

- (a) $n = 9$, $\bar{\bar{x}} = 26.7$, $\bar{R} = 5.3$.
- (b) $n = 17$, $\bar{\bar{x}} = 138.6$, $\bar{R} = 15.1$.
- (c) $n = 4$, $\bar{\bar{x}} = 84.2$, $\bar{R} = 9.6$.
- (d) $n = 22$, $\bar{\bar{x}} = 8.1$, $\bar{R} = 7.4$.

SC 10-2 Altoona Tire Company sells its ATC-50 tires with a 50,000-mile tread-life warranty. Lorrie Ackerman, a quality control engineer with the company, runs simulated road tests to monitor

the life of the output from the ATC-50 production process. From each of the last 12 batches of 1,000 tires, she has tested 5 tires and recorded the following results, with \bar{x} and R measured in thousands of miles:

Batch	1	2	3	4	5	6	7	8	9	10	11	12
\bar{x}	50.5	49.7	50.0	50.7	50.7	50.6	49.8	51.1	50.2	50.4	50.6	50.7
R	1.1	1.6	1.8	0.1	0.9	2.1	0.3	0.8	2.3	1.3	2.0	2.1

- (a) Use the data above to help Lorrie construct an \bar{x} chart.
- (b) Is the production process in-control? Explain.

Basic Concepts

- 10-11** List four types of patterns that indicate that a process is out-of-control. Give examples where each might arise.
- 10-12** For each of the following cases, find the CL, UCL, and LCL for an \bar{x} chart based on the given information.
- (a) $n = 12$, $\bar{\bar{x}} = 16.4$, $\sigma_{\bar{x}} = 1.2$.
 - (b) $n = 12$, $\bar{\bar{x}} = 16.4$, $\bar{R} = 7.6$.
 - (c) $n = 8$, $\bar{\bar{x}} = 4.1$, $\bar{R} = 1.3$.
 - (d) $n = 15$, $\bar{\bar{x}} = 141.7$, $\bar{R} = 18.6$.

Applications

- 10-13** The Wilson Piston Company manufactures pistons for LawnGuy mowers, and the diameter of each piston must be carefully monitored. Jeff Wilson, the quality control engineer, has sampled 8 pistons from each of the last 15 batches of 500 pistons and has recorded the following results, with \bar{x} and R measured in centimeters:

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
\bar{x}	15.85	15.95	15.86	15.84	15.91	15.81	15.86	15.84	15.83	15.83	15.72	15.96	15.88	15.84	15.89
R	0.15	0.17	0.18	0.16	0.14	0.21	0.13	0.22	0.19	0.21	0.28	0.12	0.19	0.22	0.24

- (a) Use the data above to help Jeff construct an \bar{x} chart.
 - (b) Is the production process in-control? Explain.
- 10-14** Dick Burney is director of 911 emergency medical services in Ann Arbor, Michigan. He is concerned about response time, the amount of time that elapses between the receipt of a call at the 911 switchboard and the arrival of a municipal rescue squad crew at the calling location. For the last 3 weeks, he has randomly sampled response times for 9 calls each day to get the following results, with \bar{x} and R measured in minutes:

Day	M	Tu	W	Th	F	Sa	Su
\bar{x}	11.6	17.4	14.8	13.8	13.9	22.7	16.6
R	14.1	19.1	22.9	18.0	14.6	23.7	21.0
Day	M	Tu	W	Th	F	Sa	Su
\bar{x}	9.5	12.7	17.7	16.3	10.5	22.5	12.6
R	12.6	17.0	12.0	15.1	22.1	24.1	21.3

Day	M	Tu	W	Th	F	Sa	Su
\bar{x}	11.4	16.0	11.0	13.3	9.3	21.5	17.9
R	12.1	21.1	13.5	20.3	16.8	20.7	23.2

- (a) Construct an \bar{x} chart to help Dick see whether the response-time process is in-control.
 (b) What aspect of the chart should disturb him? What action might he take to address this problem?
 (c) Excluding the data identified as outlying in part (b), is the process in-control? Explain.

10-15 Track Bicycle Parts manufactures precision ball bearings for wheel hubs, bottom brackets, head sets, and pedals. Seth Adams is responsible for quality control at Track. He has been checking the output of the 5-mm bearings used in front wheel hubs. For each of the last 18 hours, he has sampled 5 bearings, with the following results:

Hour	Bearing Diameters (mm)				
1	5.03	5.06	4.86	4.90	4.95
2	4.97	4.94	5.09	4.78	4.88
3	5.02	4.98	4.94	4.95	4.80
4	4.92	4.93	4.90	4.92	4.96
5	5.01	4.99	4.93	5.06	5.01
6	5.00	4.95	5.10	4.85	4.91
7	4.94	4.91	5.05	5.07	4.88
8	5.00	4.98	5.05	4.96	4.97
9	4.99	5.01	4.93	5.10	4.98
10	5.03	4.96	4.92	5.01	4.93
11	5.02	4.88	5.00	4.98	5.09
12	5.09	5.01	5.13	4.89	5.02
13	4.90	4.93	4.97	4.98	5.12
14	5.04	4.96	5.15	5.04	5.02
15	5.09	4.90	5.04	5.19	5.03
16	5.10	5.01	5.04	5.05	5.02
17	4.97	5.10	5.12	4.92	5.04
18	5.01	4.99	5.06	5.04	5.12

- (a) Construct an \bar{x} chart to help Seth determine whether the production of 5-mm bearings is in-control.

- (b) Should Seth conclude that the process is in-control? Explain.

10-16 Northern White Metals Corp. uses an extrusion process to produce various kinds of aluminum brackets. Raw aluminum ingots are forced under pressure through steel dies to produce long sections of a desired cross-sectional shape. These sections are then fed through an automatic saw, where they are cut into brackets of the desired length. NWMC operates for three shifts of 4 hours each day, and the saw is recalibrated at the beginning of each shift. This week NWMC is producing #409 brackets with a specified cut length of 4 inches. Silvia Serrano, NWMC's quality specialist, has recorded the lengths of 15 randomly chosen brackets during each half-hour of today's three shifts to get the following data:

Shift 1								
Time	0630	0700	0730	0800	0830	0900	0930	1000
\bar{x}	4.00	4.02	4.01	4.00	4.03	4.01	4.03	4.00
R	0.09	0.10	0.10	0.11	0.09	0.11	0.11	0.10
Shift 2								
Time	1030	1100	1130	1200	1230	1300	1330	1400
\bar{x}	4.03	4.06	4.04	4.06	4.04	4.03	4.06	4.05
R	0.12	0.11	0.09	0.10	0.11	0.09	0.10	0.10
Shift 3								
Time	1430	1500	1530	1600	1630	1700	1730	1800
\bar{x}	4.01	4.01	4.00	4.02	3.99	4.02	4.00	4.00
R	0.10	0.11	0.10	0.09	0.10	0.11	0.09	0.09

- (a) Help Silvia construct an \bar{x} chart to monitor the production of the #409 brackets.
 (b) What, if anything, can you see in the chart that would cause Silvia some concern? Explain. What should Silvia do to address this concern?

Worked-Out Answers to Self-Check Exercises

SC 10-1 (a) $\bar{\bar{x}} = 26.7$ $\bar{R} = 5.3$ $n = 9$ $d_2 = 2.970$

$$\text{CL} = \bar{\bar{x}} = 26.7$$

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 26.7 + \frac{3(5.3)}{2.970\sqrt{9}} = 28.5$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 26.7 - \frac{3(5.3)}{2.970\sqrt{9}} = 24.9$$

(b) $\bar{\bar{x}} = 138.6$ $\bar{R} = 15.1$ $n = 17$ $d_2 = 3.588$

$$\text{CL} = \bar{\bar{x}} = 138.6$$

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 138.6 + \frac{3(15.1)}{3.588\sqrt{17}} = 141.7$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 138.6 - \frac{3(15.1)}{3.588\sqrt{17}} = 135.5$$

(c) $\bar{\bar{x}} = 84.2$ $\bar{R} = 9.6$ $n = 4$ $d_2 = 2.059$

$$\text{CL} = \bar{\bar{x}} = 84.2$$

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 84.2 + \frac{3(9.6)}{2.059\sqrt{4}} = 91.2$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 84.2 - \frac{3(9.6)}{2.059\sqrt{4}} = 77.2$$

$$(d) \bar{x} = 8.1 \quad \bar{R} = 7.4 \quad n = 22 \quad d_2 = 3.819$$

$$CL = \bar{\bar{x}} = 8.1$$

$$UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 8.1 + \frac{3(7.4)}{3.819\sqrt{22}} = 9.3$$

$$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 8.1 - \frac{3(7.4)}{3.819\sqrt{22}} = 6.9$$

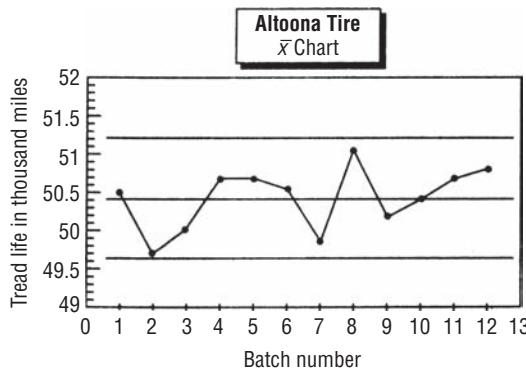
SC10-2 (a) $n = 5 \quad k = 12 \quad d_2 = 2.326$

$$\bar{\bar{x}} = \frac{\Sigma x}{k} = \frac{605.0}{12} = 50.417 \quad \bar{R} = \frac{\Sigma R}{k} = \frac{16.4}{12} = 1.367$$

$$CL = \bar{\bar{x}} = 50.417$$

$$UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 50.417 + \frac{3(1.367)}{2.326\sqrt{5}} = 51.21$$

$$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 50.417 - \frac{3(1.367)}{2.326\sqrt{5}} = 49.63$$



- (b) The production process appears to be in-control. However, there are several batches (batches 2, 7 and 8), that approach the control limits.

10.4 R CHARTS: CONTROL CHARTS FOR PROCESS VARIABILITY

Recall our discussion of quality in the first two sections of this chapter. Because quality implies consistency, reliability, and conformance to requirements, variability is the enemy of quality. Stated in a somewhat different way, the way to improve quality is to reduce variability. But before you can decide whether variability is a problem in any instance, you must be able to monitor it.

The control limits in \bar{x} charts place bounds on the amount of variability we are willing to tolerate in our sample means. However, quality concerns are addressed to individual observations (driveshaft

Monitoring variability

diameters, express-teller-line transaction times, and so on). We saw in Chapter 6 that sample means are less variable than individual observations. More precisely, Equation 6-1 tells us that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [6-1]$$

To monitor the variability in the individual observations, we use another control chart, known as an *R* chart. In *R* charts, we plot the values of the sample ranges for each of the samples. The center line for *R* charts is placed at \bar{R} . To get the control limits, we need to know something about the sampling distribution of *R*. In particular, what is its standard deviation, σ_R ? Although the derivation of the result is beyond the scope of this text, it turns out that

Center line for *R* charts

Standard Deviation of the Sampling Distribution of *R*

$$\sigma_R = d_3 \sigma \quad [10-3]$$

where

- σ = population standard deviation
- d_3 = another factor depending on n

The values of d_3 are also given in Appendix Table 9. Now we can substitute \bar{R}/d_2 for σ as we did in Equation 10-2, to compute the control limits for *R* charts:

Control limits for *R* charts

Control Limits for an *R* Chart

$$\begin{aligned} \text{UCL} &= \bar{R} + \frac{3d_3 \bar{R}}{d_2} = \bar{R} \left(1 + \frac{3d_3}{d_2} \right) \\ \text{LCL} &= \bar{R} - \frac{3d_3 \bar{R}}{d_2} = \bar{R} \left(1 - \frac{3d_3}{d_2} \right) \end{aligned} \quad [10-4]$$

To make life simple on the factory floor, these limits are often calculated as

$$\text{UCL} = \bar{R}D_4, \text{ where } D_4 = 1 + 3d_3/d_2$$

$$\text{LCL} = \bar{R}D_3, \text{ where } D_3 = 1 - 3d_3/d_2$$

The values of D_3 and D_4 can also be found in Appendix Table 9.

There is one slight wrinkle in using Equation 10-4. A sample range is always a nonnegative number (because it is the difference between the largest and smallest observations in the sample). However, when $n \leq 6$, the LCL computed by Equation 10-4 will be negative. In these cases, we set the value of LCL to zero. Accordingly, the entries for D_3 for $n \leq 6$ in Appendix Table 9 are all zeros.

Although she doesn't have any specific goals for the variability in service times on the express-teller line at the Durham office of TransCarolina Bank, Lisa Klein would like to see whether that aspect of the operation is in-control. Returning to the data in Table 10-1, she recalls that $\bar{R} = 11.85$. Using this value

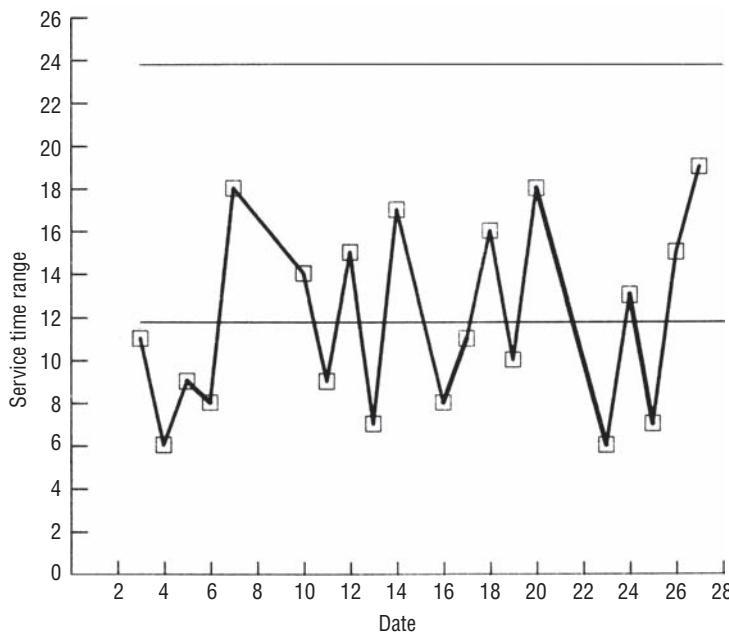


FIGURE 10-5 R CHART FOR EXPRESS-TELLER LINE AT TRANSCAROLINA BANK

in Equation 10-4, she finds the control limits for the \bar{R} chart in Figure 10-5:

$$\begin{aligned} \text{UCL} &= \bar{R} \left(1 + \frac{3d_3}{d_2} \right) = 11.85 \left(1 + \frac{3(0.848)}{2.534} \right) = 23.7 \\ \text{LCL} &= \bar{R} \left(1 - \frac{3d_3}{d_2} \right) = 11.85 \left(1 - \frac{3(0.848)}{2.534} \right) = 0 \end{aligned} \quad [10-4]$$

Although Figure 10-5 seems to indicate that the variability in service times on the express-teller line is in-control, Lisa knows that a teller trainee was at work on Fridays (the 7th, 14th, 21st, and 28th of the month). The effect of this can be seen on the R chart, because Fridays have the most variability (the highest sample ranges) during each of the 4 weeks in the sample.

Noticing a pattern in the R chart

Just as she did when looking at the process mean in Figures 10-3 and 10-4, Lisa now excludes the four Fridays to monitor the variability in service times on the express-teller line when the experienced teller is providing the service. Now $\bar{R} = 10.3$, and the control limits are

$$\begin{aligned} \text{UCL} &= \bar{R} \left(1 + \frac{3d_3}{d_2} \right) = 10.3 \left(1 + \frac{3(0.848)}{2.534} \right) = 20.6 \\ \text{LCL} &= \bar{R} \left(1 - \frac{3d_3}{d_2} \right) = 10.3 \left(1 - \frac{3(0.848)}{2.534} \right) = 0 \end{aligned} \quad [10-4]$$

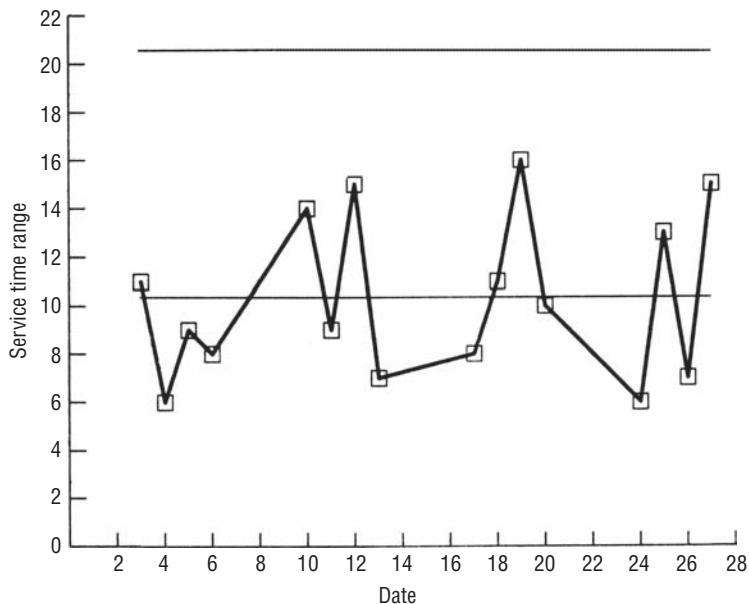


FIGURE 10-6 R CHART FOR EXPRESS-TELLER LINE AT TRANSCAROLINA BANK, WITH FRIDAYS EXCLUDED

The final R chart in Figure 10-6 shows that the experienced teller has service-time variability well in-control. There is nothing evident in the control chart to indicate the presence of any other assignable variation.

HINTS & ASSUMPTIONS

Warning: The range we plot in an R chart is only a convenient substitute for the variability of the process we are studying. Its chief advantages are that it is easy to calculate, plot, and understand. But we need to remember from Chapter 3 that the range considers only the highest and lowest values in a distribution and omits all other observations in the data set. Thus, it can ignore the nature of the variation among all of the other observations and is heavily influenced by extreme values. Also, because it measures only two values, the range can change significantly from one sample to the next in a given population.

EXERCISES 10.4

Self-Check Exercises

SC 10-3 For each of the following cases, find the CL, UCL, and LCL for an R chart based on the given information.

- $n = 9, \bar{x} = 26.7, \bar{R} = 5.3$.
- $n = 17, \bar{x} = 138.6, \bar{R} = 15.1$.

- (c) $n = 4$, $\bar{x} = 84.2$, $\bar{R} = 9.6$.
 (d) $n = 22$, $\bar{x} = 8.1$, $\bar{R} = 7.4$.

SC 10-4 Construct an R chart for the data given in Exercise SC 10-2. Is the variability in the tread life of the ATC-50 in control? Explain.

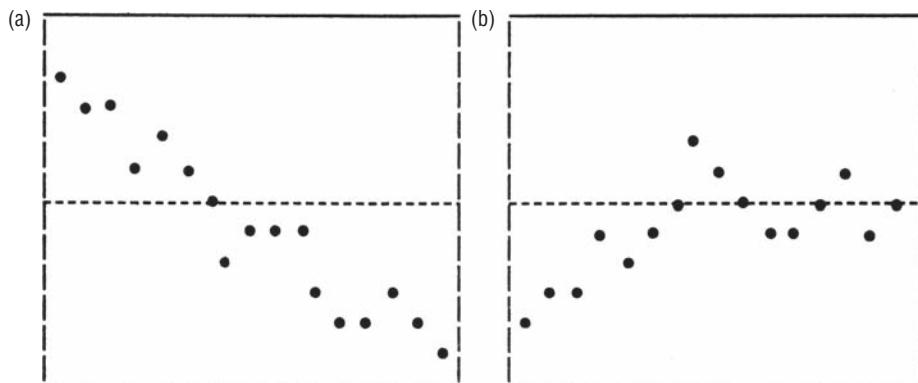
Basic Concepts

10-17 For each of the following cases, find the CL, UCL, and LCL for an R chart based on the given information:

- (a) $n = 3$, $\bar{x} = 18.4$, $\bar{R} = 3.1$.
 (b) $n = 19$, $\bar{x} = 16.2$, $\bar{R} = 6.9$.
 (c) $n = 8$, $\bar{x} = 141.7$, $\bar{R} = 18.2$.
 (d) $n = 24$, $\bar{x} = 8.6$, $\bar{R} = 1.4$.
 (e) $\bar{R} = 6.0$, LCL = 3.0, find the UCL.

Applications

10-18 Ray Underhall reproduces antique chairs. His apprentices turn spindles for the chair backs on manual lathes. The beads on the spindles are to have average diameters of $\frac{7}{8}$ -inch at their widest points. Ray monitors the apprentices' work with control charts. Which of the following patterns is he likely to see on the R chart for a new apprentice? Explain.



10-19 Construct an R chart for the data given in Exercise 10-13. Is the variability in the piston diameter under control? Explain.

10-20 Consider the emergency medical service data given in Exercise 10-14.

- (a) Construct an R chart for these data.
 (b) When he looked at the \bar{x} chart for these data, Dick Burney noted that the three Saturdays were outliers. Closer investigation revealed that this happened because the number of calls coming in was higher on Saturdays than on any other day of the week. Does the R chart you constructed in part (a) show any pattern that could be attributed to the same cause? Explain.
 (c) Exclude the 3 Saturdays and construct a new R chart. Does this chart exhibit any patterns that Dick should be concerned about? Explain.

- 10-21** Construct an R chart for the data given in Exercise 10-15. Are there any patterns in this chart that should concern Seth Adams, or does the variability in the process appear to be in-control? Explain.
- 10-22** Construct an R chart for the data given in Exercise 10-16. Are there any patterns in this chart that should concern Silvia Serrano, or does the variability in the process appear to be in-control? Explain.

Worked-Out Answers to Self-Check Exercises

SC 10-3 (a) $n = 9 \quad \bar{R} = 5.3 \quad D_4 = 1.816 \quad D_3 = 0.184$

$$\text{CL} = \bar{R} = 5.3$$

$$\text{UCL} = \bar{R}D_4 = 5.3(1.816) = 9.62$$

$$\text{LCL} = \bar{R}D_3 = 5.3(0.184) = 0.98$$

(b) $n = 17 \quad \bar{R} = 15.1 \quad D_4 = 1.622 \quad D_3 = 0.378$

$$\text{CL} = \bar{R} = 15.1$$

$$\text{UCL} = \bar{R}D_4 = 15.1(1.622) = 24.49$$

$$\text{LCL} = \bar{R}D_3 = 15.1(0.378) = 5.71$$

(c) $n = 4 \quad \bar{R} = 9.6 \quad D_4 = 2.282 \quad D_3 = 0$

$$\text{CL} = \bar{R} = 9.6$$

$$\text{UCL} = \bar{R}D_4 = 9.6(2.282) = 21.91$$

$$\text{LCL} = \bar{R}D_3 = 9.6(0) = 0$$

(d) $n = 22 \quad \bar{R} = 7.4 \quad D_4 = 1.566 \quad D_3 = 0.434$

$$\text{CL} = \bar{R} = 7.4$$

$$\text{UCL} = \bar{R}D_4 = 7.4(1.566) = 11.59$$

$$\text{LCL} = \bar{R}D_3 = 7.4(0.434) = 3.21$$

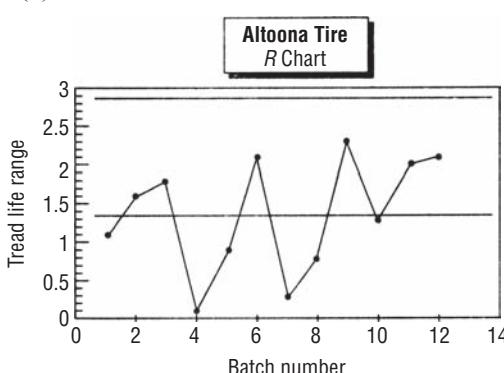
SC 10-4 $n = 5 \quad D_4 = 2.114 \quad D_3 = 0$

$$\bar{R} = 1.367$$

$$\text{CL} = \bar{R} = 1.367$$

$$\text{UCL} = \bar{R}D_4 = 1.367(2.114) = 2.89$$

$$\text{LCL} = \bar{R}D_3 = 1.367(0) = 0$$



Distinct cycling in the values of R shows that the process is out-of-control.

10.5 p CHARTS: CONTROL CHARTS FOR ATTRIBUTES

\bar{x} charts and R charts are control charts for *quantitative* variables, which take on *numerical* values. Quantitative variables are ‘measured’ (for example, heights, IQs, or speeds) or counted (for example, numbers of employees, phone calls per hour, or points scored in a basketball game). But not all the variables we encounter are quantitative. Variables such as marital status, heads or tails in a coin toss, or winning or losing a basketball game are *categorical*, or *qualitative*.

In the area of statistical process control, a qualitative variable that can take on only two values is called an *attribute*. Recalling, once again, that quality is conformity to requirements, it should not surprise you to learn that the attribute most frequently discussed in SPC is that of conformance or nonconformance of units of output to the process specifications.

Consider the case of Golden Guernsey Dairies. Harry Galloway is in charge of the milk bottling operations at GGD, an integrated dairy farm and milk packager near Sheboygan, Wisconsin. (Although cartons have long since replaced milk bottles, Harry still refers to the operations as bottling.) There is some variation in the output from GGD’s bottling machinery, so Harry monitors the process to be sure that the average half-gallon container is filled with 64.1 ounces of milk. He has long used \bar{x} charts, based on hourly samples of 100 cartons (taken 10 times each day, from 6 A.M. to 3 P.M.), to monitor the bottling operation, and the process is well under control. The Wisconsin State Department of Agriculture recently instituted a new requirement that not only must half-gallon cartons contain at least 64 ounces on average, but in addition, no more than 3 percent of them can contain less than 63.5 ounces.

The attribute that concerns Harry is whether any particular carton contains at least 63.5 ounces or less than that amount. To monitor the output, he has been keeping a record of the proportions of underfilled cartons (the fraction of cartons not conforming to the Department of Agriculture’s 63.5-ounce standard) in his hourly samples for the past week. These data are given in Table 10-2.

Because the fraction underfilled in the total sample of 7,000 cartons (7 days, 10 samples per day, 100 cartons per sample) is 0.0306, Harry is reasonably confident that GGD is meeting the new requirement. A formal test of the hypothesis $H_0: p = 0.03$, against the alternative $H_1: p > 0.03$, supports his confidence. The standard deviation of the sample proportion is

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{pq}{n}} \\ &= \sqrt{\frac{0.03(0.97)}{7,000}} = 0.0020\end{aligned}\quad [7-4]$$

Using this value to convert the observed sample fraction (0.0306) to a standard z score,

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.0306}{0.0020} = 0.3$$

TABLE 10-2 FRACTION OF UNDERFILLED HALF-GALLON CARTONS IN HOURLY SAMPLES AT GOLDEN GUERNSEY DAIRIES

Day	Hour	Fraction Underfilled	Day	Hour	Fraction Underfilled
Sunday	6	0.02	Wednesday (<i>Contd.</i>)	11	0.05
	7	0.01		12	0.05
	8	0.03		1	0.04
	9	0.03		2	0.05
	10	0.04		3	0.04
	11	0.02		6	0.01
	12	0.03		7	0.03
	1	0.03		8	0.02
	2	0.03		9	0.02
	3	0.03		10	0.03
Monday	6	0.01	Thursday	11	0.03
	7	0.01		12	0.03
	8	0.03		1	0.06
	9	0.03		2	0.05
	10	0.03		3	0.05
	11	0.02		6	0.02
	12	0.02		7	0.02
	1	0.04		8	0.03
	2	0.03		9	0.03
	3	0.05		10	0.02
Tuesday	6	0.02	Friday	11	0.03
	7	0.03		12	0.04
	8	0.02		1	0.04
	9	0.02		2	0.05
	10	0.03		3	0.03
	11	0.02		6	0.02
	12	0.02		7	0.02
	1	0.04		8	0.03
	2	0.03		9	0.03
	3	0.05		10	0.02
Wednesday	6	0.02	Saturday	11	0.03
	7	0.03		12	0.04
	8	0.01		1	0.05
	9	0.03		2	0.03
	10	0.06		3	0.04
	11	0.02		6	0.01
	12	0.02		7	0.02
	1	0.03		8	0.03
	2	0.05		9	0.02
	3	0.06		10	0.04

We find from Appendix Table 1 that the prob value for our test is $0.5000 - 0.1179 = 0.3821$. With such a large prob value, Harry can be quite confident in accepting H_0 . The fraction of half-gallon cartons being underfilled is not significantly greater than 3 percent; GGD is meeting the Department of Agriculture's new standard.

It is met!

However, because he has the overall sample broken down into $k = 70$ hourly samples of size $n = 100$ over the course of the week, there is more information available for Harry to look at. He can plot the hourly sample fractions in a control chart known as a p chart. Because

 p charts give more information

$$\mu_{\bar{p}} = p \quad [7-3]$$

and

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} \quad [7-4]$$

the center line and control limits of p charts are at

Center Line for a p Chart

$$CL = \mu_{\bar{p}} = p \quad [10-5]$$

Control Limits for a p Chart

$$UCL = \mu_{\bar{p}} + 3\sigma_{\bar{p}} = p + 3\sqrt{\frac{pq}{n}} \quad [10-6]$$

$$LCL = \mu_{\bar{p}} - 3\sigma_{\bar{p}} = p - 3\sqrt{\frac{pq}{n}}$$

Center line and control limits
for p charts

If there is a known or targeted value of p , that value should be used in Equations 10-5 and 10-6. However, if no such value of p is available, then you should estimate p by the overall sample fraction

Estimate of p

$$\bar{\bar{p}} = \frac{\sum \bar{p}_j}{k} \quad [10-7]$$

where

- \bar{p}_j = sample fraction in the j th hourly sample
- k = total number of hourly samples

Recall the slight wrinkle in using Equation 10-4 for the LCL of an R chart: Ranges cannot be negative; so if Equation 10-4 gave an LCL below 0, we replaced it by 0. In the same way, Equation 10-6 can produce a UCL above 1 or an LCL below 0

Make sure that $LCL \geq 0$ and
 $UCL \leq 1$

for a p chart. Because p is always between 0 and 1, we will replace a negative LCL by 0 and a UCL above 1 by 1.

Because Harry has a target value of $p = 0.03$ (and because he is quite confident that his filling operation is coming close to this target), he uses that value to find the center line and control limits for his p chart:

$$CL = p = 0.03 \quad [10-5]$$

$$UCL = p + 3\sqrt{\frac{pq}{n}} = 0.03 + 3\sqrt{\frac{0.03(0.97)}{100}} = 0.081 \quad [10-6]$$

$$LCL = p - 3\sqrt{\frac{pq}{n}} = 0.03 - 3\sqrt{\frac{0.03(0.97)}{100}} = -0.021$$

Harry corrects the LCL to 0 and then plots the p chart in Figure 10-7.

All of the observations on the control chart fall within the control limits, but there is a distinct pattern in the chart that repeats every day. The proportion of underfilled cartons tends to start out low in the morning and finish up high in the afternoon. Harry immediately realizes the cause of this pattern. The bottling machinery is cleaned and calibrated each morning and then runs for the entire day. Even though the

Noting a pattern, finding its cause and taking action to correct it

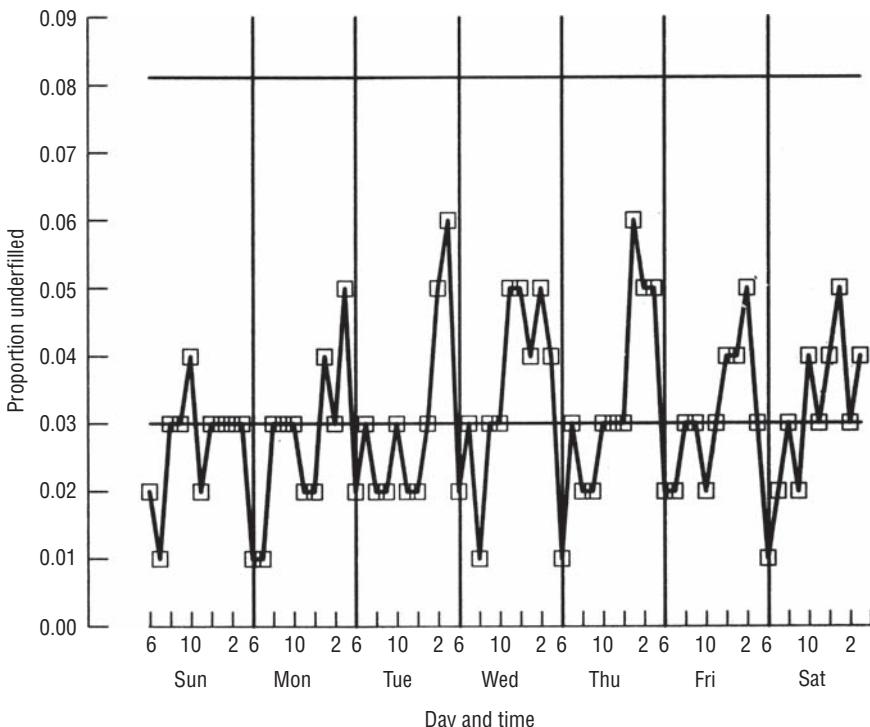


FIGURE 10-7 p CHART FOR BOTTLING MACHINERY AT GOLDEN GUERNSEY DAIRIES

fraction of underfilled cartons meets the state standard on average, Harry is unhappy with his finding. Fortunately, cleaning and calibrating are quick and easy, so Harry decides to stop the line briefly at 10 A.M. each day to clean and recalibrate the machinery for the second half of the day.

HINTS & ASSUMPTIONS

Charts of means and ranges help us control quantitative variables that can be measured, such as length of a part, life (in hours) of an engine, or the width of lumber. But many variables take on only two values, such as acceptable part/unacceptable part, fits/does not fit, or fast enough/not fast enough. In statistical process control, such a variable is called an *attribute* and we control attributes with *p* charts. Hint: Think of an attribute in terms of hair color; you are either a redhead or not, you have it or you do not. Warning: If there is a target value for *p*, you should use it for the center line of the *p* chart. If no such value is available, then use the overall sample fraction for the center line. Remember that probabilities are between 0 and 1; lower control limits below 0 or upper control limits above 1 are incorrect.

EXERCISES 10.5

Self-Check Exercises

SC 10-5 For each of the following cases, find the CL, UCL, and LCL for a *p* chart based on the given information.

- (a) $n = 144, \bar{\bar{p}} = 0.10$.
- (b) $n = 60, \bar{\bar{p}} = 0.9$.
- (c) $n = 125, 0.36$ is the target value for *p*.
- (d) $n = 48, 0.75$ is the target value for *p*.

SC 10-6 Todd Olmstead is the Meals-on-Wheels dispatcher for the Atlanta metropolitan area. He wants meals delivered to clients within 30 minutes of leaving the kitchens. Meals with longer delivery times tend to be too cold when they arrive. Each of his 10 volunteer drivers is responsible for delivering 15 meals daily. Over the past month, Todd has recorded the percentage of each day's 150 meals that were delivered on-time

Day	1	2	3	4	5	6	7	8
% on-time	89.33	81.33	95.33	88.67	96.00	86.67	98.00	84.00
Day	9	10	11	12	13	14	15	16
% on-time	90.67	80.67	88.00	86.67	96.67	85.33	78.67	89.33
Day	17	18	19	20	21	22	23	24
% on-time	89.33	78.67	94.00	94.00	99.33	95.33	94.67	92.67
Day	25	26	27	28	29	30		
% on-time	81.33	89.33	99.33	90.67	92.00	88.00		

- (a) Help Todd construct a *p* chart from these data.
- (b) How does your chart show that the attribute “fraction of meals delivered on-time” is out-of-control?
- (c) What action do you recommend for Todd?

Basic Concepts

- 10-23** Which of the following qualitative variables are attributes?
- Gender of nouns in German.
 - Gender of nouns in French.
 - Course grades under a Pass /Fail grading scheme.
 - Course grades under an A, B, C, D, F grading scheme.
- 10-24** For each of the following cases, find the CL, UCL, and LCL for a *p* chart based on the given information.
- $n = 30$, $\bar{p} = 0.25$.
 - $n = 65$, $\bar{p} = 0.15$.
 - $n = 82$, $\bar{p} = 0.05$.
 - $n = 97$, 0.42 is the target value for *p*.
 - $n = 124$, 0.63 is the target value for *p*.

Applications

- 10-25** After finding out his luggage arrived in San Antonio while his destination was Omaha, Will Richardson, a statistician for USA Airlines, decided to do some research. For the last 3 weeks, Will has sampled 200 passengers daily and determined the percentage of luggage delivered to the expected destination with the results given below

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Percent correct	0.89	0.91	0.93	0.95	0.94	0.96	0.92	0.91	0.93	0.90	0.88	0.94	0.97	0.94	0.95	0.92	0.93	0.92	0.91	0.93	0.89

- (a) Help Will construct a *p* chart from these data.
(b) Is the luggage delivery process in-control? Explain.
(c) What recommendations, if any, can you make?
- 10-26** BioAssist, Inc., manufactures high potency vitamin supplements. C-Assist, a 1,000-mg capsule of vitamin C, is BioAssist's best seller. Sherry Cohen is responsible for monitoring the quality of C-Assist. The capsules are supposed to contain between 999 and 1,001 mg of vitamin C, and BioAssist wants no more than 1.5 percent of them to fail to meet this specification. Every quarter-hour, Sherry samples 500 capsules and records the percentage failing to meet the specification (percent bad). She has gotten the following results for the last 8 hours of production:

Time	0915	0930	0945	1000	1015	1030	1045	1100	1115	1130	1145
% bad	2.4	1.8	1.6	0.6	1.0	1.4	2.0	2.8	2.4	1.6	1.0
Time	1200	1215	1230	1245	1300	1315	1330	1345	1400	1415	1430
% bad	0.4	0.6	1.6	2.2	2.6	2.2	1.6	1.0	0.4	1.2	1.6
Time	1445	1500	1515	1530	1545	1600	1615	1630	1645	1700	
% bad	2.2	2.8	1.8	1.6	0.8	0.4	1.2	1.4	2.0	2.8	

- (a) Consider all 16,000 capsules Sherry has sampled. Can she be sure that the percentage bad is not significantly greater than 1.5 percent? State and test the appropriate hypotheses.
(b) Use the data above to help Sherry construct a *p* chart.

- (c) Is there anything in the p chart about which Sherry should worry? If not, why not? If so, what should she do?

- 10-27** Andie Duvall is a finance major who has been studying the stock market for her senior honors thesis. On each of the last 100 trading days, she has randomly sampled 100 companies listed on the New York Stock Exchange and recorded the fraction whose share prices increased that day. Andie believes that there is a 50–50 chance that any given stock will increase on any given day. Explain how she can use a p chart based on her 100 days' worth of data to see if her belief is reasonable or not.
- 10-28** Ross Darrow is a flight operations analyst for Spacious Skies, Unltd. He has been assigned the task of monitoring flights at the company's hub airport in the southeast. Each day, Spacious Skies has 240 takeoffs scheduled from this hub. Ross has been concerned about the fraction of flights with late departures, and four weeks ago he instituted procedures designed to reduce that fraction. Use the data for the last 30 week-days to construct a p chart to see whether his new procedures have been successful. What further action, if any, should Ross consider?

Weeks 1 & 2

Day	M	T	W	Th	F	M	T	W	Th	F
# late	26	19	26	22	24	19	19	20	18	18

Weeks 3 & 4

Day	M	T	W	Th	F	M	T	W	Th	F
# late	17	9	13	10	12	14	14	13	9	10

Weeks 5 & 6

Day	M	T	W	Th	F	M	T	W	Th	F
# late	12	15	14	15	16	18	17	16	18	17

Worked-Out Answers to Self-Check Exercises

SC 10-5 (a) $CL = \bar{\bar{p}} = 0.10$

$$UCL = \bar{\bar{p}} + 3\sqrt{\frac{pq}{n}} = 0.10 + 3\sqrt{\frac{0.10(0.90)}{144}} = 0.175$$

$$LCL = \bar{\bar{p}} - 3\sqrt{\frac{pq}{n}} = 0.10 - 3\sqrt{\frac{0.10(0.90)}{144}} = 0.025$$

(b) $CL = \bar{\bar{p}} = 9$

$$UCL = \bar{\bar{p}} + 3\sqrt{\frac{pq}{n}} = 0.9 + 3\sqrt{\frac{0.9(0.1)}{60}} = 1.016, \text{ so the UCL} = 1$$

$$LCL = \bar{\bar{p}} - 3\sqrt{\frac{pq}{n}} = 0.9 - 3\sqrt{\frac{0.9(0.1)}{60}} = 0.784$$

(c) $CL = p = 0.36$

$$UCL = p + 3\sqrt{\frac{pq}{n}} = 0.36 + 3\sqrt{\frac{0.36(0.64)}{125}} = 0.489$$

$$\text{LCL} = p - 3\sqrt{\frac{pq}{n}} = 0.36 - 3\sqrt{\frac{0.36(0.64)}{125}} = 0.231$$

(d) $\text{CL} = p = 0.75$

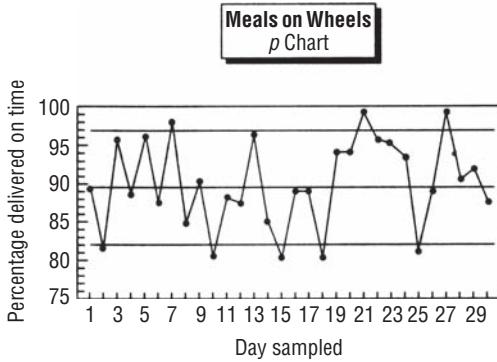
$$\text{UCL} = p + 3\sqrt{\frac{pq}{n}} = 0.75 + 3\sqrt{\frac{0.75(0.25)}{48}} = 0.938$$

$$\text{LCL} = p - 3\sqrt{\frac{pq}{n}} = 0.75 - 3\sqrt{\frac{0.75(0.25)}{48}} = 0.563$$

SC 10-6 (a) $n = 150 \quad \bar{p} = \frac{\sum \bar{p}}{k} = \frac{26.94}{30} = 0.898$

$$\text{UCL} = p + 3\sqrt{\frac{pq}{n}} = 0.898 + 3\sqrt{\frac{0.898(0.102)}{150}} = 0.972$$

$$\text{LCL} = p - 3\sqrt{\frac{pq}{n}} = 0.898 - 3\sqrt{\frac{0.898(0.102)}{150}} = 0.824$$



- (b) Five of the 30 days sampled had values of “fraction on-time” that were below the lower control limit. (Being above the upper control limit is not worrisome in this context.)
- (c) Because the percentage of meals delivered on-time is out-of-control, Todd might investigate the reasons behind the 5 days that are out-of-control. It might be a particular driver, or those days may provide heavier traffic. He might replace or train the volunteer(s) based on his findings.

10.6 TOTAL QUALITY MANAGEMENT

Statistical Process Control is very useful for continuous processes such as oil refineries and mass-production facilities. However, many managers feel that their businesses are altogether

Some processes are too complicated for control charts

too complicated to have their important aspects captured and monitored by control charts. Suppose you were the manager of a regional hub airport and were asked to reduce takeoff delays. Although delays are easy to identify, their causes are harder to pin down. Takeoffs can be delayed by weather, equipment problems, lateness of incoming crews, holiday traffic, and so on. At first glance, you wouldn't know what things to measure in order to control delays.

Total Quality Management (TQM) is a set of approaches that enable the managers of complex systems to match the firm's products to customers' expectations. The airport manager can use TQM to reduce delays so that planes match the schedules that their passengers expect. Because so many factors are involved, and because different workers have responsibility for these factors, successful use of TQM requires commitments at all levels of the firm in order to be successful. In particular, top-level management must provide strong leadership for quality, and workers at all levels must be empowered to identify problems and make changes in the system.

TQM requires companywide commitment

Fishbone Diagrams: Identifying and Grouping Causes

The TQM approach to complex businesses begins with the realization that all errors, defects, and problems have causes, and that there is only a finite number of these. The first step is to identify and discriminate between *things gone right* and *things gone wrong*. In our airport example, some of the planes do leave on time (*things gone right*). When you look at the late departures (*things gone wrong*), you can begin to build up a list of causes behind their delays.

Identify what's right and what's wrong

Even in complicated systems, causes of problems can be gathered into logical groups. For example, in our airport delays case, some of the late takeoffs are due to problems with the aircraft themselves, others result from baggage handling, and so on. As you collect the various reasons for departure delays into logical groups, it becomes clear that there are cause-and-effect relationships among them. These relationships can be captured pictorially in a *cause-and-effect* diagram such as Figure 10-8. Such cause-and-effect diagrams are sometimes called *Ishikawa diagrams*, after their Japanese developer, Kaoru Ishikawa. But, because of their appearance, these diagrams are most often called *fishbone diagrams*.

Gather causes into logical groups

The fishbone diagram takes an unstructured list of factors that contribute to delayed takeoffs and organizes that list in two major ways. First, it gathers the factors into logical groups. And then, within the groups, it indicates how the various factors feed into one another in cause-and-effect relationships. Because of this, you can see how the complex system hangs together and recognize that many factors may need to be addressed in order to resolve the problems.

The fishbone diagram also points out why personnel at all levels must be involved if TQM is to be successful. Baggage handlers are much more likely than top management or consultants to be able to identify a complete list of baggage problems that contribute to takeoff delays. In addition, because of their familiarity with details of the baggage-handling operation, they are also very likely to be able to suggest ways to improve that operation.

Successful use of TQM involves personnel at all levels

However, unless they are empowered to identify problems and make changes, they are unlikely to be willing to do so.

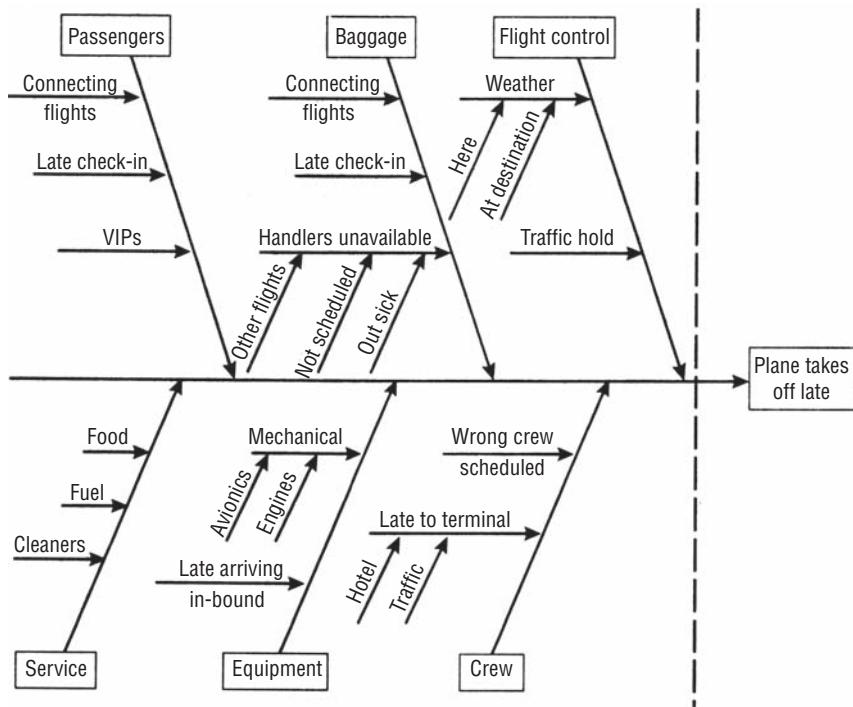


FIGURE 10-8 A FISHBONE DIAGRAM: CAUSE-AND-EFFECT REASONS FOR AIRPORT TAKEOFF DELAYS

Slay the Dragons First

In any quality improvement process, as we have seen, there are likely to be a very large number of causes for defects and errors.

Concentrate first on common causes

Looking at all the possible things that can go wrong, even if they are organized, into a neat fishbone diagram, can lead even well-motivated people to despair that “this problem is bigger than any of us can handle!” Joseph Juran’s important contribution was to insist that TQM companies distinguish between the *vital few* and the *trivial many*. In our airport example, if most of the delays are due to baggage handling, and only one delay a year is attributable to a freak hail storm, it makes good sense to start by looking at ways to improve baggage handling. In TQM parlance, companies must **slay the dragons first** in working to improve the quality of their goods or services.

A *Pareto chart* is a bar graph showing groups of error causes arranged by their frequencies of occurrence. It’s constructed by simply counting data from observations of things gone wrong. The results are usually ordered in a sequence from most common to least common, with a residual “other” category at the end. These charts are named after Vilfredo Pareto (1848–1923), an Italian economist who studied the distribution of wealth. Just as Pareto found that most of the wealth in a society is held by relatively few people, Juran noted that in most complex systems, *80 percent of defects and errors can be attributed to 20 percent of the causes*. Looking at the Pareto chart for late departures in Figure 10-9, you can see that about $\frac{2}{3}$ of the delays (45 of 68 observations) were caused by baggage-handling and equipment problems. The airport manager should begin system-improvement efforts by concentrating on these two areas.

Pareto charts

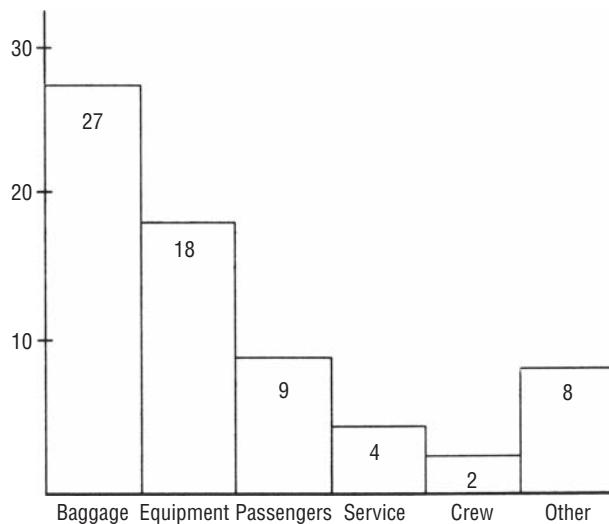


FIGURE 10-9 A PARETO DIAGRAM OF REASONS FOR AIRPORT TAKEOFF DELAYS

Continuous Quality Improvement

Once the causes of errors and defects have been identified, resources can be devoted to making changes to improve the quality of the systems' goods or services. This can sometimes involve the institution of SPC methods in the process, but more often it requires reconfiguration of the system or the reallocation of resources within the system. Improved baggage handling could require a fix as simple as hiring more baggage handlers or as complex as installing scanners that read bar-coded destination labels on pieces of luggage to facilitate their automatic routing between connecting flights or to passenger pickup areas when reaching a final destination.

When TQM efforts are successful, it is not uncommon for the leading cause of errors to drop to zero on the Pareto chart. This means that another cause becomes the “dragon,” and management attention now will shift to another part of the system. This constant attention to the identification and resolution of problems is known as *Continuous Quality Improvement* (CQI).

HINTS & ASSUMPTIONS

In the typical complex process we study, we find *many* possible causes of failure. Warning: Unless you use an organized, systematic method to look at all of these causes, you run a high risk of missing something that's important. Fishbone diagrams and Pareto charts are very effective ways of focusing and guiding your analysis of quality problems so that everything that affects quality is examined, nothing is overlooked, and the most important things get looked at *first*. A hint learned from many years of quality control experience is that Total Quality Control programs work only when you have strong top management leadership that involves line employees in the responsibility for controlling their own processes.

EXERCISES 10.6

Self-Check Exercise

SC 10-7 Northway Computers has just begun a TQM program to manage the quality of the personal computers it assembles. A careful analysis of 25,000 computer systems located the following faults:

Component	Number of Faults
CPU	25
Floppy disk drives	106
Hard disk drive	237
I/O ports	36
Keyboard	60
Monitor	42
Power supply	186
RAM memory	30
ROM BIOS	7
Video adapter	47
Other	163

Construct a Pareto chart for Northway. Northway's President, Ted White, is going to set up a series of meetings with his component suppliers. With whom should the first meetings be?

Basic Concepts

- 10-29** Explain why successful application of TQM requires the participation of employees at all levels of an organization.
- 10-30** After hearing a lecture on TQM, Joe Smithies said, "Once you've identified and slain the dragon, then you can forget TQM and get on with business as usual." Comment on Joe's understanding of TQM.

Applications

- 10-31** *The News and Reporter* has a long-standing TQM policy, and it is time to analyze this quarter's complaints and problems. The following problems have been traced by the quality control engineer:

Problem	Department	Number of Occurrences
Omitted advertisement	Classified	18
Incorrect special instructions	Classified	37
Typographical error in a news story	Reporting	14
Advertisement in the wrong section	Classified	16
Incorrectly priced advertisement	Classified	8
Factual error in news story	Reporting	16
Late delivery of all papers	Printing	3
Advertisement placed on incorrect date	Classified	6
Typographical error in commercial advertisement	Advertising	8

(Continued)

(Contd.)

Problem	Department	Number of Occurrences
Failure to respond to news report	Reporting	16
Editorialized factual story	Reporting	2
Misquoted news story	Reporting	4
Incorrect size of advertisement	Classified	7
Incorrect phone number in advertisement	Classified	9
Incorrect address in advertisement	Classified	3

Construct two Pareto charts for *The News and Reporter*. The first chart should identify which department is in need of most attention and the second chart should identify which area that department should focus on. What's the first order of business for the TQM team?

10-32

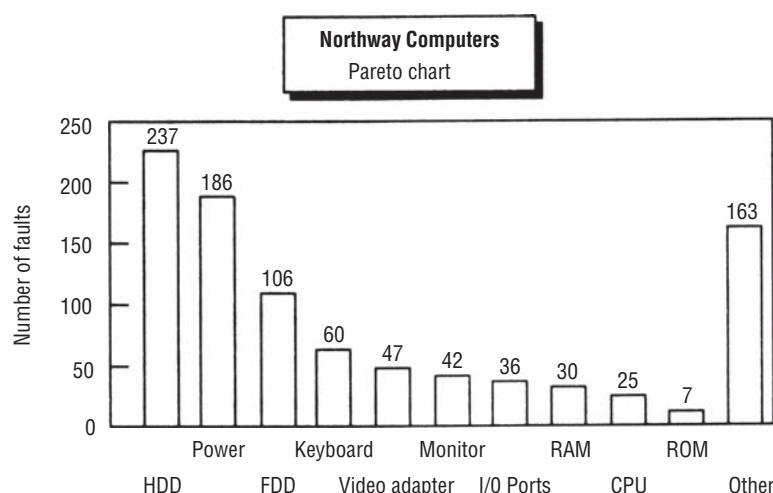
Zippy Cola is bottled in several plants around the country. Brand manager Tim Harnett has been keeping track of customer complaints about variability in the drink's flavor. Use the data below to help Tim construct a Pareto chart and decide which plants should first be visited by Zippy Cola's production specialists.

City	Number of Complaints
Atlanta	267
Boston	23
Chicago	37
Houston	175
Milwaukee	19
New Orleans	78
San Francisco	28
Seattle	43

10-33

Construct a fishbone diagram to organize the reasons why you are late to your first class of the day.

Worked-Out Answers to Self-Check Exercise

SC 10-7

The first meetings should be with the suppliers of the hard disk drives and the power supplies.

10.7 ACCEPTANCE SAMPLING

Adoption of TQM techniques implies a goal that the inputs to each stage of an operation should be defect-free because the operations at the preceding stage are under control. But manufacturers often have to accept raw materials and components from suppliers. To be sure that the results of their own operations are of high quality, they must often test inputs to make sure that they conform to requirements. In most production situations, complete inspection of an entire batch of input is impractical because of time and cost considerations. Instead, a sample of the batch is inspected, and the decision to accept or reject the entire batch is based on the quality of the sample.

Testing inputs for conformance to requirements

You may feel that reliance on sampling to ensure the quality of inputs is just moving the old-time white-coated inspectors from the end of the production line to the beginning. Many experts in quality engineering would agree with you. The whole process of inspection implies that some materials will be rejected and that amounts to a waste of materials and time. However, acceptance sampling can be an effective way to motivate suppliers to improve the quality of their outputs. In fact, it can even be more effective than inspection of the entire batch. Let's look more carefully at this apparently paradoxical assertion.

Suppose you inspect an entire batch of components sent to you by a supplier. You sort the individual units into two groups, acceptable and unacceptable. Then you send the latter units back to the supplier for replacements. If only 5 percent of the units are rejected, you have imposed a large cost on yourself and a small cost on the supplier. And to boot, you have saved the supplier the cost of being responsible for the quality of its output! On the other hand, suppose you test a small sample from the batch, find 5 percent unsatisfactory units in the sample, and on the basis of the sample send *the entire batch* back for replacement. This imposes a small cost on you and a large cost on the supplier. The supplier may resent the fact that you are sending the acceptable units back along with the unacceptable ones. However, if the supplier values your business, it is ultimately going to take responsibility for ensuring the quality of its output. And if the supplier does not value your business, then you are well served by learning this and seeking another supplier.

Acceptance sampling can motivate suppliers to improve quality

The statistical techniques used in *acceptance sampling* will be familiar to you as applications of the sampling and hypothesis-testing ideas we discussed in Chapters 6, 8 and 9. Much of the original work in acceptance sampling was done in the 1920s and 1930s by Harold F. Dodge and Harry G. Romig, who, like Walter Shewhart, did their research at Bell Labs. They discussed *single-sampling* and *double-sampling* schemes:

In single sampling, two numbers are specified: a sample size, n , and an *acceptance number*, c , the maximum number of allowable pieces with defects. A sample of size n is taken, and the lot is accepted if there are c or fewer defective pieces in the sample, but rejected if the number of defective pieces is greater than c .

Single sampling

Double sampling is more complicated, and depends on four specified numbers, n_1 , n_2 , c_1 , and c_2 ($>c_1$), which are used as follows:

Double sampling

- First a sample of size n_1 is taken. Let b_1 (b for *bad*) be the number of defective pieces in this sample:
 - If $b_1 \leq c_1$, the lot is accepted.
 - If $b_1 > c_2$, the lot is rejected.
 - If $c_1 < b_1 \leq c_2$, an additional n_2 units are sampled.

Let b_2 be the *total number* of defective pieces in the combined sample of $n_1 + n_2$ units:

- If $b_2 \leq c_2$, the lot is accepted.
- If $b_2 > c_2$, the lot is rejected.

As you can imagine, the analysis of double-sampling schemes is considerably more complicated than that of single-sampling schemes. Although double-sampling schemes are more powerful and more widely used in practice, we shall restrict our discussion to single sampling. This will enable you to learn the concepts without getting bogged down in the details.

An Example of Acceptance Sampling

Consider a problem faced by Maureen Brennan, the quality control engineer at Northway Computers, a manufacturer of personal computers. Northway is negotiating a contract for batches of 1,000 3½-inch disk drives with Drives Unlimited. Drives Unlimited has a reputation as a supplier of high-quality drives, but its output is not perfect. It claims that it can produce drives with rates of defects below 1 percent, a level that is acceptable to Maureen Brennan. This 1 percent level is called the *acceptable quality level* (AQL). Loosely speaking, it defines how high a defect level still constitutes a “good” lot.

Acceptable quality level

Now, what happens when Maureen chooses values of n and c for her sampling scheme? For instance, suppose she picks $n = 100$ and $c = 1$. If p is Drives Unlimited’s true rate of defects, the probability that any batch will be rejected can be computed using the binomial distribution. This is because Maureen’s random sample of 100 taken from a batch of 1,000 drives is also a random sample taken from Drives Unlimited’s total output stream. Now, with $n = 100$ and $p = 0.01$,

$$\begin{aligned} P(r = 0 \text{ defects}) &= \frac{n!}{r!(n-r)!} p^r q^{n-r} \\ &= \frac{100!}{0!100!} (0.01)^0 (0.99)^{100} \\ P(r = 1 \text{ defect}) &= \frac{100!}{1!99!} (0.01)^1 (0.99)^{99} \end{aligned} \quad [5-1]$$

Hence, the probability a batch will be rejected is $1 - 0.3660 - 0.3697 = 0.2643$. This probability is called the *producer’s risk*. It is the chance of rejecting a batch even when Drives Unlimited’s true rate of defects is only 1 percent. This corresponds to a Type I error in hypothesis testing.

Producer’s risk: a Type I error

The corresponding Type II error leads to *consumer’s risk* (a buyer’s risk). Suppose that the minimum defect rate Northway would like to reject in a batch of diskette drives is 2 percent. This 2 percent level is called the *lot tolerance percent defective* (LTPD). Loosely speaking, it defines how low a defect level still constitutes a “bad” lot. Suppose that a batch of 1,000 drives with 20 defective units is received by Northway. What is the probability that this batch will be accepted because Maureen’s sample of 100 contains no more than one defective unit? This probability is the consumer’s risk.

Consumer’s risk: a Type II error

Lot tolerance percent defective

Because she is sampling without replacement, the binomial distribution is not the correct distribution for computing this probability. The correct distribution is a relative of the binomial, known as the *hypergeometric distribution*. It is common to use the binomial distribution to approximate consumer's risk. This approximation always *overestimates* the true **Approximating consumer's risk** value of the consumer's risk whenever that risk is less than 0.5. With Maureen's sampling scheme, the approximate binomial probability of accepting a batch of 1,000 units with 20 defective units is computed using Equation 5-1, with $n = 100$ and $p = 0.02$:

$$\begin{aligned} P(r = 0 \text{ defects}) &= \frac{n!}{r!(n-r)!} p^r q^{n-r} \\ &= \frac{100!}{0!100!} (0.02)^0 (0.98)^{100} = 0.1326 \\ P(r = 1 \text{ defect}) &= \frac{100!}{1!99!} (0.02)^1 (0.98)^{99} = 0.2707 \end{aligned} \quad [5-1]$$

Hence, the approximate probability the batch will be accepted is $0.1326 + 0.2707 = 0.4033$. The exact hypergeometric probability of accepting a batch with 20 defective units is 0.3892, so the approximation is fairly good (the error is only 141/3, 892, or about 3.6 percent). In general, the smaller the fraction of the batch that is sampled, the better the job the binomial distribution does to approximate the hypergeometric. This is analogous to the situation we encountered in Chapter 6 (p. 314), where we saw that the finite population multiplier had little effect on the calculation of the standard error of the mean if the sampling fraction was less than 0.05.

Maureen is unwilling to accept such a high level of risk. She can reduce her risk by lowering c to 0 and rejecting lots in which any defective units show up in her sample of 100. This will reduce her risk to exactly 0.1326 (0.1190 approximate), but it will increase the producer's risk to 0.6340, which Drives Unlimited is unwilling to accept. Is there any way to reduce both the producer's and the consumer's risks? Yes, by increasing the sample size. Suppose she increases her sample size to $n = 250$, and allows 1.2 percent defects in the sample by setting $c = 3$. Then Northway's consumer's risk is now reduced to 0.2225 exact, 0.2622 approximate, and Drives Unlimited's producer's risk is reduced to 0.2419. Of course, this will increase the cost of the inspections that Maureen will have to make. Similar results can be achieved with double sampling without such a drastic increase in total sample size.

Tradeoffs between the two risks

Increasing n to decrease both risks

Acceptance Sampling in Practice: Tables and Computer Programs

As you can see from our example, the relationships between sample size (n), acceptance number (c), and the two types of risk are very complex. Extensive tables exist for helping quality engineers to choose appropriate acceptance sample schemes.*

The Dodge–Romig tables

* For example, see *Sampling Inspection Tables—Single and Double Sampling*, by H. F. Dodge and H. G. Romig, John Wiley, New York, 1959.

A	B	C	D	E	F	G	H	
1 ACCEPTSA	ACCEPTANCE SAMPLING BY ATTRIBUTES				Alt A View AOQ curve			
2					Alt O View OC curve			
3 INPUT:				OUTPUT:				
4 Lot Size		1000			Producer's risk (Prob. lot with 0.2642 defects = AQL will be rejected)			
5 Sample size		100			Consumer's risk (Prob. lot with 0.5578 defects = LTPD will be accepted)			
6 Acceptance number		1						
7 AQL		1.00%						
8 LTPD		1.50%						
9 -----								
10 # defects			Incoming quality (lot % defective)					
11 in sample	X!		0.0%	1.0%	2.5%	3.0%	4.0%	5.0%
12 0	1	1.0000	0.3679	0.2231	0.0498	0.0183	0.0067	
13 1	1	0.0000	0.3679	0.3347	0.1494	0.0733	0.0337	
14 2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
15 3	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
16 4	24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
17 5	120	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
18		-----	-----	-----	-----	-----	-----	
19 Prob. accept. lot		1.0000	0.7358	0.5578	0.1991	0.0916	0.0404	
20 AOQ		0.00%	0.66%	0.75%	0.54%	0.33%	0.18%	

FIGURE 10-10 A SPREADSHEET TO EVALUATE ACCEPTANCE-SAMPLING SCHEMES

As an alternative to looking up sampling schemes in tables, there are many computer programs available for evaluating choices of n and c . A particularly easy one to use is a Lotus 1-2-3 spreadsheet template, developed by Everette S. Gardner, Jr.,* and used with his permission. Figure 10-10 shows the application of that spreadsheet to evaluate Maureen Brennan's original ($n = 100$, $c = 1$) sampling scheme.

In cells C4 to C8 (shaded in color), Maureen entered the lot size (1,000), sample size ($n = 100$), acceptance number ($c = 1$), AQL (0.01), and LTPD (0.015). In cells H4 and H7 (shaded in color), the template calculates the producer's risk (0.2642) and the binomial approximation to the consumer's risk (0.5578). (These figures are slightly different from those we calculated earlier because Gardner uses the Poisson approximation to the binomial—see p. 243 in his spreadsheet calculations.)

The bottom part of the template calculates various probabilities that give Maureen more information about the behavior of her sampling scheme. Cells C11 to H11 originally contained incoming qualities ranging from 0 to 5 percent, by single percentage points. Because our LTPD was 1.5 percent, we replaced the original 2 percent in cell E11 (shaded in color) by 1.5 percent. The additional information can be seen most easily in two graphs that the template can produce. Maureen can get an *operating characteristic* (OC) graph (Figure 10-11). The height of the OC curve tells her the consumer's risk, the probability that her sampling scheme will *accept* a lot from a production process with an input quality read on the horizontal axis. Subtracting that probability from one gives the producer's risk as the input quality varies. As you would expect, the probability that a lot will be accepted falls as production quality becomes worse.

A spreadsheet template**The OC curve: consumer's risk as a function of input quality**

* *The Spreadsheet Operations Manager*, McGraw-Hill, New York, 1992.

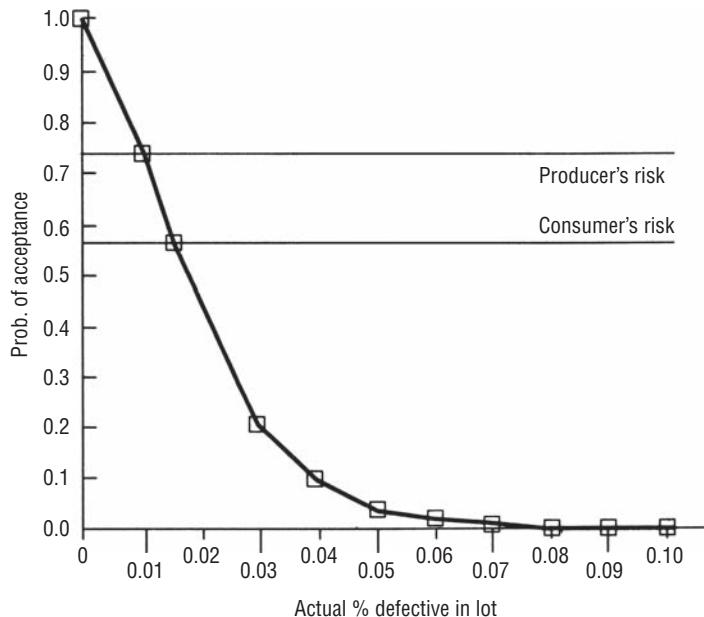


FIGURE 10-11 THE OPERATING CHARACTERISTIC (OC) CURVE FOR MAUREEN'S ACCEPTANCE-SAMPLING SCHEME

Maureen can also get an *average outgoing quality* (AOQ) graph (Figure 10-12). The height of the AOQ curve tells her how the long-run average fraction of defective units in lots accepted by her sampling scheme varies as a function of the quality of the drives supplied to Northway by Drives Unlimited. You can see from that graph that the worst long-run average quality would be 0.75 percent, or about 7.5 defective drives per accepted batch of 1,000. Of course, because AOQ is an average, some accepted batches will have more defective drives.

The AOQ curve: average outgoing quality as a function of input quality

HINTS & ASSUMPTIONS

Warning: Making up or changing your sampling plan as you go generally leads to failure. Carefully planning your sampling plan using sound statistical analysis and then adhering to the plan makes it much less likely that you will be misled by random patterns. Hint: If a municipality tests 200 street light-bulbs from a shipment of 10,000, finds that the first 100 work perfectly, and quits sampling, it can get into serious trouble. Most acceptance situations like this are looking for very low-probability defects, say 1 in 100. Because you know that random events are not uniformly distributed, you should not be swayed by the absence of defects in the first 100 and you should stick with the sampling plan you first designed if you want to benefit from the power of statistical quality control.

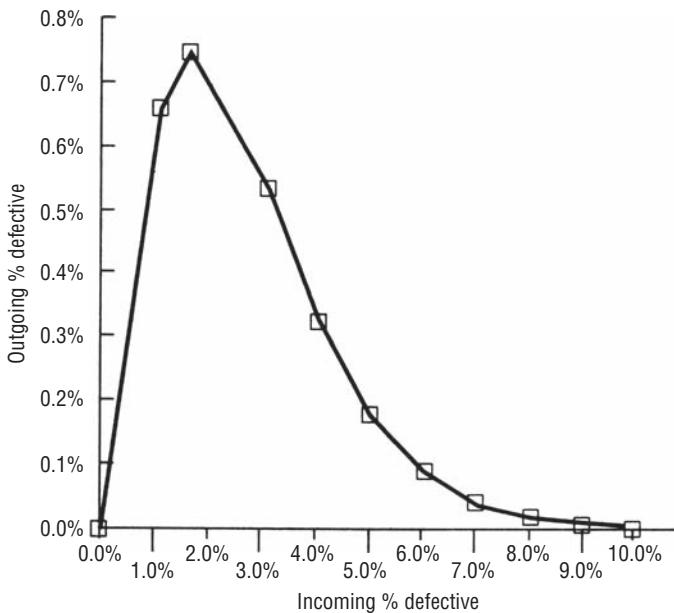


FIGURE 10-12 THE AVERAGE OUTGOING QUALITY (AOQ) CURVE FOR MAUREEN'S ACCEPTANCE-SAMPLING SCHEME

EXERCISES 10.7

Self-Check Exercises

SC 10-8 Compute the producer's risks for the following single-sampling schemes from batches of 2,000 items, with AQL = 0.005.

- (a) $n = 150, c = 1$.
- (b) $n = 150, c = 2$.
- (c) $n = 200, c = 1$.
- (d) $n = 200, c = 2$.

SC 10-9 Use the binomial distribution to approximate the consumer's risks in the sampling schemes in Exercise SC 10-8 if LTPD = 0.01.

Basic Concepts

10-34 Why is it impractical to inspect an entire batch of input from a supplier?

10-35 What is the significance of the acceptance number, c , in single sampling?

Applications

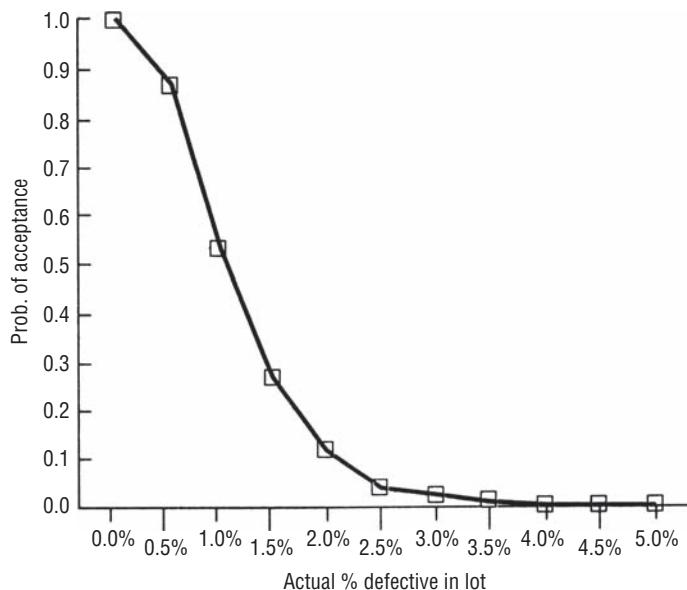
10-36 Compute the producer's risks for the following single sampling schemes from batches of 1,500 items, with AQL = 0.02.

- (a) $n = 175, c = 3$.
- (b) $n = 175, c = 5$.

- (c) $n = 250, c = 3$.
 (d) $n = 250, c = 5$.

10-37 Use the binomial distribution to approximate the consumer's risks in the sampling schemes in Exercise 10-36 if LTPD = 0.03.

10-38 The graph below, is an OC curve for a single-sampling scheme from batches of 2,500 with $n = 250$ and $c = 2$. Find the producer's risk if the AQL is
 (a) 0.005.
 (b) 0.010.
 (c) 0.015.



10-39 For the single-sampling scheme in Exercise 10-38, use the OC curve to find the consumer's risk if the LTPD is
 (a) 0.010.
 (b) 0.015.
 (c) 0.020.

Worked-Out Answers to Self-Check Exercises

SC 10-8 AQL = 0.005

$$(a) \quad n = 150 \quad c = 1 \\ r = 0:$$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{0!(150)!} (0.005)^0 (0.995)^{150} = 0.4715$$

$$r = 1:$$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{1!(149)!} (0.005)^1 (0.995)^{149} = 0.3554$$

$$1 - 0.4715 - 0.3554 = 0.1731, \text{ the producer's risk.}$$

(b) $n = 150 \quad c = 2$
 $r = 2:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{2!(148)!} (0.005)^2 (0.995)^{148} = 0.1330$$

$1 - 0.4715 - 0.3554 - 0.1330 = 0.0401$ is the producer's risk.

(c) $n = 200 \quad c = 1$
 $r = 0:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{0!(200)!} (0.005)^0 (0.995)^{200} = 0.3670$$

$r = 1:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{1!(199)!} (0.005)^1 (0.995)^{199} = 0.3688$$

$1 - 0.3670 - 0.1330 = 0.2642$, the producer's risk.

(d) $n = 200 \quad c = 2$
 $r = 2:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{2!(198)!} (0.005)^2 (0.995)^{198} = 0.1844$$

$1 - 0.3670 - 0.3688 - 0.1844 = 0.0798$, the producer's risk.

SC 10-9 LTPD = 0.01

(a) $n = 150 \quad c = 1$
 $r = 0:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{0!(150)!} (0.01)^0 (0.99)^{150} = 0.2215$$

$r = 1:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{1!(149)!} (0.01)^1 (0.99)^{149} = 0.3355$$

$0.2215 + 0.3355 = 0.557$, the consumer's risk.

(b) $n = 150 \quad c = 2$
 $r = 2:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{150!}{2!(148)!} (0.01)^2 (0.99)^{148} = 0.2525$$

$0.2215 + 0.3355 + 0.2525 = 0.8095$, the consumer's risk.

(c) $n = 200 \quad c = 1$
 $r = 0:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{0!(200)!} (0.01)^0 (0.99)^{200} = 0.1340$$

$r = 1:$

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{1!(199)!} (0.01)^1 (0.99)^{199} = 0.2707$$

$0.1340 + 0.2707 = 0.4047$, the consumer's risk.

(d) $n = 200$ $c = 2$

$r = 2$:

$$\frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{200!}{2!(198)!} (0.01)^2 (0.99)^{198} = 0.2720$$

$0.1340 + 0.2707 + 0.2720 = 0.6767$, the consumer's risk.

STATISTICS AT WORK

Loveland Computers

Case 10: Quality and Quality Control Walter Azko prided himself on his open-door policy, and any member of the firm was welcome to stop by with ideas. The only difficulty was finding Walter in his office, or indeed, in the country. He still traveled frequently to Pacific Rim countries in search of new suppliers and better prices.

But Walter was in town—and in his office—going over budget projections with Lee when Jeff Cohen from Purchasing and Harry Patel, the firm's financial controller, dropped by. Jeff and Harry were the only two CPAs in the firm so they were often found deep in conversation.

“Boss, we both went to a seminar put on by the State CPA Association,” Harry began.

“So we wanted to talk with you about quality initiatives at Loveland Computers,” added Jeff. The two were known for finishing each other's sentences.

“I'm not going to pay good money for a bunch of high-priced consultants to come in and preach to us,” Walter greeted their enthusiasm with skepticism. “In any case, I've always told you—at this end of the market we compete on price, not on quality. Our customers only care whether a Loveland Computer works and whether we have it in stock when they want to order it. And if it doesn't work, they just have to ship it back to us and we'll send them out a new one.”

“Right. And how much is that costing us?” Harry asked.

“You have the figures—the money we write off on computers that we have to scrap is very small compared to our volume. You know that.”

“Well, after that seminar, I'm sure that we don't really capture all the costs of a failure,” the controller disputed.

“Anyway, we test all the computers overnight at the end of the line before we ship them. What do you want me to do—run them for a week before we ship them out?” Walter remained unconvinced that thinking about quality could change the way Loveland did business.

“Doesn't this relate to customer satisfaction?” Lee interjected. “I read where J. D. Power—the company that reports on what automobile customers think about new models—is going to start rating PCs.”

“There's much more to quality than more testing before we ship things, in fact, if we do things right, I'm convinced we'd need less testing on the production line,” Jeff added. “Let Harry and me buy you lunch and tell you what we learned at the seminar.”

Study Questions: What arguments will Jeff and Harry make against Walter's assertion that Loveland competes only on price? What are the total costs of replacing a machine that Harry refers to? How does quality relate to customer satisfaction? Why would Loveland need less end-of-the-line testing if it adopted quality control measures? Does it matter whether Walter Azko ends up with a better understanding of quality by the end of lunch?

CHAPTER REVIEW

Terms Introduced in Chapter 10

Acceptable Quality Level (AQL) The average quality level promised by a producer; the maximum number or percentage of defective pieces in a “good” lot.

Acceptance Number The maximum number of defective pieces with which a lot will still be accepted.

Acceptance Sampling Procedures for deciding whether to accept or reject a batch of input materials based on the quality of a sample taken from that batch

Assignable Variation Nonrandom, systematic variability in a process. It usually can be corrected without redesigning the entire process.

Attributes Qualitative variables with only two categories.

Average Outgoing Quality (AOQ) Curve A graph showing how the long-run average fraction of defective units in lots accepted by a sampling scheme varies as a function of the input quality of the lots.

Cause-and-Effect Diagram Another name for a *fish-bone diagram*.

Common Variation Random variability inherent in a process. It usually cannot be reduced without re-designing the entire process.

Consumer’s Risk The chance that a “bad” lot will be accepted.

Continuous Quality Improvement (CQI) Constant attention to the identification and resolution of problems in TQM.

Control Charts A plot of some parameter of interest (such as \bar{x} , R , or p) over time, used to identify assignable variations and to make adjustments to the process being monitored.

Control Limits Upper and lower bounds on control charts. For the process to be in-control, all observations must fall within these limits.

Fishbone Diagram A pictorial device for organizing cause-and-effect relationships among the factors causing problems in complex systems.

Hypergeometric Distribution The correct distribution for computing consumer’s risk; it is often approximated by the binomial distribution.

Inherent Variation Another name for *common variation*.

Ishikawa Diagram Another name for a *fishbone diagram*.

Lot Tolerance Percent Defective (LTPD) The minimum number or percentage of defective pieces in a “bad” lot.

Operating Characteristic (OC) Curve A graph showing the probability an acceptance-sampling scheme will accept a batch as a function of the input quality of the batch.

Outliers Observations falling outside the control limits on a control chart.

Out-of-Control A process exhibiting outliers on a control chart, or showing nonrandom patterns even though there are no outliers.

p Charts Control charts for monitoring the proportion of items in a batch that meet specifications.

Pareto Chart A bar graph showing groups of error causes arranged by their frequencies of occurrence.

Producer’s Risk The chance that a “good” lot will be rejected.

Qualitative Variables Variables whose values are categorical rather than numerical.

Quality Fitness for use or conformance to requirements.

Quantitative Variables Variables with numerical values resulting from measuring or counting.

R Charts Control charts for monitoring process variability.

Special Cause Variation Another name for *assignable variation*.

Statistical Process Control (SPC) Shewhart’s system of using control charts to track variation and identify its causes.

Total Quality Management (TQM) A set of approaches that enables the managers of complex systems to match the firm's products to customers' expectations.

\bar{x} Chart Control charts for monitoring process means.

Equations Introduced in Chapter 10

$$10-1 \quad \bar{\bar{x}} = \frac{\sum x}{n \times k} = \frac{\sum \bar{x}}{k} \quad \text{p. 487}$$

To compute the grand mean ($\bar{\bar{x}}$) from several (k) samples of the same size (n), either sum all the original observations ($\sum x$) and divide by the total number of observations ($n \times k$), or else sum the means from each of the samples ($\sum \bar{x}$) and divide by the number of samples (k). Then use $\bar{\bar{x}}$ for the center line (CL) of an \bar{x} chart.

$$10-2 \quad \begin{aligned} \text{UCL} &= \bar{\bar{x}} + \frac{3\bar{R}}{d_2 \sqrt{n}} \\ \text{LCL} &= \bar{\bar{x}} - \frac{3\bar{R}}{d_2 \sqrt{n}} \end{aligned} \quad \text{p. 489}$$

To compute the control limits for an \bar{x} chart, multiply the average sample range ($\bar{R} = \Sigma R/k$) by 3, and then divide by the product of d_2 (from Appendix Table 9) and \sqrt{n} ; the result is then added to and subtracted from $\bar{\bar{x}}$. Alternatively, you can compute these limits as $\bar{\bar{x}} \pm A_2 \bar{R}$, where $A_2 (= 3/d_2 \sqrt{n})$ can also be found in Appendix Table 9.

$$10-3 \quad \sigma_R = d_3 \sigma \quad \text{p. 496}$$

To get the standard deviation of the sampling distribution of R , multiply the population standard deviation, σ , by d_3 , another factor that is also given in Appendix Table 9.

$$10-4 \quad \begin{aligned} \text{UCL} &= \bar{R} + \frac{3d_3 \bar{R}}{d_2} = \bar{R} \left(1 + \frac{3d_3}{d_2} \right) \\ \text{LCL} &= \bar{R} - \frac{3d_3 \bar{R}}{d_2} = \bar{R} \left(1 - \frac{3d_3}{d_2} \right) \end{aligned} \quad \text{p. 496}$$

To compute the control limits for an R chart, multiply the average sample range ($\bar{R} = \Sigma R/k$) by $1 \pm 3d_3/d_2$. Alternatively, you can compute these limits as

$$\begin{aligned} \text{UCL} &= \bar{R} D_4, \text{ where } D_4 = 1 + 3d_3/d_2 \\ \text{LCL} &= \bar{R} D_3, \text{ where } D_3 = 1 - 3d_3/d_2 \end{aligned}$$

Values of D_3 and D_4 are also given in Appendix Table 9. Because ranges are always nonnegative, D_3 and the LCL are taken to be 0 when $n \leq 6$.

$$10-5 \quad \text{CL} = \mu_{\bar{p}} = p \quad \text{p. 503}$$

$$10-6 \quad \begin{aligned} \text{UCL} &= \mu_{\bar{p}} + 3\sigma_{\bar{p}} = p + 3\sqrt{\frac{pq}{n}} \\ \text{LCL} &= \mu_{\bar{p}} - 3\sigma_{\bar{p}} = p - 3\sqrt{\frac{pq}{n}} \end{aligned} \quad \text{p. 503}$$

If there is a known or targeted value of p , that value should be used in Equations 10-5 and 10-6 to get the center line and control limits for a p chart. However, if no such value of p is available, then you should use the overall sample fraction

$$10-7 \quad \bar{p} = \frac{\sum \bar{p}_j}{k} \quad \text{p. 503}$$

where

- \bar{p}_j = sample fraction in the j th sample
- k = total number of samples

Review and Application Exercises

- 10-40** R&H Bloch is a large accounting firm specializing in the preparation of individual federal tax returns. The firm is very conservative in its practices and tries to avoid having more than 2 percent of its clients audited. As part of a summer internship, Jane Bloch has been asked to see whether this goal is being met on a consistent basis. For each week during a 16-week interval centered on April 15 of last year, she has randomly selected 125 returns prepared by the firm. (Those filed after April 15 had paid their estimated taxes due and requested an extension.) Her data follow:

Week Ending	2/25	3/04	3/11	3/18	3/25	4/01	4/08	4/15
# Audited	2	1	2	3	5	4	5	6
# Week Ending	4/22	4/29	5/06	5/13	5/20	5/27	6/03	6/10
# Audited	3	1	1	3	2	2	3	2

- (a) Are significantly more than 2 percent of R&H Bloch's clients being audited? State and test appropriate hypotheses using all 2,000 clients in Jane's sample.
- (b) Notwithstanding your result in part (a), construct a p chart based on Jane's data. Is there anything evident in the chart that Jane should bring to the attention of the partners in the firm? Explain.
- 10-41** When slaying dragons, should you be concerned with the "trivial many" or the "vital few"? Explain.
- 10-42** If marital status is coded as "currently married" or "never married," then marital status is an attribute. However, if it is coded as "single," "married," "widowed," or "divorced," then it is not an attribute. Explain this apparent inconsistency.
- 10-43** The amount of time a bank teller needs to process a deposit depends on how many items the customer has. Is this inherent or special cause variation? Explain.
- 10-44** All checks drafted on accounts at Global Bank are returned to the bank's check-processing center. There each check is encoded with optically scannable characters that indicate the amount for which it is drawn. The encoded checks are then scanned so that payment can be made and the accounts on which they have been drawn can be debited. Shih-Hsing Liu has been monitoring the encoding operation, and has counted the number of checks processed in 10 randomly chosen 2-minute periods during each hour of the last two 8-hour shifts. She has

recorded the following data:

Shift 1 Time	0700	0800	0900	1000	1100	1200	1300	1400
\bar{x}	49.4	49.9	48.8	50.1	49.7	48.1	48.6	48.7
R	4	7	7	4	7	10	7	10
Shift 2 time	1500	1600	1700	1800	1900	2000	2100	2200
\bar{x}	50.7	51.3	51.1	51.6	50.0	50.5	51.4	50.1
R	6	4	9	9	6	7	4	7

- (a) Help Shih-Hsing construct an \bar{x} chart from the data.
 (b) Is the process in-control? Does anything in the chart indicate that Shih-Hsing should examine the process more closely? Explain.
- 10-45** (a) Use Shih-Hsing Liu's data from Exercise 10-44 to construct an R chart.
 (b) Does anything in the chart indicate that Shih-Hsing should examine the process more closely? Explain.
- 10-46** Security Construction uses many subcontractors for the condominium apartments it builds throughout the American sunbelt. Dawn Locklear, Security's customer service representative, has been reviewing the "punch lists" submitted by the purchasers of 500 condos. A punch list is a list of problems noted when the owner moves into the apartment. Security does not receive final payment until the items on the list have been corrected. Dawn has categorized the items on the lists according to the responsible subcontractor. Use her information to construct a Pareto chart to identify which subcontractors require additional supervision.
- | Subcontractor | Number of Problems |
|---------------|--------------------|
| Electrical | 257 |
| Flooring | 23 |
| Heating/AC | 35 |
| Painting | 19 |
| Plumbing | 22 |
| Roofing | 31 |
| Tile | 51 |
| Wallboard | 303 |
| Windows | 16 |
| Other | 68 |
- 10-47** Compute the producer's risks for the following single-sampling schemes from batches of 2,500 items, with AQL = 0.01.
 (a) $n = 200, c = 1$.
 (b) $n = 200, c = 2$.
 (c) $n = 250, c = 1$.
 (d) $n = 250, c = 2$.
- 10-48** Use the binomial distribution to approximate the consumer's risks in the sampling schemes in Exercise 10-47 if LTPD = 0.015.
- 10-49** In service operations (as opposed to manufacturing) can the principle *Variation is the enemy of quality* be applied? Aren't all customers different?

- 10-50** Deshawn Jackson is the quality supervisor for Reliance Storage Media, a manufacturer of diskettes for personal computers. The company has been concerned about the quality of their Reliant economy-grade 3½" diskettes, and has completely revamped the production process. Reliant diskettes consist of a cobalt-enhanced iron-oxide coating deposited on a polyethylene terephthalate substrate. The nominal thickness of the coating is 75.0 microns (0.075 mm), but a deviation of ± 3.0 microns is acceptable. The diskettes are manufactured in batches of 2,500. In order to evaluate the new production process, Deshawn has sampled two dozen diskettes from each of the last 20 batches and recorded the following data:

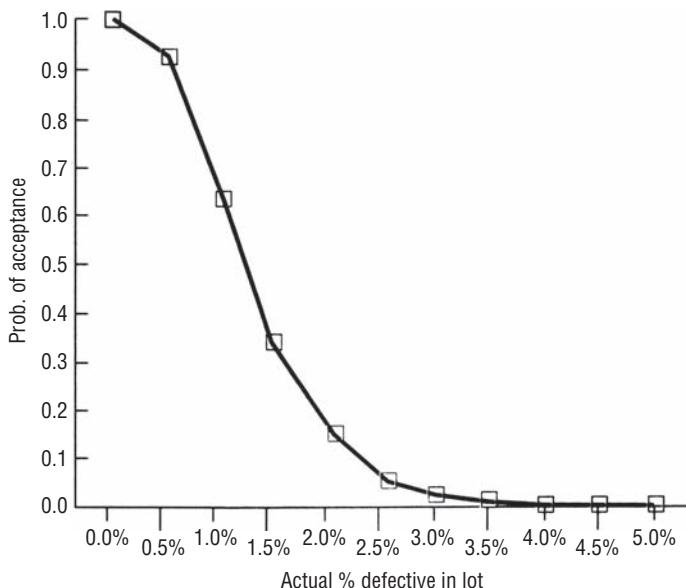
Batch	1	2	3	4	5	6	7	8	9	10
\bar{x}	75.3	75.0	74.8	75.0	75.3	74.8	74.8	74.9	74.6	74.9
R	3.2	3.3	3.6	3.5	3.8	3.7	3.4	3.3	3.4	3.1
Batch	11	12	13	14	15	16	17	18	19	20
\bar{x}	75.2	75.1	74.8	74.9	74.9	75.1	75.0	74.9	74.9	75.1
R	3.1	3.0	3.1	2.9	2.8	2.8	2.7	2.9	2.8	2.9

- (a) Use the data to construct an \bar{x} chart.
 (b) Is the process in-control?
 (c) Deshawn looks at the \bar{x} chart and says, "The last 10 batches have means that appear to be less variable than the means of the first 10 batches." Is this observation valid? Explain. Should Deshawn be concerned? Explain.
- 10-51** Consider the data Deshawn Jackson collected for Exercise 10-50:
 (a) Construct an R chart.
 (b) Should Deshawn worry about the obvious pattern in the chart? Explain.
 (c) Is there any relationship between the pattern in the R chart and the one Deshawn noticed in the \bar{x} chart? (See Exercise 10-50(c).) Is this good news or bad news for Deshawn? Explain.
- 10-52** Photomatic prints customers' 35 mm film using automated equipment. This high-volume, low-cost approach works for most typical situations, but variation in the input can lead to poor results. For example, if a customer's film has been left in a hot car, it may be printable with special handling, but the results from the automated process are unacceptable. When prints are rejected by customers, Photomatic must reprint by hand—a process that costs more than the price charged—so each "defect" is a loss for the firm. The equipment supplier notes that sophisticated light measuring circuitry should produce acceptable print quality with no more than one defect per thousand. Quality engineer B. J. Nighthorse randomly sampled 2,000 prints from each of the last 20 production runs and recorded the following information:

Run	1	2	3	4	5	6	7	8	9	10
# Defective	3	1	2	4	4	2	2	3	1	0
Run	11	12	13	14	15	16	17	18	19	20
# Defective	2	2	1	3	2	2	0	4	2	0

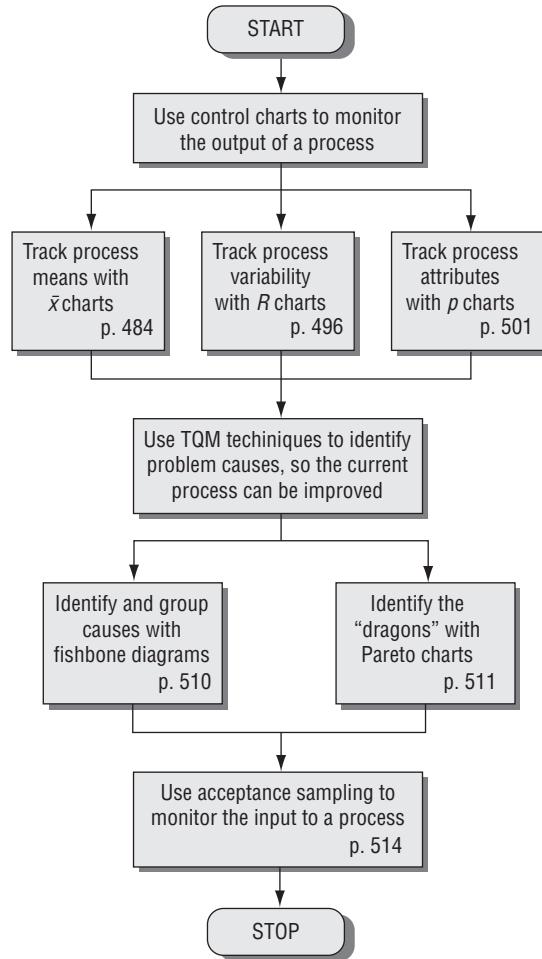
Construct a p chart to see whether the equipment is performing within the manufacturer's specifications and whether the process is in-control.

- 10-53** Explain how producer's and consumer's risks in acceptance sampling correspond to Type I and Type II errors in hypothesis testing.
- 10-54** A 14-ounce box of soda crackers will almost never weigh exactly 14 ounces. What sources of common and special cause variation might explain this observation?
- 10-55** The graph below is an OC curve for a single-sampling scheme from batches of 3,000 with $n = 300$ and $c = 3$. Find the producer's risk if the AQL is:
- 0.005.
 - 0.010.
 - 0.015.



- 10-56** For the single-sampling scheme in Exercise 10-55, use the OC curve to find the consumer's risk if the LTPD is:
- 0.010.
 - 0.015.
 - 0.020.
- 10-57** Connie Rodrigues, the Dean of Students at Midstate College, is wondering about grade inflation at the school. She has randomly selected 200 students from each of the last 20 graduating classes and has looked up their grade-point averages. In addition, for each year's sample she has calculated the percentage of A and B grades for all 200 students as a group. Explain how she can use control charts to analyze whether Midstate has been experiencing grade inflation.
- 10-58** Explain how acceptance sampling can be more effective in the long run than sampling entire batches of input.
- 10-59** Education seems to be a difficult field in which to use quality techniques. One possible outcome measure for colleges is the graduation rate (the percentage of students matriculating who graduate on time). Would you recommend using p or R charts to examine graduation rates at a school? Would this be a good measure of quality?

Flow Chart: Quality and Quality Control



This page is intentionally left blank.

Chi-Square and Analysis of Variance

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To recognize situations requiring the comparison of more than two means or proportions
- To introduce the chi-square and F distributions and learn how to use them in statistical inferences
- To use the chi-square distribution to see whether two classifications of the same data are independent of each other
- To use a chi-square test to check whether a particular collection of data is well described by a specified distribution
- To use the chi-square distribution for confidence intervals and testing hypotheses about a single population variance
- To compare more than two population means using analysis of variance
- To use the F distribution to test hypotheses about two population variances

CHAPTER CONTENTS

11.1 Introduction	532
11.2 Chi-Square as a Test of Independence	533
11.3 Chi-Square as a Test of Goodness of Fit: Testing the Appropriateness of a Distribution	548
11.4 Analysis of Variance	555
11.5 Inferences about a Population Variance	584
11.6 Inferences about Two Population Variances	589

■ Statistics at Work	597
■ Terms Introduced in Chapter 11	598
■ Equations Introduced in Chapter 11	599
■ Review and Application Exercises	601
■ Flow Chart: Chi-Square and Analysis of Variance	608

The training director of a company is trying to evaluate three different methods of training new employees. The first method assigns each to an experienced employee for individual help in the factory. The second method puts all new employees in a training room separate from the factory, and the third method uses training films and programmed learning materials. The training director chooses 16 new employees assigned at random to the three training methods and records their daily production after they complete the programs:

Method 1	15	18	19	22	11
Method 2	22	27	18	21	17
Method 3	18	24	19	16	22

The director wonders whether there are differences in effectiveness among the methods. Using techniques learned in this chapter, we can help answer that question. ■

11.1 INTRODUCTION

In Chapters 8 and 9, we learned how to test hypotheses using data from either one or two samples. We used one-sample tests to determine whether a mean or a proportion was significantly different from a hypothesized value. In the two-sample tests, we examined the difference between either two means or two proportions, and we tried to learn whether this difference was significant.

Suppose we have proportions from five populations instead of only two. In this case, the methods for comparing proportions described in Chapter 9 do *not* apply; we must use the *chi-square test*, the subject of the first portion of this chapter. Chi-square tests enable us to test whether *more* than two population proportions can be considered equal.

Actually, chi-square tests allow us to do a lot more than just test for the equality of several proportions. If we classify a population into several categories with respect to two attributes (such as age and job performance), we can then use a chi-square test to determine whether the two attributes are independent of each other.

Managers also encounter situations in which it is useful to test for the equality of more than two population means. Again, we cannot apply the methods introduced in Chapter 9 because they are limited to testing for the equality of only two means. The *analysis of variance*, discussed in the fourth section of this chapter, will enable us to test whether *more* than two population means can be considered equal.

It is clear that we will not always be interested in means and proportions. There are many managerial situations where we will be concerned about the variability in a population. Section 11.5 shows how to use the chi-square distribution to form confidence intervals and test hypotheses about a population variance. In Section 11.6, we show that hypotheses comparing the variances of two populations can be tested using the *F* distribution.

Uses of the chi-square test

Function of analysis of variance

Inferences about population variances

EXERCISES 11.1

- 11-1 Why do we use a chi-square test?
- 11-2 Why do we use analysis of variance?

- 11-3** In each of the following situations, state whether a chi-square test, analysis of variance, or inference about population variances should be done.
- We want to see whether the variance in spring temperatures is the same on the east and west coasts.
 - We want to see whether the average speed on Interstate 95 differs depending on the day of the week.
 - We want to see whether long-term stock performance on Wall Street (classified as good, average, or poor) is independent of the size of the company (classified as small, medium, or large).
 - Before testing whether $\mu_1 = \mu_2$, we want to test whether the assumption that $\sigma_1^2 = \sigma_2^2$ is reasonable.
- 11-4** Answer true or false and explain your answers.
- After reading this chapter, you should know how to make inferences about two or more population variances.
 - After reading this chapter, you should know how to make inferences about two or more population means.
 - After reading this chapter, you should know how to make inferences about two or more population proportions.
- 11-5** To help remember which distribution or technique is used, complete the following table with either the name of a distribution or the technique involved. The row classification refers to the number of parameters involved in a test, and the column classification refers to the type of parameter involved. Some cells may not have an entry; others may have more than one possible entry.

Number of Parameters Involved	Type of Parameter		
	μ	σ	P
1			
2			
3 or more			

11.2 CHI-SQUARE AS A TEST OF INDEPENDENCE

Many times, managers need to know whether the differences they observe among several sample proportions are significant or only due to chance. Suppose the campaign manager for a presidential candidate studies three geographically different regions and finds that 35, 42, and 51 percent, respectively, of the voters surveyed in the three regions recognize the candidate's name. If this difference is significant, the manager may conclude that location will affect the way the candidate should act. But if the difference is not significant (that is, if the manager concludes that the difference is solely due to chance), then he may decide that the place chosen to make a particular policy-making speech will have no effect on its reception. To run the campaign successfully, then, the manager needs to determine whether location and name recognition are dependent or independent.

Sample differences among proportions: Significant or not?

TABLE 11-1 SAMPLE RESPONSE CONCERNING REVIEW SCHEDULES FOR NATIONAL HEALTH CARE HOSPITAL EMPLOYEES

	Northeast	Southeast	Central	West Coast	Total
Number who prefer present method	68	75	57	79	279
Number who prefer new method	32	45	33	31	141
Total employees sampled in each region	100	120	90	110	420

Contingency Tables

Suppose that in four regions, the National Health Care Company samples its hospital employees' attitudes toward job-performance reviews. Respondents are given a choice between the present method (two reviews a year) and a proposed new method (quarterly reviews). Table 11-1, which illustrates the response to this question from the sample polled, is called a *contingency table*. A table such as this is made up of rows and columns; rows run horizontally, columns vertically. Notice that the four columns in Table 11-1 provide one basis of classification—geographical regions—and that the two rows classify the information another way: preference for review methods. Table 11-1 is called a 2×4 contingency table because it consists of two rows and four columns. We describe the dimensions of a contingency table by first stating the number of rows and then the number of columns. The “total” column and the “total” row are not counted as part of the dimensions.

Describing a contingency table

Observed and Expected Frequencies

Suppose we now symbolize the true proportions of the total population of employees who prefer the present plan as

Setting up the problem symbolically

- p_N ← Proportion in Northeast who prefer present plan
- p_S ← Proportion in Southeast who prefer present plan
- p_C ← Proportion in Central region who prefer present plan
- p_W ← Proportion in West Coast region who prefer present plan

Using these symbols, we can state the null and alternative hypotheses as follows:

$$H_0: p_N = p_S = p_C = p_W \leftarrow \text{Null hypothesis}$$

$$H_1: p_N, p_S, p_C, \text{ and } p_W \text{ are not all equal} \leftarrow \text{Alternative hypothesis}$$

If the null hypothesis is true, we can combine the data from the four samples and then estimate the proportion of the total workforce (the total population) that prefers the present review method:

$$\begin{aligned} &\text{Combined proportion who prefer} \\ &\text{present method assuming the null} \\ &\text{hypothesis of no difference is true} \\ &= \frac{68 + 75 + 57 + 79}{100 + 120 + 90 + 110} \\ &= \frac{279}{420} \\ &= 0.6643 \end{aligned}$$

TABLE 11-2 PROPORTION OF SAMPLED EMPLOYEES IN EACH REGION EXPECTED TO PREFER THE TWO REVIEW METHODS

	Northeast	Southeast	Central	West Coast
Total number sampled	100	120	90	110
Estimated proportion who prefer present method	$\times 0.6643$	$\times 0.6643$	$\times 0.6643$	$\times 0.6643$
Number <i>expected</i> to prefer present method	66.43	79.72	59.79	73.07
Total number sampled	100	120	90	110
Estimated proportion who prefer new method	$\times 0.3357$	$\times 0.3357$	$\times 0.3357$	$\times 0.3357$
Number <i>expected</i> to prefer new method	33.57	40.28	30.21	36.93

Obviously, if the value 0.6643 estimates the population proportion expected to prefer the present compensation method, then 0.3357 ($=1 - 0.6643$) is the estimate of the population proportion expected to prefer the proposed new method. Using 0.6643 as the *estimate* of the population proportion who prefer the present review method and 0.3357 as the *estimate* of the population proportion who prefer the new method, we can estimate the number of sampled employees in each region whom we would expect to prefer each of the review methods. The calculations are done in Table 11-2.

Determining expected frequencies

Table 11-3 combines all the information from Tables 11-1 and 11-2. It illustrates both the actual, or observed, frequency of the employees sampled who prefer each method of job-review and the theoretical, or expected, frequency of sampled employees preferring each method. Remember that the *expected frequencies*, those in color, were estimated from our combined proportion estimate.

Comparing expected and observed frequencies

To test the null hypothesis, $p_N = p_S = p_C = p_w$, we must compare the frequencies that were observed (the black ones in Table 11-3) with the frequencies we would expect if the null hypothesis is true. If the sets of observed and expected frequencies are nearly alike, we can reason intuitively that we will accept the null hypothesis. If there is a large difference between these frequencies, we may intuitively reject the null hypothesis and conclude that there are significant differences in the proportions of employees in the four regions preferring the new method.

Reasoning intuitively about chi-square tests

TABLE 11-3 COMPARISON OF OBSERVED AND EXPECTED FREQUENCIES OF SAMPLED EMPLOYEES

	Northeast	Southeast	Central	West Coast
FREQUENCY PREFERRING PRESENT METHOD:				
Observed (actual) frequency	68	75	57	79
Expected (theoretical) frequency	66.43	79.72	59.79	73.07
FREQUENCY PREFERRING NEW METHOD:				
Observed (actual) frequency	32	45	33	31
Expected (theoretical) frequency	33.57	40.28	30.21	36.93

The Chi-Square Statistic

To go beyond our intuitive feelings about the observed and expected frequencies, we can use the chi-square statistic, which is calculated this way:

Calculating the chi-square statistic

Chi-Square Statistic	
$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	[11-1]

An observed frequency
An expected frequency
Chi-square
Symbol meaning “the sum of”

This formula says that chi-square, or χ^2 , is the sum we will get if we

1. Subtract f_e from f_o for each of the eight values in Table 11-3.
2. Square each of the differences.
3. Divide each squared difference by f_e .
4. Sum all eight of the answers.

Numerically, the calculations are easy to do using a table such as Table 11-4, which shows the steps.

Interpreting the chi-square statistic

The answer of 2.764 is the value for chi-square in our problem comparing preferences for review methods. If this value were as large as, say, 20, it would indicate a substantial difference between our observed values and our

TABLE 11-4 CALCULATION OF χ^2 (CHI-SQUARE) STATISTIC FROM DATA IN TABLE 11-3

<i>f_o</i>	<i>f_e</i>	Step 1 <i>f_o – f_e</i>	Step 2 <i>(f_o – f_e)²</i>	Step 3 $\frac{(f_o - f_e)^2}{f_e}$
68	66.43	1.57	2.46	0.0370
75	79.72	-4.72	22.28	0.2795
57	59.79	-2.79	7.78	0.1301
79	73.07	5.93	35.16	0.4812
32	33.57	-1.57	2.46	0.0733
45	40.28	4.72	22.28	0.5531
33	30.21	2.79	7.78	0.2575
31	36.93	-5.93	35.16	0.9521
				2.7638

Step 4 $\sum \frac{(f_o - f_e)^2}{f_e} = 2.764 \leftarrow \chi^2$ (chi-square)

expected values. A chi-square of zero, on the other hand, indicates that the observed frequencies exactly match the expected frequencies. The value of chi-square can never be negative because the differences between the observed and expected frequencies are always *squared*.

The Chi-Square Distribution

If the null hypothesis is true, then the sampling distribution of the chi-square statistic, χ^2 , can be closely approximated by a continuous curve known as a *chi-square distribution*. As in the case of the *t* distribution, there is a different chi-square distribution for each different number of degrees of freedom. Figure 11-1 shows the three different chi-square distributions that would correspond to 1, 5, and 10 degrees of freedom. For very small numbers of degrees of freedom, the chi-square distribution is severely skewed to the right. As the number of degrees of freedom increases, the curve rapidly becomes more symmetrical until the number reaches large values, at which point the distribution can be approximated by the normal.

The chi-square distribution is a probability distribution. Therefore, the total area under the curve in each chi-square distribution is 1.0. Like the *t* distribution, so many different chi-square distributions are possible that it is not practical to construct a table that illustrates the areas under the curve for all possible values of the area. Appendix Table 5 illustrates only the areas in the tail most commonly used in significance tests using the chi-square distribution.

Determining Degrees of Freedom

To use the chi-square test, we must calculate the number of degrees of freedom in the contingency table by applying Equation 11-2:

Describing a chi-square distribution

Finding probabilities when using a chi-square distribution

Calculating degrees of freedom

Degrees of Freedom in a Chi-Square Test of Independence

$$\text{Number of degrees of freedom} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

[11-2]

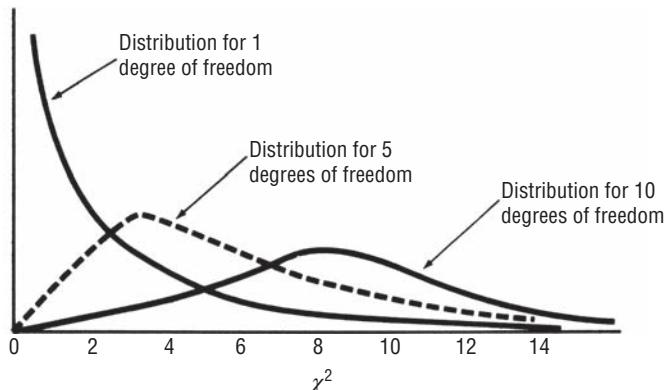


FIGURE 11-1 CHI-SQUARE DISTRIBUTIONS WITH 1, 5, AND 10 DEGREES OF FREEDOM

	Column 1	Column 2	Column 3	Column 4	
Row 1	✓	✓	✓	○	RT_1
Row 2	✓	✓	✓	○	RT_2
Row 3	○	○	○	*	RT_3
	CT_1	CT_2	CT_3	CT_4	
Column totals					

✓ Values that can be freely specified
○ Values that cannot be freely specified
*

Row totals

FIGURE 11-2 A 3×4 CONTINGENCY TABLE ILLUSTRATING DETERMINATION OF THE NUMBER OF DEGREES OF FREEDOM

Let's examine the appropriateness of this equation. Suppose we have a 3×4 contingency table like the one in Figure 11-2. We know the row and column totals that are designated RT_1 , RT_2 , RT_3 , and CT_1 , CT_2 , CT_3 , CT_4 . As we discussed in Chapter 7, the number of degrees of freedom is equal to the number of values that we can freely specify.

Look now at the first row of the contingency table in Figure 11-2. Once we specify the first three values in that row (denoted by checks in the figure), the fourth value in that row (denoted by a circle) is already determined; we are not free to specify it because we know the row total.

Likewise, in the second row of the contingency table in Figure 11-2, once we specify the first three values (denoted again by checks), the fourth value is determined and cannot be freely specified. We have denoted this fourth value by a circle.

Turning now to the third row, we see that its first entry is determined because we already know the first two entries in the first column and the column total; again, we have denoted this entry with a circle. We can apply this same reasoning to the second and third entries in the third row, both of which have been denoted by circles, too.

Turning finally to the last entry in the third row (denoted by a star), we see that we can not freely specify its value because we have already determined the first two entries in the fourth column. By counting the number of checks in the contingency table in Figure 11-2, you can see that the number of values we are free to specify is 6 (the number of checks). This is equal to 2×3 , or (the number of rows – 1) times (the number of columns – 1).

This is exactly what we have in Equation 11-2. Table 11-5 illustrates the row-and-column dimensions of three more contingency tables and indicates the appropriate degrees of freedom in each case.

TABLE 11-5 DETERMINATION OF DEGREES OF FREEDOM IN THREE CONTINGENCY TABLES

Contingency Table	Number of Rows (r)	Number of Columns (c)	$r - 1$	$c - 1$	Degrees of Freedom $(r - 1)(c - 1)$
A	3	4	$3 - 1 = 2$	$4 - 1 = 3$	$(2)(3) = 6$
B	5	7	$5 - 1 = 4$	$7 - 1 = 6$	$(4)(6) = 24$
C	6	9	$6 - 1 = 5$	$9 - 1 = 8$	$(5)(8) = 40$

Using the Chi-Square Test

Returning to our example of job-review preferences of National Health Care hospital employees, we use the chi-square test to determine whether attitude about reviews is independent of geographical region. If the company wants to test the null hypothesis at the 0.10 level of significance, our problem can be summarized:

$$H_0: p_N = p_S = p_C = p_W \leftarrow \text{Null hypothesis}$$

$$H_1: p_N, p_S, p_C \text{ and } p_W \text{ are not equal} \leftarrow \text{Alternative hypothesis}$$

$$\alpha = 0.10 \leftarrow \text{Level of significance for testing these hypotheses}$$

Stating the problem symbolically

Because our contingency table for this problem (Table 11-1) has two rows and four columns, the appropriate number of degrees of freedom is

Calculating degrees of freedom

Number of Degrees of Freedom

$$(r - 1)(c - 1)$$

Number of rows Number of columns

[11-2]

$$= (2 - 1)(4 - 1)$$

$$= (1)(3)$$

$$= 3 \leftarrow \text{Degrees of freedom}$$

Figure 11-3 illustrates a chi-square distribution with 3 degrees of freedom, showing the significance level in color. In Appendix Table 5, we can look under the 0.10 column and move down to the 3 degrees of freedom row. There we find the value of the chi-square statistic, 6.251. We can interpret this to mean that with 3 degrees of freedom, the region to the right of a chi-square value of 6.251 contains 0.10 of the area under the curve. Thus, the acceptance region for the null hypothesis in Figure 11-3 goes from the left tail of the curve to the chi-square value of 6.251.

Illustrating the hypothesis test

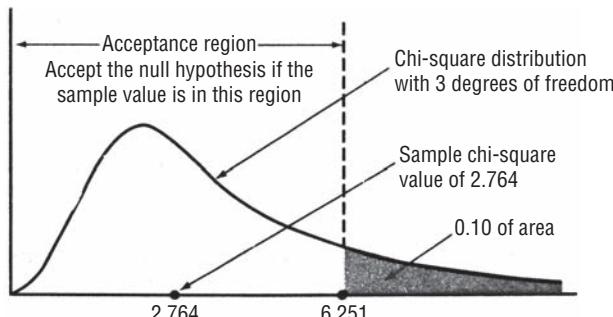


FIGURE 11-3 CHI-SQUARE HYPOTHESIS TEST AT THE 0.10 LEVEL OF SIGNIFICANCE, SHOWING ACCEPTANCE REGION AND SAMPLE CHI-SQUARE VALUE OF 2.764

As we can see from Figure 11-3, the sample chi-square value of 2.764 that we calculated in Table 11-4 falls within the acceptance region. Therefore, we accept the null hypothesis that there is no difference between the attitudes about job interviews in the four geographical regions. In other words, we conclude that attitude about performance reviews is independent of geography.

Interpreting the results

Contingency Tables with More Than Two Rows

Mr. George McMahon, president of National General Health Insurance Company, is opposed to national health insurance. He argues that it would be too costly to implement, particularly since the existence of such a system would, among other effects, tend to encourage people to spend more time in hospitals. George believes that lengths of stays in hospitals are dependent on the types of health insurance that people have. He asked Donna McCloud, his staff statistician, to check the matter. Donna collected data on a random sample of 660 hospital stays and summarized them in Table 11-6.

Are hospital stay and insurance coverage independent?

Table 11-6 gives observed frequencies in the nine different length-of-stay and type-of-insurance categories (or “cells”) into which we have divided the sample. Donna wishes to test the hypotheses:

$$H_0: \text{length of stay and type of insurance are independent}$$

Stating the hypotheses

$$H_1: \text{length of stay depends on type of insurance}$$

$$\alpha = 0.01 \leftarrow \text{Level of significance for testing these hypotheses}$$

We will use a chi-square test, so we first have to find the expected frequencies for each of the nine cells. Let's demonstrate how to find them by looking at the cell that corresponds to stays of less than 5 days and insurance covering less than 25 percent of costs.

Finding expected frequencies

A total of 180 of the 660 stays in Table 11-6 had insurance covering less than 25 percent of costs. So we can use the figure 180/660 to *estimate* the proportion in the population having insurance covering less than 25 percent of the costs. Similarly, 110/660 *estimates* the proportion of all hospital stays that last fewer than 5 days. If length of stay and type of insurance really are independent, we can use Equation 4-4 to *estimate* the proportion in the first cell (less than 5 days and less than 25 percent coverage).

Estimating the proportions in the cells

TABLE 11-6 HOSPITAL-STAY DATA CLASSIFIED BY THE TYPE OF INSURANCE COVERAGE AND LENGTH OF STAY

		Days in Hospital			
		<5	5-10	>10	Total
Fraction of costs covered by insurance	<25%	40	75	65	180
	25–50%	30	45	75	150
	>50%	40	100	190	330
Total		110	220	330	660

We let

- A = the event “a stay corresponds to someone whose insurance covers less than 25 percent of the costs”
- B = the event “a stay lasts less than 5 days”

Then,

$$\begin{aligned} P(\text{first cell}) &= P(A \text{ and } B) & [4-4] \\ &= P(A) \times P(B) \\ &= \left(\frac{180}{660} \right) \left(\frac{110}{660} \right) \\ &= 1/22 \end{aligned}$$

Because 1/22 is the expected *proportion* in the first cell, the expected *frequency* in that cell is

$$(1/22)(660) = 30 \text{ observations}$$

In general, we can calculate the expected frequency for any cell with Equation 11-3:

Calculating the expected frequencies for the cells

Expected Frequency For Any Cell

$$f_e = \frac{RT \times CT}{n} \quad [11-3]$$

where

- f_e = expected frequency in a given cell
- RT = row total for the row containing that cell
- CT = column total for the column containing that cell
- n = total number of observations

Now we can use Equations 11-3 and 11-1 to compute all of the expected frequencies and the value of the chi-square statistic. The computations are done in Table 11-7.

Figure 11-4 illustrates a chi-square distribution with 4 degrees of freedom (number of rows – 1 = 2 × (number of columns – 1 = 2), showing the 0.01 significance level in color. Appendix Table 5 (in the 0.01 column and the 4 degrees of freedom row) tells Donna that for her problem, the region to the right of a chi-square value of 13.277 contains 0.01 of the area under the curve. Thus, the acceptance region for the null hypothesis in Figure 11-4 goes from the left tail of the curve to the chi-square value of 13.277.

As Figure 11-4 shows Donna, the sample chi-square value of 24.315 she calculated in Table 11-7 is not within the acceptance region. Thus, Donna must reject the null hypothesis and inform Mr. McMahan that the evidence supports his belief that length of hospital stay and insurance coverage are dependent on each other.

Interpreting the results of the test

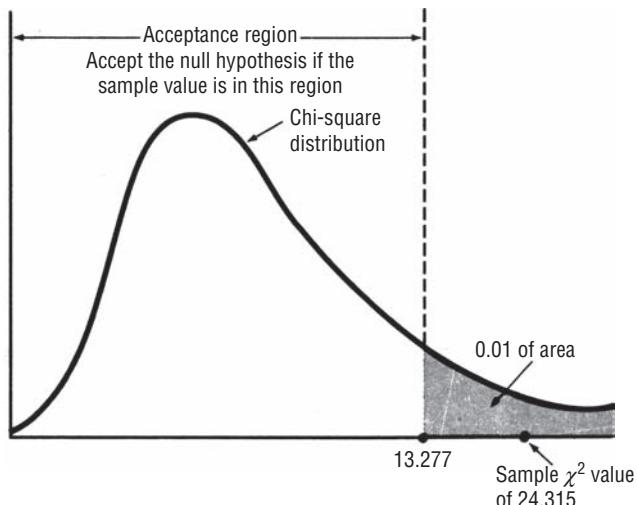
Precautions about Using the Chi-Square Test

To use a chi-square hypothesis test, we must have a sample size large enough to guarantee the similarity between the theoretically correct distribution and our sampling distribution of χ^2 , the chi-square statistic. When the expected

Use large sample sizes

TABLE 11-7 CALCULATION OF EXPECTED FREQUENCIES AND CHI-SQUARE FROM DATA IN TABLE 11-6

Row	Column	f_o	f_e	=	$\frac{RT \times CT}{n}$	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
1	1	40	30		$\frac{180 \times 110}{660}$	10	100	3.333
1	2	75	60		$\frac{180 \times 220}{660}$	15	225	3.750
1	3	65	90		$\frac{180 \times 330}{660}$	-25	625	6.944
2	1	30	25		$\frac{150 \times 110}{660}$	5	25	1.000
2	2	45	50		$\frac{150 \times 220}{660}$	-5	25	0.500
2	3	75	75		$\frac{150 \times 330}{660}$	0	0	0.000
3	1	40	55		$\frac{330 \times 110}{660}$	-15	225	4.091
3	2	100	110		$\frac{330 \times 220}{660}$	-10	100	0.909
3	3	190	165		$\frac{330 \times 330}{660}$	25	625	3.788
$[11-1] \sum \frac{(f_o - f_e)^2}{f_e} = 24.315 \leftarrow \chi^2 \text{ (chi-square)}$								

**FIGURE 11-4** CHI-SQUARE HYPOTHESIS TEST AT THE 0.01 LEVEL OF SIGNIFICANCE, SHOWING ACCEPTANCE REGION AND SAMPLE CHI-SQUARE VALUE OF 24.315

frequencies are too small, the value of χ^2 will be overestimated and will result in too many rejections of the null hypothesis. **To avoid making incorrect inferences from χ^2 hypothesis tests, follow the general rule that an expected frequency of less than 5 in one cell of a contingency table is too small to use.*** When the table contains more than one cell with an expected frequency of less than 5, we can combine these in order to get an expected frequency of 5 or more. But in doing this, we reduce the number of categories of data and will gain less information from the contingency table.

This rule will enable us to use the chi-square hypothesis test properly, but unfortunately, each test can only reflect (and not improve) the quality of the data we feed into it. So far, we have rejected the

Use carefully collected data

null hypothesis if the difference between the observed and expected frequencies—that is, the computed chi-square value—is too large. In the case of the job-review preferences, we would reject the null hypothesis at a 0.10 level of significance if our chi-square value was 6.251 or more. **But if the chi-square value was zero, we should be careful to question whether absolutely no difference exists between observed and expected frequencies.** If we have strong feelings that some difference *ought* to exist, we should examine either the way the data were collected or the manner in which measurements were taken, or both, to be certain that existing differences were not obscured or missed in collecting sample data.

In the 1860s, experiments with the characteristics of peas led the monk Gregor Mendel to propose the existence of genes. Mendel's experimental results were astoundingly close to those predicted by his theory. A century later, statisticians looked at Mendel's "pea data," performed a chi-square test, and concluded that chi-square was too small; that is, Mendel's reported experimental data were so close to what was expected that they could only conclude that he had fudged the data.

Mendel's pea data

Chi-Square Test Using SPSS

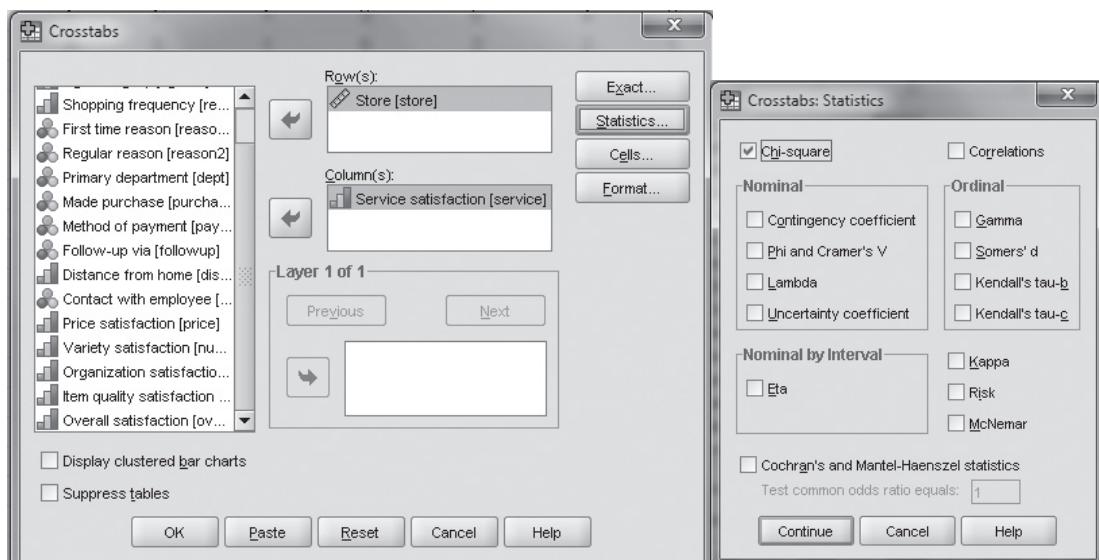
	gender	agecat	regular	reason1	reason2	dept	purchase	payment	followup	distance	store	contact	price	numitems	org	service	c
1	0	3	1	.	2	6	1	3	1	2	2	1	4	5	1	2	
2	0	1	2	.	2	7	1	5	2	2	4	1	4	5	4	1	
3	1	3	1	.	1	1	1	3	5	4	3	0	4	2	2	2	
4	0	3	1	.	2	3	1	5	3	2	1	0	3	2	2	1	
5	1	4	2	.	2	3	1	2	5	2	1	1	2	1	3	2	
6	0	3	2	.	2	7	1	2	3	2	1	1	2	4	4	2	
7	1	4	3	.	1	3	0	4	1	2	3	1	2	3	3	3	
8	0	3	3	.	1	1	1	5	1	3	3	1	5	5	5	4	
9	1	4	3	.	1	7	0	3	5	2	1	1	3	2	3	4	
10	0	3	1	.	2	6	1	4	5	1	1	1	3	4	4	5	
11	0	3	0	2	.	1	1	4	5	4	4	1	3	5	5	5	
12	0	4	2	.	2	6	1	4	1	1	2	1	5	5	3	5	
13	1	4	1	.	2	7	1	3	5	3	1	1	1	2	3	1	
14	0	3	2	.	3	7	1	4	1	4	2	1	4	1	3	1	
15	1	3	1	.	3	4	1	3	5	5	4	1	2	1	3	2	
16	0	3	3	.	1	3	1	4	5	2	2	1	2	1	1	1	
17	1	3	4	.	2	7	0	5	2	4	4	0	5	5	1	5	
18	1	2	2	.	4	4	1	3	5	2	4	1	5	5	5	5	
19	1	4	1	.	2	4	0	2	5	5	4	0	4	3	5	5	
20	1	2	3	.	4	7	1	3	4	1	3	1	3	2	4	2	
21	0	3	2	.	2	4	1	3	1	1	3	1	4	3	4	5	
22	1	2	1	.	2	5	1	1	1	5	4	0	1	1	2	3	
23	1	5	2	.	2	5	0	3	5	4	2	1	5	5	5	4	
24	1	5	2	.	1	3	1	3	5	3	3	0	5	4	4	4	
25	1	4	2	.	1	3	0	1	5	2	2	1	1	2	2	1	
26	1	4	2	.	1	4	0	3	5	2	2	0	4	3	1	5	

The above data are used for chi-square test.

*Statisticians have developed correction factors that, in some cases, allow us to use cells with expected frequencies of less than 5. The derivation and use of these correction factors are beyond the scope of this book.

In order to determine customer satisfaction rates, a retail company conducted surveys of 582 customers at 4 store locations. From the survey results, you found that the quality of customer service was the most important factor to a customer's overall satisfaction. Given this information, you want to test whether each of the store locations provides a similar and adequate level of customer service.

For χ^2 test go to **Analyze > Descriptive statistics > Crosstabs > Select rows and columns > Select Chi-square > OK**.



		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Store	* Service satisfaction	582	100.0%	0	0%	582	100.0%

Store * Service satisfaction Crosstabulation						
Count		Service satisfaction				
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive
Store	Store 1	25	20	38	30	33
	Store 2	26	30	34	27	19
	Store 3	15	20	41	33	29
	Store 4	27	35	44	22	34
Total		93	105	157	112	115
						582

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.293*	12	.178
Likelihood Ratio	17.012	12	.149
Linear-by-Linear Association	.084	1	.772
N Valid Cases	582		

a. 0 cells (0%) have expected count less than 5. The minimum expected count is 21.73.

SPSS Processor is ready

HINTS & ASSUMPTIONS

Warning: The rows and columns of a chi-square contingency table *must* be mutually exclusive categories that exhaust *all* of the possibilities of the sample. Hint: Think of the cells as little boxes and each member of the sample as a marble. Each marble must be put in a box and there can be no leftover marbles if you want the test to be valid. For example, a survey of voters that has contingency table cells for just Democrats and Republicans ignores the opinions of unaffiliated voters. Hint: The categories “car owner” and “bicycle owner” don’t allow for people who own both.

EXERCISES 11.2

Self-Check Exercises

- SC 11-1** A brand manager is concerned that her brand’s share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographic regions, a random sampling of 100 consumers in each region was surveyed, with the following results:

	REGION				
	NE	NW	SE	SW	TOTAL
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	100	100	100	100	400

Develop a table of observed and expected frequencies for this problem.

- SC 11-2** For Exercise SC 11-1:

- Calculate the sample χ^2 value.
- State the null and alternative hypotheses.
- At $\alpha = 0.05$, test whether brand share is the same across the four regions.

Basic Concepts

- 11-6** Given the following dimensions for contingency tables, how many degrees of freedom will the chi-square statistic for each have?
- 5 rows, 4 columns.
 - 6 rows, 2 columns.
 - 3 rows, 7 columns.
 - 4 rows, 4 columns.

Applications

- 11-7** An advertising firm is trying to determine the demographics for a new product. They have randomly selected 75 people in each of 5 different age groups and introduced the product to them. The results of the survey are given in the following table:

Future Activity	Age Group				
	18–29	30–39	40–49	50–59	60–69
Purchase frequently	12	18	17	22	32
Seldom purchase	18	25	29	24	30
Never purchase	45	32	29	29	13

Develop a table of observed and expected frequencies for this problem.

11-8

For Exercise 11-7:

- (a) Calculate the sample χ^2 value.
- (b) State the null and alternative hypotheses.
- (c) If the level of significance is 0.01, should the null hypothesis be rejected?

11-9

To see whether silicon chip sales are independent of where the U.S. economy is in the business cycle, data have been collected on the weekly sales of Zippy Chippy, a Silicon Valley firm, and on whether the U.S. economy was rising to a cycle peak, at a cycle peak, falling to a cycle trough, or at a cycle trough. The results are:

Economy	WEEKLY CHIP SALES			Total
	High	Medium	Low	
At Peak	20	7	3	30
At Trough	30	40	30	100
Rising	20	8	2	30
Falling	30	5	5	40
Total	100	60	40	200

Calculate a table of observed and expected frequencies for this problem.

11-10

For Exercise 11-9:

- (a) State the null and alternative hypotheses.
- (b) Calculate the sample χ^2 value.
- (c) At the 0.10 significance level, what is your conclusion?

11-11

A financial consultant is interested in the differences in capital structure within different firm sizes in a certain industry. The consultant surveys a group of firms with assets of different amounts and divides the firms into three groups. Each firm is classified according to whether its total debt is greater than stockholders' equity or whether its total debt is less than stockholders' equity. The results of the survey are:

	Firm Asset Size (in \$ thousands)			Total
	<500	500–2,000	2,000+	
Debt less than equity	7	10	8	25
Debt greater than equity	10	18	9	37
Total	17	28	17	62

Do the three firm sizes have the same capital structure? Use the 0.10 significance level.

- 11-12** A newspaper publisher, trying to pinpoint his market's characteristics, wondered whether newspaper readership in the community is related to readers' educational achievement. A survey questioned adults in the area on their level of education and their frequency of readership. The results are shown in the following table.

Frequency of Readership	Level of Educational Achievement				Total
	Professional or Postgraduate	College Graduate	High School Grad	Did Not Complete High School	
Never	10	17	11	21	59
Sometimes	12	23	8	5	48
Morning or evening	35	38	16	7	96
Both editions	28	19	6	13	66
Total	85	97	41	46	269

At the 0.10 significance level, does the frequency of newspaper readership in the community differ according to the readers' level of education?

- 11-13** An educator has the opinion that the grades high school students make depend on the amount of time they spend listening to music. To test this theory, he has randomly given 400 students a questionnaire. Within the questionnaire are the two questions: "How many hours per week do you listen to music?" "What is the average grade for all your classes?" The data from the survey are in the following table. Using a 5 percent significance level, test whether grades and time spent listening to music are independent or dependent.

Hours Spent Listening to Music	Average Grade					TOTAL
	A	B	C	D	F	
<5 hrs.	13	10	11	16	5	55
5–10 hrs.	20	27	27	19	2	95
11–20 hrs.	9	27	71	16	32	155
>20 hrs.	8	11	41	24	11	95
Total	50	75	150	75	50	400

Worked-Out Answers to Self-Check Exercises

SC 11-1

	Region			
	NE	NW	SE	SW
Purchasers				
Observed	40	55	45	50
Expected	47.5	47.5	47.5	47.5
Nonpurchasers				
Observed	60	45	55	50
Expected	52.5	52.5	52.5	52.5

SC 11-2 (a)

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
40	47.5	-7.5	56.25	1.184
55	47.5	7.5	56.25	1.184
45	47.5	-2.5	6.25	0.132
50	47.5	2.5	6.25	0.132
60	52.5	7.5	56.25	1.071
45	52.5	-7.5	56.25	1.071
55	52.5	2.5	6.25	0.119
50	52.5	-2.5	6.25	0.119
				$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 5.012$

(b) Two ways, either acceptable:

(1) H_0 : Region is independent of purchasing

H_1 : Region is related to purchasing (dependent)

(2) H_0 : $p_{ne} = p_{nw} = p_{se} = p_{sw}$

H_1 : Not all the proportions are equal

(c) With $1 \times 3 = 3$ degrees of freedom and $\alpha = 0.05$, the critical value of χ^2 is 7.815, so don't reject H_0 , because $5.012 < 7.815$. Brand share doesn't differ significantly by region.

11.3 CHI-SQUARE AS A TEST OF GOODNESS OF FIT: TESTING THE APPROPRIATENESS OF A DISTRIBUTION

In the preceding section, we used the chi-square test to decide whether to accept a null hypothesis that was a hypothesis of independence between two variables. In our example, these two variables were attitude toward job performance reviews and geographical region.

The chi-square test can also be used to decide whether a particular probability distribution, such as the binomial, Poisson, or normal, is the *appropriate* distribution. This is an important ability because as decision makers using statistics, we will need to choose a certain probability distribution to represent the distribution of the data we happen to be considering. We will need the ability to question how far we can go from the assumptions that underlie a particular distribution before we must conclude that this distribution is no longer applicable. **The chi-square test enables us to ask this question and to test whether there is a significant difference between an observed frequency distribution and a theoretical frequency distribution.** In this manner, we can determine the *goodness of fit* of a theoretical distribution (that is, how well it fits the distribution of data that we have actually observed). Thus, we can determine whether we should believe that the observed data constitute a sample drawn from the hypothesized theoretical distribution.

Function of a goodness-of-fit test

Calculating Observed and Expected Frequencies

Suppose that the Gordon Company requires that college seniors who are seeking positions with it be interviewed by three different executives. This enables the company to obtain a consensus evaluation

TABLE 11-8 INTERVIEW RESULTS OF 100 CANDIDATES

Possible Positive Ratings from Three Interviews	Number of Candidates Receiving Each of These Ratings
0	18
1	47
2	24
3	11
	100

of each candidate. Each executive gives the candidate either a positive or a negative rating. Table 11-8 contains the interview results of the last 100 candidates.

For staffing purposes, the director of recruitment for this company thinks that the interview process can be approximated by a binomial distribution with $p = 0.40$, that is, with a 40 percent chance of any candidate receiving a positive rating on any one interview. If the director wants to test this hypothesis at the 0.20 level of significance, how should he proceed?

H_0 : A binomial distribution with $p = 0.40$ is a good description of the interview process ← Null hypothesis

Stating the problem symbolically

H_1 : A binomial distribution with $p = 0.40$ is *not* a good description of the interview process ← Alternative hypothesis

$\alpha = 0.20$ ← Level of significance for testing these hypotheses

Calculating the binomial probabilities

To solve this problem, we must determine whether the discrepancies between the observed frequencies and those we would expect (if the binomial distribution *is* the proper model to use) should be ascribed to chance. We can begin by determining what the binomial probabilities would be for this interview situation. For three interviews, we would find the probability of success in the Binomial Distribution Table (Appendix Table 3) by looking for the column labeled $n = 3$ and $p = 0.40$. The results are summarized in Table 11-9.

Now we can use the theoretical binomial probabilities of the outcomes to compute the expected frequencies. By comparing these expected frequencies with our observed frequencies using the χ^2 test,

TABLE 11-10 OBSERVED FREQUENCIES, APPROPRIATE BINOMIAL PROBABILITIES, AND EXPECTED FREQUENCIES FOR THE INTERVIEW PROBLEM

Possible Positive Ratings from Three Interviews	Observed Frequency of Candidates Receiving These Ratings	Binomial Probability of Possible Outcomes	Number of Candidates Interviewed	Expected Frequency of Candidates Receiving These Ratings
0	18	0.2160	×	100
1	47	0.4320	×	100
2	24	0.2880	×	100
3	11	0.0640	×	100
	100	1.0000		100.0

TABLE 11-11 CALCULATION OF THE χ^2 STATISTIC FROM THE INTERVIEW DATA LISTED IN TABLE 11-10

Observed Frequency (f_o)	Expected Frequency (f_e)	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
18	21.6	-3.6	12.96	0.6000
47	43.2	3.8	14.44	0.3343
24	28.8	-4.8	23.04	0.8000
11	6.4	4.6	21.16	3.3063
				5.0406
$\sum \frac{(f_o - f_e)^2}{f_e} = 5.0406 \leftarrow \chi^2$				

we can examine the extent of the difference between them. Table 11-10 lists the observed frequencies, the appropriate binomial probabilities from Table 11-9, and the expected frequencies for the sample of 100 interviews.

Calculating the Chi-Square Statistic

To compute the chi-square statistic for this problem, we can use Equation 11-1:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad [11-1]$$

and the format we introduced in Table 11-4. This process is illustrated in Table 11-11.

Determining Degrees of Freedom in a Goodness-of-Fit Test

Before we can calculate the appropriate number of degrees of freedom for a chi-square goodness-of-fit test, we must count the number of classes (symbolized k) for which we have compared the observed and expected frequencies. Our interview problem contains four such classes: 0, 1, 2, and 3 positive ratings. Thus, we begin with 4 degrees of freedom. Yet because the four observed frequencies must sum to 100, the total number of observed frequencies we can freely specify is only $k - 1$, or 3. The fourth is determined because the total of the four has to be 100.

First, count the number of classes

To solve a goodness-of-fit problem, we may be forced to impose additional restrictions on the calculation of the degrees of freedom. Suppose we are using the chi-square test as a goodness-of-fit test to determine whether a normal distribution fits a set of observed frequencies. If we have six classes of observed frequencies ($k = 6$), then we would conclude that we have only $k - 1$, or 5 degrees of freedom. If, however, we also have to use the sample mean as an estimate of the population mean, we will have to subtract an additional degree of freedom, which leaves us with only 4. And, third, if we have to use the sample standard deviation to estimate the population standard deviation, we will have to subtract *one more* degree of freedom, leaving us with 3. Our general rule in these cases is, **first employ the ($k - 1$) rule and then subtract an additional degree of freedom for each population parameter that has to be estimated from the sample data.**

Then, subtract degrees of freedom lost from estimating population parameters

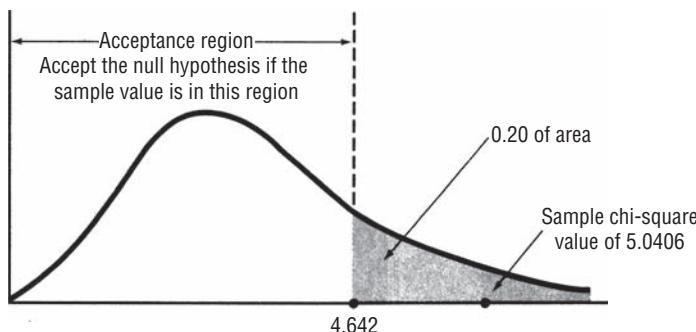


FIGURE 11-7 GOODNESS-OF-FIT TEST AT THE 0.20 LEVEL OF SIGNIFICANCE, SHOWING ACCEPTANCE REGION AND SAMPLE CHI-SQUARE VALUE OF 5.0406

In the interview example, we have four classes of observed frequencies. As a result, $k = 4$, and the appropriate number of degrees of freedom is $k - 1$, or 3. We are not required to estimate any population parameter, so we need not reduce this number further.

Using the Chi-Square Goodness-of-Fit Test

In the interview problem, the company desires to test the hypothesis of goodness of fit at the 0.20 level of significance. In Appendix Table 5, then, we must look under the 0.20 column and move down to the row labeled 3 degrees of freedom. There we find that the value of the chi-square statistic is 4.642. We can interpret this value as follows: With 3 degrees of freedom, the region to the right of a chi-square value of 4.642 contains 0.20 of the area under the curve.

Finding the limit of the acceptance region

Figure 11-7 illustrates a chi-square distribution with 3 degrees of freedom, showing in color a 0.20 level of significance. Notice that the acceptance region for the null hypothesis (the hypothesis that the sample data came from a binomial distribution with $p = 0.4$) extends from the left tail to the chi-square value of 4.642. Obviously, the sample chi-square value of 5.0406 falls outside this acceptance region. Therefore, we reject the null hypothesis and conclude that the binomial distribution with $p = 0.4$ fails to provide a good description of our observed frequencies.

Illustrating the problem

The chi-square distribution with 3 degrees of freedom has a total area of 1.00 under its curve. The acceptance region, which is the portion of the curve to the left of the critical value of 4.642, contains 0.80 of the area under the curve. The rejection region, which is the portion of the curve to the right of 4.642, contains 0.20 of the area under the curve.

Interpreting the results

HINTS & ASSUMPTIONS

Lots of folks know that a chi-square test can be used as a test of goodness of fit, and most of them can do the calculations. But fewer of them can explain the logic in using the test for this purpose in common-sense terms. Hint: If we have a distribution that we *think* may be normal, but we're not sure, we use a known normal distribution to generate the *expected values* and then using chi-square methods, we see how much difference there is between these expected values and the values we observed in a sample taken from the distribution we think is normal. If the difference is too large, our distribution isn't normal.

EXERCISES 11.3

Self-Check Exercises

- SC 11-3** At the 0.10 level of significance, can we conclude that the following 400 observations follow a Poisson distribution with $\lambda = 3$?

Number of arrivals per hour	0	1	2	3	4	5 or more
Number of hours	20	57	98	85	78	62

- SC 11-4** After years of working at a weighing station for trucks, Jeff Simpson feels that the weight per truck (in thousands of pounds) follows a normal distribution with $\mu = 71$ and $\sigma = 15$. In order to test this assumption, Jeff collected the following data one Monday, recording the weight of each truck that entered his station.

85	57	60	81	89	63	52	65	77	64
89	86	90	60	57	61	95	78	66	92
50	56	95	60	82	55	61	81	61	53
63	75	50	98	63	77	50	62	79	69
76	66	97	67	54	93	70	80	67	73

If Jeff used a chi-square goodness-of-fit test on these data, what would he conclude about the trucks' weight distribution? (Use a 0.10 significance level and be sure to state the hypotheses of interest.) (*Hint:* Use five equally probable intervals.)

Basic Concepts

- 11-14** Below is an observed frequency distribution. Using a normal distribution with $\mu = 5$ and $\sigma = 1.5$,
- Find the probability of falling in each class.
 - From part (a), compute the expected frequency of each category.
 - Calculate the chi-square statistic.
 - At the 0.10 level of significance, does this frequency distribution seem to be well described by the suggested normal distribution?

Observed value of the variable	<2.6	2.6–3.79	3.8–4.99	5–6.19	6.2–7.39	≥7.4
Observed frequency	6	30	41	52	12	9

- 11-15** At the 0.05 level of significance, can we conclude the following data follow a Poisson distribution with $\lambda = 5$?

Number of calls per minute	0	1	2	3	4	5	6	7 or more
Frequency of occurrences	4	15	42	60	89	94	52	80

Applications

- 11-16** Louis Armstrong, salesman for the Dillard Paper Company, has five accounts to visit per day. It is suggested that the variable, sales by Mr. Armstrong, may be described by the binomial distribution, with the probability of selling each account being 0.4. Given the following frequency distribution of Armstrong's number of sales per day, can we conclude that the data do in fact follow the suggested distribution? Use the 0.05 significance level.

Number of sales per day	0	1	2	3	4	5
Frequency of the number of sales	10	41	60	20	6	3

- 11-17** The computer coordinator for the business school believes the amount of time a graduate student spends reading and writing e-mail each weekday is normally distributed with mean $\mu = 14$ and standard deviation $\sigma = 5$. In order to examine this belief, the coordinator collected data one Wednesday, recording the amount of time in minutes each graduate student spent checking e-mail. Using a chi-square goodness-of-fit test on these data, what would you conclude about the distribution of e-mail times? (Use a 0.05 significance level and clearly state your hypotheses.) (*Hint:* Use five equally probable intervals.)

8.2	7.4	9.6	12.8	22.4	6.2	8.7	9.7	12.4	10.6
1.2	18.6	3.3	15.7	18.4	12.4	15.9	19.4	12.8	20.4
12.3	11.3	10.9	18.4	14.3	16.2	6.7	13.9	18.3	19.2
14.3	14.9	16.7	11.3	18.4	18.8	20.4	12.4	18.1	20.1

- 11-18** In order to plan how much cash to keep on hand in the vault, a bank is interested in seeing whether the average deposit of a customer is normally distributed. A newly hired employee hoping for a raise has collected the following information:

Deposit	\$0–\$999	\$1,000–\$1,999	\$2,000 or more
Observed frequency	20	65	25

- (a) Compute the expected frequencies if the data are normally distributed with mean \$1,500 and standard deviation \$600.
- (b) Compute the chi-square statistic.
- (c) State explicit null and alternative hypotheses.
- (d) Test your hypotheses at the 0.10 level and state an explicit conclusion.

- 11-19** The post office is interested in modeling the mangled-letter problem. It has been suggested that any letter sent to a certain area has a 0.15 chance of being mangled. Because the post office is so big, it can be assumed that two letters' chances of being mangled are independent. A sample of 310 people was selected and two test letters were mailed to each of them. The number of people receiving zero, one, or two mangled letters was 260, 40, and 10, respectively. At the 0.10 level of significance, is it reasonable to conclude that the number of mangled letters received by people follows a binomial distribution with $p = 0.15$?

- 11-20** A state lottery commission claims that for a new lottery game, there is a 10 percent chance of getting a \$1 prize, a 5 percent chance of \$100, and an 85 percent chance of getting nothing. To test whether this claim is correct, a winner from the last lottery went out and bought 1,000 tickets for the new lottery. He had 87 one-dollar prizes, 48 hundred-dollar prizes, and 865 worthless tickets. At the 0.05 significance level, is the state's claim reasonable?

- 11-21** Dennis Barry, a hospital administrator, has examined past records from 210 randomly selected 8-hour shifts to determine the frequency with which the hospital treats fractures. The numbers of days in which zero, one, two, three, four, or five or more patients with broken bones were treated were 25, 55, 65, 35, 20, and 10, respectively. At the 0.05 level of significance, can we reasonably believe that the incidence of broken-bone cases follows a Poisson distribution with $\lambda = 2$?

- 11-22** A large city fire department calculates that for any given precinct, during any given 8-hour shift, there is a 30 percent chance of receiving at least one fire alarm. Here is a random sampling of 60 days:

Number of shifts during which alarms were received	0	1	2	3
Number of days	16	27	11	6

At the 0.05 level of significance, do these fire alarms follow a binomial distribution? (*Hint:* Combine the last two groups so that all expected frequencies are greater than 5.)

- 11-23** A diligent statistics student wants to see whether it is reasonable to assume that some sales data have been sampled from a normal population before performing a hypothesis test on the mean sales. She collected some sales data, computed $\bar{x} = 78$ and $s = 9$, and tabulated the data as follows:

Sales level	≤ 65	66–70	71–75	76–80	81–85	≥ 86
Number of observations	10	20	40	50	40	40

- (a) Is it important for the statistics student to check whether the data are normally distributed? Explain.
- (b) State explicit null and alternative hypotheses for checking whether the data are normally distributed.
- (c) What is the probability (using a normal distribution with $\mu = 78$ and $\sigma = 9$) that sales will be less than or equal to 65.5, between 65.5 and 70.5, between 70.5 and 75.5, between 75.5 and 80.5; between 80.5 and 85.5, and greater than or equal to 85.5?
- (d) At the 0.05 level of significance, does the observed frequency distribution follow a normal distribution?

- 11-24** A supermarket manager is keeping track of the arrival of customers at checkout counters to see how many cashiers are needed to handle the flow. In a sample of 500 five-minute time periods, there were 22, 74, 115, 95, 94, 80, and 20 periods in which zero, one, two, three, four, five, or six or more customers, respectively, arrived at a checkout counter. Are these data consistent at the 0.05 level of significance with a Poisson distribution with $\lambda = 3$?

- 11-25** A professional baseball player, Lon Dakestraw, was at bat five times in each of 100 games. Lon claims that he has a probability of 0.4 of getting a hit each time he goes to bat. Test his claim at the 0.05 level by seeing whether the following data are distributed binomially ($p = 0.4$). (*Note:* Combine classes if the expected number of observations is less than 5).

Number of Hits per Game	Number of Games with That Number of Hits
0	12
1	38
2	27
3	17
4	5
5	1

Worked-Out Answers to Self-Check Exercises

SC 11-3 H_0 : Poisson with $\lambda = 3$

H_1 : Something else

Test at $\alpha = 0.10$, with $6 - 1 = 5$ degrees of freedom.

Arrivals/hour	0	1	2	3	4	5+
Poisson prob.	0.0498	0.1494	0.2240	0.2240	0.1680	0.1848
Observed	20	57	98	85	78	62
Expected	19.92	59.76	89.60	89.60	62.20	73.92
$\frac{(f_o - f_e)^2}{f_e}$	0.000	0.127	0.788	0.236	1.736	1.922
$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 4.809$						

With 5 degrees of freedom and $\alpha = 0.10$, the critical value of χ^2 is 9.236, so don't reject H_0 , because $4.809 < 9.236$. The data are well described by a Poisson distribution with $\lambda = 3$.

- SC 11-4** 5 equiprobable intervals; 0.2 probability for each interval, $50 \times 0.2 = 10$ trucks expected per interval.

z	$-\infty$	-0.84	-0.25	0.25	0.84	$+\infty$
$x = 71 + 15z$	$-\infty$	58.40	67.25	74.75	83.60	$+\infty$
Observed	10	16	3	10	11	
Expected	10	10	10	10	10	
$\frac{(f_o - f_e)^2}{f_e}$	0.0	3.6	4.9	0.0	0.1	
$\chi^2 = 8.6$						

H_0 : Truck weights are distributed normally with $\mu = 71$ and $\sigma = 15$

H_1 : The weights are distributed differently (either normal with a different μ and/or σ or a non-normal distribution)

With $5 - 1 = 4$ degrees of freedom and $\alpha = 0.10$, the critical value of χ^2 is 7.779, so reject H_0 , because $8.6 > 7.779$. The data are not well described by a normal distribution with $\mu = 71$ and $\sigma = 15$. Jeff is wrong.

11.4 ANALYSIS OF VARIANCE

Earlier in this chapter, we used the chi-square test to examine the differences among more than two sample proportions and to make inferences about whether such samples are drawn from populations each having the same proportion. In this section, we will learn a technique known as **analysis of variance** (often abbreviated ANOVA) that will enable us to test for the significance of the differences among more than two sample means. Using analysis of variance, we will be able to make inferences about whether our samples are drawn from populations having the same mean.

Function of analysis of variance

Analysis of variance is useful in such situations as comparing the mileage achieved by five different brands of gasoline, testing which of four different training methods produces the fastest learning

Situations where we can use ANOVA

TABLE 11-12 DAILY PRODUCTION OF 16 NEW EMPLOYEES

Method 1	Method 2	Method 3
—	—	18
15	22	24
18	27	19
19	18	16
22	21	22
11	17	15
85	105	114
$\div 5$	$\div 5$	$\div 6$
$17 = \bar{x}_1$	$21 = \bar{x}_2$	$19 = \bar{x}_3$ ← Sample means
$n_1 = 5$	$n_2 = 5$	$n_3 = 6$ ← Sample sizes

record, or comparing the first-year earnings of the graduates of half a dozen different business schools. In each of these cases, we would compare the means of *more* than two samples.

Statement of the Problem

In the training director's problem that opened this chapter, she wanted to evaluate three different training methods to determine whether there were any differences in effectiveness.

After completion of the training period, the company's statistical staff chose 16 new employees assigned at random to the three training methods.* Counting the production output by these 16 trainees, the staff has summarized the data and calculated the mean production of the trainees (see Table 11-12). Now if we wish to determine the *grand mean*, or \bar{x} (the mean for the entire group of 16 trainees), we can use one of two methods:

$$\begin{aligned} 1. \bar{x} &= \frac{15 + 18 + 19 + 22 + 11 + 22 + 27 + 18 + 21 + 17 + 18 + 24 + 19 + 16 + 22 + 15}{16} \\ &= \frac{304}{16} \\ &= 19 \leftarrow \text{Grand mean using all data} \end{aligned}$$

$$2. \bar{\bar{x}} = (5/16)(17) + (5/16)(21) + (6/16)(19)$$

$$\begin{aligned} &= \frac{304}{16} \\ &= 19 \leftarrow \text{Grand mean as a weighted average of the sample means, using the relative sample sizes as the weights} \end{aligned}$$

Calculating the grand mean

Statement of the Hypotheses

In this case, our reason for using analysis of variance is to decide whether these three samples (a *sample* is the small group of employees trained by any one method) were drawn from populations (a *population* is the total number of employees who could be trained by that method) having the same means. Because we are testing the effectiveness of the three training methods, we must determine whether the three

* Although in real practice, 16 trainees would not constitute an adequate statistical sample, we have limited the number here to be able to demonstrate the basic techniques of analysis of variance and to avoid tedious calculations.

samples, represented by the sample means $\bar{x}_1 = 17$, $\bar{x}_2 = 21$, and $\bar{x}_3 = 19$, could have been drawn from populations having the same mean, μ . A formal statement of the null and alternative hypotheses we wish to test would be

$$H_0: \mu_1 = \mu_2 = \mu_3 \leftarrow \text{Null hypothesis}$$

$$H_1: \mu_1, \mu_2, \text{ and } \mu_3 \text{ are not all equal} \leftarrow \text{Alternative hypothesis}$$

If we can conclude from our test that the sample means do not differ significantly, we can infer that the choice of training method does not influence the productivity of the employee. On the other hand, if we find differences among the sample means that are too large to attribute to chance sampling error, we can infer that the method used in training *does* influence the productivity of the employee. In that case, we would adjust our training program accordingly.

Stating the problem symbolically

Interpreting the results

Analysis of Variance: Basic Concepts

In order to use analysis of variance, we must assume that each of the samples is drawn from a normal population and that each of these populations has the same variance, σ^2 . However, if the sample sizes are large enough, we do not need the assumption of normality.

Assumptions made in analysis of variance

In our training-methods problems, our null hypothesis states that the three populations have the same mean. If this hypothesis is true, classifying the data into three columns in Table 11-12 is unnecessary and the entire set of 16 measurements of productivity can be thought of as a sample from one population. This overall population also has a variance of σ^2 .

Analysis of variance is based on a comparison of two different estimates of the variance, σ^2 , of our overall population. In this case, we can calculate one of these estimates by examining **the variance among the three sample means**, which are 17, 21, and 19. The other estimate of the population variance is determined by **the variation within the three samples** themselves, that is, (15, 18, 19, 22, 11), (22, 27, 18, 21, 17), and (18, 24, 19, 16, 22, 15). Then we compare these two estimates of the population variance. Because both are estimates of σ^2 , they should be approximately equal in value *when the null hypothesis is true*. If the null hypothesis is *not* true, these two estimates will differ considerably. The three steps in analysis of variance, then, are

1. Determine one estimate of the population variance from the variance *among the sample means*. **Steps in analysis of variance**
2. Determine a second estimate of the population variance from the variance *within the samples*.
3. Compare these two estimates. If they are approximately equal in value, *accept* the null hypothesis.

In the remainder of this section, we shall learn how to calculate these two estimates of the population variance, how to compare these two estimates, and how to perform a hypothesis test and interpret the results. As we learn how to do these computations, however, keep in mind that all are based on the above three steps.

Calculating the Variance among the Sample Means

Step 1 in analysis of variance indicates that we must obtain one estimate of the population variance from the variance among the three sample means. In statistical language, this estimate is called the *between-column variance*.

Finding the first estimate of the population variance

In Chapter 3, we used Equation 3-17 to calculate the sample variance:

$$\text{Sample variance} \longrightarrow s^2 = \frac{\sum(x - \bar{x})^2}{n-1} \quad [3-17]$$

Now, because we are working with three sample means and a grand mean, let's substitute \bar{x} for x , $\bar{\bar{x}}$ for \bar{x} , and k (the number of samples) for n to get a formula for the variance among the sample means:

First, find the variance among sample means

Variance among the Sample Means

$$s_{\bar{x}}^2 = \frac{\sum(\bar{x} - \bar{\bar{x}})^2}{k-1} \quad [11-4]$$

Next, we can return for a moment to Chapter 6, where we defined the standard error of the mean as the standard deviation of all possible samples of a given size. The formula to derive the standard error of the mean is Equation 6-1:

Then, find the population variance using the variance among sample means

$$\begin{array}{c} \text{Standard error of the mean} \\ (\text{standard deviation of all possible sample means from a given sample size}) \end{array} \longrightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \begin{array}{l} \text{Population standard deviation} \\ \text{Square root of the sample size} \end{array} \quad [6-1]$$

We can simplify this equation by cross-multiplying the terms and then squaring both sides in order to change the population standard deviation, σ , into the population variance, σ^2 :

Population Variance

$$\sigma^2 = \sigma_{\bar{x}}^2 \times n \quad [11-5]$$

↖ Standard error squared
(this is the variance among the sample means)

For our training-method problem, we do not have all the information we need to use this equation to find σ^2 . Specifically, we do not know $\sigma_{\bar{x}}^2$. We could, however, calculate the variance among the three sample means, $s_{\bar{x}}^2$, using Equation 11-4. So why not substitute $s_{\bar{x}}^2$ for $\sigma_{\bar{x}}^2$ in Equation 11-5 and calculate an estimate of the population variance? This will give us

$$\hat{\sigma}^2 = s_{\bar{x}}^2 \times n = \frac{\sum n(\bar{x} - \bar{\bar{x}})^2}{k-1}$$

There is a slight difficulty in using this equation as it stands. In Equation 6-1, n represents the sample size, but *which* sample size should we use when the different samples have different sizes? We solve this problem with Equation 11-6, where each $(\bar{x}_j - \bar{\bar{x}})^2$ is multiplied by its own appropriate n_j .

Which sample size to use

TABLE 11-13 CALCULATION OF THE BETWEEN-COLUMN VARIANCE

n	\bar{x}	$\bar{\bar{x}}$	$\bar{x} - \bar{\bar{x}}$	$(\bar{x} - \bar{\bar{x}})^2$	$n(\bar{x} - \bar{\bar{x}})^2$
5	17	19	$17 - 19 = -2$	$(-2)^2 = 4$	$5 \times 4 = 20$
5	21	19	$21 - 19 = 2$	$(2)^2 = 4$	$5 \times 4 = 20$
6	19	19	$19 - 19 = 0$	$(0)^2 = 0$	$6 \times 0 = 0$
					$\sum n_j (\bar{x}_j - \bar{\bar{x}})^2 = 40$

$$\hat{\sigma}_b^2 = \frac{\sum n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1} = \frac{40}{3-1} \quad [11-6]$$

$$= \frac{40}{2}$$

$$= 20 \text{ Between-column variance}$$

Estimate of Between-Column Variance

First estimate of the population variance $\longrightarrow \hat{\sigma}_b^2 = \frac{\sum n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$ [11-6]

where

- $\hat{\sigma}_b^2$ = our first estimate of the population variance based on the variance among the sample means (the *between-column variance*)
- n_j = size of the j th sample
- \bar{x}_j = sample mean of the j th sample
- $\bar{\bar{x}}$ = grand mean
- k = number of samples

Now we can use Equation 11-6 and the data from Table 11-12 to calculate the between column variance. Table 11-13 shows how to do these calculations.

Calculating the Variance within the Samples

Step 2 in ANOVA requires a second estimate of the population variance based on the variance within the samples. In statistical terms, this can be called the *within-column variance*. Our employee-training problem has three samples of five or six items each. We can calculate the variance within each of these three samples using Equation 3-17:

$$\text{Sample variance} \longrightarrow s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \quad [3-17]$$

Finding the second estimate
of the population variance

Because we have assumed that the variances of our three populations are the same, we could use any one of the three sample variances (s_1^2 or s_2^2 or s_3^2) as the second estimate of the population variance. Statistically, we can get a better estimate of the population variance by using a weighted average of all three sample variances. The general formula for this second estimate of σ^2 is

Estimate of Within-Column Variance

$$\text{Second estimate of the population variance} \rightarrow \hat{\sigma}_w^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) s_j^2 \quad [11-7]$$

where

- $\hat{\sigma}_w^2$ = our second estimate of the population variance based on the variances within the samples (the *within-column variance*)
- n_j = size of the j th sample
- s_j^2 = sample variance of the j th sample
- k = number of samples
- $n_T = \sum n_j$ = total sample size

This formula uses all the information we have at our disposal, not just a portion of it. Had there been seven samples instead of three, we would have taken a weighted average of all seven. The weights used in Equation 11-7 will be explained shortly. Table 11-14 illustrates how to calculate this second estimate of the population variance using the variances within all three of our samples.

Using all the information at our disposal

TABLE 11-14 CALCULATION OF VARIANCES WITHIN THE SAMPLES AND THE WITHIN-COLUMN VARIANCE

Training Method 1 Sample Mean: $\bar{x} = 17$		Training Method 2 Sample Mean: $\bar{x} = 21$		Training Method 3 Sample Mean: $\bar{x} = 19$	
$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$
$15 - 17 = -2$	$(-2)^2 = 4$	$22 - 21 = 1$	$(1)^2 = 1$	$18 - 19 = -1$	$(-1)^2 = 1$
$18 - 17 = 1$	$(1)^2 = 1$	$27 - 21 = 6$	$(6)^2 = 36$	$24 - 19 = 5$	$(5)^2 = 25$
$19 - 17 = 2$	$(2)^2 = 4$	$18 - 21 = -3$	$(-3)^2 = 9$	$19 - 19 = 0$	$(0)^2 = 0$
$22 - 17 = 5$	$(5)^2 = 25$	$21 - 21 = 0$	$(0)^2 = 0$	$16 - 19 = -3$	$(-3)^2 = 9$
$11 - 17 = -6$	$(-6)^2 = 36$	$17 - 21 = -4$	$(-4)^2 = 16$	$22 - 19 = 3$	$(3)^2 = 9$
$\Sigma(x - \bar{x})^2 = 70$		$\Sigma(x - \bar{x})^2 = 62$		$15 - 19 = -4$	$(-4)^2 = 16$
				$\Sigma(x - \bar{x})^2 = 60$	

$$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{70}{5-1} \\ = \frac{70}{4}$$

$$\text{Sample variance} \rightarrow s_1^2 = 17.5$$

$$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{62}{5-1} \\ = \frac{62}{4}$$

$$\text{Sample variance} \rightarrow s_2^2 = 15.5$$

$$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{60}{6-1} \\ = \frac{60}{5}$$

$$\text{Sample variance} \rightarrow s_3^2 = 12.0$$

And:

$$\hat{\sigma}^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) s_j^2 = (4/13)(17.5) + (4/13)(15.5) + (5/13)(12.0) \quad [11-7]$$

$$= \frac{192}{13} \\ = 14.769 \leftarrow$$

Second estimate of the population variance based on the variances within the samples (the within-column variance)

The *F* Hypothesis Test: Computing and Interpreting the *F* Statistic

Step 3 in ANOVA compares these two estimates of the population variance by computing their ratio, called *F*, as follows:

$$F = \frac{\begin{matrix} \text{first estimate of the population variance} \\ \text{based on the variance among the sample means} \end{matrix}}{\begin{matrix} \text{second estimate of the population variance} \\ \text{based on the variances within the samples} \end{matrix}} \quad [11-8]$$

If we substitute the statistical shorthand for the numerator and denominator of this ratio, Equation 11-8 becomes

F Statistic

$$F = \frac{\text{between-column variance}}{\text{within-column variance}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_w^2} \quad [11-9]$$

Now we can find the *F ratio* for the training-method problem with which we have been working:

$$\begin{aligned} F &= \frac{\text{between-column variance}}{\text{within-column variance}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_w^2} \\ &= \frac{20}{14.769} \\ &= 1.354 \leftarrow F \text{ ratio} \end{aligned} \quad [11-9]$$

Having found this *F* ratio of 1.354, how can we interpret it? *Interpreting the F ratio*
First, examine the denominator, which is based on the variance within the samples. The denominator is a good estimator of σ^2 (the population variance) whether the null hypothesis is true or not. What about the numerator? If the null hypothesis that the three methods of training have equal effects is true, then the numerator, or the variation among the sample means of the three methods, is also a good estimate of σ^2 (the population variance). As a result, **the denominator and numerator should be about equal if the null hypothesis is true**. The nearer the *F* ratio comes to 1, then the more we are inclined to accept the null hypothesis. Conversely, as the *F* ratio becomes larger, we will be more inclined to reject the null hypothesis and accept the alternative (that a difference does exist in the effects of the three training methods).

Shortly, we shall learn a more formal way of deciding when to accept or reject the null hypothesis. But even now, you should understand the basic logic behind the *F statistic*. **When populations are not the same, the between-column variance (which was derived from the variance among the sample means) tends to be larger than the within-column variance (which was derived from the variances within the samples), and the value of *F* tends to be large. This leads us to reject the null hypothesis.**

The *F* Distribution

Like other statistics we have studied, if the null hypothesis is true, then the *F* statistic has a particular sampling distribution. Like the

Describing an F distribution

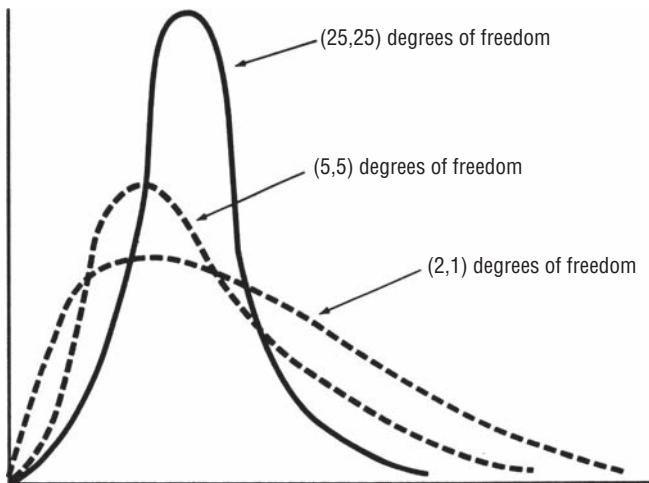


FIGURE 11-8 THREE F DISTRIBUTIONS (FIRST VALUE IN PARENTHESES EQUALS NUMBER OF DEGREES OF FREEDOM IN THE NUMERATOR OF THE F RATIO; SECOND EQUALS NUMBER OF DEGREES OF FREEDOM IN THE DENOMINATOR)

t and chi-square distributions, this *F* distribution is actually a whole family of distributions, three of which are shown in Figure 11-8. Notice that each is identified by a *pair* of degrees of freedom, unlike the *t* and chi-square distributions, which have only one value for the number of degrees of freedom. **The first number is the number of degrees of freedom in the numerator of the *F* ratio; the second is the degrees of freedom in the denominator.**

As we can see in Figure 11-8, the *F* distribution has a single mode. The specific shape of an *F* distribution depends on the number of degrees of freedom in both the numerator and the denominator of the *F* ratio. But, in general, the *F* distribution is skewed to the right and tends to become more symmetrical as the numbers of degrees of freedom in the numerator and denominator increase.

Using the *F* Distribution: Degrees of Freedom

As we have mentioned, each *F* distribution has a pair of degrees of freedom, one for the numerator of the *F* ratio and the other for the denominator. How can we calculate both of these?

First, think about the numerator, the between-column variance. In Table 11-13, we used three values of $\bar{x} - \bar{\bar{x}}$, one for each sample, to calculate $\sum n_j (\bar{x}_j - \bar{\bar{x}})^2$. Once we knew two of these $\bar{x} - \bar{\bar{x}}$ values, the third was *automatically determined* and could not be freely specified. Thus, one degree of freedom is lost when we calculate the between-column variance, and the number of degrees of freedom for the numerator of the *F* ratio is always one fewer than the number of samples. The rule, then, is

Calculating degrees of freedom

Finding the numerator degrees of freedom

Numerator Degrees of Freedom

Number of degrees of freedom in the numerator of the *F* ratio = (number of samples – 1)

[11-10]

Now, what of the denominator? Look at Table 11-14 for a moment. There we calculated the variances within the samples, and we used all three samples. For the j th sample, we used n_j values of $(x - \bar{x}_j)$ to calculate the $\sum(x - \bar{x}_j)^2$ for that sample. Once we knew all but one of these $(x - \bar{x}_j)$ values, the last was *automatically determined* and could not be freely specified. Thus, we lost 1 degree of freedom in the calculations for *each* sample, leaving us with 4, 4, and 5 degrees of freedom in the samples. Because we had three samples, we were left with $4 + 4 + 5 = 13$ degrees of freedom (which could also be calculated as $5 + 5 + 6 - 3 = 13$). We can state the rule like this:

Finding the denominator degrees of freedom

Denominator Degrees of Freedom

Number of degrees of freedom in the <i>denominator</i> of the F ratio	$= \sum(n_j - 1) = n_T - k$	[11-11]
--	-----------------------------	---------

where

- n_j = size of the j th sample
- k = number of samples
- $n_T = \sum n_j$ = total sample size

Now we can see that the weight assigned to s_j^2 in Equation 11-7 on p. 560 was just its fraction of the total number of degrees of freedom in the denominator of the F ratio.

Using the F Table

To do F hypothesis tests, we shall use an F table in which the columns represent the number of degrees of freedom for the numerator and the rows represent the degrees of freedom for the denominator. Separate tables exist for each level of significance.

Suppose we are testing a hypothesis at the 0.01 level of significance, using the F distribution. Our degrees of freedom are 8 for the numerator and 11 for the denominator. In this instance, we would turn to Appendix Table 6(b). In the body of that table, the appropriate value for 8 and 11 degrees of freedom is 4.74. If our calculated sample value of F exceeds this table value of 4.74, we would reject the null hypothesis. If not, we would accept it.

Testing the Hypothesis

We can now test our hypothesis that the three different training methods produce identical results, using the material we have developed to this point. Let's begin by reviewing how we calculated the F ratio:

Finding the F statistic and the degrees of freedom

$$\begin{aligned}
 F &= \frac{\text{first estimate of the population variance}}{\text{second estimate of the population variance}} \\
 &\quad \text{based on the variance among the sample means} \\
 &\quad \text{based on the variances within the samples} \\
 &= \frac{20}{14.769} \\
 &= 1.354 \leftarrow F \text{ statistic}
 \end{aligned} \tag{11-8}$$

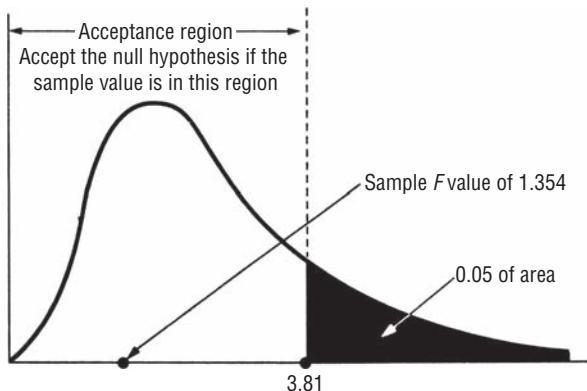


FIGURE 11-9 HYPOTHESIS TEST AT THE 0.05 LEVEL OF SIGNIFICANCE, USING THE F DISTRIBUTION AND SHOWING THE ACCEPTANCE REGION AND THE SAMPLE F VALUE

Next, calculate the number of degrees of freedom in the numerator of the F ratio, using Equation 11-10 as follows:

$$\begin{aligned} \text{Number of degrees of freedom} \\ \text{in the } \textit{numerator} \text{ of the } F \text{ ratio} &= (\text{number of samples} - 1) & [11-10] \\ &= 3 - 1 \\ &= 2 \leftarrow \text{Degrees of freedom in the numerator} \end{aligned}$$

And we can calculate the number of degrees of freedom in the denominator of the F ratio by use of Equation 11-11:

$$\begin{aligned} \text{Number of degrees of freedom} \\ \text{in the } \textit{denominator} \text{ of the } F \text{ ratio} &= \sum(n_j - 1) = n_j - k & [11-11] \\ &= (5 - 1) + (5 - 1) + (6 - 1) \\ &= 4 + 4 + 5 \\ &= 13 \leftarrow \text{Degrees of freedom in the denominator} \end{aligned}$$

Suppose the director of training wants to test at the 0.05 level the hypothesis that there are no differences among the three training methods. We can look in Appendix Table 6(a) for 2 degrees of freedom in the numerator and 13 in the denominator. The value we find there is 3.81. Figure 11-9 shows this hypothesis test graphically. The colored region represents the level of significance. The table value of 3.81 sets the upper limit of the acceptance region. Because the calculated sample value for F of 1.354 lies within the acceptance region, we would accept the null hypothesis and conclude that, according to the sample information we have, there are no significant differences in the effects of the three training methods on employee productivity.

Finding the limit of the acceptance region

Interpreting the results

Precautions about Using the F Test

As we stated earlier, our sample sizes in this problem are too small for us to be able to draw valid inferences about the effectiveness of the various training methods. We chose small samples so that we could explain the logic of analysis of

Use large sample sizes

variance without tedious calculations. In actual practice, our methodology would be the same, but our samples would be larger.

In our example, we have assumed the absence of many factors that might have affected our conclusions. We accepted as given, for example, the fact that all the new employees we sampled had the same demonstrated aptitude for learning, which may or may not be true. We assumed that all the instructors of the three training methods to manage, which may not be true. And we assumed that the company data on productivity during work periods that were similar in terms of the year, and so on. To be able to make significant decisions based to be certain that all these factors are effectively controlled.

Control all factors but the one being tested

Finally, notice that we have discussed only *one-way*, or one-factor, analysis of variance. Our problem examined the effect of the type of training method on employee productivity, nothing else. Had we wished to measure the effect of two factors, such as the training program and the age of the employee, we would need the ability to use two-way analysis of variance, a statistical method best saved for more advanced textbooks.

A test for one factor only

ANOVA: One Way Classification (Shortcut Method)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_i = \dots = \mu_m$$

Observation Table

Where

$$\bar{x}_i = \frac{\sum_{j=i}^{n_i} x_{ij}}{n_i}$$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^m T_i}{\sum n_i}$$

$$= \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}}{\sum n_i}$$

$$N = \sum_{i=1}^m n_i = \text{Total Number of observations}$$

$$\text{Raw Sum of Square (Raw S.S.)} = \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}^2$$

$$\text{Correction Factor (C.F.)} = \frac{G^2}{N}$$

$$\text{Total Sum of Squares (T.S.S.)} = \text{Raw S.S.} - \text{C.F.}$$

$$\text{Between Groups Sum of Squares (B.S.S.)} = \sum_{i=1}^m \left(\frac{\bar{T}_i^2}{n_i} \right) - \text{C.F.}$$

$$\text{Within Groups Sum of Squares (W.S.S.)} = \text{T.S.S.} - \text{B.S.S.}$$

[W.S.S. is also known as Sum of Squares due to error or Error Sum of Squares (S.S.E.)]

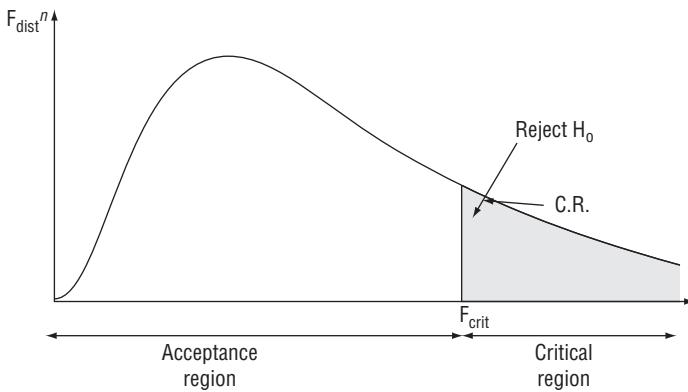
ANOVA Table

Sources of Variation	Degree of Freedom (d.f.)	Sum of Square (S.S.)	Mean Sum of Square (M.S.S.)	F-Ratio
Between Groups	$m - 1$	B.S.S.	$\text{MSB} = \frac{\text{BSS}}{m - 1}$	
Within Groups	$N - m$	W.S.S.	$\text{MSW} = \frac{\text{BSS}}{N - m}$	$F = \frac{\text{MSB}}{\text{MSW}}$
Total Variation	$N - 1$	T.S.S.	—	

$$F_{\text{crit}} = F_{\alpha, d.f.}$$

α = level of significance

$$\text{d.f.} = (m - 1), (N - m)$$



ANOVA: One Way Using MS-Excel

MS-Excel can be used to test the significance of difference between more than two samples through Analysis of Variance (One-way Classification). For this purpose, first arrange the data in form of groups, displayed in the column form, with no gap. Then the path used would be: **Data > Data Analysis > Anova: Single Factor**.

The screenshot shows a Microsoft Excel window titled "Case_Normal-Data.xls [Compatibility Mode] - Microsoft Excel". The ribbon at the top has the "Data" tab selected. A data table is displayed in the worksheet, starting from cell G15. The table has four columns labeled "Northeast", "Eastern", "Western", and "Southern". The data consists of 22 rows of numerical values. The status bar at the bottom indicates "Select destination and press ENTER or choose Paste".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																			
2		Northern	Eastern	Western	Southern														
3		74	65	64	73														
4		66	77	55	72														
5		63	86	66	71														
6		62	72	62	75														
7		57	69	80	76														
8		60	85	52	75														
9		73	80	76															
10		64	70	55															
11		71	69	61															
12		75	63	76															
13		58	85	75															
14		65	58	79															
15		69	72	67															
16		78	70	59															
17		65	71																
18		75	60																
19		55	75																
20		67	86																
21		74	68																
22		59	75																
23																			

The screenshot shows the same Microsoft Excel window as above, but now the "Data" tab is active and the "Analysis" button in the ribbon is being used to open the "Data Analysis" dialog box. This dialog box lists various statistical tools, with "Anova: Single Factor" selected. The status bar at the bottom indicates "Ready".

Data Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test: Two-Sample for Variances
- Fourier Analysis
- Histogram

When **Anova: Single Factor** dialogue box opens, enter the entire sample-data range (of all groups) simultaneously in the **Input: Input Range**, check **Label in first row**. Level of significance (Alpha) can be changed from 5% level of significance, if situation demands. Press **OK**.

The screenshot shows a Microsoft Excel spreadsheet titled "Case_Normal-Data (1).xls". A data table is present in rows 4 to 25, with columns labeled Northern, Eastern, Western, and Southern. An "Anova: Single Factor" dialog box is open, prompting for input range (\$C\$4:\$F\$24), grouping by columns, and labels in the first row. The alpha level is set to 0.05, and the output is directed to a new worksheet.

The output sheet will be displayed as below. If the P-value is small, Null Hypothesis (H_0) of equality of sample means can be rejected, showing the significant difference among sample means.

The screenshot shows the ANOVA output from the previous dialog. The summary table shows the following data:

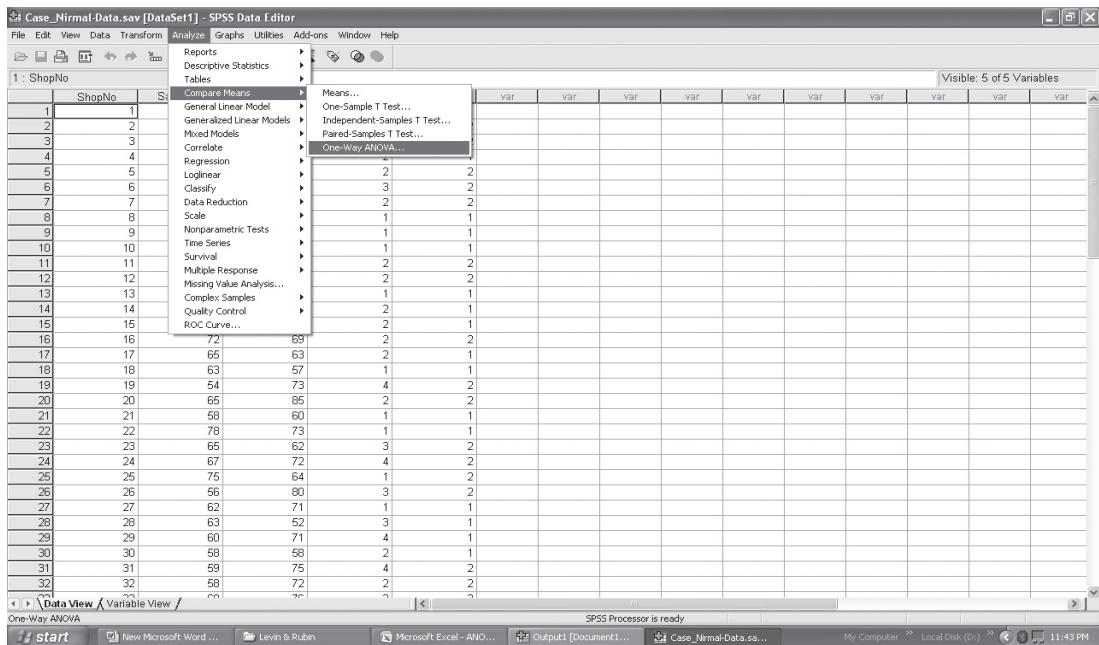
Groups	Count	Sum	Average	Variance
Northern	20	1330	66.5	47.105
Eastern	20	1476	73.8	73.958
Western	14	928	66.286	88.989
Southern	6	442	73.667	3.8667

The ANOVA table shows the following results:

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	798.01	3	266	4.285	0.0086	2.769431
Within Groups	3476.4	56	62.078			
Total	4274.4	59				

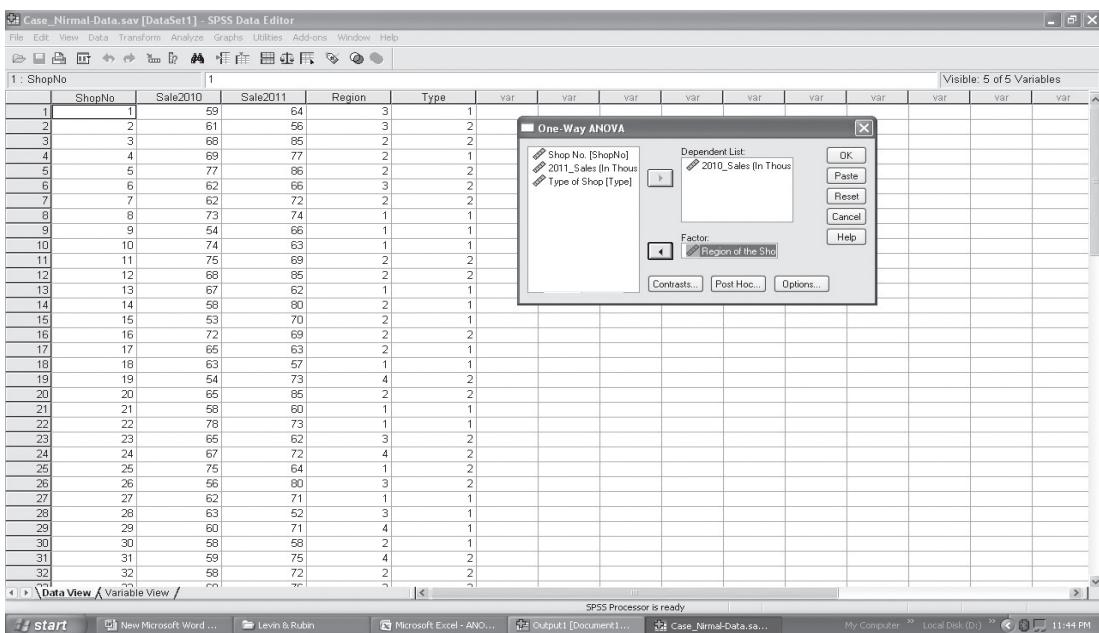
ANOVA: One Way Using SPSS

For testing the significance of difference between more than two samples through Analysis of Variance (One-way Classification), using SPSS. The path will be: **Analyze > Compare Means > One Way ANOVA**.



In the **One Way ANOVA** dialogue box, enter variable containing sample-data series in **Dependent List** drop box and the variable containing the groups into **Factor** drop box. Then press **Options** tab.

The **One Way ANOVA: Options Box** sub-dialogue box will be opened. Check **Statistics: Descriptive** check-box. Then press **Continue** button to go back to main dialogue box. Then press **OK**.



Case_Nirmal-Datas.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

Visible: 5 of 5 Variables

ShopNo	ShopNo	Sale2010	Sale2011	Region	Type	var								
1	1	59	64	3	1									
2	2	61	56	3	2									
3	3	68	85	2	2									
4	4	69	77	2	1									
5	5	77	86	2	2									
6	6	62	66	3	2									
7	7	62	72	2	2									
8	8	73	74	1	1									
9	9	54	66	1	1									
10	10	74	63	1	1									
11	11	75	69	2	2									
12	12	68	85	2	2									
13	13	67	62	1	1									
14	14	56	60	2	1									
15	15	63	70	2	1									
16	16	72	69	2	2									
17	17	65	63	2	1									
18	18	63	57	1	1									
19	19	54	73	4	2									
20	20	66	65	2	2									
21	21	58	60	1	1									
22	22	78	73	1	1									
23	23	65	62	3	2									
24	24	67	72	4	2									
25	25	75	64	1	2									
26	26	56	60	3	2									
27	27	62	71	1	1									
28	28	63	52	3	1									
29	29	60	71	4	1									
30	30	58	58	2	1									
31	31	59	75	4	2									
32	32	58	72	2	2									

One Way ANOVA

Dependent List: Shop No. [ShopNo]

Factor: Region of the Shop [Region]

OK Cancel Help

One-Way ANOVA: Options

Statistics

Descriptive

Fixed and random effects

Homogeneity of variance test

Brown-Forsythe

Welch

Means plot

Missing Values

Exclude cases analysis by analysis

Exclude cases listwise

SPSS Processor is ready

start Levin & Rubin Chat 7 doc - Microsoft... Chat 11_ANOVA.doc... Output1[Document1] Case_Nirmal-Datas... 11:54 PM

Output1 [Document1] - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

GET
FILE='E:\ddrive\Levin & Rubin\Case_Nirmal-Datas.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
ONEWAY
Sale2010 BY Region
/STATISTICS DESCRIPTIVES
/MISSING ANALYSIS .

→ Oneway

[DataSet1] E:\ddrive\Levin & Rubin\Case_Nirmal-Datas.sav

Descriptives

2010_Sales (In Thousand of Rs)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Northern Region	20	66.80	6.363	1.423	63.62	69.78	54	78
Eastern Region	20	63.90	6.688	1.495	60.77	67.03	53	77
Western Region	14	60.07	3.772	1.000	57.69	62.25	51	65
Southern Region	6	65.00	10.139	4.139	54.36	75.64	54	83
Total	60	64.00	6.761	.073	62.34	65.03	51	83

ANOVA

2010_Sales (In Thousand of Rs)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	378.655	3	126.218	3.048	.036
Within Groups	2317.929	56	41.392		
Total	2696.580	59			

SPSS Processor is ready

start New Microsoft Word... Levin & Rubin Microsoft Excel - ANOVA... Output1[Document1] Case_Nirmal-Datas... My Computer Local Disk (C:) 11:46 PM

The output sheet would be displayed as above. If the Sig. is small, Null Hypothesis (H_0) of equality of sample means can be rejected, showing the significant difference among sample means.

Analysis of Variance (ANOVA): Two Way

Analysis of Variance (Two Way Classification with one observation per cell)

Some situations demand that the experiment should be planned in such a manner so as to study the effects of two factors simultaneously. For each factor, there will be a number of classes/levels or categories.

The two factors are: Factor A and Factor B. The factor A has ' m ' levels/classes:

$$A_1, A_2, \dots, A_i, \dots, A_m$$

Factor B has ' n ' levels/classes:

$$B_1, B_2, \dots, B_j, \dots, B_n$$

Let x_{ij} be the observation under the i th level of Factor A and j th level of the factor B.

Here, observations should be so taken that there should be one observation per cell (x_{ij}) of the bivariate (Cross Tabulation) table. The corresponding analysis of variance is known as Analysis of Variance (Two Way Classification with one observation per cell).

The second option is that the observations should be taken in such a manner that cells of the bivariate table contain more than one observation per cell. This option facilitates the estimation and testing of the interaction effect. Interaction effect is an effect peculiar to the combination (A_i, B_j). If the joint effect of A_i and B_j is different from the sum of the effects due to A_i and B_j taken individually, then the interaction effect is said to be present.

Analysis of Variance (Two Way Classification with One Observation Per Cell)

Short-Cut Method

1st Factor (A): $H_{01}: \alpha_1 = \alpha_2 = \dots = \alpha_i = \dots = \alpha_m$

H_{11} : At least two α_i 's are different.

2nd Factor (B): $H_{02}: \beta_1 = \beta_2 = \dots = \beta_i = \dots = \beta_n$

H_{12} : At least two β_j 's are different.

Observation Table

Factor B				
Factor A		$B_1, B_2, \dots, B_j, \dots, B_n$	Total	Mean
A_1		$x_{11}, x_{12}, \dots, x_{1j}, \dots, x_{1n}$	$T_{1\cdot}$	$\bar{x}_{1\cdot}$
A_2		$x_{21}, x_{22}, \dots, x_{2j}, \dots, x_{2n}$	$T_{2\cdot}$	$\bar{x}_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots
A_i		$x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in}$	$T_{i\cdot}$	$\bar{x}_{i\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots
A_m		$x_{m1}, x_{m2}, \dots, x_{mj}, \dots, x_{mn}$	$T_{m\cdot}$	$\bar{x}_{m\cdot}$
Total		$T_{\cdot 1}, T_{\cdot 2}, \dots, T_{\cdot j}, \dots, T_{\cdot n}$	G	
Mean		$\bar{x}_{\cdot 1}, \bar{x}_{\cdot 2}, \dots, \bar{x}_{\cdot j}, \dots, \bar{x}_{\cdot n}$		$\bar{\bar{x}}_{\cdot ..}$

Where

$$\begin{aligned} T_{i \cdot} &= \sum_{j=1}^n x_{ij} & \bar{x}_{i \cdot} &= \frac{T_{i \cdot}}{n} = \frac{\sum_{j=1}^n x_{ij}}{n} \\ T_{\cdot j} &= \sum_{i=1}^m x_{ij} & \bar{x}_{\cdot j} &= \frac{T_{\cdot j}}{m} = \frac{\sum_{i=1}^m x_{ij}}{m} \\ G &= \sum_{i=1}^m \sum_{j=1}^n x_{ij} & \bar{x} &= \frac{G}{N} = \frac{\sum_{i=1}^m \sum_{j=1}^n x_{ij}}{N} \end{aligned}$$

$N = m \times n$ (Total Number of observations)

$$\text{Raw Sum of Squares (R.S.S.)} = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2$$

$$\text{Correction Factor (C.F.)} = \frac{G^2}{N}$$

$$\text{Total Sum of Square (T.S.S.)} = \text{Raw S.S.} - \text{C.F.}$$

$$\text{Sum of Squares due to factor A (SSA)} = \sum_{i=1}^m \left(\frac{T_{i \cdot}^2}{n} \right) - \text{C.F.}$$

$$\text{Sum of Squares due to factor B (SSB)} = \sum_{j=1}^n \left(\frac{T_{\cdot j}^2}{m} \right) - \text{C.F.}$$

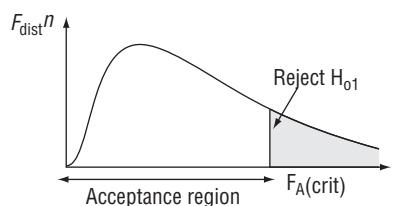
$$\text{Sum of Squares due to error (SSE)} = \text{T.S.S.} - \text{S.S.A} - \text{S.S.B}$$

ANOVA Table

Sources of Variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	F-Ratio
Between the levels of A (due to factor A)	$(m - 1)$	SSA	$MSA = \frac{SSA}{m-1}$	$F_A = \frac{MSA}{MSE}$
Between the levels of B (due to factor B)	$(n - 1)$	SSB	$MSB = \frac{SSB}{n-1}$	
Error (Residual)	$(m - 1) \times (n - 1)$	SSE	$MSE = \frac{SSE}{(m-1) \times (n-1)}$	$F_B = \frac{MSB}{MSE}$
Total	$N - 1$	TSS	-	

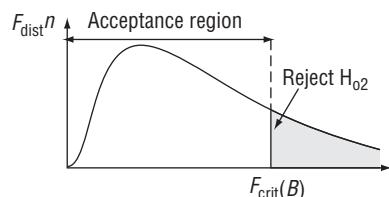
$$F_{A(\text{crit})} = F_{\alpha, d.f.(A)}$$

$$d.f.(A) = (m - 1), (m - 1) \cdot (n - 1)$$



$$F_{B(\text{crit})} = F_{\alpha, \text{d.f.}(B)}$$

$$\text{d.f.}(B) = (n - 1), (m - 1) \cdot (n - 1)$$



ANOVA: Two Way Using MS-Excel

MS-Excel can be used for performing Analysis of Variance (Two-way Classification: Without Replication i.e., one observation per cell). For this purpose, first arrange the data in form of a cross tabulation format, one source of variation in rows and the other in columns. Then the path used would be: **Data > Data Analysis > Anova: Two Factor Without Replication.**

When **Anova: Two Factor without replication** dialogue box opens enter the entire sample-data range (in the form of cross table) simultaneously in the **Input: Input Range**, check **Label**. Level of significance (Alpha) can be changed from 5% level of significance, if situation demands. Press **OK**.

The output sheet would be displayed as under. If the P-value is small for a particular source of variation then that source of variation will show the significant effect i.e. there will be significant difference between sample-means corresponding to that source of variation.

The screenshot shows an Excel spreadsheet titled "ANOVA-a-2Way.xls" in Compatibility Mode. The data is organized into two main sections: "SUMMARY" and "ANOVA".

SUMMARY:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Anova: Two-Factor Without Replication															
2																
3	SUMMARY	Count	Sum	Average	Variance											
4	0-2	4	70	17.5	91.66667											
5	2-5	4	95	23.75	422.9167											
6	5-10	4	80	20	66.66667											
7	10-15	4	95	23.75	22.91667											
8	More than 15	4	125	31.25	106.25											
9																
10	Training 1	5	90	18	57.5											
11	Training 2	5	150	30	50											
12	Training 3	5	65	13	45											
13	Training 4	5	160	32	170											
14																
15																
16	ANOVA															
17	Source of Variation	SS	df	MS	F	P-value	F crit									
18	Rows	432.5	4	108.125	1.51312	0.259817	3.259167									
19	Columns	1273.75	3	424.5833333	5.941691	0.010063	3.490295									
20	Error	857.5	12	71.45833333												
21	Total	2563.75	19													
22																
23																

ANOVA:

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	432.5	4	108.125	1.51312	0.259817	3.259167
Columns	1273.75	3	424.5833333	5.941691	0.010063	3.490295
Error	857.5	12	71.45833333			
Total	2563.75	19				

ANOVA: Two Way Using SPSS

For performing Analysis of Variance (Two-way Classification), using SPSS, the path will be: **Analyze > General Linear Model > Univariate.**

The screenshot shows the SPSS Data Editor window with the title "Case_Nirmal>Data.sav [DataSet1] - SPSS Data Editor". The menu bar is visible at the top, and the "Analyze" menu is currently selected. A sub-menu path "General Linear Model > Univariate..." is highlighted, indicating the active step in the process of performing a two-way ANOVA.

In the **Univariate** dialogue box, enter the variable containing sample-data series in **Dependent Variable** drop box and the two variables containing the possible sources of variation into **Fixed Factors** drop box. Then press **OK**.

The output sheet would be displayed as under.

Between-Subjects Factors		
	Value Label	N
Region	1 Northern Region	20
	2 Eastern Region	20
3 Western Region	14	
	4 Southern Region	6
Type of Shop	1 Urban Shop	31
	2 Rural Shop	29

Tests of Between-Subjects Effects					
Dependent Variable: 2010_Sales (In Thousand of Rs)					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	739.976*	7	105.711	2.809	.015
Intercept	167998.706	1	167998.706	4464.842	.000
Region	446.439	3	148.813	3.955	.013
Type	.981	1	.981	.026	.872
Region * Type	281.075	3	93.958	2.497	.070
Error	1956.605	52	37.627		
Total	249097.000	60			

HINTS & ASSUMPTIONS

The focus of analysis of variance is to test whether three or more samples have been drawn from populations having the same mean. Analysis of variance is important in research such as the evaluation of new drugs, where we need to examine the effects of dose, frequency of medication, effects of other drugs, and patient differences in a single study. Analysis of variance compares two estimates of the population variance. One estimate comes from the variance among the sample means, the other from the variance within the samples themselves. If they are approximately equal, the chances are high that the samples came from the same population. Warning: It's vital not to abandon common sense when interpreting results. While it may be true that a study can identify differences in brand preferences for instant coffee that apply to coffee purchases made on weekday mornings, it's hard to say what a coffee company should do with this information.

EXERCISES 11.4**Self-Check Exercises**

SC 11-5 A study compared the effects of four 1-month point-of-purchase promotions on sales. The unit sales for five stores using all four promotions in different months follow.

Free sample	78	87	81	89	85
One-pack gift	94	91	87	90	88
Cents off	73	78	69	83	76
Refund by mail	79	83	78	69	81

- (a) Compute the mean unit sales for each promotion and then determine the grand mean.
- (b) Estimate the population variance using the between-column variance (Equation 11-6).
- (c) Estimate the population variance using the within-column variance computed from the variance within the samples.
- (d) Calculate the F ratio. At the 0.01 level of significance, do the promotions produce different effects on sales?

SC 11-6 A research company has designed three different systems to clean up oil spills. The following table contains the results, measured by how much surface area (in square meters) is cleared in 1 hour. The data were found by testing each method in several trials. Are the three systems equally effective? Use the 0.05 level of significance

System A	55	60	63	56	59	55
System B	57	53	64	49	62	
System C	66	52	61	57		

Applications

11-26 A study compared the number of hours of relief provided by five different brands of antacid administered to 25 different people, each with stomach acid considered strong. The results are given in the following table:

Brand	A	B	C	D	E
	4.4	5.8	4.8	2.9	4.6
	4.6	5.2	5.9	2.7	4.3
	4.5	4.9	4.9	2.9	3.8
	4.1	4.7	4.6	3.9	5.2
	3.8	4.6	4.3	4.3	4.4

- (a) Compute the mean number of hours of relief for each brand and determine the grand mean.
 (b) Estimate the population variance using the between-column variance (Equation 11-6).
 (c) Estimate the population variance using the within-column variance computed from the variance within the samples.
 (d) Calculate the F ratio. At the 0.05 level of significance, do the brands produce significantly different amounts of relief to people with strong stomach acid?

11-27 Three training methods were compared to see whether they led to greater productivity after training. The following are productivity measures for individuals trained by each method.

Method 1	45	40	50	39	53	44
Method 2	59	43	47	51	39	49
Method 3	41	37	43	40	52	37

At the 0.05 level of significance, do the three training methods lead to different levels of productivity?

11-28 The following data show the number of claims processed per day for a group of four insurance company employees observed for a number of days. Test the hypothesis that the employees' mean claims per day are all the same. Use the 0.05 level of significance.

Employee 1	15	17	14	12
Employee 2	12	10	13	17
Employee 3	11	14	13	15
Employee 4	13	12	12	14
			10	9

11-29 Given the measurements in the four samples that follow, can we conclude that they come from populations having the same mean value? Use the 0.01 level of significance.

Sample 1	16	21	24	28	29
Sample 2	29	18	20	19	30
Sample 3	14	15	21	19	28
Sample 4	21	28	20	22	18

11-30 The manager of an assembly line in a clock manufacturing plant decided to study how different speeds of the conveyor belt affect the rate of defective units produced in an 8-hour shift. To examine this, he ran the belt at four different speeds for five 8-hour shifts each and measured the number of defective units found at the end of each shift. The results of the study follow:

Defective Units per Shift			
Speed 1	Speed 2	Speed 3	Speed 4
37	27	32	35
35	32	36	27
38	32	33	33
36	34	34	31
34	30	40	29

- (a) Calculate the mean number of defective units, \bar{x} , for each speed; then determine the grand mean, $\bar{\bar{x}}$.
- (b) Using Equation 11-6, estimate the population variance (the between-column variance).
- (c) Calculate the variances *within* the samples and estimate the population variance based upon these variances (the within-column variance).
- (d) Calculate the F ratio. At the 0.05 level of significance, do the four conveyor-belt speeds produce the same mean rate of defective clocks per shift?

- 11-31** We are interested in testing for differences in the palatability of three spicy salsas: A, B, and C. For each product, a sample of 25 men was chosen. Each rated the product from -3 (terrible) to +3 (excellent). The following SAS output was produced.

ANALYSIS OF VARIANCE PROCEDURE			
DEPENDENT VARIABLE:	SCORE (-3 TO +3)		
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	2	15.68	7.84
ERROR	72	94.4	1.31111111
CORRECTED TOTAL	74	110.08	
MODEL F =	5.98		PR > F = 0.004

- (a) State explicit null and alternative hypotheses.
 (b) Test your hypotheses with the SAS output. Use $\alpha = 0.05$.
 (c) State an explicit conclusion.

- 11-32** The supervisor of security at a large department store would like to know whether the store apprehends relatively more shoplifters during the Christmas holiday season than in the weeks before or after the holiday. He gathered data on the number of shoplifters apprehended in the store during the months of November, December, and January over the past 6 years. The information follows:

	Number of Shoplifters					
	43	37	59	55	38	48
November	43	37	59	55	38	48
December	54	41	48	35	50	49
January	36	28	34	41	30	32

At the 0.05 level of significance, is the mean number of apprehended shoplifters the same during these 3 months?

- 11-33** An Introduction to Economics course is offered in 3 sections, each with a different instructor. The final grades from the spring term are presented below. Is there a significant difference in the average grades given by the instructors? State and test appropriate hypotheses at $\alpha = 0.01$.

Section 1	Section 2	Section 3
98.4	97.6	94.5
97.6	99.2	92.3
84.7	82.6	92.4
88.5	81.2	82.3

(continued)

(contd.)

Section 1	Section 2	Section 3
77.6	64.5	62.6
84.3	82.3	68.6
81.6	68.4	92.7
88.4	75.6	82.3
95.1		91.2
90.4		92.6
89.4		87.4
65.6		
94.5		
99.4		
68.7		
83.4		

- 11-34** The manufacturer of silicon chips requires so-called clean rooms, where the air is specially filtered to keep the number of dust particles at a minimum. The Outel Corporation wants to make sure that each of its five clean rooms has the same number of dust particles. Five air samples have been taken in each room. The “dust score,” on a scale of 1 (low) to 10 (high), was measured. At the 0.05 level of significance, do the rooms have the same average dust score?

Dust Score (1 to 10)					
Room 1	5	6.5	4	7	6
Room 2	3	6	4	4.5	3
Room 3	1	1.5	3	2.5	4
Room 4	8	9.5	7	6	7.5
Room 5	1	2	3.5	1.5	3

- 11-35** A lumber company is concerned about how rising interest rates are affecting the new housing starts in the area. To explore this question, the company has gathered data on new housing starts during the past three quarters for five surrounding counties. This information is presented in the following table. At the 0.05 level of significance, are there any differences in the number of new housing starts during the three quarters?

Quarter 1	41	53	54	55	43
Quarter 2	45	51	48	43	39
Quarter 3	34	44	46	45	51

- 11-36** Genes-and-Jeans, Inc., offers clones of such popular jeans as Generic, DNA, RNA, and Oops. The store wants to see whether there are differences in the number of pairs sold of different brands. The manager has counted the number of pairs sold for each brand on

several different days. At the 0.05 significance level, are the sales of the four brands the same?

	Pairs of Jeans Sold				
Generic	17	21	13	27	12
DNA	27	13	29	9	
RNA	13	15	17	23	10
Oops	18	25	15	27	12

- 11-37** The Government Accounting Office (GAO) is interested in seeing whether similar-sized offices spend similar amounts on personnel and equipment. (Offices spending more are targeted for special auditing.) Monthly expenses for three offices have been examined: one office in the Agriculture Department, one in the State Department, and one in the Interior Department. The data follow. At the 0.01 significance level, are there differences in expenses for the different offices?

	Monthly Office Expenses (\$ thousands) for Some Past Months				
Agriculture	10	8	11	9	12
State	15	9	8	10	13
Interior	8	16	12		

- 11-38** In Bigville, a fast-food chain feels it is gaining a bad reputation because it takes too long to serve the customers. Because the chain has four restaurants in this town, it is concerned with whether all four restaurants have the same average service time. One of the owners of the fast-food chain has decided to visit each of the stores and monitor the service time for five randomly selected customers. At his four noontime visits, he records the following service times in minutes:

Restaurant 1	3	4	5.5	3.5	4
Restaurant 2	3	3.5	4.5	4	5.5
Restaurant 3	2	3.5	5	6.5	6
Restaurant 4	3	4	5.5	2.5	3

- (a) Using a 0.05 significance level, do all the restaurants have the same mean service time?
- (b) Based on his results, should the owner make any policy recommendations to any of the restaurant managers?

- 11-39** LWP is a large multinational company, having more than 2000 employees under its payroll. The management of the LWP company has introduced an intensive and comprehensive training programme for its managerial level employees. There are four different types of training programmes, currently under consideration. Employees were divided into four groups and different groups were given different training programmes. Now the management of LWP wants to examine the effectiveness of these training programmes i.e. whether these programmes have been resulting into same managerial skill enhancement or their degree of effectiveness is different. Simultaneously, the management also wants to examine whether there is some change in the effectiveness – degree with the level of experience of employees. Management believes that these two factors are independent. There are 5 categories of experience-levels, as envisaged by the HR department of LWP. For this dual purpose, 20 employees were randomly

chosen, according to a particular scheme. Five employees were chosen from each training programme and in each subgroup, 5 employees belong to 5 different experience categories. The employees thus selected were given a managerial-skill aptitude examination to examine their managerial aptitude and their scores are tabulated in form of following table:

Experience Category (Years)	Training Programme			
	Training 1	Training 2	Training 3	Training 4
0–2	10	20	10	30
2–5	10	30	5	50
5–10	20	30	10	20
10–15	25	30	20	20
More than 15	25	40	20	40

Analyze the data and examine the following perceptions:

- (a) The four training programmes are equally effective in enhancing the managerial-skill of the employees.
- (b) There is no significant aptitude-difference among employees of different experience-categories.

Comment on the results.

Worked-Out Answers to Self-Check Exercises

SC 11-5 (a)	Free	Gift	Cents	Refund
	78	94	73	79
	87	91	78	83
	81	87	69	78
	89	90	83	69
	85	88	76	81
Σx	<u>420</u>	<u>450</u>	<u>379</u>	<u>390</u>
n	5	5	5	5
\bar{x}	84	90	75.8	78
Σx^2	35,360	40,530	28,839	30,536
s^2	20	7.5	27.7	29
Grand mean = $\bar{\bar{x}} = \frac{420 + 450 + 379 + 390}{20} = 81.95$				

$$(b) \hat{\sigma}_b^2 = \frac{\sum n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1} = \frac{5[84 - 81.95]^2 + (90 - 81.95)^2 + (75.8 - 81.95)^2 + (78 - 81.95)^2]}{4-1}$$

$$= \frac{612.15}{3} = 204.05$$

$$(c) \hat{\sigma}_w^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) s_j^2 = \frac{4(20 + 7.5 + 27.7 + 29)}{20 - 4} = \frac{336.8}{16} = 21.05$$

$$(d) \quad F = \frac{204.05}{21.05} = 9.69$$

With 3 degrees of freedom in the numerator, 16 degrees of freedom in the denominator, and $\alpha=0.01$, the critical value of F is 5.29, so reject H_0 , because $9.69 > 5.29$. The promotions have significantly different effects on sales.

SC 11-6	n	\bar{x}	s^2
System A	6	58	10.4000
System B	5	57	38.5000
System C	4	59	35.3333

$$\bar{\bar{x}} = \frac{6(58) + 5(57) + 4(59)}{6+5+4} = 57.9333$$

$$\hat{\sigma}_b^2 = \frac{\sum n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

$$= \frac{6(58 - 57.9333)^2 + 5(57 - 57.9333)^2 + 4(59 - 57.9333)^2}{3-1}$$

$$= \frac{8.9333}{2} = 4.4667$$

$$\sigma_w^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) s_j^2 = \frac{5(10.4) + 4(38.5) + 3(35.3333)}{15 - 3} = \frac{312}{12} = 26$$

$$F = \frac{\hat{\sigma}_w^2}{\hat{\sigma}_b^2} = \frac{4.4667}{26} = 0.17$$

With 2 degrees of freedom in the numerator, 12 degrees of freedom in the denominator, and $\alpha=0.05$, the critical value of F is 3.89, so don't reject H_0 , because $0.17 < 3.89$. The systems do not have significantly different effectiveness.

11.5 INFERENCES ABOUT A POPULATION VARIANCE

In Chapters 7–9, we learned how to form confidence intervals and test hypotheses about one or two population means or proportions. Earlier in this chapter, we used chi-square and F tests to make inferences about more than two means or proportions. But we are not always interested in means and proportions. In many situations, responsible decision makers have to make inferences about the variability in a population. In order to schedule the labor force at harvest time, a peach grower needs to know not only the mean time to maturity of the peaches, but also their variance around that mean. A sociologist investigating the effect of education on earning power wants to know whether the incomes of college graduates are more variable than those of high school graduates. Precision instruments used in laboratory work must be quite accurate on the average but in addition, repeated measurements should show very little variation. In this section, we shall see how to make inferences about a single population variance. The next section looks at problems involving the variances of two populations.

**Need to make decisions
about variability in a
population**

TABLE 11-15 DELIVERY TIME (IN HOURS) FOR LETTERS GOING BETWEEN NEW YORK AND CHICAGO

Time x	\bar{x}	$x - \bar{x}$	$(x - \bar{x})^2$
50	59	-9	81
45	59	-14	196
27	59	-32	1,024
66	59	7	49
43	59	-16	256
96	59	37	1,369
45	59	-14	196
90	59	31	961
69	59	10	100
$\Sigma x = 531$			$\Sigma(x - \bar{x})^2 = 4,232$
$s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{4,232}{8}$			[3-17]
= 59 hours			= 529 hours squared
		$s = \sqrt{s^2} = \sqrt{529}$	[3-18]
		= 23 hours	

The Distribution of the Sample Variance

In response to a number of complaints about slow mail delivery, the Postmaster General initiates a preliminary investigation. An investigator follows nine letters from New York to Chicago, to estimate the standard deviation in time of delivery. Table 11-15 gives the data and computes \bar{x} , s^2 , and s . As we saw in Chapter 7, we use s to estimate σ .

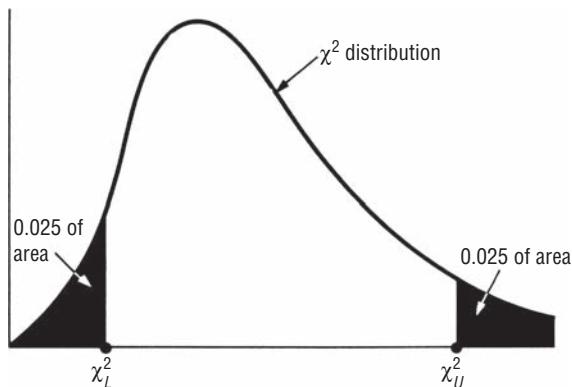
We can tell the Postmaster General that the *population* standard deviation, as estimated by the *sample* standard deviation, is approximately 23 hours. But he also wants to know how accurate that estimate is and what uncertainty is associated with it. In other words, he wants a confidence interval, not just a point estimate of σ . In order to find such an interval, we must know the sampling distribution of s . It is traditional to talk about s^2 rather than s , but this will cause us no trouble, because we can always go from s^2 and σ^2 to s and σ by taking square roots; we can go in the other direction by squaring.

Determining the uncertainty attached to estimates of the population standard deviation

Chi-Square Statistic for Inferences about One Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad [11-12]$$

If the population variance is σ^2 , then the statistic has a chi-square distribution with $n - 1$ degrees of freedom. This result is exact if the population is normal, but even for samples from nonnormal populations,

**FIGURE 11-14 CONSTRUCTING A CONFIDENCE INTERVAL FOR σ^2 .**

it is often a good approximation. We can now use the chi-square distribution to form confidence intervals and test hypotheses about σ^2 .

Confidence Intervals for the Population Variance

Suppose we want a 95 percent confidence interval for the variance in our mail-delivery problem. Figure 11-14 shows how to begin constructing this interval.

We locate two points on the χ^2 distribution: χ_u^2 cuts off 0.025 of the area in the upper tail of the distribution, and χ_L^2 cuts off 0.025 of the area in the lower tail. (For a 99 percent confidence interval, we would put 0.005 of the area in each tail, and similarly for other confidence levels.) The values of χ_L^2 and χ_u^2 , can be found in Appendix Table 5. In our mail problem, with $9 - 1 = 8$ degrees of freedom, $\chi_L^2 = 2.180$ and $\chi_u^2 = 17.535$.

Now Equation 11-12 gives χ^2 in terms of s^2 , n , and σ^2 . To get a confidence interval for σ^2 , we solve Equation 11-12 for σ^2 :

Constructing a confidence Interval for a Variance

Upper and lower limits for the confidence interval

$$\sigma^2 = \frac{(n-1)s^2}{\chi^2} \quad [11-13]$$

and then our confidence interval is given by

Confidence Interval for σ^2

$$\sigma_L^2 = \frac{(n-1)s^2}{\chi_u^2} \leftarrow \text{Lower confidence limit} \quad [11-14]$$

$$\sigma_u^2 = \frac{(n-1)s^2}{\chi_L^2} \leftarrow \text{Upper confidence limit}$$

Notice that because χ^2 appears in the denominator in Equation 11-13, we can use χ^2_u to find σ_L^2 and χ^2_L to find σ_u^2 . Continuing with the Postmaster General's problem, we see he can be 95 percent confident that the population variance lies between 241.35 and 1,941.28 hours squared:

$$\sigma_L^2 = \frac{(n-1)s^2}{\chi^2_L} = \frac{8(529)}{17.535} = 241.35 \quad [11-14]$$

$$\sigma_u^2 = \frac{(n-1)s^2}{\chi^2_u} = \frac{8(529)}{2.180} = 1,941.28$$

So a 95 percent confidence interval for σ would be from $\sqrt{241.35}$ to $\sqrt{1,941.28}$ hours, that is, from 15.54 to 44.06 hours.

A Two-Tailed Test of a Variance

A management professor has given careful thought to the design of examinations. In order for him to be reasonably certain that an exam does a good job of distinguishing the differences in achievement shown by the students, the standard deviation of scores on the examination cannot be too small. On the other hand, if the standard deviation is too large, there will tend to be a lot of very low scores, which is bad for student morale. Past experience has led the professor to believe that a standard deviation of about 13 points on a 100-point exam indicates that the exam does a good job of balancing these two objectives.

Testing hypotheses about a variance: Two-tailed tests

The professor just gave an examination to his class of 31 freshmen and sophomores. The mean score was 72.7 and the sample standard deviation was 15.9. Does this exam meet his goodness criterion? We can summarize the data:

$$\begin{aligned}\sigma_{H_0} &= 13 && \leftarrow \text{Hypothesized value of the population standard deviation} \\ s &= 15.9 && \leftarrow \text{Sample standard deviation} \\ n &= 31 && \leftarrow \text{Sample size}\end{aligned}$$

If the professor uses a significance level of 0.10 in testing his hypothesis, we can symbolically state the problem:

$$\begin{aligned}H_0 : \sigma &= 13 && \leftarrow \text{Null hypothesis: The true standard deviation is 13 points} \\ H_1 : \sigma &\neq 13 && \leftarrow \text{Alternative hypothesis: The true standard deviation is not 13 points} \\ \alpha &= 0.10 && \leftarrow \text{Level of significance for testing these hypotheses}\end{aligned}$$

The first thing we do is to use Equation 11-12 to calculate the χ^2 statistic:

$$\begin{aligned}\chi^2 &= \frac{(n-1)s^2}{\sigma^2} \\ &= \frac{30(15.9)^2}{(13)^2} \\ &= 44.88\end{aligned} \quad [11-12]$$

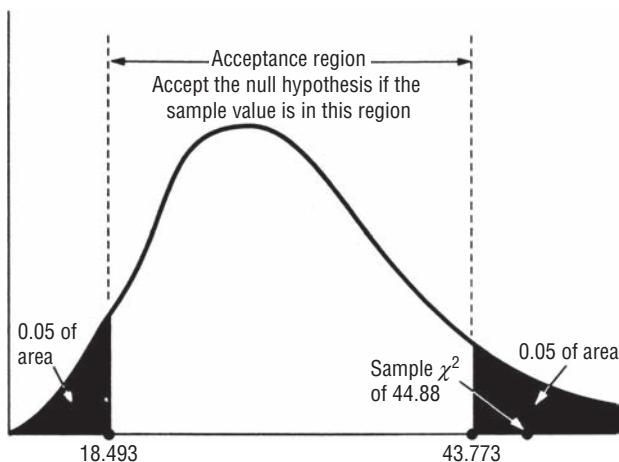


FIGURE 11-15 TWO-TAILED HYPOTHESIS TEST AT THE 0.10 LEVEL OF SIGNIFICANCE, SHOWING ACCEPTANCE REGION AND SAMPLE χ^2

This statistic has a χ^2 distribution with $n - 1$ ($=30$ in this case) degrees of freedom. We will accept the null hypothesis if χ^2 is neither too big nor too small. From the χ^2 distribution table (Appendix Table 5), we can see that the appropriate χ^2 values for 0.05 of the area to lie in each tail of the curve are 18.493 and 43.773. These two limits of the acceptance region and the observed sample statistic ($\chi^2 = 44.88$) are shown in Figure 11-15. We see that the sample value of χ^2 is not in the acceptance region, so the professor should reject the null hypothesis; this exam does not meet his goodness criterion.

Interpreting the results

A One-Tailed Test of a Variance

Precision Analytics manufactures a wide line of precision instruments and has a fine reputation in the field for quality of its instruments. In order to preserve that reputation, it maintains strict quality control on all of its output. It will not release an analytic balance for sale, for example, unless that balance shows a standard deviation significantly below one microgram (at $\alpha = 0.01$) when weighing quantities of about 500 grams. A new balance has just been delivered to the quality control division from the production line.

Testing hypotheses about a variance: One-tailed tests

The new balance is tested by using it to weigh the same 500-gram standard weight 30 different times. The sample standard deviation turns out to be 0.73 microgram. Should this balance be sold? We summarize the data:

$$\sigma_{H_0} = 1 \quad \leftarrow \text{Hypothesized value of the population standard deviation}$$

$$s = 0.73 \quad \leftarrow \text{Sample standard deviation}$$

$$n = 30 \quad \leftarrow \text{Sample size}$$

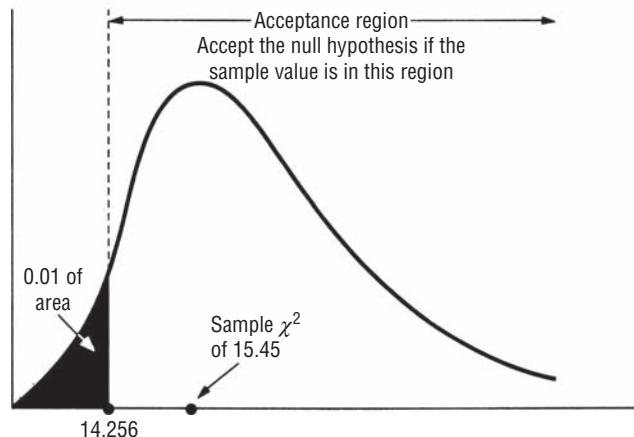


FIGURE 11-16 ONE-TAILED HYPOTHESIS TEST AT THE 0.01 SIGNIFICANCE LEVEL, SHOWING ACCEPTANCE REGION AND SAMPLE χ^2

and state the problem:

Stating the problem symbolically

$$H_0: \sigma = 1 \quad \leftarrow \text{Null hypothesis: The true standard deviation is 1 microgram}$$

$$H_1: \sigma < 1 \quad \leftarrow \text{Alternative hypothesis. The true standard deviation is less than 1 microgram}$$

$$\alpha = 0.01 \quad \leftarrow \text{Level of significance for testing these hypotheses}$$

We begin by using Equation 11-12 to calculate the χ^2 statistic:

Calculating the χ^2 statistic

$$\begin{aligned} \chi^2 &= \frac{(n-1)s^2}{\sigma^2} & [11-12] \\ &= \frac{29(0.73)^2}{(1)^2} \\ &= 15.45 \end{aligned}$$

We will reject the null hypothesis and release the balance for sale if *Interpreting the results* this statistic is sufficiently small. From Appendix Table 5, we see that with 29 degrees of freedom ($30 - 1$), the value of χ^2 that leaves an area of 0.01 in the lower tail of the curve is 14.256. The acceptance region and the observed value of χ^2 are shown in Figure 11-16. We see that we cannot reject the null hypothesis. The balance will have to be returned to the production line for adjusting.

HINTS & ASSUMPTIONS

Up to this point, we've seen how to make inferences about one, two, or several *means or proportions*. But we're also interested in making inferences about population *variability*. For one population, we do this by using the sample variance and the chi-square distribution. Warning: Chi-square tests can be one-tailed or two-tailed. Hint: If the question to be answered is worded *less than, more than, less than or equal to, or greater than or equal to*, use a one-tailed test. If the question concerns *different from, or changed from*, use a two-tailed test.

EXERCISES 11.5

Self-Check Exercises

SC 11-7 Given a sample variance of 127 from a set of nine observations, construct a 95 percent confidence interval for the population variance.

SC 11-8 A production manager feels that the output rate of experienced employees is surely greater than that of new employees, but he does not expect the variability in output rates to differ for the two groups. In previous output studies, it has been shown that the average unit output per hour for new employees at this particular type of work is 20 units per hour with a variance of 56 units squared. For a group of 20 employees with 5 years' experience, the average output for this same type of work is 30 units per hour, with a sample variance of 28 units squared. Does the Variability in output appear to differ at the two experience levels? Test the hypotheses at the 0.05 significance level.

Basic Concepts

- 11-39** A sample of 20 observations from a normal distribution has a mean of 37 and a variance of 12.2. Construct a 90 percent confidence interval for the true population variance.
- 11-40** The standard deviation of a distribution is hypothesized to be 50. If an observed sample of 30 yields a sample standard deviation of 57, should we reject the null hypothesis that the true standard deviation is 50? Use the 0.05 level of significance.
- 11-41** Given a sample standard deviation of 6.4 from a sample of 15 observations, construct a 90 percent confidence interval for the population variance.

Applications

- 11-42** A telescope manufacturer wants its telescopes to have standard deviations in resolution to be significantly below 2 when focusing on objects 500 light-years away. When a new telescope is used to focus on an object 500 light-years away 30 times, the sample standard deviation turns out to be 1.46. Should this telescope be sold?
- State explicit null and alternative hypotheses.
 - Test your hypotheses at the $\alpha = 0.01$ level.
 - State an explicit conclusion.
- 11-43** MacroSwift has designed a new operating system that will revolutionize the computing industry. The only problem is, the company expects the average amount of time required to learn the software to be 124 hours. Even though this is a long educational time, the company is truly concerned with the variance of the learning time. Preliminary data indicate the variance is 171 hours squared. Recent testing of 25 people found an average learning time of 123 hours and a sample variance of 196.5 hours squared. Do these data indicate the variability in learning time is different from the previous estimate? Test your hypotheses at the 0.02 significance level.
- 11-44** A psychologist is aware of studies showing that the variability of attention spans of 5-year-olds can be summarized by $\sigma^2 = 64$ minutes squared. She wonders whether the attention span of 6-year-olds is different. A sample of twenty 6-year-olds gives $s^2 = 28$ minutes squared.
- State explicit null and alternative hypotheses.
 - Test your hypotheses at the $\alpha = 0.05$ level.
 - State an explicit conclusion.

- 11-45** In checking its cars for adherence to emissions standards set by the government, an automaker measured emissions of 30 cars. The average number of particles of pollutants emitted was found to be within the required levels, but the sample variance was 50. Find a 90 percent confidence interval for the variance in emission particles for these cars.
- 11-46** A bank is considering ways to reduce the costs associated with passbook savings accounts. The bank has found that the variance in the number of days between account transactions for passbook accounts is 80 days squared. The bank wants to reduce the variance by discouraging the present use of accounts for short-term storage of cash. Therefore, after implementing a new policy that penalizes the customer with a service charge for withdrawals more than once a month, the bank decides to test for a change in the variance of days between account transactions. From a sample of 25 savings accounts, the bank finds the variance between transactions to be 28 days squared. Is the bank justified in claiming that the new policy reduces the variance of days between transactions? Test the hypotheses at the 0.05 level of significance.
- 11-47** Sam Bogart, the owner of the Play-It-Again Stereo Company, offers 1-year warranties on all the stereos his company sells. For the 30 stereos that were serviced under the warranty last year, the average cost to fix a stereo was \$75 and sample standard deviation was \$15. Calculate a 95 percent confidence interval for the true standard deviation of the cost of repair. Sam has decided that unless the true standard deviation is less than \$20, he will buy his stereos from a different wholesaler. Help Sam test the appropriate hypotheses, using a significance level of 0.01. Should he switch wholesalers?

Worked-Out Answers to Self-Check Exercises

SC 11-7 For a 95 percent confidence interval with 8 degrees of freedom:

$$\sigma_L^2 = \frac{(n-1)s^2}{\chi_u^2} = \frac{8(127)}{17.535} = 57.941$$

$$\sigma_u^2 = \frac{(n-1)s^2}{\chi_L^2} = \frac{8(127)}{2.180} = 466.055$$

Thus, the confidence interval is (57.941, 466.055).

SC 11-8 For testing $H_0: \sigma^2 = 56$ versus $H_1: \sigma^2 \neq 56$ at $\alpha = 0.05$, the limits of the acceptance region are

$$\chi^2 = 8.907 \quad \text{and} \quad \chi^2 = 32.852$$

The observed $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19(28)}{56} = 9.5$, so we do not reject H_0 ; the variability is not significantly different.

11.6 INFERENCES ABOUT TWO POPULATION VARIANCES

In Chapter 9, we saw several situations in which we wanted to compare the means of two different populations. Recall that we did this by looking at the *difference* of the means of two samples drawn from those populations. Here, we want to compare the variances of two populations. However, rather than looking at the *difference* of the two sample variances, it turns out to be more convenient if we look at their *ratio*. The next two examples show how this is done.

**Comparing the variances
of two populations**

A One-Tailed Test of Two Variances

A prominent sociologist at a large midwestern university believes that incomes earned by college graduates show much greater variability than the earnings of those who did not attend college. In order to test this theory, she dispatches two research assistants to Chicago to look at the earnings of these two populations. The first assistant takes a random sample of 21 college graduates and finds that their earnings have a sample standard deviation of $s_1 = \$17,000$. The second assistant samples 25 nongraduates and obtains a standard deviation in earnings of $s_2 = \$7,500$. The data of our problem can be summarized as follows:

$s_1 = 17,000$	← Standard deviation of first sample
$n_1 = 21$	← Size of first sample
$s_2 = 7,500$	← Standard deviation of second sample
$n_2 = 25$	← Size of second sample

Data for the problem

Because the sociologist theorizes that the earnings of college graduates are *more* variable than those of people not attending college, a one-tailed test is appropriate. She wishes to verify her theory at the 0.01 level of significance. We can formally state her hypotheses:

Why a one-tailed test is appropriate

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ (or } \sigma_1^2/\sigma_2^2 = 1\text{)} \quad \leftarrow \text{Null hypothesis: the two variances are the same}$$

Statement of the hypotheses

$$H_1: \sigma_1^2 > \sigma_2^2 \text{ (or } \sigma_1^2/\sigma_2^2 > 1\text{)} \quad \leftarrow \text{Alternative hypothesis: earnings of college graduates have more variance}$$

$$\alpha = 0.01 \quad \leftarrow \text{Level of significance for testing these hypotheses}$$

Description of the F statistic

F Ratio for Inferences about Two Variances

$$F = \frac{s_1^2}{s_2^2} \quad [11-15]$$

has an F distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_1 - 2$ degrees of freedom in the denominator.

In the earnings problem, we calculate the sample F statistic:

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ &= \frac{(17,000)^2}{(7,500)^2} \\ &= \frac{289,000,000}{56,250,000} \\ &= 5.14 \end{aligned} \quad [11-15]$$

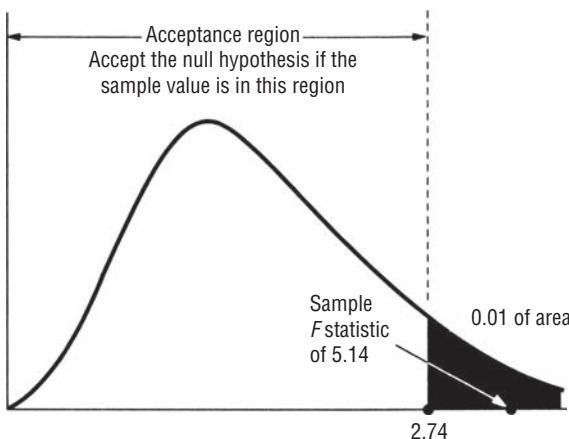


FIGURE 11-17 ONE-TAILED HYPOTHESIS TEST AT THE 0.01 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE SAMPLE F STATISTIC

For 20 degrees of freedom ($21 - 1$) in the numerator and 24 degrees of freedom ($25 - 1$) in the denominator, Appendix Table 6 tells us that the critical value separating the acceptance and rejection regions is 2.74. Figure 11-17 shows the acceptance region and the observed F statistic of 5.14. Our sociologist rejects the null hypothesis, and the sample data support her theory.

Interpreting the results

A word of caution about the use of Appendix Table 6 is necessary at this point. You will notice that the table gives values of the F statistic that are appropriate only for *upper-tailed* tests. How can we handle alternative hypotheses of the form $\sigma_1^2 < \sigma_2^2$ (or $\sigma_1^2/\sigma_2^2 < 1$)? This is easily done if we notice that $\sigma_1^2/\sigma_2^2 < 1$ is equivalent to $\sigma_2^2/\sigma_1^2 > 1$. Thus, all we need to do is calculate the ratio s_2^2/s_1^2 , which also has an F distribution (but with $n_2 - 1$ numerator degrees of freedom and $n_1 - 1$ denominator degrees of freedom), and then we can use Appendix Table 6. There is another way to say the same thing: **Whenever you are doing a one-tailed test of two variances, number the populations so that the alternative hypothesis has the form**

Handling lower-tailed tests in Appendix Table 6

$$H_1 : \sigma_1^2 > \sigma_2^2 \text{ (or } \sigma_1^2/\sigma_2^2 > 1\text{)}$$

and then proceed as we did in the earnings example.

A Two-Tailed Test of Two Variances

The procedure for a two-tailed test of two variances is similar to that for a one-tailed test. The only problem arises in finding the critical value in the lower tail. This is related to the problem about lower-tailed tests discussed in the last paragraph, and we will resolve it in a similar way.

Finding the critical value in a two-tailed test

One criterion in evaluating oral anesthetics for use in general dentistry is the variability in the length of time between injection and complete loss of sensation in the patient. (This is called the effect delay time.) A large pharmaceutical firm has just developed two new oral anesthetics, which it will market under the names Oralcaine and Novasthetic. From similarities in the chemical structure of the

two compounds, it has been predicted that they should show the same variance in effect delay time. Sample data from tests of the two compounds (which controlled other variables such as age and weight) are given in Table 11-16.

The company wants to test at a 2 percent significance level whether the two compounds have the same variance in effect delay time. Symbolically, the hypotheses are

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ (or } \sigma_1^2/\sigma_2^2 = 1\text{)}$$

$$H_1: \sigma_1^2 \neq \sigma_2^2 \text{ (or } \sigma_1^2/\sigma_2^2 \neq 1\text{)}$$

$$\alpha = 0.02$$

TABLE 11-16 EFFECT DELAY TIMES FOR TWO ANESTHETICS

Anesthetic	Sample Size (n)	Sample Variance (Seconds Squared) (s^2)
Oralcaine	31	1,296
Novasthetic	41	784

Statement of the hypotheses

← Null hypothesis: the two variances are the same

← Alternative hypothesis: the two variances are different

← Significance level of the test

To test these hypotheses, we again use Equation 11-15:

$$F = \frac{S_1^2}{S_2^2} = \frac{1,296}{784} = 1.65 \quad [11-15]$$

Calculating the F statistic

This statistic comes from an F distribution with $n - 1$ degrees of freedom in the numerator (30, in this case) and $n_2 - 1$ degrees of freedom in the denominator (40, in this case). Let us use the notation

$$F(n, d, \alpha)$$

Same useful notation for the test

to denote that value of F with n numerator degrees of freedom, d denominator degrees of freedom, and an area of α in the upper tail. In our problem, the acceptance region extends from $F(30, 40, 0.99)$ to $F(30, 40, 0.01)$, as illustrated in Figure 11-18.

We can get the value of $F(30, 40, 0.01)$ directly from Appendix Table 6; it is 2.20. However, the value of $F(30, 40, 0.99)$ is not in the table. Now $F(30, 40, 0.99)$ will correspond to a small value of s_1^2/s_2^2 ,

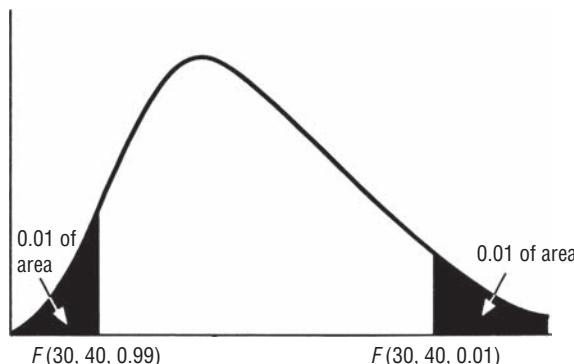


FIGURE 11-18 TWO-TAILED TEST OF HYPOTHESES AT THE 0.02 SIGNIFICANCE LEVEL

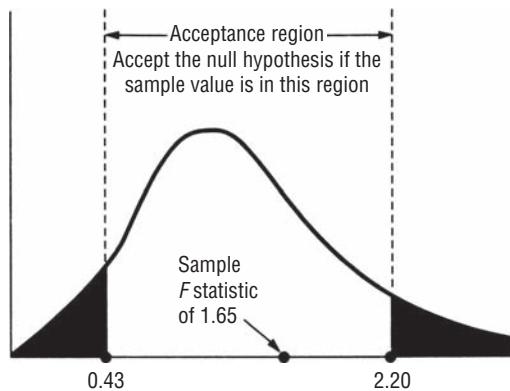


FIGURE 11-19 TWO-TAILED HYPOTHESIS TEST AT THE 0.02 LEVEL OF SIGNIFICANCE, SHOWING ACCEPTANCE REGION AND THE SAMPLE F STATISTIC

and hence to a *large* value of, s_2^2/s_1^2 , which is just the reciprocal of s_1^2/s_2^2 . Given the discussion on p. 591 about lower-tailed tests, we might suspect that

Lower-Tail Value of F for Two-Tailed Tests

$$F(n, d, \alpha) = \frac{1}{F(d, n, 1 - \alpha)} \quad [11-16]$$

and this turns out to be true. We can use this equation to find $F(30, 40, 0.99)$:

$$\begin{aligned} F(30, 40, 0.99) &= \frac{1}{F(40, 30, 0.01)} \\ &= \frac{1}{2.30} \\ &= 0.43 \end{aligned}$$

In Figure 11-19 we have illustrated the acceptance region for this hypothesis test and the observed value of F . We see there that the null hypothesis is accepted, so we conclude that the observed difference in the sample variances of effect delay times for the two anesthetics is not statistically significant.

Interpreting the results

HINTS & ASSUMPTIONS

This section has been about using an F test to compare the variances of two *populations* by looking at the ratio of the variances from two *samples*. Warning: Appendix Table 6 gives values of F that are appropriate for *upper-tailed* tests only. Hint: If you want to do a *lower-tailed* test, be sure to convert it to an upper-tailed test as shown on p. 591. And if you want to do a *two-tailed* test, use Equation 11-16 to convert an upper-tailed value from the table into the lower-tailed value needed for your test.

EXERCISES 11.6

Self-Check Exercises

- SC 11-9** A quality control supervisor for an automobile manufacturer is concerned with uniformity in the number of defects in cars coming off the assembly line. If one assembly line has significantly more variability in the number of defects, then changes have to be made. The supervisor has collected the following data:

	Number of Defects	
	Assembly Line A	Assembly Line B
Mean	10	11
Variance	9	25
Sample size	20	16

Does assembly line B have significantly more variability in the number of defects? Test at the 0.05 significance level.

- SC 11-10** Techgene, Inc., is concerned about variability in the number of bacteria produced by different cultures. If the cultures have significantly different variability in the number of bacteria produced, then experiments are messed up and some strange things get produced. (The management of the company gets understandably anxious when the scientists produce strange things.) The following data have been collected:

	Number of Bacteria (in thousands)									
Culture Type A	91	89	83	101	93	98	144	118	108	125
Culture Type B	62	76	90	75	88	99	110	140	145	130

- (a) Compute s_A^2 and s_B^2 .
 (b) State explicit null and alternative hypotheses and then test at the 0.02 significance level.

Basic Concepts

- 11-48** For two populations thought to have the same variance, the following information was found. A sample of 16 from population 1 exhibited a sample variance of 3.75, and a sample of 10 from population 2 had a variance of 5.38.
- (a) Calculate the F ratio for the test of equality of variances.
 - (b) Find the critical F value for the upper tail, using the 0.10 significance level.
 - (c) Find the corresponding F value for the lower tail.
 - (d) State the conclusion of your test.
- 11-49** In our study of comparisons between the means of two groups, it was noted that the most common form of the two-group t -test for the difference between two means assumes that the population variances for the two groups are the same. One experimenter, using a control condition and an experimental condition in his study of drug reaction, wished to verify that this assumption held, that is, that the treatment administered affected only the mean, not the variance of the variable under study. From his data, he calculated the variance of the experimental group to be 25.8 and that of the control group to be 20.6. The experimental group had 25 subjects, and the control group had 31. Can he proceed to use the t -test, which assumes equal variances for the two groups? Use $\alpha = 0.10$.

- 11-50** From a sample of 25 observations, the estimate of the standard deviation of the population was found to be 15.0. From another sample of 14 observations, the estimate was found to be 9.7. Can we accept the hypothesis that the two samples come from populations with equal variances, or must we conclude that the variance of the second population is smaller? Use the 0.01 level of significance.

Applications

- 11-51** Mr. Raj, an investor, has narrowed his search for a mutual fund down to the Oppy fund or the MLPFS fund. Oppy's rate of return is lower, but seems to be more stable than MLPFS's. If Oppy's variability in rate of return is significantly lower than MLPFS's, then he will invest his money there. If there is no significant difference in variability, he'll go with MLPFS. To make a decision, Mr. Raj has taken a sample of 21 monthly rates of return for both firms. For Oppy, the *standard deviation* was 2, and for MLPFS, the standard deviation was 3. Which firm should Mr. Raj invest in? Test at the $\alpha = 0.05$ level.
- 11-52** An insurance company is interested in the length of hospital-stays for various illnesses. The company has randomly selected 20 patients from hospital A and 25 from hospital B who were treated for the same ailment. The amount of time spent in hospital A had an average of 2.4 days with a standard deviation of 0.6 day. The treatment time in hospital B averaged 2.3 days with a standard deviation of 0.9 day. Do patients at hospital A have significantly less variability in their recovery time? Test at a 0.01 significance level.
- 11-53** Nation's Broadcasting Company is interested in the number of people who tune in to their hit shows *Buddies* and *Ride to Nowhere*; more importantly, the company is very concerned in the variability in the number of people who watch the shows. Advertisers want consistent viewers in hopes that consistent prolonged advertising will help to sell a product. Data are given below (in millions of viewers) for the past few months.

Number of Viewers (in millions)										
Buddies	57.4	62.6	54.6	52.4	60.5	61.8	71.4	67.5	62.6	58.4
Ride to Nowhere	64.5	58.2	39.5	24.7	40.2	41.6	38.4	33.6	34.4	37.8

- (a) Compute $S_{BUDDIES}^2$ and S_{RIDE}^2 .
- (b) State explicit hypotheses to determine whether the variability is the same between the two populations. Test at a 0.10 significance level.
- 11-54** The HAL Corporation is about to unveil a new, faster personal computer, PAL, to replace its old model, CAL. Although PAL is faster than CAL on average, PAL's processing speed seems more variable. (Processing speed depends on the program being run, the amount of input, and the amount of output.) Two samples of 25 runs, covering the range of jobs expected, were submitted to PAL and CAL (one sample to each). The results were as follows:

	Processing Time (in hundredths of a second)	
	PAL	CAL
Mean	50	75
Standard deviation	20	10

At the 0.05 level of significance, is PAL's processing speed significantly more variable than CAL's?

- 11-55** Two brand managers were in disagreement over the issue of whether urban homemakers had greater variability in grocery shopping patterns than did rural homemakers. To test their conflicting ideas, they took random samples of 70 homemakers from urban areas and 60 homemakers from rural areas. They found that the variance in days squared between shopping visits for urban homemakers was 14 and the sample variance for the rural homemakers was 3.5. Is the difference between the variances in days between shopping visits significant at the 0.01 level?
- 11-56** Two competing ice cream stores, Yum-Yum and Goody, both advertise quarter-pound scoops of ice cream. There is some concern about the variability in the serving sizes, so two members of a local consumer group have sampled 25 scoops of Yum-Yum's ice cream and 11 scoops of Goody's. Of course, both members now have stomachaches, so you must help them out. Is there a difference in the variance of ice cream weights between Yum-Yum and Goody? The following data have been collected. Test at the 0.10 level.

Scoop Weight (in hundredths of a pound)		
	Yum-Yum	Goody
Mean	25	25
Variance	16	10

Worked-Out Answers to Self-Check Exercises

SC 11-9 $H_0: \sigma_B^2 = \sigma_A^2$

$H_1: \sigma_B^2 > \sigma_A^2$

Observed $F = \frac{s_B^2}{s_A^2} = \frac{25}{9} = 2.778$

$F_{CRIT} = F_{0.05}(15, 19) = 2.23$

Thus, we reject H_0 ; assembly line B does have significantly more variability in the number of defects, so some changes have to be made. (Note: We are just checking for uniformity here; the cars could be uniformly bad.)

SC 11-10 (a) $s_A^2 = 423.4$ $s_B^2 = 755.818$

(b) $H_0: \sigma_A^2 = \sigma_B^2$

$H_1: \sigma_A^2 \neq \sigma_B^2$

Observed $F = \frac{s_B^2}{s_A^2} = \frac{423.4}{755.818} = 0.56$

$F_{0.01}(10, 10) = 4.85$

$F_{0.99}(10, 10) = \frac{1}{F(10, 10, 0.01)} = \frac{1}{4.85} = 0.21$

Thus, accept H_0 ; management doesn't have to worry about strange things in the laboratory.

STATISTICS AT WORK

Loveland Computers

Case 11: Chi-Square and Anova Tom Hodges had been supervisor of Loveland Computers' technical support team for a little over a year. Like many computer suppliers, Loveland contracted with a nationwide service company to provide 1 year of on-site repair. This guarantee was important in inducing customers to buy computers by phone. But Loveland had found that more than 90 percent of customers' problems could be solved by simply reading the instruction manuals that were packed with each machine, and 95 percent of all problems could be "talked through" if customers were encouraged to call customer service before seeking on-site repair. To save on warranty costs, Loveland had invested heavily in its customer-support center, where as many as 24 staff members would respond to customers' calls.

The customer-support staff were of two types. Most of the staff did not have much background in computers. These first-level support staff had been recruited for their telephone skills and had been trained internally to run through a routine checklist for common problems. When a customer's problem couldn't be corrected with the standard protocol, or when a customer called in with a "difficult" question, the call was transferred to a technician. Some of the technicians were full-time employees, but Hodges had found that plenty of part-time help could be found by recruiting students from the local engineering and computer science graduate programs. To suit their class schedule, most were scheduled to work a late shift, beginning at 4 P.M.

Examples of the kinds of problems that first-level support staff handled included talking customers through loading programs from floppy disks onto the hard drive and helping them check cable connections. The technicians handled problems such as the incompatibility of some "memory-resident" programs, and how to recover "lost" data.

The heads of several departments were meeting together to plan a strategy for improving telephone support. Loveland's support rating had slipped from "excellent" to "good" in a recent poll conducted by a marketing research firm. Walter Azko sent Lee along to "sit in on the meeting and see if you can be of any help."

Margot Derby, head of marketing, began the meeting with an air of finality: "Tom, the problem's obvious. When we call people back who've written complaint letters, they say they can never get through to a technician. They talk with the first-level support staff and then hold forever. It's obvious that it's business customers who are most likely to have a 'difficult' question that's beyond the scope of the first-level staff. You just need to schedule more technicians on the early shift."

Hodges replied, "On the contrary, Margot. It's the home users who need to talk with the technicians, so most of those calls come in on the late shift. They come up with these rocket-scientist questions while they're playing with their machines after work. In any case, the technicians are keeping busy on the late shift—I get a printout of their total time on the phone."

"Yes, but I'll bet that if you look at the average call time, it goes way up in the evening. I think your technicians are just chatting with the customers to fill in time."

"Well, we clearly need to know when the 'difficult' calls typically come in," said Lee, hoping to turn the discussion in a more productive direction. "Because no one ever talks to a technician without talking to a first-level service rep, we can have the first-level support staff assign each question to the easy or difficult category and gather data for each shift. Then, we can test to see if there really are more technical questions on the day shift or the late shift."

"Don't forget that it's my business customers who have more of the difficult questions," said Margot.

"I still think you're wrong on that. And, by the way, I have a gut feeling that the day of the week makes things different," added Tom. "We get a lot of technician-level calls early in the week but not toward the weekend."

Study Questions: In what format should the data be tabulated? Which statistical test might be useful if Lee just focuses on the shift issue (and sets aside the comments about business customers and the day of the week)? And which technique would be most useful for examining the effects of customer type, shift, and day of the week? What might distort the data that Lee asks the customer-support group to collect?

CHAPTER REVIEW

Terms Introduced in Chapter 11

Analysis of Variance (ANOVA) A statistical technique used to test the equality of three or more sample means and thus make inferences as to whether the samples come from populations having the same mean.

Between-Column Variance An estimate of the population variance derived from the variance among the sample means.

Chi-Square Distribution A family of probability distributions, differentiated by their degrees of freedom, used to test a number of different hypotheses about variances, proportions, and distributional goodness of fit.

Contingency Table A table having R rows and C columns. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body of the table are the frequencies with which each variable combination occurred.

Expected Frequencies The frequencies we would expect to see in a contingency table or frequency distribution if the null hypothesis is true.

F Distribution A family of distributions differentiated by two parameters (df-numerator, df-denominator), used primarily to test hypotheses regarding variances.

F Ratio A ratio used in the analysis of variance, among other tests, to compare the magnitude of two estimates of the population variance to determine whether the two estimates are approximately equal; in ANOVA, the ratio of between-column variance to within-column variance is used.

Goodness-of-Fit Test A statistical test for determining whether there is a significant difference between an observed frequency distribution and a theoretical probability distribution hypothesized to describe the observed distribution.

Grand Mean The mean for the entire group of subjects from all the samples in the experiment.

Test of Independence A statistical test of proportions of frequencies to determine whether membership in categories of one variable is different as a function of membership in the categories of a second variable.

Within-Column Variance An estimate of the population variance based on the variances within the k samples, using a weighted average of the k sample variances.

Equations Introduced in Chapter 11

11-1 $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ p. 536

This formula says that the *chi-square statistic* (χ^2) is equal to the sum (Σ) we will get if we

1. Subtract the expected frequencies, f_e , from the observed frequencies, f_o , for each category of our contingency table.
2. Square each of the differences.
3. Divide each squared difference by f_e .
4. Sum all the results of step 3.

11-2 Number of degrees of freedom = (number of rows – 1) (number of columns – 1) p. 537
To calculate number of *degrees of freedom in a chi-square test of independence*, multiply the number of rows (less 1) times the number of columns (less 1).

11-3 $f_e = \frac{RT \times CT}{n}$ p. 541

With this formula, we can calculate the expected frequency for any cell in a contingency table. RT is the row total for the row containing the cell, CT is the column total for the column containing the cell, and n is the total number of observations.

11-4 $s_x^2 = \frac{\sum (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$ p. 558

To calculate the *variance among the sample means*, use this formula.

11-5 $\sigma^2 = \sigma_x^2 \times n$ p. 558

The *population variance* is equal to the product of the square of the standard error of the mean and the sample size.

11-6 $\hat{\sigma}_b^2 = \frac{\sum n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$ p. 559

One estimate of the population variance (the between-column variance) can be obtained by using this equation. We obtain this equation by first substituting s_x^2 for σ_x^2 in Equation 11-5, and then by weighting each $(\bar{x}_j - \bar{\bar{x}})^2$ by its own appropriate sample size (n_j).

11-7 $\hat{\sigma}_w^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) s_j^2$ p. 560

A second estimate of the population variance (the within-column variance) can be obtained from this equation. This equation uses a weighted-average of all the sample variances. In this formulation, $n_T = \sum n_j$, the total sample size.

11-8 $F = \frac{\text{first estimate of the population variance}}{\text{based on the variance among the sample means}} \quad \frac{\text{second estimate of the population variance}}{\text{based on the variances within the samples}}$ p. 561

This ratio is the way we can compare the two estimates of the population variance, which we calculated in Equations 11-6 and 11-7. In a hypothesis test based on an F distribution, we are more likely to accept the null hypothesis if this *F ratio* or *F statistic* is near to the value of 1. As the F ratio increases, the more likely it is that we will reject the null hypothesis.

11-9 $F = \frac{\text{between-column variance}}{\text{within-column variance}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_w^2}$ p. 561

This restates Equation 11-8, using statistical shorthand for the numerator and the denominator of the F ratio.

11-10 Number of degrees of freedom in the *numerator* of the F ratio = (number of samples – 1) p. 562

To do an analysis of variance, we calculate the number of *degrees of freedom in the between-column variance* (the numerator of the F ratio) by subtracting 1 from the number of samples collected.

11-11 Number of degrees of freedom in the *denominator* of the F ratio = $\sum(n_j - 1) = n_T - k$ p. 563

We use this equation to calculate the number of degrees of freedom in the denominator of the F ratio. This turns out to be the total sample size, n_T , minus the number of samples, k .

11-12 $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ p. 583

With a population variance of σ^2 , the χ^2 statistic given by this equation has a chi-square distribution with $n - 1$ degrees of freedom. This result is exact if the population is normal, but even in samples from nonnormal populations, it is often a good approximation.

11-13 $\sigma^2 = \frac{(n-1)s^2}{\chi^2}$ p. 584

To get a confidence interval for σ^2 , we solve Equation 11-12 for σ^2 .

11-14 $\sigma_L^2 = \frac{(n-1)s^2}{\chi_u^2} \leftarrow \text{Lower confidence limit}$ p. 584

$$\sigma_u^2 = \frac{(n-1)s^2}{\chi_L^2} \leftarrow \text{Upper confidence limit}$$

These formulas give the lower and upper confidence limits for a confidence interval for σ^2 . (Notice that because χ^2 appears in the denominator, we use χ_u^2 to find σ_u^2 and χ_L^2 to find σ_L^2 .)

11-15 $F = \frac{s_1^2}{s_2^2}$ p. 590

This ratio has an F distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator. (This assumes that the two populations are reasonably well described by normal distributions.) It is used to test hypotheses about two population variances.

11-16 $F(n, d, \alpha) = \frac{1}{F(d, n, 1 - \alpha)}$ p. 593

Appendix Table 6 gives values of F for upper-tailed tests only, but this equation enables us to find appropriate values of F for lower-tailed and two-tailed tests.

Review and Application Exercises

- 11-57** The post office is concerned about the variability in the number of days it takes a letter to go from the east coast to the west coast. A sample of letters was mailed from the east coast, and the time taken for the letters to arrive at their address on the west coast was recorded. The following data were collected:

Mailing Time (in days)									
2.2	1.7	3.0	2.9	1.9	3.1	4.2	1.5	4.0	2.5

Find a 90 percent confidence interval for the variance in mailing times.

- 11-58** For the following contingency table, calculate the observed and expected frequencies and the chi-square statistic. State and test the appropriate hypotheses at the 0.05 significance level.

Attitude Toward Social Legislation			
Occupation	Favor	Neutral	Oppose
Blue-collar	19	16	37
White-collar	15	22	46
Professional	24	11	32

- 11-59** Marketers know that tastes differ in various regions of the country. In the rental car business, an industry expert has given the opinion that there are strong regional preferences for size of car and quotes the following data in support of that view:

Region of Country				
Preferred Car Type	Northeast	Southeast	Northwest	Southwest
Full-size	105	120	105	70
Intermediate	120	100	130	150
All other	25	30	15	30

- (a) State the appropriate null and alternative hypotheses.
- (b) Do the data support the expert's opinion at the 0.05 significance level?
- (c) What about at the 0.20 significance level?

- 11-60** What probability distribution is used in each of these types of statistical tests?

- (a) Comparing two population proportions.
- (b) Value of a single population variance.
- (c) Comparing three or more population means.
- (d) Comparing two population means from small, dependent samples.

- 11-61** What probability distribution is used in each of these types of statistical tests?

- (a) Comparing the means of two small samples from populations with unknown variances.
- (b) Comparing two population variances.
- (c) Value of a single population mean based on large samples.
- (d) Comparing three or more population proportions.

- 11-62** Retail stores set prices, but manufacturers have an interest in final retail price as this is part of their promotion strategy. The marketing manager for Brand C ballpoint pens complains that excessive price-cutting by stores results in the perception of Brand C as an "off brand." The sales manager replies that, "Everyone discounts—all the brands—to some extent." During sales calls, they collected data on the final sales price for four brands of pens, including their

own, from five different stores. At the 0.05 confidence level, is there significant variation in price between the brands?

Price (in cents)			
Brand A	Brand B	Brand C	Brand D
61	52	47	67
55	58	52	63
57	54	49	68
60	55	49	59
62	58	57	65

- 11-63** An outdoor advertising company must know whether significantly different traffic volumes pass three billboard locations in Newark because the company charges different rates for different traffic volumes. The company measures the volume of traffic at the three locations during randomly selected 5-minute intervals. The table shows the data gathered. At the 0.05 level of significance, are the volumes of traffic passing the three billboards the same?

Volume of Traffic								
Billboard 1	30	45	26	44	18	38	42	29
Billboard 2	29	38	36	21	36	18	17	30
Billboard 3	32	44	40	43	24	28	18	32

- 11-64** An investor is interested in seeing whether there are significant differences in the rates of return on stocks, bonds, and mutual funds. He has taken random samples of each type of investment and has recorded the following data.

Rate of Return (percent)						
Stocks	2.0	6.0	2.0	2.1	6.2	2.9
Bonds	4.0	3.1	2.2	5.3	5.9	
Mutual funds	3.5	3.1	2.9	6.0		

- (a) State null and alternative hypotheses.
- (b) Test your hypotheses at the 0.05 significance level.
- (c) State an explicit conclusion.

- 11-65** For the following contingency table:

- (a) Construct a table of observed and expected frequencies.
- (b) Calculate the chi-square statistic.
- (c) State the null and alternative hypotheses.
- (d) At a 0.05 level of significance, should the null hypothesis be rejected?

Church	Income Level		
	Attendance	Low	Middle
Attendance	Low	Middle	High
Never	27	48	15
Occasional	25	63	14
Regular	22	74	12

- 11-66** Quick Logistic Company (QLC) is a national level logistic company. QLC uses three modes of transportation for fetching goods to the desired destinations: Heavy-Duty Trucks, Buses and Medium Capacity Mobile Vans. QLC divides the order received for shipping into two categories: ‘charted’ and ‘contracted’, depending upon the choice of the customers. The differences between the two categories are the cost of transportation, reliability and the locking period of the couriered goods. ‘Charted’ mode involves higher cost; longer locking period but chances of the safe delivery of the couriered goods is much higher. QLC collected the data regarding frequency of the shipments sent last month. Analyze the data at 10% level of significance and comment whether three types of shipments are equally likely to be used for the category of ‘charted’.

	Heavy Duty Trucks	Mobile Vans	Buses
Charted	12	13	11
Contracted	18	7	4

- 11-67** For the following contingency table:
- Construct a table of observed and expected frequencies.
 - Calculate the chi-square statistic.
 - State the null and alternative hypotheses.
 - At a 0.01 level of significance, should the null hypothesis be rejected?

Type of Car Driven	Age Group			
	16–21	22–30	31–45	46+
4 × 4 Off Road	19	23	15	2
Sports car	9	14	11	7
Compact	6	8	7	9
Midsize	11	13	19	24
Full size	9	13	22	26

- 11-68** Swami Zhami claims to be psychic. He says he can correctly guess the suit (diamonds, clubs, hearts, spades) of a randomly chosen card with probability 0.5. Because the cards are chosen randomly from a big pile, we assume that Zhami’s guesses are independent. On 100 randomly chosen days, Zhami made 10 guesses, and the number of correct guesses was recorded. We want to see whether the number of correct guesses is binomially distributed with $n = 10$, $p = 0.5$. The following data have been collected:

Number of correct guesses per day	0–2	3–5	6–8	9–10
Frequency of number of correct guesses	50	47	2	1

- State explicit null and alternative hypotheses.
- Test your hypotheses. Use $\alpha = 0.10$.
- If Zhami has no psychic power, then he should have a probability of 0.25 of guessing a card correctly. (Why?) See whether the number of correct guesses is distributed binomially with $n = 10$, $p = 0.25$.

- 11-69** There has been some sociological evidence that women as a group are more variable than men in their attitudes and beliefs. A large private research organization has conducted a survey of men’s attitudes on a certain issue and found the standard deviation on this attitude scale to be

16 points. A sociologist gave the same scale to a group of 30 women and found that the sample variance was 400 points squared. At the 0.01 significance level, is there reason to believe that women do indeed show greater variability on this attitude scale?

- 11-70** Jim Kreeg makes predictions about the number of baskets that will be made by his favorite basketball team. We are interested in testing whether his errors are normally distributed with mean 0 and variance 16. Using the following data, state explicit null and alternative hypotheses and test them at the $\alpha = 0.05$ level.

Error	≤ -7	-6 to 0	1 to 6	≥ 7
Number of predictions	5	45	45	5

- 11-71** Psychologists have often wondered about the effects of stress and anxiety on test performance. An aptitude test was given to two randomly chosen groups of 18 college students, one group in a nonstressful situation and the other in a stressful situation. The experimenter expects the stress treatment to increase the variance of scores on the test because he feels some students perform better under stress while others experience adverse reactions to stress. The variances computed for the two groups are $s_1^2 = 23.9$ for the non-stress group and $s_2^2 = 81.2$ for the stress group. Was his hypothesis confirmed? Use the 0.05 level of significance to test the hypotheses.
- 11-72** In order to determine how women respond to brands of business attire, On the Job, an area boutique, surveyed groups of realtors, secretaries, entrepreneurs, and account executives about what fashion style they wore most often (A, B, C, D). The following data were collected:

Occupation	Style			
	A	B	C	D
Realtor	5	7	6	8
Secretary	10	15	12	8
Entrepreneur	8	12	21	25
Account Executive	12	14	20	25

At the 0.10 level of significance, test whether the style a woman prefers depends on her occupation.

- 11-73** In the development of new drugs for the treatment of anxiety, it is important to check the drugs' effects on various motor functions, one of which is driving. The Confab Pharmaceutical Company is testing four different tranquilizing drugs for their effects on driving skill. Subjects take a simulated driving test, and their scores reflect their errors. More severe errors lead to higher scores. The results of these tests produced the following table:

Drug 1	245	258	239	241
Drug 2	277	276	263	274
Drug 3	215	232	225	247
Drug 4	241	253	237	246
				226

At the 0.05 level of significance, do the four drugs affect driving skill differently?

- 11-74** James Clark has just purchased two paper mills and is concerned that they have significantly different variability in output, even though both plants produce about the same average

amount of paper each day. The following information was gathered to see whether Mr. Clark's concerns are justified. At the $\alpha = 0.02$ level of significance, do the two plants show the same variance in output?

Plant	<i>n</i>	s^2
Number 1	31	984 tons squared
Number 2	41	1,136 tons squared

- 11-75** Fuel costs are important to profitability in the airline business. A small regional carrier has been operating three types of aircraft and has collected the following cost data from its 14 planes, expressed as fuel cost (in cents) per available seat mile:

Type A	7.3	8.3	7.6	6.8	8.0
Type B	5.6	7.6	7.2		
Type C	7.9	9.5	8.7	8.3	9.4

At the 0.01 level of significance, can we conclude that there is no true difference between plane types in fuel costs?

- 11-76** Different news papers contain a special section "classified", having advertisements on different categories of products on Sundays. It is expected that readers will have more free-time to go through these advertisements. Mr. Rahman regularly follows this section in different news papers like Morning Times (MT), News Express (NE) and Voice of Nation (VON). On a particular Sunday, he observed the following numbers of for sale advertisements for Petrol Cars, Diesel Cars and Passenger Vans.

	MT	NE	VON
Petrol Cars	27	36	11
Diesel Cars	44	22	15
Passenger Vans	29	22	24

- (a) Test whether the proportions of three types of advertisements vary significantly among the three newspapers.
 (b) In your conclusion in the part (a) helpful in deciding which newspaper should be followed if you are interested in getting a good bargain for a car at a level of significance of 5%?
- 11-77** Sagar Krishna is the Chief Consultant of TXI Solutions, a firm specialized in providing consultation to business organizations on business solutions and strategic planning. Recently, he is working on a project related to the conflict situations resulting due to interactions among different departments/wings in an organization. He wants to check the perception that "Planning Issues" take more time to be resolved after deliberations than issues related to other functions. He has collected relevant data.

The following values are the number of weeks spent by different departments to arrive at a acceptable solution, in some of the successful organizations.

Using level of significance as 5%, analyze the data and comment on the research of Mr Sagar.

Planning Issues	Implementation Issues	Evaluation Issues
3.5	3.0	1.0
4.8	5.5	2.5
3.0	6.0	2.0
6.5	4.0	1.5
7.5	4.0	1.5
8.0	4.5	6.0
2.0	6.0	3.8
6.0	2.0	4.5
5.5	9.0	0.5
6.5	4.5	2.0
7.0	5.0	3.5
9.0	2.5	1.0
5.0	7.0	2.0
10.0		
6.0		

- 11-78** The following table shows the price/earnings ratios for 26 companies belonging to the different sector of the Indian Market. These 26 companies can be further classified into three categories: Marketing Companies, Financial Services Companies and Banking Companies. They are coded as industry code 1, 2 and 3 respectively. Analyze the data and comment which type of industry (Marketing Companies, Financial Services Companies and Banking Companies) dominates the market on the basis of price/earnings ratios.

Company	Industry Code	P/E	Company	Industry Code	P/E
A	1	21	N	2	17
B	1	12	O	2	15
C	1	23	P	2	21
D	1	15	Q	2	20
E	1	16	R	2	16
F	1	14	S	3	20
G	1	20	T	3	17
H	1	15	U	3	21
I	1	17	V	3	18
J	1	16	W	3	10
K	1	18	X	3	15
L	1	15	Y	3	20
M	2	21	Z	3	13

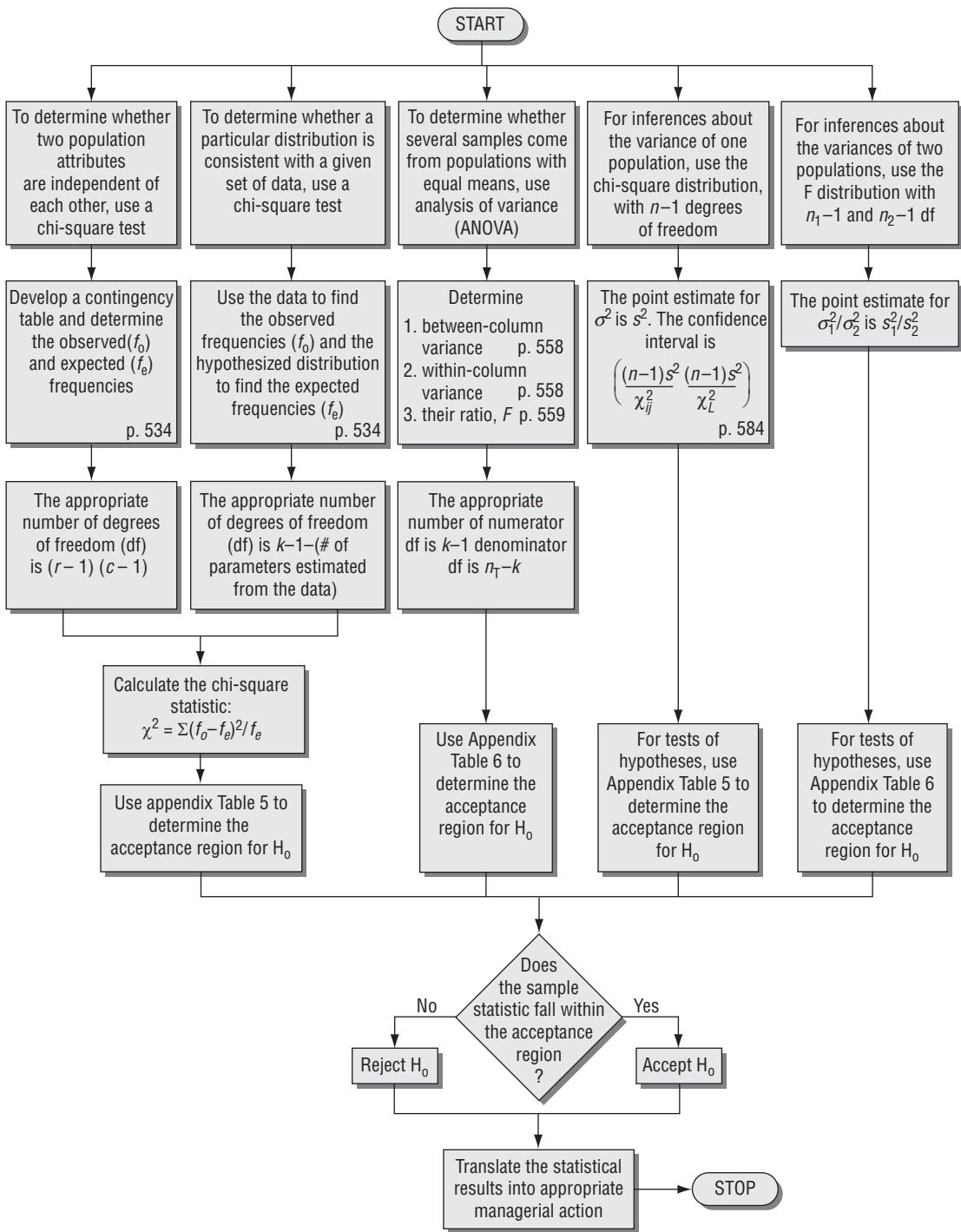


Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different educational groups. (Q9 & Q17)
2. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different professions. (Q9 & Q18)
3. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different age groups. (Q9 & Q14)
4. Test the hypothesis that the satisfaction level of the respondent with respect to the e-banking services provided by their bank depends upon the age of the respondent. (Q9 & Q14)
5. Test the hypothesis that the perception of the respondents towards the reliability of the e-banking services provided by their banks dependent upon the profession of the respondents. (Q8b & Q18)



Flow Chart: Chi-Square and Analysis of Variance



Simple Regression and Correlation

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn how many business decisions depend on knowing the specific relationship between two or more variables
 - To use scatter diagrams to visualize the relationship between two variables
 - To use regression analysis to estimate the relationship between two variables
 - To use the least-squares estimating equation to predict future values of the dependent variable
 - To learn how correlation analysis describes the degree to which two variables are linearly related to each other
 - To understand the coefficient of determination as a measure of the strength of the relationship between two variables
 - To learn limitations of regression and correlation analyses and caveats about their use
-

CHAPTER CONTENTS

12.1 Introduction	610
12.2 Estimation Using the Regression Line	617
12.3 Correlation Analysis	643
12.4 Making Inferences about Population Parameters	657
12.5 Using Regression and Correlation Analyses: Limitations, Errors, and Caveats	664

■ Statistics at Work	667
■ Terms Introduced in Chapter 12	667
■ Equations Introduced in Chapter 12	668
■ Review and Application Exercises	670
■ Flow Chart: Regression and Correlation	676

The vice president for research and development of a large chemical and fiber manufacturing company believes that the firm's annual profits depend on the amount spent on R&D. The new chief executive officer does not agree and has asked for evidence. Here are data for 6 years:

Year	Amount Spent on Research and Development (Millions)	Annual Profit (Millions)
1990	2	20
1991	3	25
1992	5	34
1993	4	30
1994	11	40
1995	5	31

The vice president for R&D wants an equation for predicting annual profits from the amount budgeted for R&D. With methods in this chapter, we can supply such a decision-making tool and tell him something about the accuracy he can expect in using it to make decisions. ■

12.1 INTRODUCTION

Every day, managers make personal and professional decisions that are based on predictions of future events. To make these forecasts, they rely on the relationship (intuitive and calculated) between what is already known and what is to be estimated. If decision makers can determine how the known is related to the future event, they can aid the decision-making process considerably. That is the subject of this chapter: how to determine the *relationship between variables*.

Relationship between variables

In Chapter 11, we used chi-square tests of independence to determine whether a statistical relationship existed between two variables. The chi-square test tells us *whether* there is such a relationship, but it does not tell us *what* that relationship is. **Regression and correlation analyses show us how to determine both the nature and the strength of a relationship between two variables.** We will learn to predict, with some accuracy, the value of an unknown variable based on past observations of that variable and others.

Difference between chi-square and topics in this chapter

The term *regression* was first used as a statistical concept in 1877 by Sir Francis Galton. Galton made a study that showed that the height of children born to tall parents tends to move back, or "regress," toward the mean height of the population. He designated the word *regression* as the name of the general process of predicting one variable (the height of the children) from another (the height of the parent). Later, statisticians coined the term *multiple regression* to describe the process by which several variables are used to predict another.

Origin of terms regression and multiple regression

In *regression analysis*, we shall develop an *estimating equation*—that is, a mathematical formula that relates the known variables to the unknown variable. Then, after we have learned the pattern of this relationship, we can apply *correlation analysis* to determine the degree to which the variables are related. Correlation analysis, then, tells us how well the estimating equation actually describes the relationship.

Development of an estimating equation

Types of Relationships

Regression and correlation analyses are based on the relationship, or association, between two (or more) variables. The known variable (or variables) is called the *independent variable(s)*. The variable we are trying to predict is the *dependent variable*.

Independent and dependent variables

Scientists know, for example, that there is a relationship between the annual sales of aerosol spray cans and the quantity of fluorocarbons released into the atmosphere each year. If we studied this relationship, “the number of aerosol cans sold each year” would be the independent variable and “the quantity of fluorocarbons released annually” would be the dependent variable.

Let’s take another example. Economists might base their predictions of the annual gross domestic product, or GDP, on the final consumption spending within the economy. Thus, “the final consumption spending” is the independent variable and “the GDP” is the dependent variable.

In regression, we can have only one dependent variable in our estimating equation. However, we can use more than one independent variable. Often when we add independent variables, we improve the accuracy of our prediction. Economists, for example, often add a second independent variable, “the level of investment spending,” to improve their estimate of the nation’s GDP.

Our two examples of fluorocarbons and GDP are illustrations of direct associations between independent and dependent variables. As the independent variable increases, the dependent variable also increases. In like manner, we expect the sales of a company to increase as the advertising budget increases. We can graph such a *direct relationship*, plotting the independent variable on the X -axis and the dependent variable on the Y -axis. We have done this in Figure 12-1(a). Notice how the line slopes up as X takes on larger and larger values. The slope of this line is said to be *positive* because Y increases as X increases.

Direct relationship between X and Y

Relationships can also be *inverse* rather than direct. In these cases, the dependent variable decreases as the independent variable increases. The government assumes that such an inverse association exists between a company’s increased annual expenditures for pollution-abatement devices and decreased pollution emissions. This type of relationship is illustrated in Figure 12-1(b), and is characterized by a *negative slope* (the dependent variable Y decreases as the independent variable X increases).

Inverse relationship between X and Y

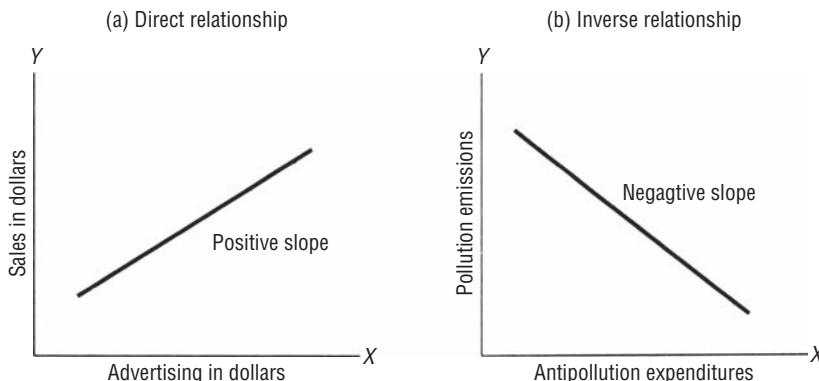


FIGURE 12-1 DIRECT AND INVERSE RELATIONSHIPS BETWEEN INDEPENDENT VARIABLE X AND DEPENDENT VARIABLE Y

We often find a causal relationship between variables; that is, the independent variable “causes” the dependent variable to change. This is the case in the antipollution example above. But in many cases, other factors cause the changes in both the dependent and the independent variables. We might be able to predict the sales of diamond earrings from the sales of new Cadillacs, but we could not say that one is caused by the other. Instead, we realize that the sales levels of both Cadillacs and diamond earrings are caused by another factor, such as the level of disposable income.

For this reason, it is important that you consider the relationships found by regression to be relationships of association but not necessarily of cause and effect. Unless you have specific reasons for believing that the values of the dependent variable are caused by the values of the independent variable(s), do not infer causality from the relationships you find by regression.

Relationships of association, not cause and effect

Scatter Diagrams

The first step in determining whether there is a relationship between two variables is to examine the graph of the observed (or known) data. This graph, or chart, is called a *scatter diagram*.

Scatter diagram

A scatter diagram can give us two types of information. Visually, we can look for patterns that indicate that the variables are related. Then, if the variables are related, we can see what kind of line, or estimating equation, describes this relationship.

We are going to develop and use a specific scatter diagram. Suppose a university admissions director asks us to determine whether any relationship exists between a student’s scores on an entrance examination and that student’s cumulative grade-point average (GPA) upon graduation. The administrator has accumulated a random sample of data from the records of the university. This information is recorded in Table 12-1.

To begin, we should transfer the information in Table 12-1 to a graph. Because the director wishes to use examination scores to predict success in college, we have placed the cumulative GPA (the dependent variable) on the vertical or *Y*-axis and the entrance examination score (the independent variable) on the horizontal or *X*-axis. Figure 12-2 shows the completed scatter diagram.

Transfer tabular information to a graph

At first glance, we can see why we call this a scatter diagram. The pattern of points results because each pair of data from Table 12-1 has been recorded as a single point. When we view all these points together, we can visualize the relationship that exists between the two variables. As a result, we can draw, or “fit,” a straight line through our scatter diagram to represent the relationship. We have done this in Figure 12-3. It is common to try to draw these lines so that an equal number of points lies on either side of the line.

Drawing, or “fitting,” a straight line through a scatter diagram

TABLE 12-1 STUDENT SCORES ON ENTRANCE EXAMINATIONS AND CUMULATIVE GRADE-POINT AVERAGES AT GRADUATION

Student	A	B	C	D	E	F	G	H
Entrance examination scores (100 = maximum possible score)	74	69	85	63	82	60	79	91
Cumulative GPA (4.0 = A)	2.6	2.2	3.4	2.3	3.1	2.1	3.2	3.8

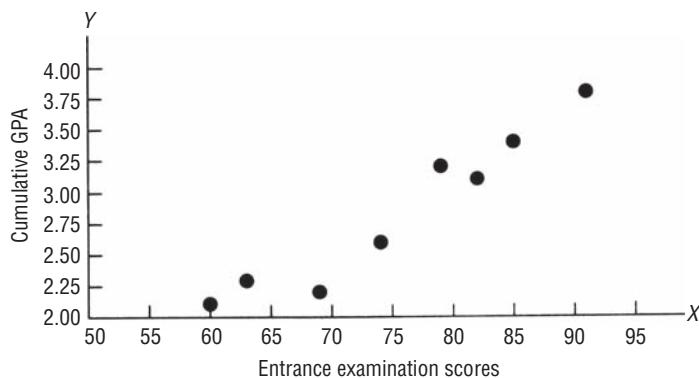


FIGURE 12-2 SCATTER DIAGRAM OF STUDENT SCORES ON ENTRANCE EXAMINATIONS PLOTTED AGAINST CUMULATIVE GRADE-POINT AVERAGES

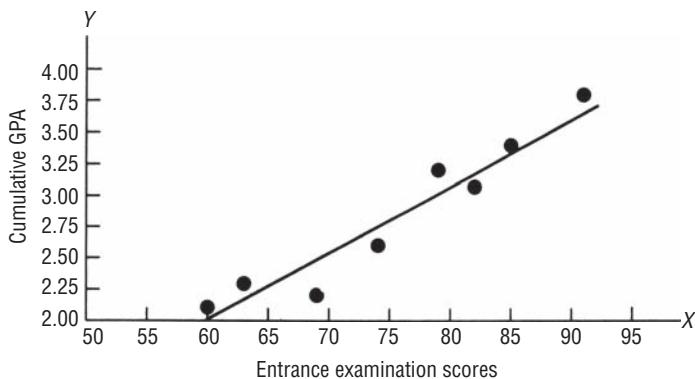


FIGURE 12-3 SCATTER DIAGRAM WITH STRAIGHT LINE REPRESENTING THE RELATIONSHIP BETWEEN X AND Y "FITTED" THROUGH IT

In this case, the line drawn through our data points represents a direct relationship because Y increases as X increases. Because the data points are relatively close to this line, we can say that there is a high degree of association between the examination scores and the cumulative GPA. In Figure 12-3, we can see that the relationship described by the data points is well described by a straight line. Thus, we can say that it is a *linear* relationship.

Interpreting our straight line

The relationship between X and Y variables can also take the form of a curve. Statisticians call such a relationship *curvilinear*. The employees of many industries, for example, experience what is called a "learning curve"; that is, as they produce a new product, the time required to produce one unit is reduced by some fixed proportion as the total number of units doubles. One such industry is aviation. Manufacturing time per unit for a new aircraft tends to decrease by 20 percent each time the total number of completed new planes doubles. Figure 12-4 illustrates the curvilinear relationship of this "learning curve" phenomenon.

Curvilinear relationships

The direction of the curve can indicate whether the curvilinear relationship is direct or inverse. The curve in Figure 12-4 describes an inverse relationship because Y decreases as X increases.

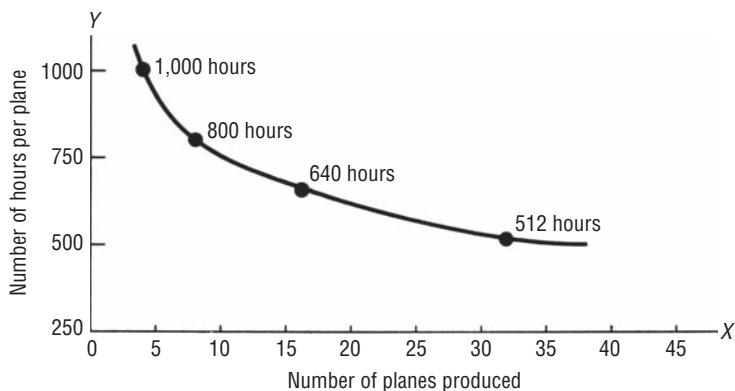


FIGURE 12-4 CURVILINEAR RELATIONSHIP BETWEEN NEW-AIRCRAFT CONSTRUCTION TIME AND NUMBER OF UNITS PRODUCED

To review the relationships possible in a scatter diagram, examine the graphs in Figure 12-5. Graphs (a) and (b) show direct and inverse linear relationships. Graphs (c) and (d) are examples of curvilinear relationships that demonstrate direct and inverse associations between variables, respectively. Graph (e) illustrates an inverse linear relationship with a widely scattered pattern of points. The wider scattering indicates that there is a lower degree of association between the

Review of possible relationships

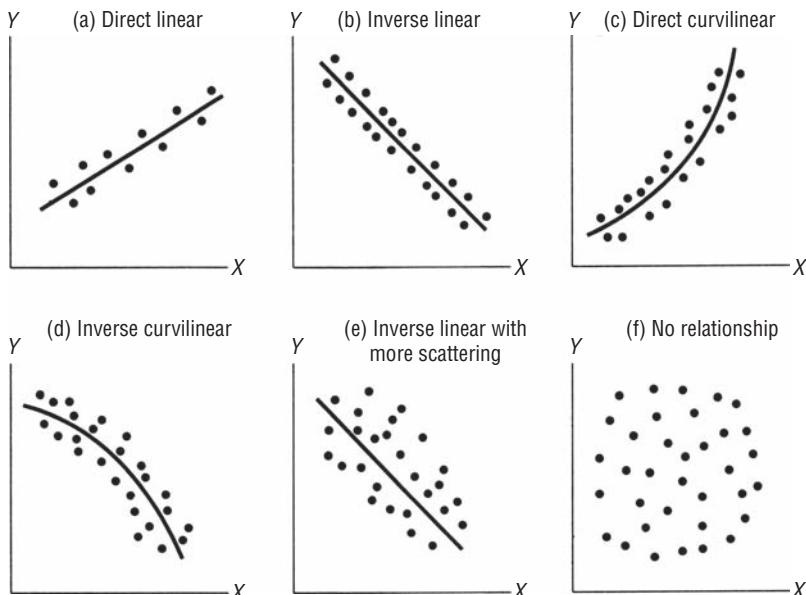


FIGURE 12-5 POSSIBLE RELATIONSHIPS BETWEEN X AND Y IN SCATTER DIAGRAMS

independent and dependent variables than there is in graph (b). The pattern of points in graph (f) seems to indicate that there is no linear relationship between the two variables; therefore, knowledge of the past concerning one variable does not allow us to predict future occurrences of the other.

EXERCISES 12.1

Self-Check Exercise

SC 12-1 An instructor is interested in finding out how the number of students absent on a given day is related to the mean temperature that day. A random sample of 10 days was used for the study. The following data indicate the number of students absent (ABS) and the mean temperature (TEMP) for each day.

ABS	8	7	5	4	2	3	5	6	8	9
TEMP	10	20	25	30	40	45	50	55	59	60

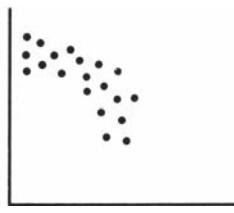
- (a) State the dependent (Y) variable and the independent (X) variable.
- (b) Draw a scatter diagram of these data.
- (c) Does the relationship between the variables appear to be linear or curvilinear?
- (d) What type of curve could you draw through the data?
- (e) What is the logical explanation for the observed relationship?

Basic Concepts

- 12-1** What is regression analysis?
- 12-2** In regression analysis, what is an estimating equation?
- 12-3** What is the purpose of correlation analysis?
- 12-4** Define direct and inverse relationships.
- 12-5** To what does the term *causal relationship* refer?
- 12-6** Explain the difference between linear and curvilinear relationships.
- 12-7** Explain why and how we construct a scatter diagram.
- 12-8** What is multiple-regression analysis?
- 12-9** For each of the following scatter diagrams, indicate whether a relationship exists and, if so, whether it is direct or inverse and linear or curvilinear.



(a)



(b)



(c)

Applications

- 12-10** A professor is trying to show his students the importance of quizzes even though 90 percent of the final grade is determined by exams. He believes that the higher the quiz grade, the higher

the final grade. A random sample of 15 students in his class was selected with the data given below:

	Quiz Average	Final Average
	59	65
	92	84
	72	77
	90	80
	95	77
	87	81
	89	80
	77	84
	76	80
	65	69
	97	83
	42	40
	94	78
	62	65
	91	90

- (a) State the dependent (Y) variable and the independent (X) variable.
- (b) Draw a scatter diagram of these data.
- (c) Does the relationship between the variables appear to be linear or curvilinear?
- (d) Does the professor's belief appear to be justified? Explain your reasoning.

12-11

William Hawkins, VP of personnel for International Motors, is working on the relationship between a worker's salary and absentee rate. Hawkins divided the salary range of International into twelve grades or levels (1 being the lowest grade, 12 the highest) and then randomly sampled a group of workers. He determined the salary grade for each worker and the number of days that employee had missed over the last 3 years.

Salary ranking	11	10	8	5	9	9	7	3
Absences	18	17	29	36	11	26	28	35
Salary ranking	11	8	7	2	9	8	6	3
Absences	14	20	32	39	16	26	31	40

Construct a scatter diagram for these data and indicate the type of relationship.

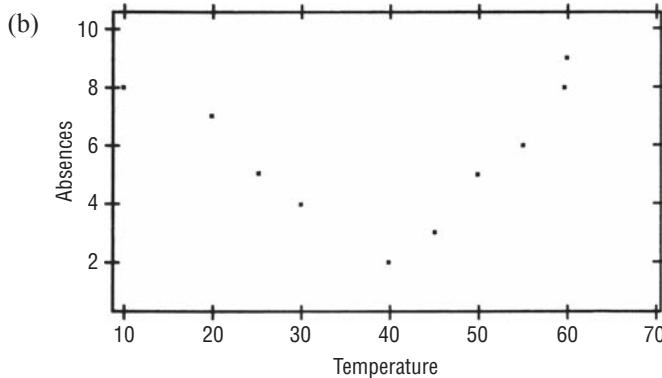
12-12

The National Institute of Environmental Health Sciences (NIEHS) has been studying the statistical relationships between many different variables and the common cold. One of the variables being examined is the use of facial tissues (X) and the number of days that cold symptoms were exhibited (Y) by seven people over a 12-month period. What relationship, if any seems to hold between the two variables? Does this indicate any causal effect?

X	2,000	1,500	500	750	600	900	1,000
Y	60	40	10	15	5	25	30

Worked-Out Answers to Self-Check Exercise

SC 12-1 (a) We want to see whether absences (ABS) depend on temperature (TEMP).



- (c) Curvilinear.
- (d) Quadratic curve (parabola).
- (e) When it is very cold and when it is very hot there are many absences. For moderate temperatures, there are not as many absences.

12.2 ESTIMATION USING THE REGRESSION LINE

In the scatter diagrams we have used to this point, the *regression lines* were put in place by fitting the lines visually among the data points. In this section, we shall learn how to calculate the regression line somewhat more precisely, using an equation that relates the two variables mathematically. Here, we examine only linear relationships involving two variables. We shall deal with relationships among more than two variables in the next chapter.

Calculating the regression line using an equation

The equation for a straight line where the dependent variable Y is determined by the independent variable X is:

Equation for a straight line

Equation for a Straight Line

$$Y = a + bX$$

[12-1]

Dependent variable → Y ← Independent variable
 Y-intercept → a ← Slope of the line bX

Using this equation, we can take a given value of X and compute the value of Y . The a is called the *Y-intercept* because its value is the point at which the regression line crosses the *Y-axis*—that is, the vertical axis. The b in Equation 12-1 is the slope of the line. It represents how much each unit change of the independent variable X changes the dependent variable Y . Both a and b are numerical *constants* because for any given straight line, their values do not change.

Interpreting the equation

Suppose we know that a is 3 and b is 2. Let us determine what Y would be for an X equal to 5. When we substitute the values of

Calculating Y from X using the equation for a straight line

a , b , and X in Equation 12-1, we find the corresponding value of Y to be

$$\begin{aligned} Y &= a + bX \\ &= 3 + 2(5) \\ &= 3 + 10 \\ &= 13 \leftarrow \text{Value for } Y \text{ given } X = 5 \end{aligned} \quad [12-1]$$

Using the Estimation Equation for a Straight Line

How can we find the values of the numerical constants, a and b ? Finding the values for a and b

To illustrate this process, let's use the straight line in Figure 12-6.

Visually, we can find a (the Y -intercept) by locating the point where the line crosses the Y -axis. In Figure 12-6, this happens where $a = 3$.

To find the slope of the line, b , we must determine how the dependent variable, Y , changes as the independent variable, X , changes. We can begin by picking two points on the line in Figure 12-6. Now, we must find the values of X and Y (the *coordinates*) of both points. We can call the coordinates of our first point (X_1, Y_1) and those of the second point (X_2, Y_2) . By examining Figure 12-6, we can see that $(X_1, Y_1) = (1, 5)$ and $(X_2, Y_2) = (2, 7)$. At this point, then, we can calculate the value of b using this equation:

The Slope of a Straight Line

$$b = \frac{Y_2 - Y_1}{X_2 - X_1} \quad [12-2]$$

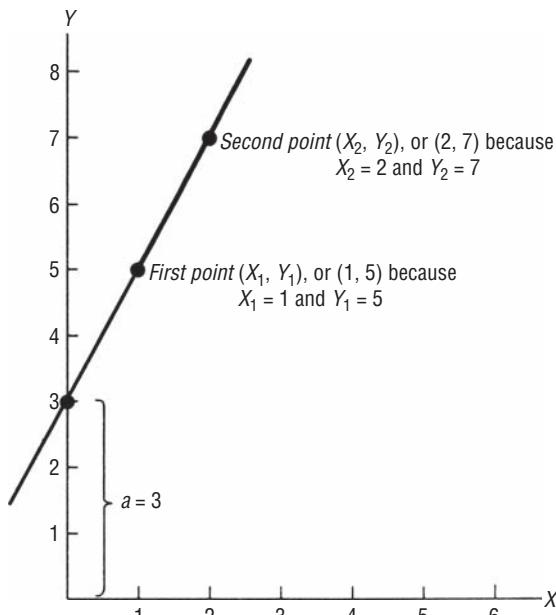


FIGURE 12-6 STRAIGHT LINE WITH A POSITIVE SLOPE, WITH THE Y -INTERCEPT AND TWO POINTS ON THE LINE DESIGNATED

$$\begin{aligned} b &= \frac{7-5}{2-1} \\ &= \frac{2}{1} \\ &= 2 \leftarrow \text{Slope of the line} \end{aligned}$$

In this manner, we can learn the values of the numerical constants, a and b , and write the equation for a straight line. The line in Figure 12-6 can be described by Equation 12-1, where $a = 3$ and $b = 2$. Thus

$$Y = a + bX \quad [12-1]$$

and

$$Y = 3 + 2X$$

Using this equation, we can determine the corresponding value of the dependent variable for any value of X . Suppose we wish to find the value of Y when $X = 7$. The answer would be

$$\begin{aligned} y &= a + bX \\ &= 3 + 2(7) \\ &= 3 + 14 \\ &= 17 \end{aligned} \quad [12-1]$$

If you substitute more values for X into the equation, you will notice that Y increases as X increases. Thus, the relationship between the variables is *direct* and the slope is *positive*.

Direct relationship; positive slope

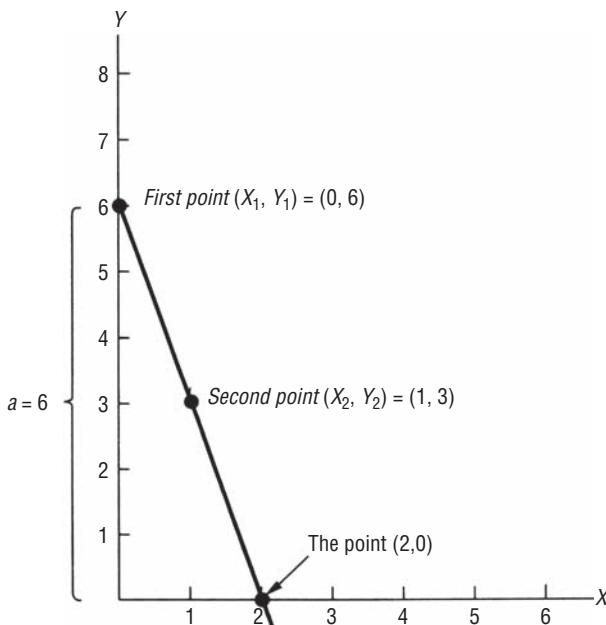
Now consider the line in Figure 12-7. We see that it crosses the Y -axis at 6. Therefore, we know that $a = 6$. If we select the two points where $(X_1, Y_1) = (0, 6)$ and $(X_2, Y_2) = (1, 3)$, we find that the slope of the line is

$$\begin{aligned} b &= \frac{Y_2 - Y_1}{X_2 - X_1} \\ &= \frac{3 - 6}{1 - 0} \\ &= \frac{-3}{1} \\ &= -3 \end{aligned} \quad [12-2]$$

Notice that when b is negative, the line represents an *inverse* relationship and the slope is *negative* (Y decreases as X increases). Now, with the numerical values of a and b determined, we can substitute them into the general equation for a straight line:

$$\begin{aligned} Y &= a + bX \\ &= 6 + (-3)X \\ &= 6 - 3X \end{aligned} \quad [12-1]$$

Inverse relationship; negative slope

**FIGURE 12-7 STRAIGHT LINE WITH A NEGATIVE SLOPE**

Assume that we wish to find the value of the dependent variable that corresponds to $X = 2$. Substituting into Equation 12-1, we get

$$\begin{aligned} Y &= 6 - (3)(2) \\ &= 6 - 6 \\ &= 0 \end{aligned}$$

Thus, when $X = 2$, Y must equal 0. If we refer to the line in Figure 12-7, we can see that the point $(2, 0)$ does lie on the line.

Finding Y given X

The Method of Least Squares

Now that we have seen how to determine the equation for a straight line, let's think about how we can calculate an equation for a line that is drawn through the middle of a set of points in a scatter diagram. How can we "fit" a line mathematically if none of the points lies on the line? To a statistician, the line will have a "good fit" if it *minimizes the error* between the estimated points on the line and the actual observed points that were used to draw it.

Fitting a regression line mathematically

Before we proceed, we need to introduce a new symbol. So far, we have used Y to represent the individual values of the observed points measured along the Y -axis. Now we should begin to use \hat{Y} (Y hat) to symbolize the individual values of the *estimated* points—that is, the points that lie on the estimating line. Accordingly, we shall write the equation for the estimating line as

Introduction of Y

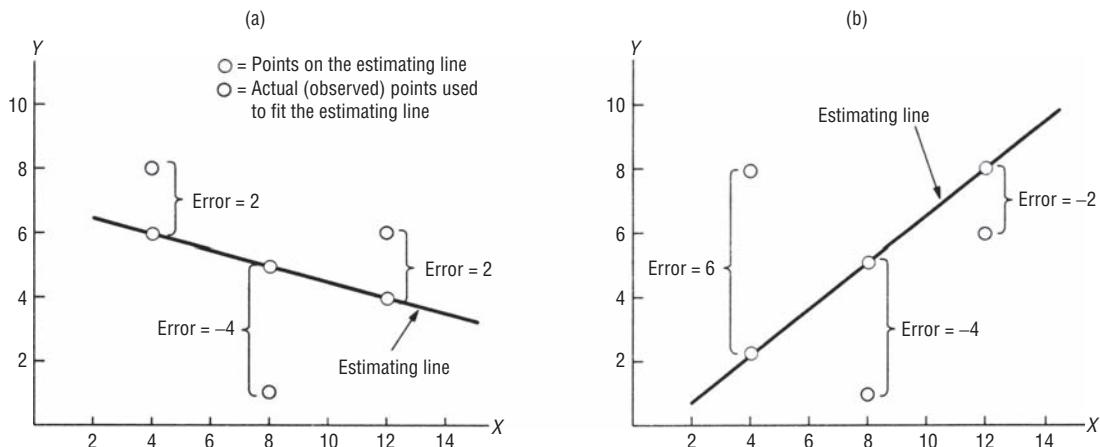


FIGURE 12-8 TWO DIFFERENT ESTIMATING LINES FITTED TO THE SAME THREE OBSERVED DATA POINTS, SHOWING ERRORS IN BOTH CASES

The Estimating Line

$$\hat{Y} = a + bX$$

[12-3]

In Figure 12-8, we have two estimating lines that have been fitted to the same set of three data points. These three given, or observed, data points are shown in black. Two very different lines have been drawn to describe the relationship between the two variables. Obviously, we need a way to decide which of these lines gives us a better fit.

One way we can “measure the error” of our estimating line is to *sum* all the individual differences, or errors, between the estimated points shown in color and the observed points shown in black. In Table 12-2, we have calculated the individual differences between the corresponding Y and \hat{Y} , and then we have found the sum of these differences.

Which line fits best?

Using total error to determine best fit

TABLE 12-2 SUMMING THE ERRORS OF THE TWO ESTIMATING LINES IN FIGURE 12-8

Graph (a) $Y - \hat{Y}$	Graph (b) $Y - \hat{Y}$
$8 - 6 = 2$	$8 - 2 = 6$
$1 - 5 = -4$	$1 - 5 = -4$
$6 - 4 = 2$	$6 - 8 = -2$
$\overline{0 \leftarrow \text{Total error}}$	$\overline{0 \leftarrow \text{Total error}}$

TABLE 12-3 SUMMING THE ABSOLUTE VALUES OF THE ERRORS OF THE TWO ESTIMATING LINES IN FIGURE 12-8

Graph (a) $ Y - \hat{Y} $	Graph (b) $ Y - \hat{Y} $
$ 8 - 6 = 2$	$ 8 - 2 = 6$
$ 1 - 5 = 4$	$ 1 - 5 = 4$
$ 6 - 4 = 2$	$ 6 - 8 = 2$
$\overline{8} \leftarrow \text{Total absolute error}$	$\overline{12} \leftarrow \text{Total absolute error}$

A quick visual examination of the two estimating lines in Figure 12-8 reveals that the line in graph (a) fits the three data points better than the line in graph (b).* However, our process of summing the individual differences in Table 12-2 indicates that both lines describe the data equally well (the total error in both cases is zero). Thus, we must conclude that the process of summing individual differences for calculating the error is not a reliable way to judge the goodness of fit of an estimating line.

The problem with adding the individual errors is the canceling effect of the positive and negative values. From this, we might deduce that the proper criterion for judging the goodness-of-fit would be to add the *absolute values* (the values without their algebraic signs) of each error. We have done this in Table 12-3. (The symbol for absolute value is two parallel vertical lines $||$.) Because the absolute error in graph (a) is smaller than the absolute error in graph (b), and because we are looking for the minimum absolute error, we have confirmed our intuitive impression that the estimating line in graph (a) is the better fit.

Using absolute value of error to measure best fit

On the basis of this success, we might conclude that minimizing the sum of the absolute values of the errors is the best criterion for finding a good fit. But before we feel too comfortable with it, we should examine a different situation.

In Figure 12-9, we again have two identical scatter diagrams with two different estimating lines fitted to the three data points. In Table 12-4, we have added the absolute values of the errors and found that the estimating line in graph (a) is a better fit than the line in graph (b). Intuitively, however, it appears that the line in graph (b) is the better fit line because it has been moved vertically to take the middle point into consideration. Graph (a) on the other hand, seems to ignore the middle point completely. So we would probably discard this second criterion for finding the best fit. Why? **The sum of the absolute values does not stress the magnitude of the error.**

It seems reasonable that the farther away a point is from the estimating line, the more serious is the error. We would rather have several small absolute errors than one large one, as we saw in the last example. **In effect, we want to find a way to “penalize” large absolute errors so that we can avoid them. We can accomplish this if we square the individual errors before we add them.** Squaring each term accomplishes two goals:

Giving more weight to farther points; squaring the error

* We can reason that this is so by noticing that whereas both estimating lines miss the second and third points (reading from left to right) by an equal distance, the line in graph (a) misses the first point by considerably less than the line in graph (b).

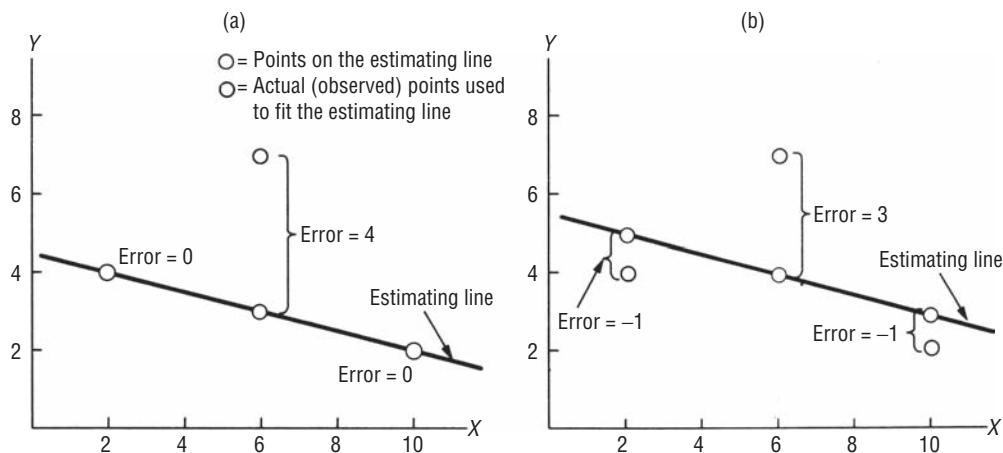


FIGURE 12-9 TWO DIFFERENT ESTIMATING LINES FITTED TO THE SAME THREE OBSERVED DATA POINTS, SHOWING ERRORS IN BOTH CASES

TABLE 12-4 SUMMING THE ABSOLUTE VALUES OF THE ERRORS OF THE TWO ESTIMATING LINES IN FIGURE 12-9

Graph (a) $ Y - \hat{Y} $	Graph (b) $ Y - \hat{Y} $
$ 4 - 4 = 0$	$ 4 - 5 = 1$
$ 7 - 3 = 4$	$ 7 - 4 = 3$
$ 2 - 4 = 0$	$ 2 - 3 = 1$
$\overline{4} \leftarrow \text{Total absolute error}$	$\overline{5} \leftarrow \text{Total absolute error}$

1. It magnifies, or penalizes, the larger errors.
2. It cancels the effect of the positive and negative values (a negative error squared is still positive).

Because we are looking for the estimating line that minimizes the sum of the squares of the errors, we call this the *least-squares method*.

Using least squares as a measure of best fit

Let's apply the least-squares criterion to the problem in Figure 12-9. After we have organized the data and summed the squares in Table 12-5, we can see that, as we thought, the estimating line in graph (b) is the better fit.

Using the criterion of least squares, we can now determine whether one estimating line is a better fit than another. But for a set of data points through which we could draw an infinite number of estimating lines, how can we tell when we have found *the best-fitting line*?

Finding the best-fitting least-squares line mathematically

Statisticians have derived two equations we can use to find the slope and the Y-intercept of the best-fitting regression line. The first formula calculates the slope:

TABLE 12-5 APPLYING THE LEAST-SQUARES CRITERION TO THE ESTIMATING LINES

Graph (a) $(Y - \hat{Y})^2$	Graph (b) $(Y - \hat{Y})^2$
$(4 - 4)^2 = (0)^2 = 0$	$(4 - 5)^2 = (-1)^2 = 1$
$(7 - 3)^2 = (4)^2 = 16$	$(7 - 4)^2 = (3)^2 = 9$
$(2 - 2)^2 = (0)^2 = 0$	$(2 - 3)^2 = (-1)^2 = 1$
$\overline{16} \leftarrow \text{Sum of the squares}$	$\overline{11} \leftarrow \text{Sum of the squares}$

Slope of the Best-Fitting Regression Line

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \quad [12-4]$$

Slope of the least-squares regression line

where

- b = slope of the best-fitting estimating line
- X = values of the independent variable
- Y = values of the dependent variable
- \bar{X} = mean of the values of the independent variable
- \bar{Y} = mean of the values of the dependent variable
- n = number of data points (that is, the number of pairs of values for the independent and dependent variables)

The second formula calculates the Y -intercept of the line whose slope we calculated using Equation 12-4:

 Y -Intercept of the Best-Fitting Regression Line

$$a = \bar{Y} - b\bar{X} \quad [12-5]$$

Intercept of the least-squares regression line

where

- a = Y -intercept
- b = slope from Equation 12-4
- \bar{Y} = mean of the values of the dependent variable
- \bar{X} = mean of the values of the independent variable

With these two equations, we can find the best-fitting regression line for any two-variable set of data points.

Using the Least-Squares Method in Two Problems

Suppose the director of the Chapel Hill Sanitation Department is interested in the relationship between the age of a garbage truck and the annual repair expense she should expect to incur. In order to determine this relationship, the director has accumulated information concerning four of the trucks the city currently owns (Table 12-6).

TABLE 12-6 ANNUAL TRUCK-REPAIR EXPENSES

Truck Number	Age of Truck in Years (X)	Repair Expense During Last Year in Hundreds of \$ (Y)
101	5	7
102	3	7
103	3	6
104	1	4

The first step in calculating the regression line for this problem is to organize the data as outlined in Table 12-7. This allows us to substitute directly into Equations 12-4 and 12-5 in order to find the slope and the Y -intercept of the best-fitting regression line.

Example of the least-squares method

With the information in Table 12-7, we can now use the equations for the slope (Equation 12-4) and the Y -intercept (Equation 12-5) to find the numerical constants for our regression line. The slope is:

$$\begin{aligned} b &= \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} & [12-4] & \text{Finding the value of } b \\ &= \frac{78 - (4)(3)(6)}{44 - (4)(3)^2} \\ &= \frac{78 - 32}{44 - 36} \\ &= \frac{6}{8} \\ &= 0.75 \leftarrow \text{The slope of the line} \end{aligned}$$

TABLE 12-7 CALCULATION OF INPUTS FOR EQUATIONS 12-4 AND 12-5

Trucks ($n = 4$) (1)	Age (X) (2)	Repair Expense (Y) (3)	XY (2) \times (3)	X^2 (2) 2
101	5	7	35	25
102	3	7	21	9
103	3	6	18	9
104	1	4	4	1
	$\sum X = 12$	$\sum Y = 24$	$\sum XY = 78$	$\sum X^2 = 44$

$$\begin{aligned} \bar{X} &= \frac{\sum X}{n} \\ &= \frac{12}{4} \\ &= 3 \leftarrow \text{Mean of the values of the independent variable} \end{aligned}$$

$$\begin{aligned} \bar{Y} &= \frac{\sum Y}{n} \\ &= \frac{24}{4} \\ &= 6 \leftarrow \text{Mean of the values of the dependent variable} \end{aligned}$$

And the Y -intercept is

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= 6 - (0.75)(3) \\ &= 6 - 2.25 \\ &= 3.75 \leftarrow \text{The } Y\text{-intercept} \end{aligned} \quad [12-5]$$

Finding the value of a

Now, to get the estimating equation that describes the relationship between the age of a truck and its annual repair expense, we can substitute the values of a and b in the equation for the estimating line:

$$\begin{aligned} \hat{Y} &= a + bX \\ &= 3.75 + 0.75X \end{aligned} \quad [12-3]$$

Determining the estimating equation

Using this estimating equation (which we could plot as a regression line if we wished), the Sanitation Department director can estimate the annual repair expense, given the age of her equipment. For example, if the city has a truck that is 4 years old, the director could use the equation to predict the annual repair expense for this truck as follows:

$$\begin{aligned} \hat{Y} &= 3.75 + 0.75(4) \\ &= 3.75 + 3 \\ &= 6.75 \leftarrow \text{Expected annual repair expense of } \$675.00 \end{aligned}$$

Using the estimating equation

Thus, the city might expect to spend about \$675 annually in repairs on a 4-year-old truck.

Now we can solve the chapter-opening problem concerning the relationship between money spent on research and development and the chemical firm's annual profits. Table 12-8 presents the information for the preceding 6 years. With this, we can determine the regression equation describing the relationship.

Another example

Again, we can facilitate the collection of the necessary information if we perform the calculations in a table such as Table 12-9.

TABLE 12-8 ANNUAL RELATIONSHIP BETWEEN RESEARCH AND DEVELOPMENT AND PROFITS

Year	Amount of Money Spent on Research and Development (\$ Millions) (X)	Annual Profit (\$ Millions) (Y)
1995	5	31
1994	11	40
1993	4	30
1992	5	34
1991	3	25
1990	2	20

TABLE 12-9 CALCULATION OF INPUTS FOR EQUATIONS 12-4 AND 12-5

Year (n = 6)	Expenditures for R&D (X)	Annual Profits (Y)	XY	X ²
1995	5	31	155	25
1994	11	40	440	121
1993	4	30	120	16
1992	5	34	170	25
1991	3	25	75	9
1990	2	20	40	4
	$\sum X = 30$	$\sum Y = 180$	$\sum XY = 1,000$	$\sum X^2 = 200$
	$\bar{X} = \frac{\sum X}{n}$ [3-2]			
	$= \frac{30}{6}$			
		$= 5 \leftarrow$ Mean of the values of the independent variable		
	$\bar{Y} = \frac{\sum Y}{n}$ [3-2]			
	$= \frac{180}{6}$			
		$= 30 \leftarrow$ Mean of the values of the dependent variable		

With this information, we are ready to find the numerical constants a and b for the estimating equation. The value of b is

$$\begin{aligned}
 b &= \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} & [12-4] \quad &\text{Finding } b \\
 &= \frac{1,000 - (6)(5)(30)}{200 - (6)(5)^2} \\
 &= \frac{1,000 - 900}{200 - 150} \\
 &= \frac{100}{50} \\
 &= 2 \leftarrow \text{The slope of the line}
 \end{aligned}$$

And the value for a is

$$\begin{aligned}
 a &= \bar{Y} - b\bar{X} & [12-5] \quad &\text{Finding } a \\
 &= 30 - (2)(5) \\
 &= 30 - 10 \\
 &= 20 \leftarrow \text{The } Y\text{-intercept}
 \end{aligned}$$

So we can substitute these values of a and b into Equation 12-3 and get

$$\begin{aligned}\hat{Y} &= a + bX \\ &= 20 + 2X\end{aligned}$$

Determining the estimating equation

[12-3]

Using this estimating equation, the vice president for research and development can predict what the annual profits will be from the amount budgeted for R&D. If the firm spends \$8 million for R&D in 1996, it can expect to earn approximately \$36 million in profits during that year:

$$\begin{aligned}\hat{Y} &= 20 + 2(8) \\ &= 20 + 16 \\ &= 36 \leftarrow \text{Expected annual profit (\$ millions)}\end{aligned}$$

Using the estimating equation to predict

Estimating equations are not perfect predictors. In Figure 12-10, which plots the points found in Table 12-8, the \$36 million estimate of profit for 1996 is only that—an estimate. Even so, the regression does give us an idea of what to expect for the coming year.

Shortcoming of the estimating equation

Checking the Estimating Equation

Now that we know how to calculate the regression line, we can learn how to check our work. A crude way to verify the accuracy of the estimating equation is to examine the graph of the sample points. As we can see from the previous problem, the regression line in Figure 12-10 does appear to follow the path described by the sample points.

Checking the estimating equation: One way

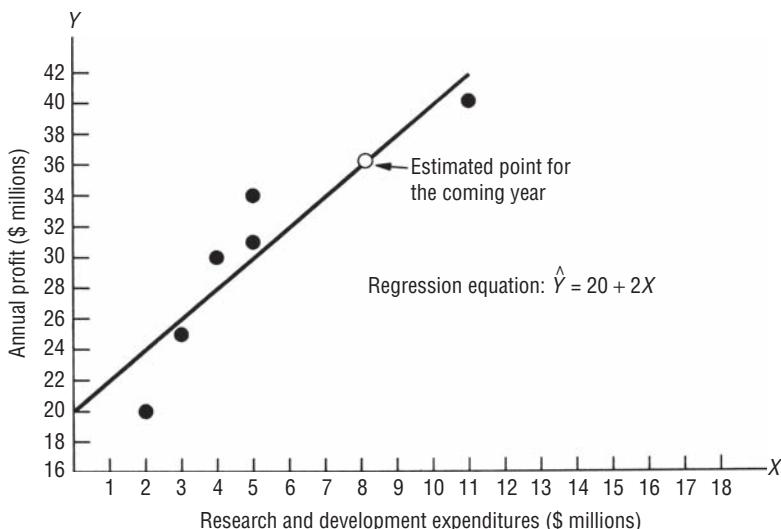


FIGURE 12-10 SCATTERING OF POINTS AROUND THE REGRESSION LINE

TABLE 12-10 CALCULATING THE SUM OF THE INDIVIDUAL ERRORS IN TABLE 12-9

<i>Y</i>		\hat{Y} (That Is, $20 + 2X$)		Individual Error
31	—	[$20 + (2)(5)$]	=	1
40	—	[$20 + (2)(11)$]	=	-2
30	—	[$20 + (2)(4)$]	=	2
34	—	[$20 + (2)(5)$]	=	4
25	—	[$20 + (2)(3)$]	=	-1
20	—	[$20 + (2)(2)$]	=	-4
				$0 \leftarrow \text{Total error}$

A more sophisticated method comes from one of the mathematical properties of a line fitted by the method of least squares; that is, the individual positive and negative errors must sum to zero. Using the information from Table 12-9, check to see whether the sum of the errors in the last problem is equal to zero. This is done in Table 12-10.

Another way to check the estimating equation

Because the sum of the errors in Table 12-10 does equal zero, and because the regression line appears to “fit” the points in Figure 12-10, we can be reasonably certain that we have not committed any serious mathematical mistakes in determining the estimating equation for this problem.

The Standard Error of Estimate

The next process we need to learn in our study of regression analysis is how to measure the reliability of the estimating equation we have developed. We alluded to this topic when we introduced scatter diagrams. There, we realized intuitively that a line is more accurate as an estimator when the data points lie close to the line (as in Figure 12-11 (a)) than when the points are farther away from the line (as in Figure 12-11 (b)).

Measuring the reliability of the estimating equation

To measure the reliability of the estimating equation, statisticians have developed the *standard error of estimate*. This standard error is symbolized s_e and is similar to the standard deviation (which we first examined in Chapter 3), in that both are measures of dispersion. You will recall that the standard deviation is used to measure the dispersion of a set of observations about the mean. **The standard error of estimate, on the other hand, measures the variability, or scatter, of the observed values around the regression line.** Even so, you will see the similarity between the standard error of estimate and the standard deviation if you compare Equation 12-6, which defines the standard error of estimate, with Equation 3-18, which defines the standard deviation:

Definition and use of the standard error of estimate

Standard Error of Estimate

$$s_e = \sqrt{\frac{\sum(X - \hat{Y})^2}{n - 2}} \quad [12-6]$$

Equation for calculating the standard error of estimate

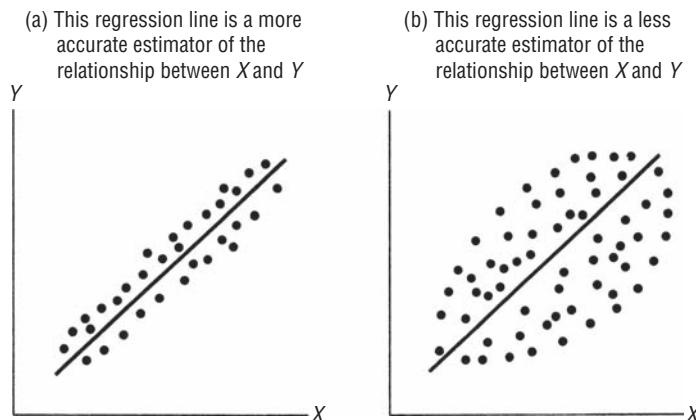


FIGURE 12-11 CONTRASTING DEGREES OF SCATTERING OF DATA POINTS AND THE RESULTING EFFECT ON THE ACCURACY OF THE REGRESSION LINE

where

- Y = values of the dependent variable
- \hat{Y} = estimated values from the estimating equation that correspond to each Y value
- n = number of data points used to fit the regression line

Notice that in Equation 12-6, the sum of the squared deviations is divided by $n - 2$, not by n . This happens because we have lost 2 degrees of freedom in estimating the regression line.

*$n - 2$ is the divisor in
Equation 12-6*

We can reason that because the values of a and b were obtained from a sample of data points, we lose 2 degrees of freedom when we use these points to estimate the regression line.

Now let's refer again to our earlier example of the Sanitation Department director who related the age of her trucks to the amount of annual repairs. We found the estimating equation in that situation to be

$$\hat{Y} = 3.75 + 0.75X$$

where X is the age of the truck and \hat{Y} is the estimated amount of annual repairs (in hundreds of dollars).

To calculate s_e for this problem, we must first determine the value of $\sum(Y - \hat{Y})^2$, that is, the numerator of Equation 12-6. We have done this in Table 12-11, using $(3.75 + 0.75X)$ for \hat{Y} whenever it was necessary. Because $\sum(Y - \hat{Y})^2$ is equal to 1.50, we can now use Equation 12-6 to find the standard error of estimate:

*Calculating the standard error
of estimate*

$$\begin{aligned}
 s_e &= \sqrt{\frac{\sum(X - \hat{Y})^2}{n - 2}} & [12-6] \\
 &= \sqrt{\frac{1.50}{4 - 2}} \\
 &= \sqrt{0.75} \\
 &= 0.866 \leftarrow \text{Standard error of estimate of \$86.60}
 \end{aligned}$$

TABLE 12-11 CALCULATING THE NUMERATOR OF THE FRACTION IN EQUATION 12-6

X (1)	Y (2)	\hat{Y} (That is, $3.75 + 0.75X$) (3)	Individual Error ($Y - \hat{Y}$) (2) - (3)	$(Y - \hat{Y})^2$ $[(2) - (3)]^2$
5	7	$3.75 + (0.75)(5)$	$7 - 7.5 = -0.5$	0.25
3	7	$3.75 + (0.75)(3)$	$7 - 6.0 = 1.0$	1.00
3	6	$3.75 + (0.75)(3)$	$6 - 6.0 = 0.0$	0.00
1	4	$3.75 + (0.75)(1)$	$4 - 4.5 = -0.5$	0.25
				$\sum(Y - \hat{Y})^2 = 1.50 \leftarrow \text{Sum of squared errors}$

Using a Short-Cut Method to Calculate the Standard Error of Estimate

To use Equation 12-6, we must do the tedious series of calculations outlined in Table 12-11. For every value of Y , we must compute the corresponding value of \hat{Y} . Then we must substitute these values into the expression $\sum(Y - \hat{Y})^2$.

Fortunately, we can eliminate some of the steps in this task by using the short cut provided by Equation 12-7, that is:

Short-Cut Method for Finding the Standard Error of Estimate

$$s_e = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n - 2}} \quad [12-7]$$

A quicker way to calculate s_e

where

- X = values of the independent variable
- Y = values of the dependent variable
- a = Y -intercept from Equation 12-5
- b = slope of the estimating equation from Equation 12-4
- n = number of data points

This equation is a short cut because, when we first organized the data in this problem so that we could calculate the slope and the Y -intercept (Table 12-7), we determined every value we need for Equation 12-7 except one: the value of $\sum Y^2$. Table 12-12 is a repeat of Table 12-7 with the Y^2 column added.

Now we can refer to Table 12-12 and our previous calculations of a and b in order to calculate s_e using the short-cut method:

$$\begin{aligned} s_e &= \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n - 2}} \\ &= \sqrt{\frac{150 - (3.75)(24) - (0.75)(78)}{4 - 2}} \end{aligned} \quad [12-7]$$

TABLE 12-12 CALCULATION OF INPUTS FOR EQUATION 12-7

Trucks ($n = 4$) (1)	Age X (2)	Repair Expense Y (3)	XY (2) \times (3)	X^2 (2) ²	Y^2 (3) ²
101	5	7	35	25	49
102	3	7	21	9	49
103	3	6	18	9	36
104	1	4	4	1	16
	$\sum X = 12$	$\sum Y = 24$	$\sum XY = 78$	$\sum X^2 = 44$	$\sum Y^2 = 150$

$$\begin{aligned}
 &= \sqrt{\frac{150 - 90 - 58.5}{2}} \\
 &= \sqrt{0.75} \\
 &= 0.866 \leftarrow \text{Standard error of } \$86.60
 \end{aligned}$$

This is the same result as the one we obtained using Equation 12-6, but think of how many steps we saved!

Interpreting the Standard Error of Estimate

As was true of the standard deviation, the larger the standard error of estimate, the greater the scattering (or dispersion) of points around the regression line. Conversely, if $s_e = 0$, we expect the estimating equation to be a “perfect” estimator of the dependent variable. In that case, all the data points would lie directly on the regression line, and no points would be scattered around it.

We shall use the standard error of estimate as a tool in the same way that we can use the standard deviation. That is to say, assuming that the observed points are normally distributed around the regression line, we can expect to find 68 percent of the points within $\pm 1s_e$ (or plus and minus 1 standard error of estimate), 95.5 percent of the points within $\pm 2s_e$, and 99.7 percent of the points within $\pm 3s_e$. Figure 12-12 illustrates these “bounds” around the regression line. **Another thing to notice in Figure 12-12 is that the standard error of estimate is measured along the Y -axis, rather than perpendicularly from the regression line.**

At this point, we should state the assumptions we are making because shortly we shall make some probability statements based on these assumptions. Specifically, we have assumed

1. The observed values for Y are normally distributed around each estimated value of \hat{Y} .
2. The variance of the distributions around each possible value of Y is the same.

If this second assumption were not true, then the standard error at one point on the regression line could differ from the standard error at another point on the line.

Interpreting and using the standard error of estimate

Using s_e to form bounds around the regression line

Assumptions we make in use of s_e

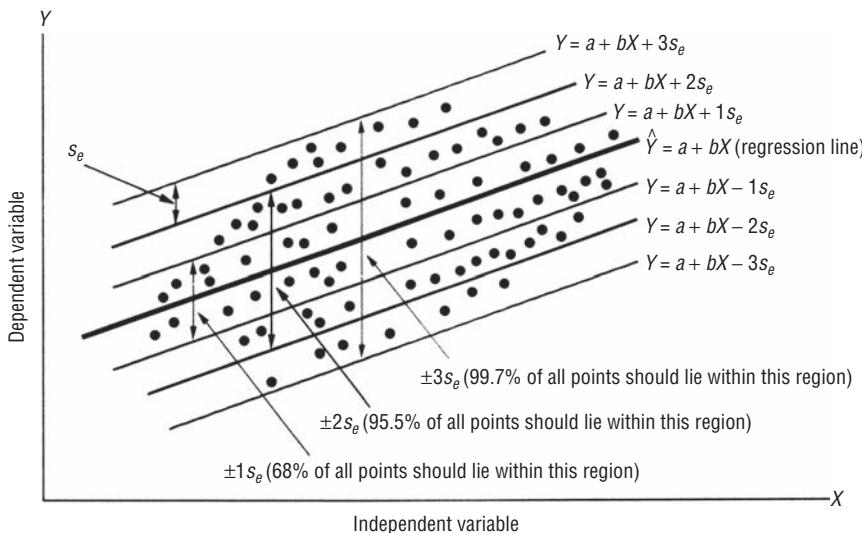


FIGURE 12-12 $\pm 1s_e$, $\pm 2s_e$ AND $\pm 3s_e$ BOUNDS AROUND THE REGRESSION LINE (MEASURED ON THE \hat{Y} AXIS)

Approximate Prediction Intervals

One way to view the standard error of estimate is to think of it as the statistical tool we can use to make a probability statement about the interval around an estimated value of \hat{Y} , within which the actual value of Y lies. We can see, for instance, in Figure 12-12 that we can be 95.5 percent certain that the actual value of Y will lie within 2 standard errors of the estimated value of \hat{Y} . We call these intervals around the estimated \hat{Y} *approximate prediction intervals*. They serve the same function as the confidence intervals did in Chapter 7.

Using s_e to generate prediction intervals

Now, applying the concept of approximate prediction intervals to the Sanitation Department director's repair expenses, we know that the estimating equation used to predict the annual repair expense is

$$\hat{Y} = 3.75 + 0.75X$$

And we know that if the department has a 4-year-old truck, we predict it will have an annual repair expense of \$675:

Applying prediction intervals

$$\begin{aligned}\hat{Y} &= 3.75 + 0.75(4) \\ &= 3.75 + 3.00 \\ &= 6.75 \leftarrow \text{Expected annual repair expense of } \$675\end{aligned}$$

Finally, you will recall that we calculated the standard error of estimate to be $s_e = 0.866$ (\$86.60). We can now combine these two pieces of information and say that we are roughly 68 percent confident that the actual repair expense will be within ± 1 standard error of estimate from \hat{Y} . We can calculate the upper and lower limits of this prediction interval for the repair

One-standard-error prediction interval

expense as follows:

$$\begin{aligned}\hat{Y} + 1s_e &= \$675 + (1)(\$86.60) \\ &= \$761.60 \leftarrow \text{Upper limit of prediction interval}\end{aligned}$$

and

$$\begin{aligned}\hat{Y} + 1s_e &= \$675 + (1)(\$86.60) \\ &= \$588.40 \leftarrow \text{Lower limit of prediction interval}\end{aligned}$$

If, instead, we say that we are roughly 95.5 percent confident that the actual repair expense will be within ± 2 standard errors of estimate from \hat{Y} , we would calculate the limits of this new prediction interval like this:

$$\begin{aligned}\hat{Y} + 2s_e &= \$675 + (2)(\$86.60) \\ &= \$848.20 \leftarrow \text{Upper limit}\end{aligned}$$

and

$$\begin{aligned}\hat{Y} + 2s_e &= \$675 - (2)(\$86.60) \\ &= \$501.80 \leftarrow \text{Lower limit}\end{aligned}$$

Keep in mind that statisticians apply prediction intervals based on the normal distribution (68 percent for $1s_e$, 95.5 percent for $2s_e$, and 99.7 percent for $3s_e$) *only* to large samples, that is, where $n > 30$. In this problem, our sample size is too small ($n = 4$). Thus, *our conclusions are inaccurate*. But the method we have used nevertheless demonstrates the principle involved in prediction intervals.

If we wish to avoid the inaccuracies caused by the size of the sample, we need to use the *t* distribution. Recall that the *t* distribution is appropriate when n is less than 30 and the population standard deviation is unknown. We meet both these conditions because $n = 4$, and s_e is an estimate rather than the known population standard deviation.

Now suppose the Sanitation Department director wants to be roughly 90 percent certain that the annual truck-repair expense will lie within the prediction interval. How should we calculate this interval? Because the *t* distribution table focuses on the probability that the parameter we are estimating will lie *outside* the prediction interval, we need to look in Appendix Table 2 under the $100\% - 90\% = 10\%$ value column. Once we locate that column, we look for the row representing 2 degrees of freedom; because $n = 4$ and because we know we lose 2 degrees of freedom (in estimating the values of a and b), then $n - 2 = 2$. Here we find the appropriate *t* value to be 2.920.

Now using this value of *t*, we can make a more accurate calculation of our prediction interval limits, as follows:

$$\begin{aligned}\hat{Y} + t(s_e) &= \$675 + (2.920)(\$86.60) \\ &= \$675 + \$252.87 \\ &= \$927.87 \leftarrow \text{Upper limit}\end{aligned}$$

Two-standard-error prediction interval

n is too small to use the normal distribution

Using the *t* distribution for prediction intervals

An example using the *t* distribution to calculate prediction intervals

and

$$\begin{aligned}\hat{Y} + t(s_e) &= \$675 - (2.920)(\$86.60) \\ &= \$675 - \$252.87 \\ &= \$422.13 \leftarrow \text{Lower limit}\end{aligned}$$

So the director can be 90 percent certain that the annual repair expense on a 4-year old truck will lie between \$422.13 and \$927.87.

We stress again that these prediction intervals are only *approximate*. In fact, statisticians can calculate the exact standard error for the prediction s_p using this formula:

$$s_p = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

where X_0 is the specific value of X at which we want to predict the value of Y .

Notice that if we use this formula, s_p is different for each value of X_0 . In particular, if X_0 is *far* from \bar{X} , then s_p is large because $(X_0 - \bar{X})^2$ is large. On the other hand, if X_0 is close to \bar{X} and n is moderately large (greater than 10), then s_p is close to s_e . This happens because $1/n$ is small and $(X_0 - \bar{X})^2$ is small. Therefore, the value under the square-root sign is close to 1, the square root is even closer to 1, and s_p is very close to s_e . This justifies our use of s_e to compute approximate prediction intervals.

HINTS & ASSUMPTIONS

Hint: Before you spend time computing a regression line for a set of data points, it makes sense to sketch the scatter diagram for those points. This lets you investigate any outlying points because some of the data you have may not represent the problem you are trying to solve. For example, the manager of a restaurant chain near college campuses who wants to examine the hypothesis that lunchtime sales are lower on hot days may find that data gathered during spring break or university holidays distort an otherwise useful regression. **Warning:** It is dangerous to pick and choose data points because they do or don't "fit" with your preconceived idea about what the conclusion should be. In regression analysis, thoughtful selection and consistent use of the best database lead to the most useful estimating equation.

Simple Linear Regression Using SPSS

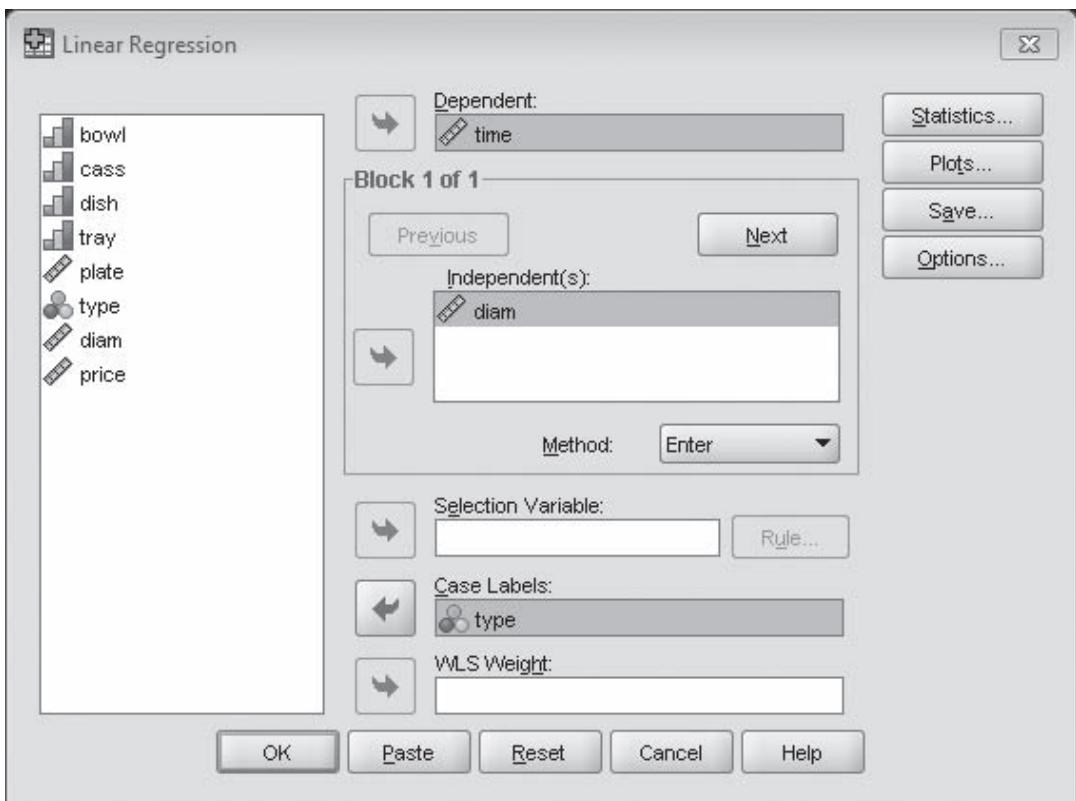
For Simple Regression Go to **Analyze>Regression>Linear Regression>Choose dependent and independent variables>Select desired statistics**

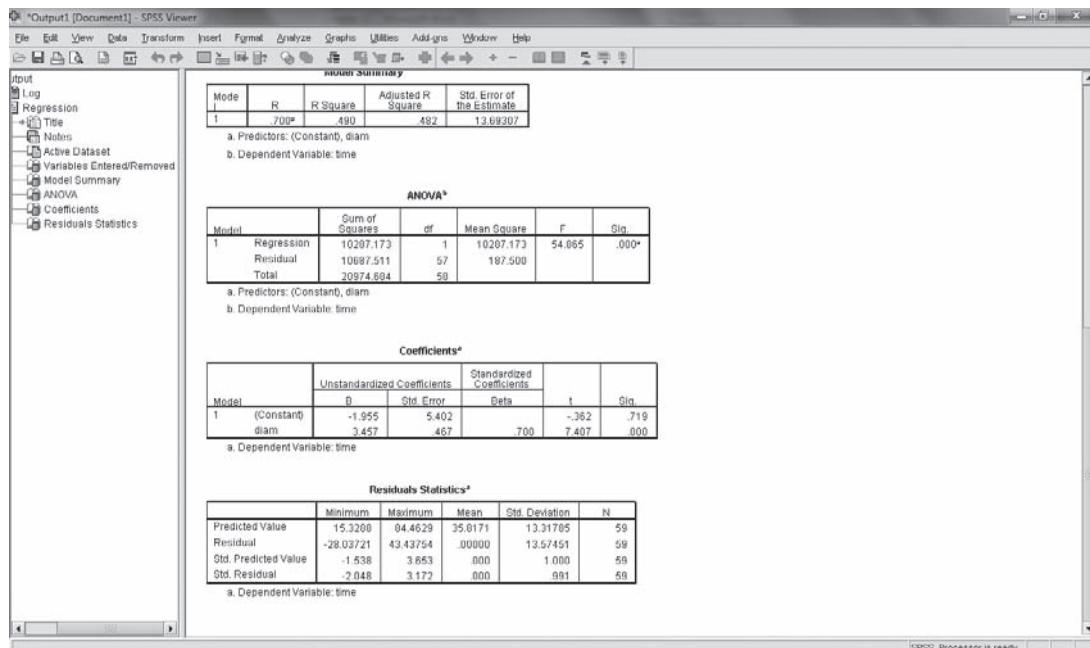
polishing.sav [DataSet1] - SPSS Data Editor

Visible: 9 of 9 Variables

	bowl	cass	dish	tray	plate	type	diam	time	price	var	var	var	var	var	var
1	0	1	0	0	0.00	2.00	10.70	47.65	144.00						
2	0	1	0	0	0.00	2.00	14.00	63.13	215.00						
3	0	1	0	0	0.00	2.00	9.00	50.76	105.00						
4	1	0	0	0	0.00	1.00	8.00	34.88	69.00						
5	0	0	1	0	0.00	3.00	10.00	55.53	134.00						
6	0	1	0	0	0.00	2.00	10.50	43.14	129.00						
7	0	0	0	1	0.00	4.00	16.00	54.86	155.00						
8	0	0	0	1	0.00	4.00	15.00	44.14	99.00						
9	0	0	1	0	0.00	3.00	6.50	17.46	38.50						
10	0	0	1	0	0.00	3.00	5.00	21.04	36.50						
11	0	0	0	1	0.00	4.00	25.00	109.38	260.00						
12	1	0	0	0	0.00	1.00	10.40	17.67	54.00						
13	1	0	0	0	0.00	1.00	7.40	16.41	39.00						
14	1	0	0	0	0.00	1.00	6.40	12.02	29.50						
15	0	1	0	0	0.00	2.00	15.40	49.48	109.00						
16	0	1	0	0	0.00	2.00	12.40	48.74	89.50						
17	1	0	0	0	0.00	1.00	6.00	23.21	42.00						
18	1	0	0	0	0.00	1.00	9.00	28.64	65.00						
19	1	0	0	0	0.00	1.00	9.00	44.95	115.00						
20	0	0	0	0	1.00	5.00	12.40	23.77	49.50						
21	1	0	0	0	0.00	1.00	7.50	20.21	36.50						
22	1	0	0	0	0.00	1.00	14.00	32.62	109.00						
23	1	0	0	0	0.00	1.00	7.00	17.84	45.00						
24	1	0	0	0	0.00	1.00	9.00	22.02	50.00						
25	1	0	0	0	0.00	1.00	12.00	29.48	89.00						
26	1	0	0	0	0.00	1.00	5.50	15.61	30.00						

Data View Variable View SPSS Processor is ready





EXERCISES 12.2

Self-Check Exercises

SC 12-2 For the following set of data:

- Plot the scatter diagram.
- Develop the estimating equation that best describes the data.
- Predict Y for $X = 10, 15, 20$.

X	13	6	14	11	17	9	13	17	18	12
Y	6.2	8.6	7.2	4.5	9.0	3.5	6.5	9.3	9.5	5.7

SC 12-3 Cost accountants often estimate overhead based on the level of production. At the Standard Knitting Co., they have collected information on overhead expenses and units produced at different plants, and want to estimate a regression equation to predict future overhead.

Overhead	191	170	272	155	280	173	234	116	153	178
Units	40	42	53	35	56	39	48	30	37	40

- Develop the regression equation for the cost accountants.
- Predict overhead when 50 units are produced.
- Calculate the standard error of estimate.

Basic Concepts

12-13 For the following data

- Plot the scatter diagram.

- (b) Develop the estimating equation that best describes the data.
 (c) Predict Y for $X = 6, 13.4, 20.5$.

X	2.7	4.8	5.6	18.4	19.6	21.5	18.7	14.3
Y	16.66	16.92	22.3	71.8	80.88	81.4	77.46	48.7
X	11.6	10.9	18.4	19.7	12.3	6.8	13.8	
Y	50.48	47.82	71.5	81.26	50.1	39.4	52.8	

12-14 Using the data given below

- (a) Plot the scatter diagram.
 (b) Develop the estimating equation that best describes the data.
 (c) Predict Y for $X = 5, 6, 7$.

X	16	6	10	5	12	14
Y	-4.4	8.0	2.1	8.7	0.1	-2.9

12-15 Given the following set of data:

- (a) Find the best-fitting line.
 (b) Compute the standard error of estimate.
 (c) Find an approximate prediction interval (with a 95 percent confidence level) for the dependent variable given that X is 44.

X	56	48	42	58	40	39	50
Y	45	38.5	34.5	46.1	33.3	32.1	40.4

Applications

12-16 Sales of major appliances vary with the new housing market: when new home sales are good, so are the sales of dishwashers, washing machines, driers, and refrigerators. A trade association compiled the following historical data (in thousands of units) on major appliance sales and housing starts:

Housing Starts (thousands)	Appliance Sales (thousands)
2.0	5.0
2.5	5.5
3.2	6.0
3.6	7.0
3.3	7.2
4.0	7.7
4.2	8.4
4.6	9.0
4.8	9.7
5.0	10.0

- (a) Develop an equation for the relationship between appliance sales (in thousands) and housing starts (in thousands).
 (b) Interpret the slope of the regression line.
 (c) Compute and interpret the standard error of estimate.

- (d) Housing starts next year may be beyond the recorded range; estimates as high as 8.0 million units have been predicted. Compute an approximate 90 percent prediction interval for appliance sales, based on the previous data and the new prediction of housing starts.

- 12-17** During recent tennis matches, Diane has noticed that her lobs have been less than totally effective because her opponents have been returning more of them. Some of the people she plays are quite tall, so she was wondering whether the height of her opponent could be used to explain the number of lobs not returned during a match. The following data were collected from five recent matches.

Opponent's Height (H)	Unreturned Lobs (L)
5.0	9
5.5	6
6.0	3
6.5	0
5.0	7

- (a) Which variable is the dependent variable?
 (b) What is the least-squares estimating equation for these data?
 (c) What is your best estimate of the number of unreturned lobs in her match tomorrow with an opponent who is 5.9 feet tall?

- 12-18** A study by the Atlanta, Georgia, Department of Transportation on the effect of bus-ticket prices on the number of passengers produced the following results:

Ticket price (cents)	25	30	35	40	45	50	55	60
Passengers per 100 miles	800	780	780	660	640	600	620	620

- (a) Plot these data.
 (b) Develop the estimating equation that best describes these data.
 (c) Predict the number of passengers per 100 miles if the ticket price were 50 cents. Use a 95 percent approximate prediction interval.

- 12-19** William C. Andrews, an organizational behavior consultant for Victory Motorcycles, has designed a test to show the company's supervisors the dangers of oversupervising their workers. A worker from the assembly line is given a series of complicated tasks to perform. During the worker's performance, a supervisor constantly interrupts the worker to assist him or her in completing the tasks. The worker, upon completion of the tasks, is then given a psychological test designed to measure the worker's hostility toward authority (a high score equals low hostility). Eight different workers were assigned the tasks and then interrupted for the purpose of instructional assistance various numbers of times (line X). Their corresponding scores on the hostility test are revealed in line Y .

X (number of times worker interrupted)	5	10	10	15	15	20	20	25
Y (worker's score on hostility test)	58	41	45	27	26	12	16	3

- (a) Plot these data.
 (b) Develop the equation that best describes the relationship between the number of times interrupted and the test score.
 (c) Predict the expected test score if the worker is interrupted 18 times.

- 12-20** The editor-in-chief of a major metropolitan newspaper has been trying to convince the paper's owner to improve the working conditions in the pressroom. He is convinced that the noise level when the presses are running creates unhealthy levels of tension and anxiety. He recently

had a psychologist conduct a test during which press operators were placed in rooms with varying levels of noise and then given a test to measure mood and anxiety levels. The following table shows the index of their degree of arousal or nervousness and the level of noise to which they were exposed (1.0 is low and 10.0 is high).

Noise level	4	3	1	2	6	7	2	3
Degree of arousal	39	38	16	18	41	45	25	38

- (a) Plot these data.
- (b) Develop an estimating equation that describes these data.
- (c) Predict the degree of arousal we might expect when the noise level is 5.

- 12-21** A firm administers a test to sales trainees before they go into the field. The management of the firm is interested in determining the relationship between the test scores and the sales made by the trainees at the end of one year in the field. The following data were collected for 10 sales personnel who have been in the field one year.

Salesperson Number	Test Score (T)	Number of Units Sold (S)
1	2.6	95
2	3.7	140
3	2.4	85
4	4.5	180
5	2.6	100
6	5.0	195
7	2.8	115
8	3.0	136
9	4.0	175
10	3.4	150

- (a) Find the least-squares regression line that could be used to predict sales from trainee test scores.
- (b) How much does the expected number of units sold increase for each 1-point increase in a trainee's test score?
- (c) Use the least-squares regression line to predict the number of units that would be sold by a trainee who received an average test score.

- 12-22** The city council of Bowie, Maryland, has gathered data on the number of minor traffic accidents and the number of youth soccer games that occur in town over a weekend.

X (soccer games)	20	30	10	12	15	25	34
Y (minor accidents)	6	9	4	5	7	8	9

- (a) Plot these data.
- (b) Develop the estimating equation that best describes these data.
- (c) Predict the number of minor traffic accidents that will occur on a weekend during which 33 soccer games take place in Bowie.
- (d) Calculate the standard error of estimate.

- 12-23** In economics, the demand function for a product is often estimated by regressing the quantity sold (Q) on the price (P). The Bamsy Company is trying to estimate the demand function for its new doll “Ma’am,” and has collected the following data:

p	20.0	17.5	16.0	14.0	12.5	10.0	8.0	6.5
q	125	156	183	190	212	238	250	276

- (a) Plot these data.
- (b) Calculate the least-squares regression line.
- (c) Draw the fitted regression line on your plot from part (a).

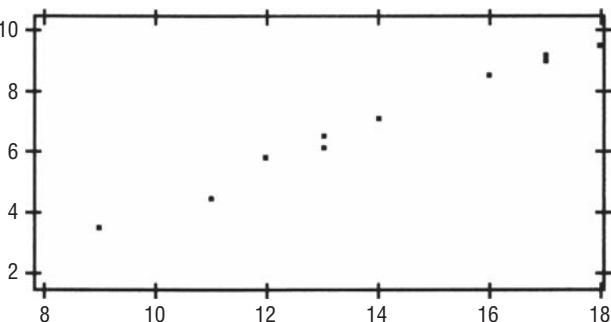
- 12-24** A tire manufacturing company is interested in removing pollutants from the exhaust at the factory, and cost is a concern. The company has collected data from other companies concerning the amount of money spent on environmental measures and the resulting amount of dangerous pollutants released (as a percentage of total emissions).

Money Spent (\$ thousands)	8.4	10.2	16.5	21.7	9.4	8.3	11.5
Percentage of Dangerous Pollutants	35.9	31.8	24.7	25.2	36.8	35.8	33.4
Money Spent (\$ thousands)	18.4	16.7	19.3	28.4	4.7	12.3	
Percentage of Dangerous Pollutants	25.4	31.4	27.4	15.8	31.5	28.9	

- (a) Compute the regression equation.
- (b) Predict the percentage of dangerous pollutants released when \$20,000 is spent on control measures.
- (c) Calculate the standard error of estimate.

Worked-Out Answers to Self-Check Exercises

SC 12-2 (a)



(b)

X	Y	XY	X^2
13	6.2	80.6	169
16	8.6	137.6	256
14	7.2	100.8	196
11	4.5	49.5	121
17	9.0	153.0	289

(Continued)

X	Y	XY	X^2
9	3.5	31.5	81
13	6.5	84.5	169
17	9.3	158.1	289
18	9.5	171.0	324
12	5.7	68.4	144
$\sum X = 140$	$\sum Y = 70.0$	$\sum XY = 1,035.0$	$\sum X^2 = 2,038$

$$\bar{X} = 140 / 10 = 14 \quad \bar{Y} = 70.0 / 10 = 7.0$$

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{1,035.0 - 10(14)(7.0)}{2,038 - 10(14)^2} = 0.7051$$

$$a = \bar{Y} - b\bar{X} = 7.0 - (0.7051)(14) = -2.8714$$

Thus, $\hat{Y} = -2.8714 + 0.7051X$. If you used a computer regression package to do your computation, you probably got

$$\hat{Y} = -2.8718 + 0.7051X.$$

This slight difference occurs because most computer packages carry their calculations to more than ten decimal places, but we rounded b to only four places before finding a . For most practical purposes, this slight difference (i.e., $a = -2.8714$ instead of -2.8718) is inconsequential.

(c) $X = 10, \hat{Y} = -2.8714 + 0.7051(10) = 4.1796$

$X = 15, \hat{Y} = -2.8714 + 0.7051(15) = 7.7051$

$X = 20, \hat{Y} = -2.8714 + 0.7051(20) = 11.2306$

SC 12-3 In this problem, Y = overhead and X = units produced.

(a)	X	Y	XY	X^2	Y^2
	40	191	7,640	1,600	36,481
	42	170	7,140	1,764	28,900
	53	272	14,416	2,809	73,984
	35	155	5,425	1,225	24,025
	56	280	15,680	3,136	78,400
	39	173	6,747	1,521	29,929
	48	234	11,232	2,304	54,756
	30	116	3,480	900	13,456
	37	153	5,661	1,369	23,409
	40	178	7,120	1,600	31,684
	$\sum X = 420$	$\sum Y = 1,922$	$\sum XY = 84,541$	$\sum X^2 = 18,228$	$\sum Y^2 = 395,024$

$$\bar{X} = \frac{420}{10} = 42 \quad \bar{Y} = \frac{1,922}{10} = 192.2$$

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{84,541 - 10(42)(192.2)}{18,228 - 10(42)^2} = 6.4915$$

$$a = \bar{Y} - b\bar{X} = 192.2 - 6.4915(42) = -80.4430$$

Thus, $\hat{Y} = -80.4430 + 6.4915X$ (Computer packages: $\hat{Y} = -80.4428 + 6.4915X$).

(b) $\hat{Y} = -80.4430 + 6.4915(50) = 244.1320$

(c) $s_e = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n-2}}$

$$= \sqrt{\frac{395,024 - (-80.4430)(1,922) - 6.4915(84,541)}{8}} = 10.2320$$

12.3 CORRELATION ANALYSIS

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another. Often, correlation analysis is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable, Y . Correlation can also be used by itself, however, to measure the degree of association between two variables.

What correlation analysis does

Statisticians have developed two measures for describing the correlation between two variables: the *coefficient of determination* and the *coefficient of correlation*. Introducing these two measures of association is the purpose of this section.

Two measures that describe correlation

The Coefficient of Determination

The coefficient of determination is the primary way we can measure the extent, or strength, of the association that exists between two variables, X and Y . Because we have used a sample of points to develop regression lines, we refer to this measure as the *sample coefficient of determination*.

Developing the sample coefficient of determination

The sample coefficient of determination is developed from the relationship between two kinds of variation: the variation of the Y values in a data set around

1. The fitted regression line
2. Their own mean

The term *variation* in both cases is used in its usual statistical sense to mean “the sum of a group of squared deviations.” By using this definition, then, it is reasonable to express the variation of the Y values around the regression line with this equation:

Variation of Y Values around the Regression Line

$$\text{Variation of the } Y \text{ values around the regression line} = \sum(Y - \hat{Y})^2$$

[12-8]

The second variation, that of the Y values around their own mean, is determined by

Variation of Y Values around Their Own Mean

$$\text{Variation of the } Y \text{ values around their own mean} = \sum(Y - \hat{Y})^2 \quad [12-9]$$

One minus the ratio between these two variations is the sample coefficient of determination, which is symbolized r^2 :

Sample Coefficient of Determination

$$r^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \quad [12-10]$$

The next two sections will show you that r^2 , as defined by Equation 12-10, is a measure of the degree of linear association between X and Y .

An Intuitive Interpretation of r^2

Consider the two extreme ways in which variables X and Y can be related. In Table 12-13, every observed value of Y lies on the estimating line, as can be verified visually by Figure 12-13. This is *perfect correlation*.

The estimating equation appropriate for these data is easy to determine. Because the regression line passes through the origin, we know that the Y -intercept is zero; because Y increases by 4 every time X increases by 1, the slope must equal 4. Thus, the regression line is

$$\hat{Y} = 4X$$

Estimating equation appropriate for perfect correlation example

TABLE 12-13 ILLUSTRATION OF PERFECT CORRELATION BETWEEN TWO VARIABLES, X AND Y

Data Point	Value of X	Value of Y
1st	1	4
2nd	2	8
3rd	3	12
4th	4	16
5th	5	20
6th	6	$\bar{Y} = \frac{144}{8} = 18 \leftarrow \text{Mean of the values of } Y$
7th	7	28
8th	8	32
$\Sigma Y = 144$		

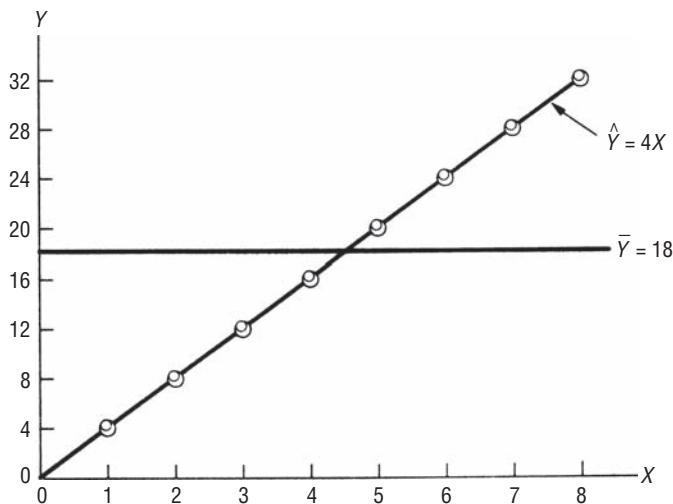


FIGURE 12-13 PERFECT CORRELATION BETWEEN X AND Y: EVERY DATA POINT LIES ON THE REGRESSION LINE

Now, to determine the sample coefficient of determination for the regression line in Figure 12-13, we first calculate the numerator of the fraction in Equation 12-10:

Determining the sample coefficient of determination for the perfect correlation example

$$\text{Variation of the } Y \text{ values around the regression line} = \sum(Y - \hat{Y})^2 \quad [12-8]$$

$$\begin{aligned}
 &= \sum(0)^2 \\
 &= 0
 \end{aligned}$$

Because every Y value is on the regression line, the difference between Y and \hat{Y} is zero in each case

Then we can find the denominator of the fraction:

$$\begin{aligned}
 &\text{Variation of the } Y \text{ values} \\
 &\text{around their own mean} = \sum(Y - \bar{Y})^2 \quad [12-9] \\
 &(4 - 18)^2 = (-14)^2 = 196 \\
 &(8 - 18)^2 = (-10)^2 = 100 \\
 &(12 - 18)^2 = (-6)^2 = 36 \\
 &(16 - 18)^2 = (-2)^2 = 4 \\
 &(20 - 18)^2 = (-2)^2 = 4 \\
 &(24 - 18)^2 = (-6)^2 = 36 \\
 &(28 - 18)^2 = (-10)^2 = 100 \\
 &(32 - 18)^2 = (-14)^2 = 196 \\
 &\underline{\mathbf{672}} \leftarrow \sum(Y - \bar{Y})^2
 \end{aligned}$$

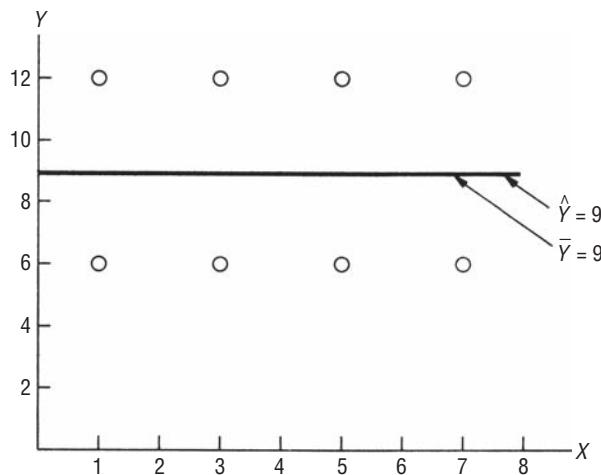


FIGURE 12-14 ZERO CORRELATION BETWEEN X AND Y: SAME VALUES OF Y APPEAR FOR DIFFERENT VALUES OF X

With these values to substitute into Equation 12-10, we can find that the sample coefficient of determination is equal to +1:

$$\begin{aligned}
 r^2 &= 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} & [12-10] \\
 &= 1 - \frac{0}{672} \\
 &= 1 - 0 \\
 &= 1 \leftarrow \text{Sample coefficient of determination} \\
 &\quad \text{when there is perfect correlation}
 \end{aligned}$$

In fact, r^2 is equal to +1 whenever the regression line is a perfect estimator.

A second extreme way in which the variables X and Y can be related is that the points could lie at equal distances on both sides of a horizontal regression line, as pictured in Figure 12-14. The data set here consists of eight points, all of which have been recorded in Table 12-14.

From Figure 12-14, we can see that the least-squares regression line appropriate for these data is given by the equation $\hat{Y} = 9$. The slope of the line is zero because the same values of Y appear for all the different values of X . Both the Y -intercept and the mean of the Y values are equal to 9.

Now we'll compute the two variations using Equations 12-8 and 12-9 so that we can calculate the sample coefficient of determination for this regression line. First, the variation of the Y values around the estimating line $\hat{Y} = 9$:

Determining the sample coefficient of determination for zero correlation

TABLE 12-14 ILLUSTRATION OF ZERO CORRELATION BETWEEN TWO VARIABLES, X AND Y

Data Point	Value of X	Value of Y
1st	1	6
2nd	1	12
3rd	3	6
4th	3	12
5th	5	6
6th	5	12
7th	7	6
8th	7	12
		$\sum Y = \underline{72}$
		$\bar{Y} = \frac{72}{8}$
		= 9 ← Mean of the values of Y

Variation of the Y values

$$\text{around the regression line} = \sum(Y - \hat{Y})^2 \quad [12-8]$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$\underline{72} \leftarrow \sum(Y - \hat{Y})^2$$

Variation of the Y values

$$\text{around their own mean} = \sum(Y - \bar{Y})^2 \quad [12-9]$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$(6 - 9)^2 = (-3)^2 = 9$$

$$(12 - 9)^2 = (3)^2 = 9$$

$$\underline{72} \leftarrow \sum(Y - \bar{Y})^2$$

Substituting these two values into Equation 12-10, we see that the sample coefficient of determination is 0:

$$\begin{aligned}
 r^2 &= 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \\
 &= 1 - \frac{72}{72} \\
 &= 1 - 1 \\
 &= 0 \leftarrow \text{Sample coefficient of determination} \\
 &\quad \text{when there is no correlation}
 \end{aligned} \tag{12-10}$$

Thus, the value of r^2 is zero when there is no correlation.

In the problems most decision makers encounter, r^2 lies somewhere between these two extremes of 1 and 0. Keep in mind, however, that an r^2 close to 1 indicates a strong correlation between X and Y , whereas an r^2 near 0 means that there is little correlation between these two variables.

Interpreting r^2 values

One point that we must emphasize strongly is that r^2 measures only the strength of a linear relationship between two variables. For example, if we had a lot of X , Y points that fell on the circumference of a circle but at randomly scattered places, clearly there would be a relationship among these points (they all lie on the same circle). But in this instance, if we computed r^2 , it would turn out to be close to zero, because the points do not have a *linear* relationship with each other.

Interpreting r^2 Another Way

Statisticians also interpret the sample coefficient of determination by looking at the *amount of the variation in Y that is explained by the regression line*. To understand this meaning of r^2 , consider the regression line (shown in color) in Figure 12-15. Here, we have singled out one observed value of Y , shown as the upper black-circle. If we use the mean of the Y values, \bar{Y} , to estimate this black-circled value of Y , then the *total deviation* of this Y from its mean would be $(Y - \bar{Y})$. Notice that if we used the regression line

Another way to interpret
the sample coefficient of
determination

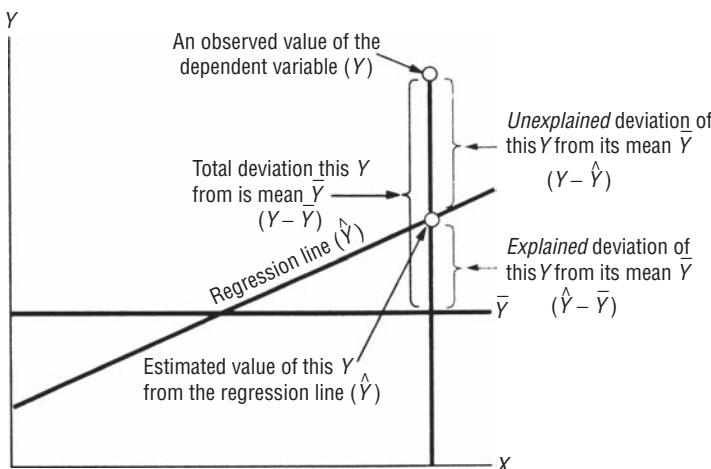


FIGURE 12-15 TOTAL DEVIATION, EXPLAINED DEVIATION, AND UNEXPLAINED DEVIATION FOR ONE OBSERVED VALUE OF Y

to estimate this black-circled value of Y , we would get a better estimate. However, even though the regression line accounts for, or explains $(\hat{Y} - Y)$ of the total deviation, the remaining portion of the total deviation, $(Y - \hat{Y})$, is still *unexplained*.

Explained and unexplained deviation

But consider a whole set of observed Y values instead of only one value. The total variation—that is, the sum of the squared total deviations—of these points from their mean would be

$$\Sigma(Y - \bar{Y})^2 \quad [12-9]$$

and the *explained* portion of the total variation, or the sum of the squared explained deviations of these points from their mean, would be

Explained and unexplained variation

$$\Sigma(\hat{Y} - \bar{Y})^2$$

The *unexplained* portion of the total variation (the sum of the squared unexplained deviations) of these points from the regression line would be

$$\Sigma(Y - \hat{Y})^2 \quad [12-8]$$

If we want to express the fraction of the total variation that remains *unexplained*, we would divide the unexplained variation, $\Sigma(Y - \hat{Y})^2$, by the total variation, $\Sigma(Y - \bar{Y})^2$, as follows

$$\frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2} \leftarrow \text{Fraction of the total variation that is unexplained}$$

Finally, if we subtract the fraction of the total variation that remains unexplained from 1, we will have the formula for finding that fraction of the total variation of Y that is explained by the regression line. That formula is

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2} \quad [12-10]$$

the same equation we have previously used to calculate r^2 . It is in this sense, then, that r^2 measures how well X explains Y , that is, the degree of association between X and Y .

One final word about calculating r^2 . To obtain r^2 using Equations 12-8, 12-9, and 12-10 requires a series of tedious calculations. To bypass these calculations, statisticians have developed a short-cut version, using values we would have determined already in the regression analysis. The formula is

Short-cut method to calculate r^2

Short-Cut Method for Finding Sample Coefficient of Determination

$$r^2 \text{ calculated by short-cut method} \rightarrow r^2 = \frac{a \sum Y + b \sum XY - n \bar{Y}^2}{\sum^2 - n \bar{Y}^2} \quad [12-11]$$

TABLE 12-15 CALCULATIONS OF INPUTS FOR EQUATION 12-11

Year (n = 6)	R&D Expense (X) (1)	Anual Profit (Y) (3)	XY (2) × (3)	X ² (2) ²	Y ² (3) ²
1995	5	31	155	25	961
1994	11	40	440	121	1,600
1993	4	30	120	16	900
1992	5	34	170	25	1,156
1991	3	25	75	9	625
1990	2	20	40	4	400
	$\sum X = 30$	$\sum Y = 180$	$\sum XY = 1,000$	$\sum X^2 = 200$	$\sum Y^2 = 5,642$
	$\bar{Y} = \frac{180}{6}$				
	$= 30 \leftarrow$ Mean of the values of the dependent variable				

where

- r^2 = sample coefficient of determination
- a = Y -intercept
- b = slope of the best-fitting estimating line
- n = number of data points
- X = values of the independent variable
- Y = values of the dependent variable
- \bar{Y} = mean of the observed values of the dependent variable

To see why this formula is a short cut, apply it to our earlier regression relating research and development expenditures to profits. In Table 12-15, we have repeated the columns from Table 12-9, adding a Y^2 column. Recall that when we found the values for a and b , the regression line for this problem was

$$\hat{Y} = 20 + 2X$$

Applying the short-cut method

Using this line and the information in Table 12-15, we can calculate r^2 as follows:

$$\begin{aligned}
 r^2 &= \frac{a \sum Y + b \sum XY - n \bar{Y}^2}{\sum Y^2 - n \bar{Y}^2} && [12-11] \\
 &= \frac{(20)(180) + (2)(1,000) - (6)(30)^2}{5,642 - (6)(30)^2} \\
 &= \frac{3,600 + 2,000 - 5,400}{5,642 - 5,400} \\
 &= \frac{200}{242} \\
 &= 0.826 \leftarrow \text{Sample coefficient of determination}
 \end{aligned}$$

Thus, we can conclude that the variation in the research and development expenditures (the independent variable X) explains 82.6 percent of the variation in the annual profits (the dependent variable Y). *Interpreting r^2*

The Coefficient of Correlation

The coefficient of correlation is the second measure that we can use to describe how well one variable is explained by another. When we are dealing with samples, the *sample coefficient of correlation* is denoted by r and is the square root of the sample coefficient of determination:

Sample coefficient of correlation

Sample Coefficient of Correlation

$$r = \sqrt{r^2}$$

[12-12]

When the slope of the estimating equation is positive, r is the positive square root, but if b is negative, r is the negative square root. Thus, **the sign of r indicates the direction of the relationship between the two variables X and Y .** If an inverse relationship exists—that is, if Y decreases as X increases—then r will fall between 0 and -1 . Likewise, if there is a direct relationship (if Y increases as X increases), then r will be a value within the range of 0 to 1. Figure 12-16 illustrates these various characteristics of r .

The coefficient of correlation is more difficult to interpret than r^2 . What does $r = 0.9$ mean? To answer that question, we must remember that $r = 0.9$ is the same as $r^2 = 0.81$. The latter tells us that 81 percent of the variation in Y is explained by the regression line. So we see that r is nothing more than the square root of r^2 , and we cannot interpret its meaning directly. *Interpreting r*

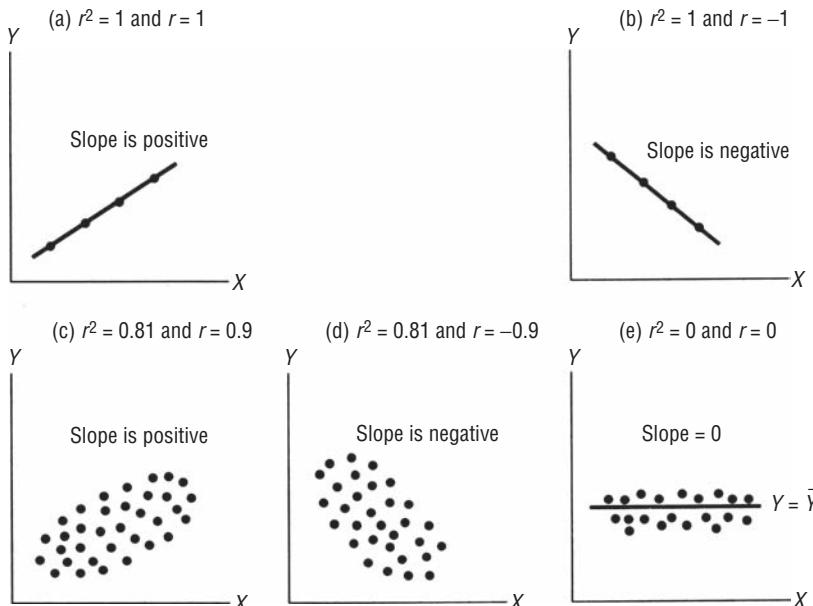


FIGURE 12-16 VARIOUS CHARACTERISTICS OF r , THE SAMPLE COEFFICIENT OF CORRELATION

Now let's find the coefficient of correlation of our problem relating research and development expenditures and annual profits. In the previous section, we found that the sample coefficient of determination is $r^2 = 0.826$, so we can substitute this value into Equation 12-12 and find that

$$\begin{aligned} r &= \sqrt{r^2} \\ &= \sqrt{0.826} \\ &= 0.909 \leftarrow \text{Sample coefficient of correlation} \end{aligned} \quad [12-12]$$

Calculating r for the research and development problem

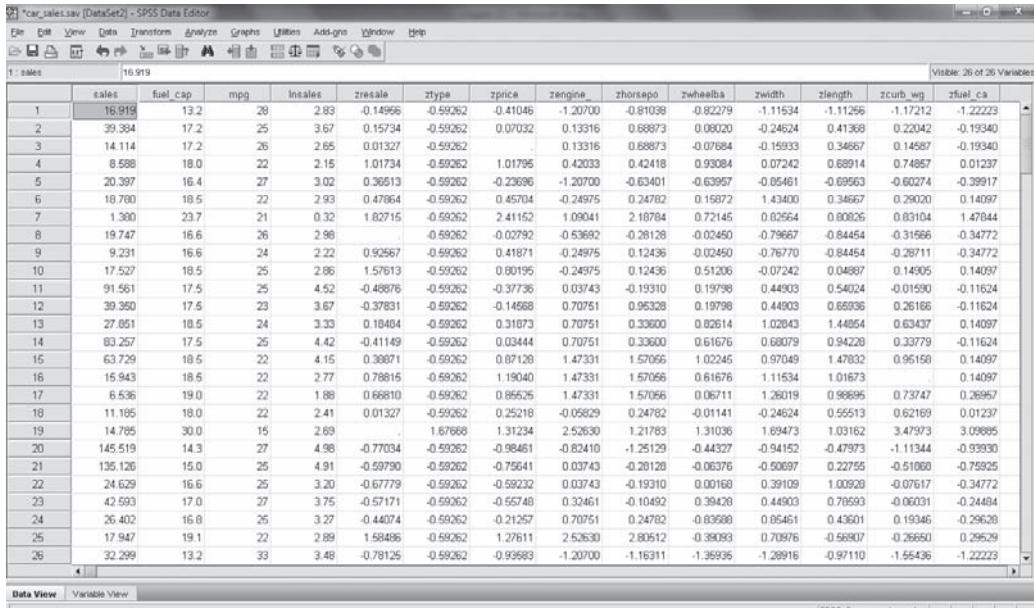
The relation between the two variables is direct and the slope is positive; therefore, the sign for r is positive.

HINTS & ASSUMPTIONS

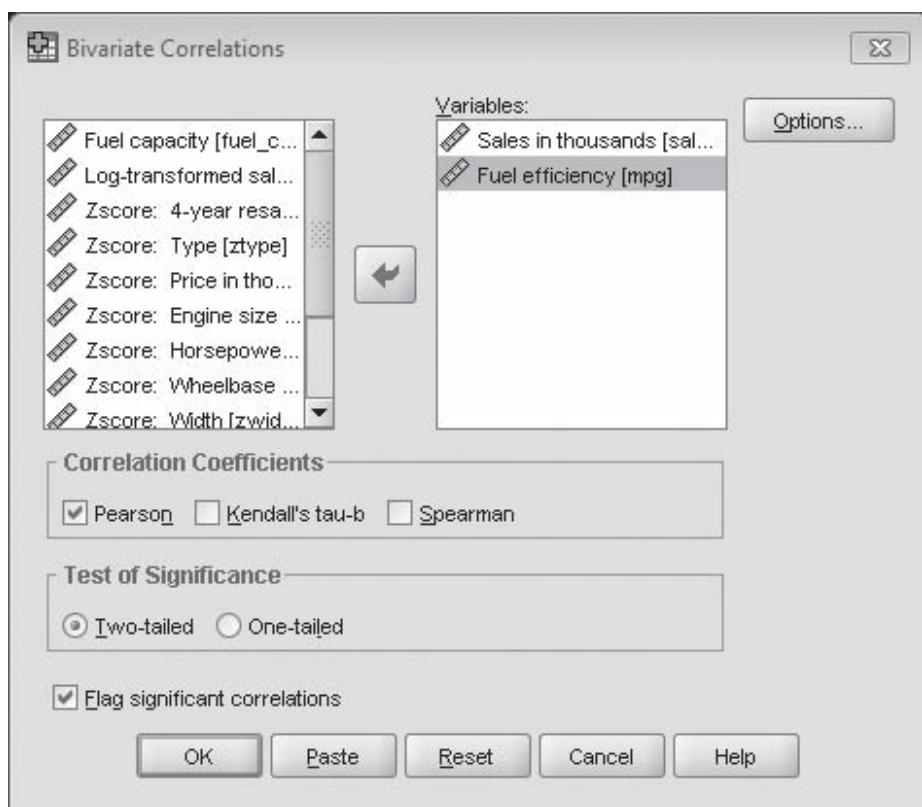
Warning: Because you know that the coefficient of determination (r^2) is the square of the coefficient of correlation, r , you should be wary of using all but the highest correlations as the basis for making decisions. Hint: If we find that the amount spent on movies correlates 0.6 with family income, that seems like a fairly strong correlation (0.6 is closer to 1.0 than it is to zero). But when you square 0.6 you see that it accounts for only $0.6 \times 0.6 = 0.36$ or 36 percent of the variation in the amount of money families spend on movies. If you designed your marketing strategy to appeal only to families with high incomes, you'd miss a lot of potential customers. Hint: Instead, try to find what else is influencing family movie decisions.

Simple Correlation Using SPSS

For Correlation go to Analyze>Correlation>Bivariate>Select variables for coorelation



The screenshot shows the SPSS Data Editor window with the title "car_sales.sav - DataSet2 - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. The toolbar has icons for opening files, saving, printing, and other functions. The status bar at the bottom right says "Visible: 26 of 26 Variables". The main area displays a correlation matrix for 26 variables. The columns and rows are labeled with variable names: sales, fuel_efficiency, mpg, Insales, zresale, ztype, zprice, zengine, zhorsepower, zwheelbase, zwidth, zlength, zcurb_wg, zfuel_ca. The matrix contains numerical values ranging from approximately -0.99 to 1.00, indicating the strength and direction of linear relationships between the variables. The diagonal elements are all 1.00, representing the correlation of each variable with itself. The first few rows of data are also visible below the matrix.



*Output1 [Document1] - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

Output Log Correlations

GET
FILE=E:\BookChapter 3.sav.
DATASET NAME DataSet0 WINDOW=FRONT.
GET
FILE=F:\Softwares\SPSS Install 16\Samples\car_sales.sav.
DATASET NAME DataSet2 WINDOW=FRONT.
DATASET CLOSE DataSet1.
CORRELATIONS
/VARIABLES=sales mpg
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.

Correlations

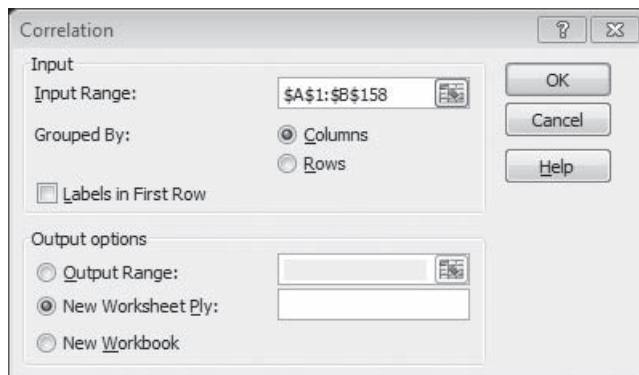
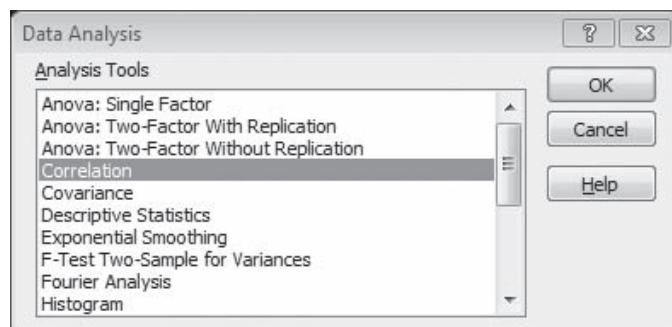
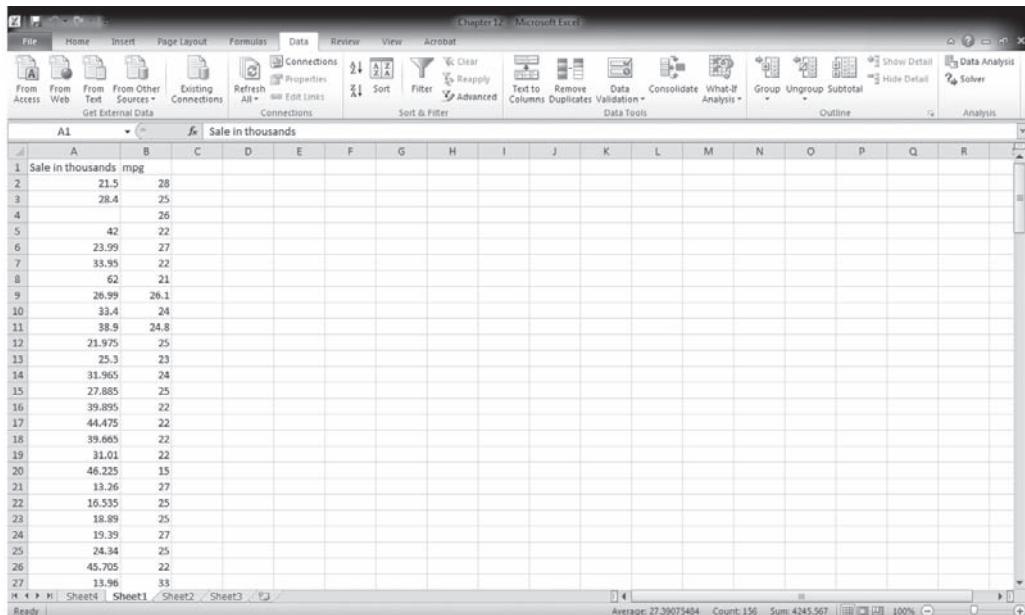
[DataSet2] F:\Softwares\SPSS Install 16\Samples\car_sales.sav

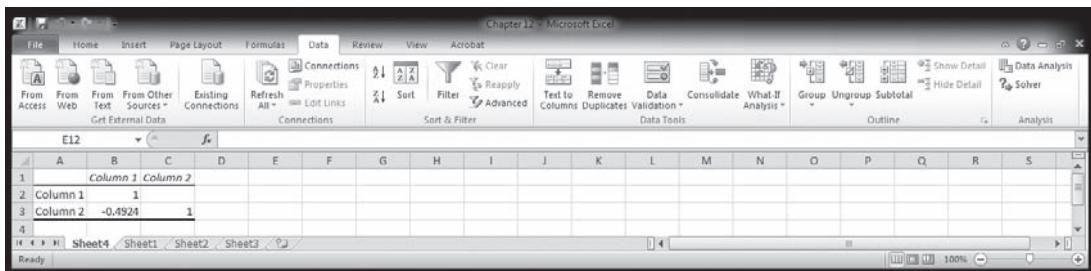
		Sales in thousands	Fuel efficiency
Sales in thousands	Pearson Correlation Sig. (2-tailed) N	1 157 154	-.017 .837 154
Fuel efficiency	Pearson Correlation Sig. (2-tailed) N	-.017 .837 154	1 154

SPSS Processor is ready

Simple Correlation Using MS Excel

For Correlation go to Data>Data Analysis>Correlation>Define Input range & variable grouping





EXERCISES 12.3

Self-Check Exercises

SC 12-4 Campus Stores has been selling the *Believe It or Not: Wonders of Statistics Study Guide* for 12 semesters and would like to estimate the relationship between sales and number of sections of elementary statistics taught in each semester. The following data have been collected:

Sales (units)	33	38	24	61	52	45
Number of sections	3	7	6	6	10	12
Sales (units)	65	82	29	63	50	79
Number of sections	12	13	12	13	14	15

- (a) Develop the estimating equation that best fits the data.
- (b) Calculate the sample coefficient of determination and the sample coefficient of correlation.

SC 12-5 Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise SC 12-3.

Basic Concepts

- 12-25** What type of correlation (positive, negative, or zero) should we expect from these variables?
- (a) Ability of supervisors and output of their subordinates.
 - (b) Age at first full-time job and number of years of education.
 - (c) Weight and blood pressure.
 - (d) College grade-point average and student's height.
- In the following exercises, calculate the sample coefficient of determination and the sample coefficient of correlation for the problems specified.
- 12-26** Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise 12-17.
- 12-27** Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise 12-18.
- 12-28** Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise 12-19.
- 12-29** Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise 12-20.
- 12-30** Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Exercise 12-21.

Applications

- 12-31** Bank of Lincoln is interested in reducing the amount of time people spend waiting to see a personal banker. The bank is interested in the relationship between waiting time (Y) in minutes and number of bankers on duty (X). Customers were randomly selected with the data given below:

X	2	3	5	4	2	6	1	3	4	3	3	2	4
Y	12.8	11.3	3.2	6.4	11.6	3.2	8.7	10.5	8.2	11.3	9.4	12.8	8.2

- (a) Calculate the regression equation that best fits the data.
(b) Calculate the sample coefficient of determination and the sample coefficient of correlation.
- 12-32** Zippy Cola is studying the effect of its latest advertising campaign. People chosen at random were called and asked how many cans of Zippy Cola they had bought in the past week and how many Zippy Cola advertisements they had either read or seen in the past week.

X (number of ads)	3	7	4	2	0	4	1	2
Y (cans purchased)	11	18	9	4	7	6	3	8

- (a) Develop the estimating equation that best fits the data.
(b) Calculate the sample coefficient of determination and the sample coefficient of correlation.

Worked-Out Answers to Self-Check Exercises

- SC 12-4** In this problem, Y = sales and X = number of sections.

(a)	X	Y	XY	X^2	Y^2
	3	33	99	9	1,089
	7	38	266	49	1,444
	6	24	144	36	576
	6	61	366	36	3,721
	10	52	520	100	2,704
	12	45	540	144	2,025
	12	65	780	144	4,225
	13	82	1,066	169	6,724
	12	29	348	144	841
	13	63	819	169	3,969
	14	50	700	196	2,500
	15	79	1,185	225	6,241
	$\sum X = 123$	$\sum Y = 621$	$\sum XY = 6,833$	$\sum X^2 = 1,421$	$\sum Y^2 = 36,059$

$$\bar{X} = 123/12 = 10.25 \quad \bar{Y} = 621/12 = 51.75$$

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{6,883 - 12(10.25)(51.75)}{1,421 - 12(10.25)^2} = 2.9189$$

$$a = \bar{Y} - b\bar{X} = 51.75 - 2.9189(10.25) = 21.8313.$$

Thus, $\hat{Y} = 21.8313 + 2.9189X$ (Computer packages: $\hat{Y} = 21.8315 + 2.9189X$).

$$(b) r^2 = \frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

$$= \frac{21.8313(621) + 2.9189(6,833) - 12(51.75)^2}{36,059 - 12(51.75)^2} = 0.3481$$

$$r = \sqrt{0.3481} = 0.5900$$

SC 12-5 From the solution to Exercise SC 12-3 on page 637, we have $n = 10$, $\sum Y = 1,922$, $\bar{Y} = 192.2$, $\sum XY = 84,541$, $\sum Y^2 = 395,024$, $a = -80.4430$, and $b = 6.4915$. Hence

$$r^2 = \frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

$$= \frac{-80.4430(1,922) + 6.4915(84,541) - 10(192.2)^2}{395,024 - 10(192.2)^2}$$

$$= 0.9673$$

$$r = \sqrt{0.9673} = 0.9835$$

12.4 MAKING INFERENCES ABOUT POPULATION PARAMETERS

So far, we have used regression and correlation analyses to relate two variables on the basis of sample information. But data from a sample represent only part of the total population. Because of this, we may think of our estimated sample regression line as an estimate of a true but unknown population regression line of the form

Relationship of sample regression line and population regression line

Population Regression Line

$$Y = A + BX$$

[12-13]

Recall our discussion of the Sanitation Department director who tried to use the age of a truck to explain its annual repair expense. That expense will probably consist of two parts:

1. Regular maintenance that does not depend on the age of the truck: tune-ups, oil changes, and lubrication. This expense is captured in the intercept term A in Equation 12-13.
2. Expenses for repairs due to aging: relining brakes, engine and transmission overhauls, and painting. Such expenses tend to increase with the age of the truck, and they are captured in the BX term of the population regression line $Y = A + BX$ in Equation 12-13.

Of course, all the brakes of all the trucks will not wear out at the same time, and some of the trucks will run for years without engine overhauls. Because of this, the individual data points will

Why data points do not lie exactly on the regression line

probably not lie exactly on the population regression line. Some will be above it; some will fall below it. So, instead of satisfying

$$Y = A + BX \quad [12-13]$$

the individual data points will satisfy the formula

Population Regression Line with a Random Disturbance

$$Y = A + BX + e \quad [12-13a]$$

where e is a random disturbance from the population regression line. On the average, e equals zero because disturbances above the population regression line are canceled out by disturbances below the line. We can denote the standard deviation of these individual disturbances by σ_e . The standard error of estimate s_e , then, is an estimate of σ_e , the standard deviation of the disturbance.

Random disturbance e and its behavior

Let us look more carefully at Equations 12-13 and 12-13a. Equation 12-13a expresses the individual values of Y (in this case, annual repair expense) in terms of the individual values of X (the age of the truck) and the random disturbance (e). Because disturbances above the population regression line are canceled out by those below the line, we know that the expected value of e is zero, and we see that if we had several trucks of the same age, X , we would expect the average annual repair expense on these trucks to be $Y = A + BX$. This shows us that the population regression line (Equation 12-13) gives the mean value of Y associated with each value of X .

Because our *sample* regression line, $\hat{Y} = a + bX$ (Equation 12-3), estimates the *population* regression line, $Y = A + BX$ (Equation 12-13), we should be able to use it to make inferences about the population regression line. In this section, then, we shall make inferences about the slope B of the “true” regression equation (the one for the entire population) that are based on the slope b of the regression equation estimated from a sample of values.

Making inferences about B from b

Slope of the Population Regression Line

The regression line is derived from a sample and not from the entire population. As a result, we cannot expect the true regression equation, $Y = A + BX$ (the one for the entire population), to be exactly the same as the equation estimated from the sample observations, or $\hat{Y} = a + bX$. Even so, we can use the value of b , the slope we calculate from a sample, to test hypotheses about the value of B , the slope of the regression line for the entire population.

Difference between true regression equation and one estimated from sample observations

The procedure for testing a hypothesis about B is similar to procedures discussed in Chapters 8 and 9, on hypothesis testing. To understand this process, return to the problem that related annual expenditures for research and development to profits. On page 627, we pointed out that $b = 2$. The first step is to find some value for B to compare with $b = 2$.

Testing a hypothesis about B

Suppose that over an extended past period of time, the slope of the relationship between X and Y was 2.1. To test whether this is still the case, we could define the hypotheses as

$$H_0: B = 2.1 \leftarrow \text{Null hypothesis}$$

$$H_1: B \neq 2.1 \leftarrow \text{Alternative hypothesis}$$

In effect, then, we are testing to learn whether current data indicate that B has changed from its historical value of 2.1.

To find the test statistic for B , it is necessary first to find the *standard error of the regression coefficient*. Here, the regression coefficient we are working with is b , so the standard error of this coefficient is denoted s_b . Equation 12-14 presents the mathematical formula for s_b :

Standard Error of b

$$s_b = \frac{s_e}{\sqrt{\sum X^2 - n\bar{X}^2}} \quad [12-14]$$

where

- s_b = standard error of the regression coefficient
- s_e = standard error of estimate
- X = values of the independent variable
- \bar{X} = mean of the values of the independent variable
- n = number of data points

Once we have calculated s_b , we can use Equation 12-15 to standardize the slope of our fitted regression equation:

Standardizing the regression coefficient

Standardized Value of b

$$t = \frac{b - B_{H_0}}{s_b} \quad [12-15]$$

where

- b = Slope of fitted regression
- B_{H_0} = actual slope hypothesized for the population
- s_b = standard error of the regression coefficient

Because the test will be based on the t distribution with $n - 2$ degrees of freedom, we use t to denote the standardized statistic.

A glance at Table 12-15 on page 650 enables us to calculate the values of $\sum X^2$ and $n\bar{X}^2$. To obtain s_e , we can take the short-cut method, as follows:

$$s_e = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n-2}} \quad [12-15]$$

Calculating s_e

$$= \sqrt{\frac{5,642 - (20)(180) - (2)(1,000)}{6-2}}$$

$$\begin{aligned}
 &= \sqrt{\frac{42}{4}} \\
 &= \sqrt{10.5} \\
 &= 3.24 \leftarrow \text{Standard error of estimate}
 \end{aligned}$$

Now we can determine the standard error of the regression coefficient:

$$s_b = \frac{s_e}{\sqrt{\sum X^2 - n\bar{X}^2}} \quad [12-14]$$

Calculating s_b

$$\begin{aligned}
 &= \frac{3.24}{\sqrt{200 - (6)(5)^2}} \\
 &= \frac{3.24}{\sqrt{50}} \\
 &= \frac{3.24}{7.07} \\
 &= 0.46 \leftarrow \text{Standard error of the regression coefficient}
 \end{aligned}$$

Now we use the standard error of the regression coefficient to calculate our standardized test statistic:

Standardizing the regression coefficient

$$\begin{aligned}
 t &= \frac{b - B_{H_0}}{s_b} \\
 &= \frac{2.0 - 2.1}{0.46} \\
 &= -0.217 \leftarrow \text{Standardized regression coefficient}
 \end{aligned} \quad [12-15]$$

Suppose we have reason to test our hypothesis at the 10 percent level of significance. Because we have six observations in our sample data, we know that we have $n - 2$ or $6 - 2 = 4$ degrees of freedom. We look in Appendix Table 2 under the 10 percent column and come down until we find the 4-degrees-of-freedom row. There we see that the appropriate t value is 2.132. Because we are concerned whether b (the slope of the sample regression line) is significantly *different* from B (the hypothesized slope of the population regression line), this is a two-tailed test, and the critical values are ± 2.132 . The standardized regression coefficient is -0.217 , which is *inside* the acceptance region for our hypothesis test. Therefore, we accept the null hypothesis that B still equals 2.1. In other words, there is not enough difference between b and 2.1 for us to conclude that B has changed from its historical value. Because of this, we feel that each additional million dollars spent on research and development still increases annual profits by about \$2.1 million, as it has in the past.

Conducting the hypothesis test

In addition to hypothesis testing, we can also construct a *confidence interval* for the value of B . In the same way that b is a point estimate of B , such confidence intervals are interval estimates of B . The problem we just completed, and for which we did a hypothesis test, will illustrate the process of constructing

a confidence interval. There we found that

$$b = 2.0$$

$$s_b = 0.46$$

$t = 2.132 \leftarrow$ 10 percent level of significance and 4 degrees of freedom

With this information, we can calculate confidence intervals like this:

Confidence interval for B

$$\begin{aligned} b + t(s_b) &= 2 + (2.132)(0.46) \\ &= 2 + 0.981 \\ &= 2.981 \leftarrow \text{Upper limit} \\ b - t(s_b) &= 2 - (2.132)(0.46) \\ &= 2 - 0.981 \\ &= 1.019 \leftarrow \text{Lower limit} \end{aligned}$$

In this situation, then, we are 90 percent confident that the true value of B lies between 1.019 and 2.981; that is, each additional million dollars spent on research and development increases annual profits by some amount between \$1.02 million and \$2.98 million.

Interpreting the confidence interval

HINTS & ASSUMPTIONS

In this section we've been using sample observations to calculate b , the slope of the *sample* regression line, which we then use to test hypotheses about B , the true slope of the *population* regression line. Hint: We use s_e to calculate the standard error of the regression coefficient just as we used the sample standard deviation in Chapter 6 to compute the standard error of the mean. Warning: Whenever you use your computer to develop a regression line, don't forget to ask, "Is this regression coefficient significantly different from zero?" If it's *not*, no matter how much good-looking computer output you have, you haven't demonstrated any significant relationship between the variables, and you need to keep looking for more useful relationships. For example, if you own a tanning salon and you have a hunch that more people come in on cloudy days, you might do a regression of "number of visits" on "hours of sunshine." If you do that and it yields a regression line with a slope that is *not* significant, keeping track of the weather is not going to help your business.

EXERCISES 12.4

Self-Check Exercises

SC 12-6 In finance, it is of interest to look at the relationship between Y , a stock's average return, and X , the overall market return. The slope coefficient computed by linear regression is called the stock's *beta* by investment analysts. A beta greater than 1 indicates that the stock is relatively sensitive to changes in the market; a beta less than 1 indicates that the stock is relatively insensitive. For the following data, compute the beta and test to see whether it is significantly less than 1. Use $\alpha = 0.05$.

$Y(\%)$	10	12	8	15	9	11	8	10	13	11
$X(\%)$	11	15	3	18	10	12	6	7	18	13

SC 12-7 In a regression problem with a sample size of 17, the slope was found to be 3.73 and the standard error of estimate 28.654. The quantity $(\sum X^2 - n\bar{X}^2) = 871.56$.

- Find the standard error of the regression slope coefficient.
- Construct a 98 percent confidence interval for the population slope.
- Interpret the confidence interval of part (b).

Basic Concepts

12-33 In a regression problem with a sample size of 25, the slope was found to be 1.12 and the standard error of estimate 8.516. The quantity $(\sum X^2 - n\bar{X}^2) = 327.52$.

- Find the standard error of the regression slope coefficient.
- Test whether the regression coefficient is different from 0 at a significance level of 0.05.
- Construct a 95 percent confidence interval for the population slope.

Applications

12-34 Ned's Beds is considering hiring an advertising firm to stimulate business. Ned's brother Fred has done some research in the bed advertising field, and he has collected the following data concerning the amount of profit (Y) a bed company earns and the amount spent on advertising (X). If Fred computes the regression equation, the slope of the line will indicate the amount of profit increase per dollar spent on advertising. Ned will advertise only if the amount of profit earned from \$1 in advertising exceeds \$1.50. Compute the slope of the regression equation and test whether it is greater than 1.50. At a significance level of 0.05, will Ned advertise?

Amount of Advertising (X), \$ hundreds	3.6	4.8	9.7	12.6	11.5	10.9
Amount of Profit (Y) hundreds	12.13	14.7	22.83	28.4	28.33	27.05
Amount of Advertising (X), \$ hundreds	14.6	18.2	3.7	9.8	12.4	16.9
Amount of Profit (Y), hundreds	33.6	40.8	9.4	24.84	30.17	34.7

12-35 A broker for a local investment firm has been studying the relationship between increases in the price of gold (X) and her customers' requests to liquidate stocks (Y). From a data set based on 15 observations, the sample slope was found to be 2.9. If the standard error of the regression slope coefficient is 0.18, is there reason to believe (at the 0.05 significance level) that the slope has changed from its past value of 3.2?

12-36 For a sample of 25, the slope was found to be 1.685 and the standard error of the regression coefficient was 0.11. Is there reason to believe that the slope has changed from its past value of 1.50? Use the 0.05 significance level.

12-37 Realtors are often interested in seeing how the appraised value of a home varies according to the size of the home. Some data on area (in thousands of square feet) and appraised value (in thousands of dollars) for a sample of 11 homes follow.

Area	1.1	1.5	1.6	1.6	1.4	1.3	1.1	1.7	1.9	1.5	1.3
Value	75	95	110	102	95	87	82	115	122	98	90

- Estimate the least-squares regression to predict appraised value from size.
- Generally, realtors feel that a home's value goes up by \$50,000 (= 50 thousands of dollars) for every additional 1,000 square feet in area. For this sample, does this relationship seem to hold? Use $\alpha = 0.10$.

- 12-38** In 1969, a government health agency found that in a number of counties, the relationship between smokers and heart-disease fatalities per 100,000 population had a slope of 0.08. A recent study of 18 counties produced a slope of 0.147 and a standard error of the regression slope coefficient of 0.032.
- Construct a 90 percent confidence interval estimate of the slope of the true regression line. Does the result from this study indicate that the true slope has changed?
 - Construct a 99 percent confidence interval estimate of the slope of the true regression line. Does the result from this study indicate that the true slope has changed?
- 12-39** The local phone company has always assumed that the average number of daily phone calls goes up by 1.5 for each additional person in a household. It has been suggested that people are more talkative than this. A sample of 64 households was taken, and the slope of the regression of Y (average number of daily phone calls) on X (size of household) was computed to be 1.8 with a standard error of the regression slope coefficient of 0.2. Test whether significantly more calls per additional person are being made than the phone company assumes, using $\alpha = 0.05$. State explicit hypotheses and an explicit conclusion.
- 12-40** College admissions officers are constantly seeking variables with which to predict grade-point averages for applicants. One commonly used variable is high school grade-point average. For one college, past data indicated that the slope was 0.85. A recent small study of 20 students found that the sample slope was 0.70 and the standard error of estimate was 0.60. The quantity $(\sum X^2 - n\bar{X}^2)$ was equal to 0.25. At the 0.01 level of significance, should the college conclude that the slope has changed?

Worked-Out Answers to Self-Check Exercises

SC 12-6

	X	Y	XY	X^2	Y^2
	11	10	110	121	100
	15	12	180	225	144
	3	8	24	9	64
	18	15	270	324	225
	10	9	90	100	81
	12	11	132	144	121
	6	8	48	36	64
	7	10	70	49	100
	18	13	234	324	169
	13	11	143	169	121
$\Sigma X = 113$	$\Sigma Y = 107$	$\Sigma XY = 1,301$	$\Sigma X^2 = 1,501$	$\Sigma Y^2 = 1,189$	

$$\bar{X} = \frac{113}{10} = 11.3 \quad \bar{Y} = \frac{107}{10} = 10.7$$

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{1,301 - 10(11.3)(10.7)}{1,501 - 10(11.3)^2} = 0.4101$$

$$a = \bar{Y} - b\bar{X} = 10.7 - 0.4101(11.3) = 6.0659$$

(Computer packages: 6.0660)

$$s_e = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n-2}}$$

$$= \sqrt{\frac{1,189 - 6.0659(107) - 0.4101(1,301)}{8}} = 0.8950$$

(Computer packages: 0.8953)

$$s_b = \frac{s_e}{\sqrt{\sum X^2 - n\bar{X}^2}} = \frac{0.8950}{\sqrt{224.1}} = 0.060$$

$$H_0: B = 1 \quad H_1: B < 1 \quad \alpha = 0.05$$

The standardized statistic is $t = \frac{b - B_{H_0}}{s_b} = \frac{0.4101 - 1}{0.06} = -9.83$. Because the critical value of $t(-1.860)$ is greater than -9.83 , we reject H_0 . Stock is insensitive to changes in the market (the slope is significantly <1).

SC 12-7 (a) $s_b = \frac{s_e}{\sqrt{\sum X^2 - n\bar{X}^2}} = \frac{28.654}{\sqrt{871.56}} = 0.9706$

(b) The 98 percent confidence interval is

$$b \pm t(s_b) = 3.73 \pm 2.602(0.9706) = 3.73 \pm 2.53 = (1.20, 6.26).$$

(c) In repeated sampling, 98 out of 100 intervals constructed as above would contain the true, unknown population slope B . For our single sample, we can say that we are 98 percent confident that our computed interval contains B .

12.5 USING REGRESSION AND CORRELATION ANALYSES: LIMITATIONS, ERRORS, AND CAVEATS

Regression and correlation analyses are statistical tools that, when properly used, can significantly help people make decisions. Unfortunately, they are often misused. As a result, decision makers often make inaccurate forecasts and less-than-desirable decisions. We'll mention the most common errors made in the use of regression and correlation in the hope that you will avoid them.

Misuse of regression and correlation

Extrapolation beyond the Range of the Observed Data

A common mistake is to assume that the estimating line can be applied over any range of values. Hospital administrators can properly use regression analysis to predict the relationship between costs per bed and occupancy levels at various occupancy levels. Some administrators, however, incorrectly use the same regression equation to predict the costs per bed for occupancy levels that are significantly higher than those that were used to estimate the

Specific limited range over which regression equation holds

regression line. Although one relationship holds over the range of sample points, an entirely different relationship may exist for a different range. As a result, these people make decisions on one set of costs and find that the costs change drastically as occupancy increases (owing to things such as overtime costs and capacity constraints). Remember that **an estimating equation is valid only over the same range as the one from which the sample was taken initially.**

Cause and Effect

Another mistake we can make when we use regression analysis is to assume that a change in one variable is caused by a change in the other variable. As we discussed earlier, **regression and correlation analyses can in no way determine cause and effect.** If we say that there is a correlation between students' grades in college and their annual earnings 5 years after graduation, we are *not* saying that one causes the other. Rather, both may be caused by other factors, such as sociological background, parental attitudes, quality of teachers, effectiveness of the job-interviewing process, and economic status of parents—to name only a few potential factors.

Regression and correlation analyses do not determine cause and effect

We have extensively used the example about research and development expenses and annual profits to illustrate various aspects of regression analysis. But it is really highly unlikely that profits in a given year are *caused* by R&D expenditures in that year. Certainly, it would be foolhardy for the VP for R&D to suggest to the chief executive that profits could be immediately increased merely by increasing R&D expenditures. Particularly in high-technology industries, the R&D activity can be used to explain profits, but a better way to do so would be to predict current profits in terms of past research and development expenditures as well as in terms of economic conditions, dollars spent on advertising, and other variables. This can be done by using the multiple-regression techniques, to be discussed in the next chapter.

Using Past Trends to Estimate Future Trends

We must take care to reappraise the historical data we use to estimate the regression equation. Conditions can change and violate one or more of the assumptions on which our regression analysis depends. Earlier in this chapter, we made the point that we assume that the variance of the disturbance e around the mean is constant. In many situations, however, this variance changes from year to year.

Conditions change and invalidate the regression equation

Another error that can arise from the use of historical data concerns the dependence of some variables on time. Suppose a firm uses regression analysis to determine the relationship between the number of employees and the production volume. If the observations used in the analysis extend back for several years, the resulting regression line may be too steep because it may fail to recognize the effect of changing technology.

Values of variables change over time

Misinterpreting the Coefficients of Correlation and Determination

The coefficient of correlation is occasionally misinterpreted as a percentage. If $r = 0.6$, it is incorrect to state that the regression equation "explains" 60 percent of the total variation in Y . Instead, if $r = 0.6$, then r^2 must be $0.6 \times 0.6 = 0.36$. Only 36 percent of the total variation is explained by the regression line.

Misinterpreting r and r^2

The coefficient of determination is misinterpreted if we use r^2 to describe the percentage of the change in the dependent variable that is *caused* by a change in the independent variable. This is wrong because r^2 is a measure only of how well one variable describes another, *not* of how much of the change in one variable is caused by the other variable.

Finding Relationships When They Do Not Exist

When applying regression analysis, people sometimes find a relationship between two variables that, in fact, have no common bond. Even though one variable does not cause a change in the other, they think that there must be some factor common to both variables. It might be possible, for example, to find a statistical relationship between a random sample of the number of miles per gallon consumed by eight different cars and the distance from earth to each of the other eight planets. But because there is absolutely no common bond between gas mileage and the distance to other planets, this “relationship” would be meaningless.

Relationships that have no common bond

In this regard, if one were to run a large number of regressions between many pairs of variables, it would probably be possible to get some rather interesting suggested “relationships.” It might be possible, for example, to find a high statistical relationship between your income and the amount of beer consumed in the United States, or even between the length of a freight train (in cars) and the weather. But in neither case is there a factor common to both variables; hence, such “relationships” are meaningless. As in most other statistical situations, it takes *both* knowledge of the inherent limitations of the technique that is used *and* a large dose of common sense to avoid coming to unwarranted conclusions.

Finding things that do not exist

HINTS & ASSUMPTIONS

Warning: Smart managers *ought* to be able to reason toward a common-sense connection between two variables even before they run a regression analysis on those variables. But computer regressions of large databases sometimes turn up surprising results in terms of unexpected relationships. That doesn’t invalidate common sense at all. What it suggests is that these same smart managers should retest these “surprises” on a new sample to see whether the “surprising” relationship continues to hold true. Hint: What you *may* have is a data problem, not a breakdown of common sense.

EXERCISES 12.5

- 12-41 Explain why an estimating equation is valid over only the range of values used for its development.
- 12-42 Explain the difference between the coefficient of determination and the coefficient of correlation.
- 12-43 Why should we be cautious in using past data to predict future trends?
- 12-44 Why must we not attribute causality in a relationship even when there is strong correlation between the variables or events?

STATISTICS AT WORK

Loveland Computers

Case 12: Simple Regression and Correlation Loveland Computers was running its production line more often to assemble computers from readily available components as the demand for high-end computers grew. Walter Azko was very clear that this was just assembly, not “real manufacturing.” He often joked that the only part that was unique to Loveland Computers was the plastic base to the keyboard—it was distinguished by the Loveland logo (an outline of the Front Range of the Rocky Mountains, just as it was visible from the window in Walt’s office). The base came in two parts that snapped together. And that was the next problem referred to Lee Azko. Nancy Rainwater, the production supervisor, explained her frustrations to Lee.

“When we started assembling this model last summer, the keyboard bases seemed to go together just fine. Now we are having to reject a lot of them because the little lugs that hold the top to the bottom break off when the operator tries to snap them together. When that happens, we have to throw out both pieces. We do not have any way to recycle that kind of plastic, and it doesn’t seem right to be sending all that to the landfill—not to mention what it must be doing to our costs.

“I’ve talked to purchasing and I had Tyronza Wilson inspect the bases when they are delivered. The lugs measure exactly within specs, and the plastics company that makes them for us did some lab work. They say there is nothing wrong with the plastic they are using.

“I noticed that we had more breakages early in the morning—so I wondered whether it just happened because people were being careless on the line. I even wondered if it was because the employees were not properly trained; but the fact is these people are more experienced now than they were last summer—we really have not had much turnover.

“Tyronza wondered if it is not happening because the plastic’s too cold. That might fit with more defects in the winter. But the warehouse has a couple of heaters, so I am not sure if that is right. And I can hardly walk around with a thermometer and check out the temperature of each set of base parts before sending them down the line, can I?”

“Maybe there is another way to figure this out,” Lee said, remembering that it had been quite simple to get weather statistics from the National Weather Service. “You did record the number of bases thrown away for each day you ran the production line, did not you?”

Study Questions: How should Lee investigate the relationship between the weather and the problem with the plastic bases? Will this “prove” whether Tyronza’s explanation is correct?

CHAPTER REVIEW

Terms Introduced in Chapter 12

Coefficient of Correlation The square root of the coefficient of determination. Its sign indicates the direction of the relationship between two variables, direct or inverse.

Coefficient of Determination A measure of the proportion of variation in Y , the dependent variable, that is explained by the regression line, that is, by Y ’s relationship with the independent variable.

Correlation Analysis A technique to determine the degree to which variables are linearly related.

Curvilinear Relationship An association between two variables that is described by a curved line.

Dependent Variable The variable we are trying to predict in regression analysis.

Direct Relationship A relationship between two variables in which, as the independent variable’s value increases, so does the value of the dependent variable.

Estimating Equation A mathematical formula that relates the unknown variable to the known variables in regression analysis.

Independent Variables The known variable or variables in regression analysis.

Inverse Relationship A relationship between two variables in which, as the independent variable increases, the dependent variable decreases.

Least-Squares Method A technique for fitting a straight line through a set of points in such a way that the sum of the squared vertical distances from the n points to the line is minimized.

Linear Relationship A particular type of association between two variables that can be described mathematically by a straight line.

Multiple Regression The statistical process by which several variables are used to predict another variable.

Regression The general process of predicting one variable from another by statistical means, using previous data.

Regression Line A line fitted to a set of data points to estimate the relationship between two variables.

Scatter Diagram A graph of points on a rectangular grid; the X and Y coordinates of each point correspond to the two measurements made on some particular sample element, and the pattern of points illustrates the relationship between the two variables.

Slope A constant for any given straight line whose value represents how much each unit change of the independent variable changes the dependent variable.

Standard Error of Estimate A measure of the reliability of the estimating equation, indicating the variability of the observed points around the regression line, that is, the extent to which observed values differ from their predicted values on the regression line.

Standard Error of the Regression Coefficient A measure of the variability of sample regression coefficient around the true population regression coefficient.

Y -Intercept A constant for any given straight line whose value represents the value of the Y variable when the X variable has a value of 0.

Equations Introduced in Chapter 12

12-1

$$Y = a + bX$$

p. 617

This is the equation for a *straight line*, where the dependent variable Y is “determined” by the independent variable X . The a is called the *Y -intercept* because its value is the point at which the line crosses the Y -axis (the vertical axis). The b is the *slope* of the line; that is, it tells how much each unit change of the independent variable X changes the dependent variable Y . Both a and b are numerical constants because for any given straight line, their values do not change.

12-2

$$b = \frac{Y_2 - Y_1}{X_2 - X_1}$$

p. 618

To calculate the numerical constant b for any given line, find the value of the coordinates, X and Y , for two points that lie on the line. The coordinates of the first point are (X_1, Y_1) and the second point (X_2, Y_2) . Remember that b is the slope of the line.

12-3

$$\hat{Y} = a + bX$$

p. 621

In regression analysis, \hat{Y} (*Y hat*) symbolizes the individual Y values of the *estimated* points, that is, the points that lie on the estimating line. Accordingly, Equation 12-3 is the equation for the estimating line.

12-4 $b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$ p. 624

The equation enables us to calculate the *slope of the best-fitting regression line* for any two-variable set of data points. We introduce two new symbols in this equation, \bar{X} and \bar{Y} , which represent the means of the values of the independent variable and the dependent variable, respectively. In addition, this equation contains n , which, in this case, represents the number of data points to which we are fitting the regression line.

12-5 $a = \bar{Y} - b\bar{X}$ p. 624

Using this formula, we can compute the *Y-intercept of the best-fitting regression line* for any two-variable set of data points.

12-6 $s_e = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}}$ p. 629

The *standard error of estimate*, s_e , measures the variability or scatter of the observed values around the regression line. In effect, it indicates the reliability of the estimating equation. The denominator is $n - 2$ because we lose 2 degrees of freedom (for the values a and b) in estimating the regression line.

12-7 $s_e = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n-2}}$ p. 631

Because Equation 12-6 requires tedious calculations, statisticians have devised this *short-cut method for finding the standard error of estimate*. In calculating the values for b and a , we have already calculated every quantity in Equation 12-7 except $\sum Y^2$, which we can do very easily.

12-8 Variation of the Y values around the regression line = $\sum(Y - \hat{Y})^2$ p. 643

The variation of the Y values in a data set around the fitted regression line is one of two quantities from which the sample coefficient of determination is developed. Equation 12-8 shows how to measure this particular dispersion, which is the *unexplained* portion of the total variation of the Y values.

12-9 Variation of the Y values around their own mean = $\sum(Y - \bar{Y})^2$ p. 644

This formula measures the *total variation* of a whole set of Y values, that is, the variation of these Y values around their own mean.

12-10 $\tilde{r}^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$ p. 644

The *sample coefficient of determination*, r^2 , gives the fraction of the total variation of Y that is explained by the regression line. It is an important measure of the degree of association between X and Y . If the value of r^2 is +1, then the regression line is a perfect estimator. If $r^2 = 0$, there is no correlation between X and Y .

12-11 $r^2 = \frac{a \sum Y + b \sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$ p. 649

This is a short-cut equation for calculating r^2

12-12

$$r = \sqrt{r^2}$$

p. 651

The *sample coefficient of correlation* is denoted by r and is found by taking the square root of the sample coefficient of determination. It is a second measure (in addition to r^2) we can use to describe how well one variable is explained by another. The sign of r is the same as the sign of b ; it indicates the direction of the relationship between the two variables X and Y .

12-13

$$Y = A + BX$$

p. 657

Each *population regression line* is of the form in Equation 12-13, where A is the Y -intercept for the population and B is the slope.

12-13a

$$Y = A + BX + e$$

p. 658

Because all the individual points in a population do not lie on the population regression line, the *individual* data points will satisfy Equation 12-13a, where e is a random disturbance from the population regression line. On the average, e equals zero because disturbances above the population regression line are canceled out by disturbances below it.

12-14

$$s_b = \frac{s_e}{\sqrt{\sum X^2 - n\bar{X}^2}}$$

p. 659

When we are dealing with a sample, we can use this formula to find the *standard error of the regression coefficient*, b .

12-15

$$t = \frac{b - B_{H_0}}{s_b}$$

p. 659

Once we have calculated s_b using Equation 12-14, we can use this equation to standardize the observed value of the regression coefficient. Then we perform the hypothesis test by comparing this standardized value with the critical value(s) from Appendix Table 2.

Review and Application Exercises

12-45

A consultant is interested in seeing how accurately a new job-performance index measures what is important for a corporation. One way to check is to look at the relationship between the job-evaluation index and an employee's salary. A sample of eight employees was taken, and information about salary (in thousands of dollars) and job-performance index (1–10; 10 is best) was collected.

Job-performance index (X)	9	7	8	4	7	5	5	6
Salary (Y)	36	25	33	15	28	19	20	22

(a) Develop an estimating equation that best describes these data.

(b) Calculate the standard error of estimate, s_e , for these data.

(c) Calculate the sample coefficient of determination, r^2 , for these data.

12-46

The Stork Foundation wishes to show with statistics that, contrary to popular belief, storks *do* bring babies. Thus, it has collected data on the number of storks and the number of babies (both in thousands) in several large cities in central Europe.

Storks	27	38	13	24	6	19	15
Babies	35	46	19	32	15	31	20

- (a) Compute the sample coefficient of determination and the sample correlation coefficient for these data.

- (b) Has statistical science disproved popular belief?

12-47 (Fill in the blanks.) Regression and correlation analyses deal with the _____ between variables. Regression analysis, through _____ equations, enables us to _____ an unknown variable from a set of known variables. The unknown variable is called the _____ variable; known variables are called _____ variables. The correlation between two variables indicates the _____ of the linear relationship between them and thus gives an idea of how well the _____ in regression describes the relationship between the variables.

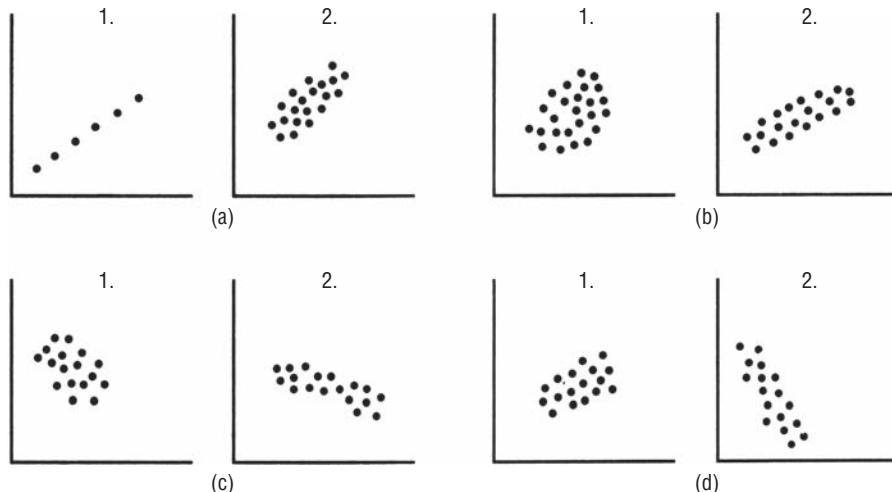
12-48 Calculate the sample coefficient of determination and the sample correlation coefficient for Exercise 12-14.

12-49 “Nothing succeeds like success” is an old adage in the advertising business. The president of a multiline auto dealership has observed that sales staff who earn the biggest end-of-year bonus are the ones who are most likely to exceed their quota for sales in the following year (and hence earn another bonus).

Last Year's Bonus (\$ thousands)	7.8	6.9	6.7	6.0	6.9	5.2
This Year's Sales Over Quota	64	73	42	49	71	46
Last Year's Bonus (\$ thousands)	6.3	8.4	7.2	10.1	10.8	7.7
This Year's Sales Over Quota	32	88	53	84	85	93

- (a) Develop the line of best fit to describe these data.
 (b) Calculate the standard error of estimate for the relationship.
 (c) Develop an approximate 90 percent confidence interval for predicting the sales over quota for a sales staff member who earned a bonus of \$9,600 last year.

12-50 For each of the following pairs of plots, state which has a higher value of r , the correlation coefficient, and state the sign of r .



- 12-51** An operations manager is interested in predicting costs C (in thousands of dollars) based on the amount of raw material input R (in hundreds of pounds) for a jeans manufacturer. If the slope is significantly greater than 0.5 in the following sample data, then there is something wrong with the production process and the assembly-line machinery should be adjusted. At the 0.05 significance level, should the machinery be adjusted? State explicit hypotheses and an explicit conclusion.

C	10	7	5	6	7	6
R	25	20	16	17	19	18

- 12-52** Calculate the sample coefficient of determination and the sample correlation coefficient for Exercise 12-13.

- 12-53** We should not extrapolate to predict values outside the range of data used in constructing the regression line. The reason (choose one):

- (a) The relationship between the variables may not be the same for different values of the variables.
- (b) The independent variable may not have the causal effect on the dependent variable for these values.
- (c) The variables' values may change over time.
- (d) There may be no common bond to explain the relationship.

- 12-54** Economists are often interested in estimating consumption functions. This is done by regressing consumption Y on income X . (For this regression, economists call the slope the *marginal propensity to consume*.) For a sample of 25 families, a slope of 0.87 and a standard error of the regression slope coefficient of 0.035 were computed. For this sample, has the marginal propensity to consume decreased below the standard of 0.94? Use $\alpha = 0.05$. State explicit hypotheses and an explicit conclusion.

- 12-55** Unlike the coefficient of determination, the coefficient of correlation (choose one)

- (a) Indicates whether the slope of the regression line is positive or negative.
- (b) Measures the strength of association between the two variables more exactly.
- (c) Can never have an absolute value greater than 1.
- (d) Measures the percentage of variance explained by the regression line.

- 12-56** Are good grades in college important for earning a good salary? A business statistics student has taken a random sample of starting salaries and college grade-point averages for some recently graduated friends of his. The data follow:

Starting salary (\$ thousands)	36	30	30	24	27	33	21	27
Grade-point average	4.0	3.0	3.5	2.0	3.0	3.5	2.5	2.5

- (a) Plot these data.
- (b) Develop the estimating equation that best describes these data.
- (c) Plot the estimating equation on the scatter plot of part (a).

- 12-57** A landlord is interested in seeing whether his apartment rents are typical. Thus, he has taken a random sample of 11 rents and apartment sizes of similar apartment complexes. The data follow:

Rent	230	190	450	310	218	185	340	245	125	350	280
Number of bedrooms	2	1	3	2	2	2	2	1	1	2	2

(a) Develop an estimating equation that best describes these data.

(b) Calculate the coefficient of determination.

(c) Predict the rent for a two-bedroom apartment.

- 12-58** Many small companies buy advertising without considering its effect. "Hamburger wars" (substantial price rivalry with special "value meals") have cut the profits of Ethiopian Burgers of Santa Cruz, California, a small regional chain. The marketing manager is trying to make the case that "you have to spend money to make money." Spending on billboard advertisements, in the manager's opinion, has a direct result on sales. There are records for 7 months:

Monthly expenditure on billboards ($\times \\$1,000$)	25	16	42	34	10	21	19
Monthly sales revenue ($\times \\$100,000$)	34	14	48	32	26	29	20

(a) Develop an estimating equation that best describes these data.

(b) Calculate the standard error of estimate for this relationship.

(c) For a month with a billboard expenditure of \$28,000, develop an approximate 95 percent confidence interval for the expected monthly sales for that month.

- 12-59** In an FAA study of airline operations, a survey of 18 companies disclosed that the relationship between the number of pilots employed and the number of planes in service has a slope of 4.3. Previous studies indicated that the slope of this relationship was 4.0. If the standard error of the regression slope coefficient has been calculated to be 0.17, is there reason to believe, at the 0.05 level of significance, that the true slope has changed?

- 12-60** Dave Proffitt, a second-year MBA student, is doing a study of companies going public for the first time. He is curious to see whether there is a significant relationship between the size of the offering (in millions of dollars) and the price per share.

(a) Given the following data, develop the estimating equation that best fits the data.

Size (\$ Millions)	Price (\$)
108.00	12.00
4.40	4.00
3.50	5.00
3.60	6.00
39.00	13.00
68.40	19.00
7.50	8.50
5.50	5.00
375.00	15.00
12.00	6.00
51.00	12.00

(Continued)

(Contd.)

66.00	12.00
10.40	6.50
4.00	3.00

- (b) Calculate the sample coefficient of determination. Should Dave use this regression equation for predictive purposes or search elsewhere for additional explanatory variables?

- 12-61** A manufacturer of cellular phones is testing two different types of batteries to see how long they last in typical use. Provisional data are in the following table:

Hours of Daily Use	Approximate Life (months)	
	Lithium	Alkaline
2.0	3.1	1.3
1.5	4.2	1.6
1.0	5.1	1.8
0.5	6.3	2.2

- (a) Develop two linear estimating equations, one to predict product life based on daily use with lithium batteries and one for alkaline batteries.
- (b) Find an approximate 90 percent confidence interval for the life (in months) with 1.25 hours of daily use, for each battery type. Can the company make any claims about which battery will provide a longer life based on these numbers?

- 12-62** A study has been proposed to investigate the relationship between the birthweight of male babies and their adult height. Using the following data, develop the least-squares estimating equation. What percentage of the variation in adult height is explained by this regression line?

Birthweight	Adult Height
5 lb, 8 oz	5'9"
7 lb	6'
6 lb, 4 oz	5'6"
7 lb, 8 oz	5'11"
8 lb, 2 oz	6'1"
6 lb, 12 oz	5'10"

- 12-63** Many college students transfer in the summer before their junior years. To aid in evaluating the academic potential of these junior transfers, Barbara Hoopes, the Dean of Admissions at Piedmont College, is conducting an analysis that compares students' grade-point averages (GPAs) during their first 2 years of college with their GPAs during their final 2 years, after transferring. Using the following data:

Freshman/sophomore GPA	1.7	3.5	2.3	2.6	3.0	2.8	2.4	1.9	2.0	3.1
Junior/senior GPA	2.4	3.7	2.0	2.5	3.2	3.0	2.5	1.8	2.7	3.7

- (a) Calculate the least-squares estimating equation Hoopes should use to predict junior/senior GPAs for students transferring to Piedmont College.
- (b) Hoopes will not admit junior-transfer applicants unless approximate 90 percent prediction intervals for their junior/senior GPAs fall entirely above 2.0. Will she admit a transfer applicant with a 2.5 freshman/sophomore GPA?

12-64 The Accounts Department of MNT Enterprises wants to decide on Travelling Allowances along with Boarding-Lodging Expenses rates for different cities. So far MNT is following a uniform rate system for all the cities but there has been a major complaint among the employees with respect to this rule. The employees' contention is that there is a major variation in this aspect in different cities. The Accounts Department has accepted this contention and ready to modify the system by introducing variable rate system as desired by the employees. The following table gives Hotel rates and the Taxi rates per day in 10 major cities. Should the accounts department consider both hotel costs and the taxi-rates, or would the hotel costs by city provide sufficient information to decide the allowance-rates?

City	Hotel Rates (Per Day)	Taxi Rates (Per Day)
A	3000	550
B	2800	400
C	1800	200
D	4500	1000
E	3000	700
F	2000	300
G	2500	450
H	2700	600
I	3600	900
J	1900	300

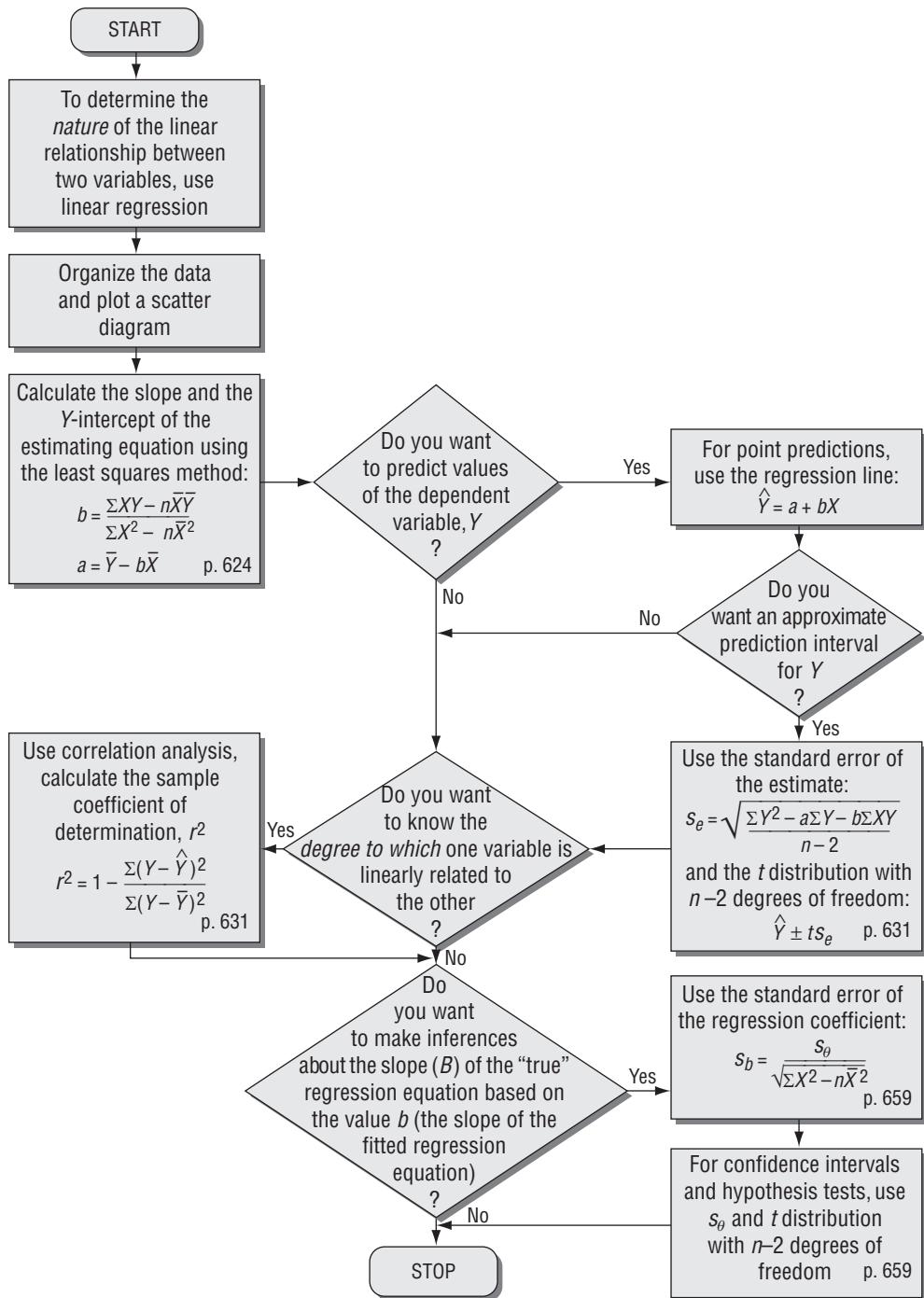


Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Build a regression model to study the impact of reliability of e-banking transactions on the level of satisfaction of customers. {Regress Q9 on 8(b)}. Interpret the results.
2. Study the impact of Bank Site congestion in performing e-transactions on the level of satisfaction of customers. {Regress Q9 on 8(f)}. Interpret the results.



Flow Chart: Regression and Correlation



Multiple Regression and Modeling

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To extend the regression techniques of the last chapter to handle more than one explanatory variable for a quantity we are trying to predict
 - To examine decision-making situations where multiple regression can be used to make predictions
 - To interpret the output from computer regression packages
 - To test hypotheses about regressions
 - To use modeling techniques to incorporate qualitative variables into regression equations
 - To learn how to fit curves to data
 - To understand the importance of residuals in regression analysis
-

CHAPTER CONTENTS

13.1	Multiple Regression and Correlation Analysis	678
13.2	Finding the Multiple-Regression Equation	679
13.3	The Computer and Multiple Regression	688
13.4	Making Inferences about Population Parameters	698
13.5	Modeling Techniques	717

■ Statistics at Work	733
■ Terms Introduced in Chapter 13	734
■ Equations Introduced in Chapter 13	735
■ Review and Application Exercises	736
■ Flow Chart: Multiple Regression and Modeling	745

A manufacturer of small office copiers and word-processing machinery pays its salespeople a small base salary plus a commission equal to a fixed percentage of the person's sales. One of the salespeople charges that this salary structure discriminates against women. Current base salaries for the firm's nine salespeople are as follows:

Salesmen		Saleswomen	
Months Employed	Base Salary (\$1,000s)	Months Employed	Base Salary (\$1,000s)
6	7.5	5	6.2
10	8.6	13	8.7
12	9.1	15	9.4
18	10.3	21	9.8
30	13.0		

The director of personnel sees that base salary depends on length of service, but she does not know how to use the data to learn whether it also depends on gender and whether there is discrimination against women. Methods in this chapter will enable her to find out. ■

13.1 MULTIPLE REGRESSION AND CORRELATION ANALYSIS

As we mentioned in Chapter 12, we can use more than one independent variable to estimate the dependent variable and, in this way, attempt to increase the accuracy of the estimate. This process is called multiple regression and correlation analysis. It is based on the same assumptions and procedures we have encountered using simple regression.

Consider the real-estate agent who wishes to relate the number of houses the firm sells in a month to the amount of her monthly advertising. Certainly, we can find a simple estimating equation that relates these two variables. Could we also improve the accuracy of our equation by including in the estimating process the number of salespeople she employs each month? The answer is probably yes. And now, because we want to use both the number of sales agents and the advertising expenditures to predict monthly house sales, we must use *multiple*, not simple, regression to determine the relationship.

The principal advantage of *multiple regression* is that it allows us to use more of the information available to us to estimate the dependent variable. Sometimes the correlation between two variables may be insufficient to determine a reliable estimating equation. Yet, if we add the data from more independent variables, we may be able to determine an estimating equation that describes the relationship with greater accuracy.

Multiple regression and correlation analysis involve a three-step process such as the one we used in simple regression. In this process, we

1. Describe the multiple-regression equation.
2. Examine the multiple-regression standard error of estimate.
3. Use multiple-correlation analysis to determine how well the regression equation describes the observed data.

Using more than one independent variable to estimate the dependent variable

Advantage of multiple regression

Steps in multiple regression and correlation

In addition, in multiple regression, we can look at each individual independent variable and test whether it contributes significantly to the way the regression describes the data.

In this chapter, we shall see how to find the best-fitting regression equation for a given set of data and how to analyze the equation we get. Although we shall show how to do multiple regression by hand or on a hand-held calculator, it will quickly become obvious to you that you would not want to do even a modest-size real-life problem by hand. Fortunately, there are available many computer packages for doing multiple regressions and other statistical analyses. These packages do the “number crunching” and leave you free to concentrate on analyzing the significance of the resulting estimating equation.

Computer regression packages

Multiple regression will also enable us to fit curves as well as lines. Using the technique of *dummy variables*, we can even include qualitative factors such as gender in our multiple regression. This technique will enable us to analyze the discrimination problem that opened this chapter. Dummy variables and fitting curves are only two of the many *modeling techniques* that can be used in multiple regression to increase the accuracy of our estimating equations.

EXERCISES 13.1

Basic Concepts

- 13-1** Why would we use multiple regression instead of simple regression in estimating a dependent variable?
- 13-2** How will dummy variables be used in our study of multiple regression?
- 13-3** To what does the word *multiple* refer in the phrase *multiple regression*?
- 13-4** The owner of a chain of stores would like to predict monthly sales from the size of city in which a store is located. After fitting a simple regression model, she decides that she wants to include the effect of season of the year in the model. Can this be done using the techniques in this chapter?
- 13-5** Describe the three steps in the process of multiple regression and correlation analysis.
- 13-6** Will the procedures used in multiple regression differ greatly from those we used in simple regression? Explain.

13.2 FINDING THE MULTIPLE-REGRESSION EQUATION

Let's see how we can compute the multiple-regression equation. For convenience, we shall use only two independent variables in the problem we work in this section. Keep in mind, however, that the same sort of technique is, in principle, applicable to any number of independent variables.

A problem demonstrating multiple regression

The Internal Revenue Service is trying to estimate the monthly amount of unpaid taxes discovered by its auditing division. In the past, the IRS estimated this figure on the basis of the expected number of field-audit labor hours. In recent years, however, field audit labour hours have become an erratic predictor of the actual unpaid taxes. As a result, the IRS is looking for another factor with which it can improve the estimating equation.

The auditing division does keep a record of the number of hours its computers are used to detect unpaid taxes. Could we combine this information with the data on field-audit labor hours and come up

TABLE 13-1 DATA FROM IRS AUDITING RECORDS DURING THE LAST 10 MONTHS

Month	X_1 Field-Audit Labor Hours (00's omitted)	X_2 Computer Hours (00's omitted)	Y Actual Unpaid Taxes Discovered (millions of dollars)
January	45	16	29
February	42	14	24
March	44	15	27
April	45	13	25
May	43	13	26
June	46	14	28
July	44	16	30
August	45	16	28
September	44	15	28
October	43	15	27

with a more accurate estimating equation for the unpaid taxes discovered each month? Table 13-1 presents these data for the last 10 months.

In simple regression, X is the symbol used for the values of the independent variable. In multiple regression, we have more than one independent variable. So we shall continue to use X , but we shall add a subscript (for example, X_1 , X_2) to distinguish between the independent variables we are using.

In this problem, we will let X_1 represent the number of field-audit labor hours and X_2 represent the number of computer hours. The dependent variable, Y , will be the actual unpaid taxes discovered.

Recall that in simple regression, the estimating equation $\hat{Y} = a + bX$ describes the relationship between the two variables X and Y . In multiple regression, we must extend that equation, adding one term for each new variable. In symbolic form, Equation 13-1 is the formula we can use when we have two independent variables:

Appropriate symbols

Defining the variables

Estimating equation for multiple regression

Estimating Equation Describing Relationship among Three Variables

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

[13-1]

where

- \hat{Y} = estimated value corresponding to the dependent variable
- a = Y -intercept
- X_1 and X_2 = values of the two independent variables
- b_1 and b_2 = slopes associated with X_1 and X_2 , respectively

We can visualize the simple estimating equation as a line on a graph; similarly, we can picture a two-variable multiple regression equation as a plane, such as the one shown in Figure 13-1. Here

Visualizing multiple regression

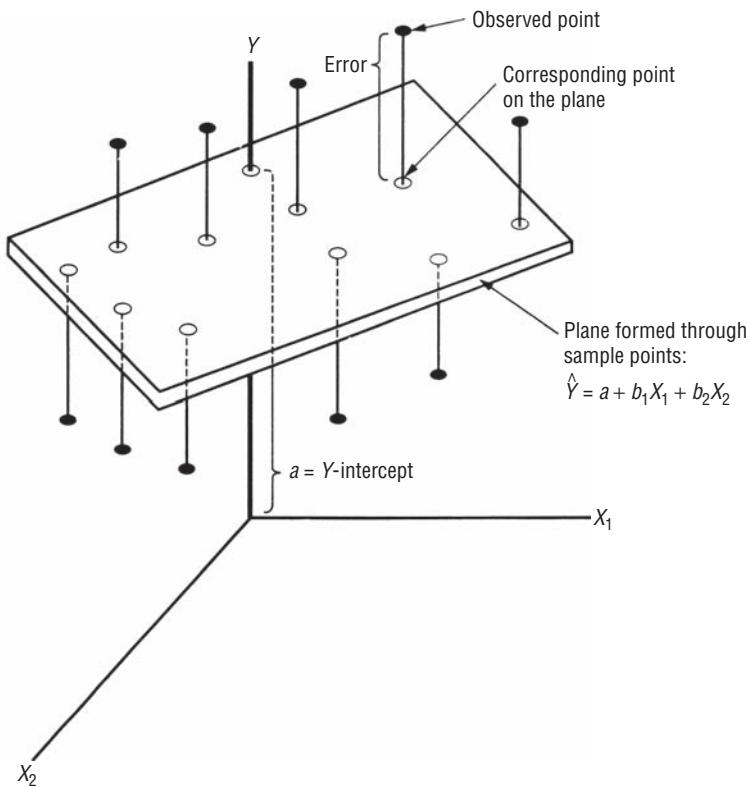


FIGURE 13-1 MULTIPLE REGRESSION PLANE FOR 10 DATA POINTS

we have a three-dimensional shape that possesses depth, length, and width. To get an intuitive feel for this three-dimensional shape, visualize the intersection of the axes, Y , X_1 , and X_2 as one corner of a room.

Figure 13-1 is a graph of 10 sample points and the plane about which these points seem to cluster. Some points lie above the plane and some fall below it—just as points lie above and below the simple regression line.

Our problem is to decide which of the possible planes that we could draw will be the best fit. To do this, we shall again use the least-squares criterion and locate the plane that minimizes the sum of the squares of the errors, that is, the distances from the points around the plane to the corresponding points *on* the plane. We use our data and the following three equations (which statisticians call the “normal equations”) to determine the values of the numerical constants, a , b_1 , and b_2 .

*Using the least-squares criterion
to fit a regression plane*

Normal Equations

$$\Sigma Y = na + b_1 \Sigma X_1 + b_2 \Sigma X_2 \quad [13-2]$$

$$\Sigma X_1 Y = a \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2 \quad [13-3]$$

$$\Sigma X_2 Y = a \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2 \quad [13-4]$$

TABLE 13-2 VALUES FOR FITTING LEAST-SQUARES PLANE, WHERE $n = 10$

Y (1)	X_1 (2)	X_2 (3)	$X_1 Y$ (2) \times (1)	$X_2 Y$ (3) \times (1)	$X_1 X_2$ (2) \times (3)	X_1^2 (2) ²	X_2^2 (3) ²	Y^2 (1) ²
29	45	16	1,305	464	720	2,025	256	841
24	42	14	1,008	336	588	1,764	196	576
27	44	15	1,188	405	660	1,936	225	729
25	45	13	1,125	325	585	2,025	169	625
26	43	13	1,118	338	559	1,849	169	676
28	46	14	1,288	392	644	2,116	196	784
30	44	16	1,320	480	704	1,936	256	900
28	45	16	1,260	448	720	2,025	256	784
28	44	15	1,232	420	660	1,936	225	784
27	43	15	1,161	405	645	1,849	225	729
272	441	147	12,005	4,013	6,485	19,461	2,173	7,428
↑	↑	↑	↑	↑	↑	↑	↑	↑
ΣY	ΣX_1	ΣX_2	$\Sigma X_1 Y$	$\Sigma X_2 Y$	$\Sigma X_1 X_2$	ΣX_1^2	ΣX_2^2	ΣY^2
$\bar{Y} = 27.2$								
$\bar{X}_1 = 44.1$								
$\bar{X}_2 = 14.7$								

Solving Equations 13-2, 13-3, and 13-4 for a , b_1 and b_2 will give us the coefficients for the regression plane. Obviously, the best way to compute all the sums in these three equations is to use a table to collect and organize the necessary information, just as we did in simple regression. This we have done for the IRS problem in Table 13-2.

Now, using the information from Table 13-2 in Equations 13-2, 13-3, and 13-4, we get three equations in the three unknown constants (a , b_1 , and b_2):

$$\begin{aligned} 272 &= 10a + 441b_1 + 147b_2 \\ 12,005 &= 441a + 19,461b_1 + 6,485b_2 \\ 4,013 &= 147a + 6,485b_1 + 2,173b_2 \end{aligned}$$

**Equation 13-2, 13-3, and 13-4
used to solve for a , b_1 , and b_2**

When we solve these three equations simultaneously, we get

$$\begin{aligned} a &= -13.828 \\ b_1 &= 0.564 \\ b_2 &= 1.099 \end{aligned}$$

Substituting these three values into the general two-variable regression equation (Equation 13-1), we get an equation that describes the relationship among the number of field-audit labor hours, the number of computer hours, and the unpaid taxes discovered by the auditing division:

$$\begin{aligned} \hat{Y} &= a + b_1 X_1 + b_2 X_2 \\ &= -13.828 + 0.564 X_1 + 1.099 X_2 \end{aligned} \quad [13-1]$$

The auditing division can use this equation monthly to estimate the amount of unpaid taxes it will discover.

Suppose the IRS wants to increase its discoveries in the coming month. Because trained auditors are scarce, the IRS does not intend to hire additional personnel. The number of field-audit labor hours, then, will remain at October's level of about 4,300 hours. But in order to increase its discoveries of unpaid taxes, the IRS expects to increase the number of computer hours to about 1,600. As a result:

$$\begin{aligned}X_1 &= 43 \leftarrow 4,300 \text{ hours of field-audit labor} \\X_2 &= 16 \leftarrow 1,600 \text{ hours of computer time}\end{aligned}$$

Substituting these values into the auditing division's regression equation, we get

$$\begin{aligned}\hat{Y} &= -13.828 + 0.564X_1 + 1.099X_2 \\&= -13.828 + (0.564)(43) + (1.099)(16) \\&= -13.828 + 24.252 + 17.584 \\&= 28.008 \leftarrow \text{Estimated discoveries of } \$28,008,000\end{aligned}$$

Therefore, in the November forecast, the audit division can indicate that it expects about \$28 million of discoveries for this combination of factors.

So far, we have referred to a as the Y -intercept and to b_1 and b_2 as the slopes of the multiple-regression plane. But to be more precise, we should say that these numerical constants are the *estimated regression coefficients*. The constant a is the value of \hat{Y} (in this case, the estimated unpaid taxes) if both X_1 and X_2 happen to be zero. The coefficients b_1 and b_2 describe how changes in X_1 and X_2 affect the value of \hat{Y} . In our IRS example, we can hold the number of field-audit labor hours, X_1 , constant and change the number of computer hours, X_2 . When we do, the value of \hat{Y} will increase \$1,099,000 for every additional 100 hours of computer time. Likewise, we can hold X_2 constant and find that, for every 100-hour increase in the number of field-audit labor hours, \hat{Y} increases by \$564,000.

Interpreting our estimate

a , b_1 , and b_2 are the estimated regression coefficients

HINTS & ASSUMPTIONS

Hint: If you have trouble picturing in your mind what multiple regression is actually doing, think back to Chapter 12 and remember that a regression *line* describes the relationship between *two* variables. In multiple regression, the regression *plane* such as the one on page 681 describes the relationship among *three* variables Y , X_1 , and X_2 . The appropriate regression plane is conceptually the same as the appropriate regression line, that is, the one that minimizes the sum of the squared vertical distances between the data points and the plane in this instance. It may help to remember that each independent variable may account for *some* of the variation in the dependent variable. Multiple regression is just a way to use several independent variables to make a better prediction of the dependent variable.

Using the multiple-regression equation to estimate

(Continued)

Assumptions: The classical linear regression model (CLRM) makes certain assumptions about the independent variables X_i 's and the error term u , which are very important for the valid interpretation of the regression estimates. The assumptions are:

1. The regression model is linear in parameters, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

2. The independent variables are assumed to be non random.
3. For the given values of X_i 's, the expected value of the disturbance term is zero, i.e.,

$$E(u/X_i) = 0$$

4. For the given values of X_i 's, the variance of u_i is identical, i.e.,

$$\text{Var}(u/X_i) = \sigma^2 \text{ (Constant)}$$

5. For any two values of X_i , X_j and X_l ($i \neq j$), the correlation between any two u_i and u_j ($i \neq j$) is zero, i.e.,

$$\text{Cov}(u_i u_j / X_i X_j) = 0$$

6. There are no exact linear relationships between the independent variables.
-

EXERCISES 13.2

Self-Check Exercises

SC 13-1 Given the following set of data

- (a) Calculate the multiple-regression plane.
- (b) Predict Y when $X_1 = 3.0$ and $X_2 = 2.7$.

Y	X_1	X_2
25	3.5	5.0
30	6.7	4.2
11	1.5	8.5
22	0.3	1.4
27	4.6	3.6
19	2.0	1.3

SC 13-2 The following information has been gathered from a random sample of apartment renters in a city. We are trying to predict rent (in dollars per month) based on the size of the apartment (number of rooms) and the distance from downtown (in miles).

Rent (\$)	Number of Rooms	Distance from Downtown
360	2	1
1,000	6	1
450	3	2
525	4	3
350	2	10
300	1	4

- (a) Calculate the least-squares equation that best relates these three variables.
 (b) If someone is looking for a two-bedroom apartment 2 miles from downtown, what rent should he expect to pay?

Basic Concepts

13-7 Given the following set of data

- (a) Calculate the multiple-regression plane.
 (b) Predict Y when $X_1 = 10.5$ and $X_2 = 13.6$

Y	X_1	X_2
11.4	4.5	13.2
16.6	8.7	18.7
20.5	12.6	19.8
29.4	19.7	25.4
7.6	2.9	22.8
13.8	6.7	17.8
28.5	17.4	14.6

13-8 For the following set of data:

- (a) Calculate the multiple-regression plane.
 (b) Predict Y for $X_1 = 28$ and $X_2 = 10$.

Y	X_1	X_2
10	8	4
17	21	9
18	21	11
26	17	20
35	36	13
8	9	28

13-9 Given the following set of data

- (a) Calculate the multiple-regression plane.
 (b) Predict Y when $X_1 = -1$ and $X_2 = 4$.

Y	X_1	X_2
6	1	3
10	3	-1
9	2	4
14	-2	7
7	3	2
5	6	-4

Applications

- 13-10** Sam Spade, owner and general manager of the Campus Stationery Store, is concerned about the sales behavior of a compact cassette tape recorder sold at the store. He realizes that there are many factors that might help explain sales, but believes that advertising and price are major determinants. Sam has collected the following data:

Sales (units sold)	Advertising (number of ads)	Price (\$)
33	3	125
61	6	115
70	10	140
82	13	130
17	9	145
24	6	140

- (a) Calculate the least-squares equation to predict sales from advertising and price.
 (b) If advertising is 7 and price is \$132, what sales would you predict?
- 13-11** A developer of food for pigs would like to determine what relationship exists among the age of a pig when it starts receiving a newly developed food supplement, the initial weight of the pig, and the amount of weight it gains in a 1-week period with the food supplement. The following information is the result of a study of eight piglets:

Piglet Number	X_1 Initial Weight (Pounds)	X_2 Initial Age (Weeks)	Y Weight Gain
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

- (a) Calculate the least-squares equation that best describes these three variables.
 (b) How much might we expect a pig to gain in a week with the food supplement if it were 9 weeks old and weighed 48 pounds?
- 13-12** A graduate student trying to purchase a used Neptune car has researched the prices. She believes the year of the car and the number of miles the car has been driven both influence the purchase price. Data are given below for 10 cars with the price (Y) in thousands of dollars, year (X_1) and miles driven (X_2) in thousands.
- (a) Calculate the least-squares equation that best relates these three variables.
 (b) The student would like to purchase a 1991 Neptune with about 40,000 miles on it. How much do you predict she will pay?

(Y) Price (\$ thousands)	X_1 Year	X_2 Miles (thousands)
2.99	1987	55.6
6.02	1992	18.4
8.87	1993	21.3
3.92	1988	46.9
9.55	1994	11.8
9.05	1991	36.4
9.37	1992	28.2
4.2	1988	44.2
4.8	1989	34.9
5.74	1991	26.4

- 13-13** The Federal Reserve is performing a preliminary study to determine the relationship between certain economic indicators and annual percentage change in the gross national product (GNP). Two such indicators being examined are the amount of the federal government's deficit (in billions of dollars) and the Dow Jones Industrial Average (the mean value over the year). Data for 6 years follow:

Y Change in GNP	X_1 Federal Deficit	X_2 Dow Jones
2.5	100	2,850
-1.0	400	2,100
4.0	120	3,300
1.0	200	2,400
1.5	180	2,550
3.0	80	2,700

- (a) Calculate the least-squares equation that best describes the data.
- (b) What percentage change in GNP would be expected in a year in which the federal deficit was \$240 billion and the mean Dow Jones value was 3,000?

Worked-Out Answers to Self-Check Exercises

SC 13-1 (a)	Y	X_1	X_2	$X_1 Y$	$X_2 Y$	$X_1 X_2$	X_1^2	X_2^2	Y^2
	25	3.5	5.0	87.5	125.0	17.50	12.25	25.00	625
	30	6.7	4.2	201.0	126.0	28.14	44.89	17.64	900
	11	1.5	8.5	16.5	93.5	12.75	2.25	72.25	121
	22	0.3	1.4	6.6	30.8	0.42	0.09	1.96	484
	27	4.6	3.6	124.2	97.2	16.56	21.16	12.96	729
	19	2.0	1.3	38.0	24.7	2.60	4.00	1.69	361
	134	18.6	24.0	473.8	497.2	77.97	84.64	131.50	3,220

Equations 13-2, 13-3, and 13-4 become

$$\begin{aligned}\sum Y &= na + b_1 \sum X_1 + b_2 \sum X_2 & 134 &= 6a + 18.6b_1 + 24.0b_2 \\ \sum X_1 Y &= a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 & 473.8 &= 18.6a + 84.64b_1 + 77.97b_2 \\ \sum X_2 Y &= a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 & 497.2 &= 24.0a + 77.97b_1 + 131.50b_2\end{aligned}$$

Solving these equations simultaneously, we get

$$a = 20.3916 \quad b_1 = 2.3403 \quad b_2 = -1.3283$$

So the regression equation is $\hat{Y} = 20.3916 + 2.3403X_1 - 1.3283X_2$.

- (b) With $X_1 = 3.0$ and $X_2 = 2.7$,

$$\hat{Y} = 20.3916 + 2.3403(3.0) - 1.3283(2.7) = 23.83.$$

SC 13-2 (a) In this problem, Y = rent, X_1 = number of rooms, X_2 = distance from downtown.

Y	X_1	X_2	$X_1 Y$	$X_2 Y$	$X_1 X_2$	X_1^2	X_2^2	Y^2
360	2	1	720	360	2	4	1	129,600
1,000	6	1	6,000	1,000	6	36	1	1,000,000
450	3	2	1,350	900	6	9	4	202,500
525	4	3	2,100	1,575	12	16	9	275,625
350	2	10	700	3,500	20	4	100	122,500
300	1	4	300	1,200	4	1	16	90,000
2,985	18	21	11,170	8,535	50	70	131	1,820,225

Equations 13-2, 13-3, and 13-4 become

$$\begin{aligned}\sum Y &= na + b_1 \sum X_1 + b_2 \sum X_2 & 2,985 &= 6a + 18b_1 + 21b_2 \\ \sum X_1 Y &= a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 & 11,170 &= 18a + 70b_1 + 50b_2 \\ \sum X_2 Y &= a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 & 8,535 &= 21a + 50b_1 + 131b_2\end{aligned}$$

Solving these equations simultaneously, we get

$$a = 96.4581 \quad b_1 = 136.4847 \quad b_2 = -2.4035$$

So the regression equation is

$$\hat{Y} = 96.4581 + 136.4847X_1 - 2.4035X_2$$

- (b) When number of rooms = 2 and distance from downtown = 2,

$$\hat{Y} = 96.4581 + 136.4847(2) - 2.4035(2) = \$365$$

13.3 THE COMPUTER AND MULTIPLE REGRESSION

In Chapter 12, and so far in this chapter, we have presented simplified problems and samples of small sizes. After the example in the last section, you have probably concluded that you are not

Impracticality of computing regressions by hand

interested in regression if you have to do the computations by hand. In fact, as sample size gets larger and the number of independent variables in the regression increases, it quickly becomes impractical to do the computations even on a hand-held calculator.

As managers, however, we will have to deal with complex problems requiring larger samples and additional independent variables. To assist us in solving these more detailed problems, we will use a computer, which allows us to perform a large number of computations in a very small period of time.

Suppose that we have not one or two independent variables, but rather that we have k of them: X_1, X_2, \dots, X_k . As before, we will let n denote the number of data points we have. The regression equation we are trying to estimate is

Multiple Regression Estimating Equation

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

[13-5]

Now we'll see how we can use a computer to estimate the regression coefficients.

To demonstrate how a computer handles multiple-regression analysis, take our IRS problem from the preceding section. Suppose the auditing division adds to its model the information concerning rewards to informants. The IRS wishes to include this third independent variable, X_3 , because it feels certain that there is some relationship between these payments and the unpaid taxes discovered. Information for the last 10 months is recorded in Table 13-3.

To solve this problem, the auditing division has used the multiple-regression procedure in Minitab. Of course, we do not yet know how to interpret the solution provided by Minitab but as we shall see, most of the numbers given in the solution correspond fairly closely to what we have already discussed in the context of simple regression.

Demonstration of multiple regression using the computer

Using Minitab to solve multiple regression problems

TABLE 13-3 FACTORS RELATED TO THE DISCOVERY OF UNPAID TAXES

Month	Field-Audit Labor Hours (00's omitted)	Computer Hours (00's omitted)	Reward to Informants (000's omitted)	Actual Unpaid Taxes Discovered (000,000's omitted)
	X_1	X_2	X_3	Y
January	45	16	71	29
February	42	14	70	24
March	44	15	72	27
April	45	13	71	25
May	43	13	75	26
June	46	14	74	28
July	44	16	76	30
August	45	16	69	28
September	44	15	74	28
October	43	15	73	27

Minitab Output

Once all the data have been entered and the independent and dependent variables chosen, Minitab computes the regression coefficients and several statistics associated with the regression equation. Let's look at the output for the IRS problem and see what all the numbers mean. The first part of the output is given in Figure 13-2.

Output from the Minitab program

1. *The regression equation.* From the numbers given in the Coef column, we can read the estimating equation:

$$\begin{aligned}\hat{Y} &= a + b_1 X_1 + b_2 X_2 + b_3 X_3 \\ &= -45.796 + 0.597X_1 + 1.177X_2 + 0.405X_3\end{aligned}\quad [13-5]$$

We can interpret this equation in much the same way that we interpreted the two-variable regression equation on page 683.

Finding and interpreting the regression equation

If we hold the number of field-audit labor hours, X_1 , and the number of computer hours, X_2 , constant and change the rewards to informants, X_3 , then the value of \hat{Y} will increase \$405,000 for each additional \$1,000 paid to informants. Similarly, holding X_1 and X_3 constant, we see that each additional 100 hours of computer time used will increase \hat{Y} by \$1,177,000. Finally, if X_2 and X_3 are held constant, we estimate that an additional 100 hours spent in the field audits will uncover an additional \$597,000 in unpaid taxes. Notice that we have rounded the values provided by the Minitab regression output in Figure 13-2.

Suppose that in November, the IRS intends to leave the field-audit labor hours and computer hours at their October levels (4,300 and 1,500) but to increase the rewards paid to informants to \$75,000. How much in unpaid taxes do they expect to discover in November? Substituting these values into the estimated regression equation, we get

$$\begin{aligned}\hat{Y} &= -45.796 + 0.597X_1 + 1.177X_2 + 0.405X_3 \\ &= -45.796 + 0.597(43) + 1.177(15) + 0.405(75) \\ &= -45.796 + 25.671 + 17.655 + 30.375 \\ &= 27.905 \leftarrow \text{Estimated discoveries of } \$27,905,000\end{aligned}$$

So the audit division expects to discover about \$28 million in unpaid taxes in November.

Regression Analysis

The regression equation is

DISCOVER = -45.8 + 0.597 AUDIT + 1.18 COMPUTER + 0.405 REWARDS

Predictor	Coef	Stdev	t-ratio	p
Constant	-45.796	4.878	-9.39	0.000
AUDIT	0.59697	0.08112	7.36	0.000
COMPUTER	1.17684	0.08407	14.00	0.000
REWARDS	0.40511	0.04223	9.59	0.000
s =	0.2861	R - sq =	98.3%	

FIGURE 13-2 MINITAB OUTPUT FOR IRS REGRESSION

2. A measure of dispersion, the standard error of estimate for multiple regression. Now that we have determined the equation that relates our three variables, we need some measure of the dispersion around this multiple-regression plane. In simple regression, the estimation becomes more accurate as the degree of dispersion around the regression gets smaller. The same is true of the sample points around the multiple-regression plane. To measure this variation, we shall again use the measure called the standard error of estimate:

Measuring dispersion around the multiple regression plane; using the standard error of estimate

Standard Error of Estimate

$$S_e = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - k - 1}} \quad [13-6]$$

where

- Y = sample values of the dependent variable
- \hat{Y} = corresponding estimated values from the regression equation
- n = number of data points in the sample
- k = number of independent variables (=3 in our example)

The denominator of this equation indicates that in multiple regression with k independent variables, the standard error has $n - k - 1$ degrees of freedom. This occurs because the degrees of freedom are reduced from n by the $k + 1$ numerical constants, a , b_1 , b_2 , ..., b_k that have all been estimated from the same sample.

To compute S_e , we look at the individual errors ($Y - \hat{Y}$) in the fitted regression plane, square them, compute their mean (dividing by $n - k - 1$ instead of n), and take the square root of the result. Because of the way it is computed, s_e is sometimes called the root-mean-square error (or root mse for short). From the Minitab output, which uses the symbol s , rather than s_e to denote the standard error of estimate, we see that the root mse in our IRS problem is 0.286, that is, \$286,000.

As was the case in simple regression, we can use the standard error of estimate and the t distribution to form an approximate confidence interval around our estimated value \hat{Y} . In the unpaid tax problem, for 4,300 field-audit labor hours, 1,500 computer hours, and \$75,000 paid to informants, our \hat{Y} is \$27,905,000 estimated unpaid taxes discovered, and our s_e is \$286,000. If we want to construct a 95 percent confidence interval around this estimate of \$27,905,000, we look in Appendix Table 2 under the 5 percent column until we locate the $n - k - 1 = 10 - 3 - 1 = 6$ degrees of freedom row. The appropriate t value for our interval estimate is 2.447. Therefore, we can calculate the limits of our confidence interval like this:

$$\begin{aligned}\hat{Y} + t(s_e) &= 27,905,000 + (2.447)(286,000) \\ &= 27,905,000 + 699,800 \\ &= 28,604,800 \leftarrow \text{Upper limit} \\ \hat{Y} - t(s_e) &= 27,905,000 - (2.447)(286,000) \\ &= 27,905,000 - 699,800 \\ &= 27,205,200 \leftarrow \text{Lower limit}\end{aligned}$$

Confidence intervals for \hat{y}

With a confidence level as high as 95 percent, the auditing division can feel certain that the actual discoveries will lie in this large interval from \$27,205,200 to \$28,604,800. If the IRS wishes to use a lower confidence level, such as 90 percent, it can narrow the range of values in estimating the unpaid taxes discovered. As was true with simple regression, we can use the standard normal distribution, Appendix Table 1, to approximate the t distribution whenever our degrees of freedom (n minus the number of estimated regression coefficients) are greater than 30.

Interpreting the confidence interval

Did adding the third independent variable (rewards to informants) make our regression better? Because s_e measures the dispersion of the data points around the regression plane, smaller values of s_e should indicate better regressions. For the two-variable regression done earlier in this chapter, s_e turns out to be 1.076. Because the addition of the third variable reduced s_e to 0.286, we see that adding the third variable *did* improve the fit of the regression in this example. **It is not true in general, however, that adding variables always reduces s_e .**

Value of additional variables

3. *The coefficient of multiple determination.* In our discussion of simple correlation analysis, we measured the strength of the relation between two variables using the sample coefficient of determination, r^2 . This coefficient of determination is the fraction of the total variation of the dependent variable Y that is explained by the estimating equation.

Similarly, in multiple correlation, we shall measure the strength of the relationship among three variables using the *coefficient of multiple determination*, R^2 , or its square root, R (the coefficient of multiple correlation). **This coefficient of multiple determination is also the proportion of the total variation of Y that is “explained” by the regression plane.**

Using the coefficient of multiple determination

Notice that the output gives the value of R^2 as 98.3 percent. This tells us that 98.3 percent of the total variation in unpaid taxes discovered is explained by the three independent variables. For the two-variable regression done earlier, R^2 is only 0.729, so 72.9 percent of the variation is explained by field-audit labor hours and computer hours. Adding in rewards to informants explains another 25.4 percent of the variation.

We still have not explained the numbers in the columns headed Stdev, t -ratio, and p in Figure 13-2. These numbers will be used to make inferences about the population regression plane, the topic of Section 13.4.

HINTS & ASSUMPTIONS

No one hand-computes regressions anymore; there are too many interesting things to do with your time. We have explained the technique using hand-computed solutions so you won't have to think of your computer as a “black box” that does lots of useful things you can't explain. Hint: The real values of using your computer to do multiple regressions are that you can handle many independent variables, thus working toward a better estimating equation; you can measure whether adding another independent variable really improved your results; and you can quickly see the behavior of R^2 , which tells you the proportion of the total variation in the dependent variable that is explained by the independent variables. The computer does all the tedious arithmetic quickly, accurately, and without complaints, freeing up your time for the more important work of understanding the results and using them to make better decisions.

Self-Check Exercise

- SC13-3** Pam Schneider owns and operates an accounting firm in Ithaca, New York. Pam feels that it would be useful to be able to predict in advance the number of rush income-tax returns during the busy March 1 to April 15 period so that she can better plan her personnel needs during this time. She has hypothesized that several factors may be useful in her prediction. Data for these factors and number of rush returns for past years are as follows:

X_1 Economic Index	X_2 Population within 1 Mile of Office	X_3 Average Income in Ithaca	Y Number of Rush Returns, March 1 to April 15
99	10,188	21,465	2,306
106	8,566	22,228	1,266
100	10,557	27,665	1,422
129	10,219	25,200	1,721
179	9,662	26,300	2,544

- (a) Use the following Minitab output to determine the best-fitting regression equation for these data:

```
The regression equation is
Y = -1275 + 17.1 X1 + 0.541 X2 - 0.174 X3
Predictor Coef Stdev t-ratio p
Constant -1275 2699 -0.47 0.719
X1 17.059 6.908 2.47 0.245
X2 0.5406 0.3144 1.72 0.335
X3 -0.1743 0.1005 -1.73 0.333
s = 396.1 R - sq = 87.2%
```

- (b) What percentage of the total variation in the number of rush returns is explained by this equation?
(c) For this year, the economic index is 169, the population within 1 mile of the office is 10,212, and the average income in Ithaca is \$26,925. How many rush returns should Pam expect to process between March 1 and April 15?

Basic Concepts

- 13-14** Given the following set of data, use whatever computer package is available to find the best-fitting regression equation and answer the following:
(a) What is the regression equation?
(b) What is the standard error of estimate?
(c) What is R^2 for this regression?
(d) What is the predicted value for Y when $X_1 = 5.8$, $X_2 = 4.2$, and $X_3 = 5.1$?

Y	X_1	X_2	X_3
64.7	3.5	5.3	8.5
80.9	7.4	1.6	2.6
24.6	2.5	6.3	4.5
43.9	3.7	9.4	8.8
77.7	5.5	1.4	3.6
20.6	8.3	9.2	2.5
66.9	6.7	2.5	2.7
34.3	1.2	2.2	1.3

- 13-15** Given the following set of data, use whatever computer package is available to find the best-fitting regression equation and answer the following:
- What is the regression equation?
 - What is the standard error of estimate?
 - What is R^2 for this regression?
 - Give an approximate 95 percent confidence interval for the value of Y when the values of X_1 , X_2 , X_3 , and X_4 are 52.4, 41.6, 35.8, and 3, respectively.

X_1	X_2	X_3	X_4	Y
21.4	62.9	21.9	-2	22.8
51.7	40.7	42.9	5	93.7
41.8	81.8	69.8	2	64.9
11.8	41.0	90.9	-4	19.2
71.6	22.6	12.9	8	55.8
91.9	61.5	30.9	1	23.1

Applications

- 13-16** Police stations across the country are interested in predicting the number of arrests they can expect to process each month so as to better schedule office employees. Historically, the average number of arrests (Y) each month is influenced by the number of officers on the police force (X_1), the population of the city in thousands (X_2), and the percentage of unemployed people in the city (X_3). Data for these factors in 15 cities are presented below.
- Using whatever computer package is available, determine the best-fitting regression equation for these data.
 - What percentage of the total variation in the number of arrests (Y) is explained by this equation?
 - The ChapelBoro police department is trying to predict the number of monthly arrests. ChapelBoro has a population of 75,000, a police force of 82, and an unemployment percentage of 10.5 percent. How many arrests do you predict for each month?

Monthly Average Number of Arrests (Y)	Number of Officers on the Force (X_1)	Size of the City (X_2) in Thousands	Percentage Unemployed (X_3)
390.6	68	81.6	4.3
504.3	94	75.1	3.9
628.4	125	97.3	5.6
745.6	175	123.5	8.7
585.2	113	118.4	11.4
450.3	82	65.4	9.6
327.8	46	61.6	12.4
260.5	32	54.3	18.3
477.5	89	97.4	4.6
389.8	67	82.4	6.7
312.4	47	56.4	8.4
367.5	59	71.3	7.6
374.4	61	67.4	9.8
494.6	87	96.3	11.3
487.5	92	86.4	4.7

- 13-17** We are trying to predict the annual demand for widgets (DEMAND) using the following independent variables.

PRICE = price of widgets (in \$)

INCOME = consumer income (in \$)

SUB = price of a substitute commodity (in \$)

(Note: A substitute commodity is one that can be substituted for another commodity. For example, margarine is a substitute commodity for butter.)

Data have been collected from 1982 to 1996:

Year	Demand	Price (\$)	Income (\$)	Sub (\$)
1982	40	9	400	10
1983	45	8	500	14
1984	50	9	600	12
1985	55	8	700	13
1986	60	7	800	11
1987	70	6	900	15
1988	65	6	1,000	16
1989	65	8	1,100	17
1990	75	5	1,200	22
1991	75	5	1,300	19
1992	80	5	1,400	20
1993	100	3	1,500	23
1994	90	4	1,600	18
1995	95	3	1,700	24
1996	85	4	1,800	21

- (a) Using whatever computer package is available, determine the best-fitting regression equation for these data.
- (b) Are the signs (+ or -) of the regression coefficients of the independent variables as one would expect? Explain briefly. (*Note:* This is not a statistical question; you just need to think about what the regression coefficients mean.)
- (c) State and interpret the coefficient of multiple determination for this problem.
- (d) State and interpret the standard error of estimate for this problem.
- (e) Using the equation, what would you predict for DEMAND if the price of widgets was \$6, consumer income was \$1,200, and the price of the substitute commodity was \$17?

13-18 Bill Buxton, a statistics professor in a leading business school, has a keen interest in factors affecting students' performance on exams. The midterm exam for the past semester had a wide distribution of grades, but Bill feels certain that several factors explain the distribution: He allowed his students to study from as many different books as they liked, their IQs vary, they are of different ages, and they study varying amounts of time for exams. To develop a predicting formula for exam grades, Bill asked each student to answer, at the end of the exam, questions regarding study time and number of books used. Bill's teaching records already contained the IQs and ages for the students, so he compiled the data for the class and ran a multiple regression with Minitab. The output from Bill's computer run was as follows:

Predictor	Coef	Stdev	t-ratio	p
Constant	-49.948	41.55	-1.20	0.268
HOURS	1.06931	0.98163	1.09	0.312
IQ	1.36460	0.37627	3.63	0.008
BOOKS	2.03982	1.50799	1.35	0.218
AGE	-1.79890	0.67332	-2.67	0.319
$s = 11.657$		$R - sq = 76.7\%$		

- (a) What is the best-fitting regression equation for these data?
- (b) What percentage of the variation in grades is explained by this equation?
- (c) What grade would you expect for a 21-year-old student with an IQ of 113, who studied 5 hours and used three different books?

13-19 Fourteen Twenty-Two Food Stores, Inc., is planning to expand its convenience store chain. To aid in selecting locations for the new stores, it has collected weekly sales data from each of its 23 stores. To help explain the variability in weekly sales, it has also collected information describing four variables that it believes are related to sales. The data that were collected follow. The variables are defined as follows:

SALES : average weekly sales for each store in thousands of dollars

AUTOS : average weekly auto traffic volume in thousands of cars

ENTRY : ease of entry/exit measured on a scale of 1 to 100

ANNINC : average annual household income for the area in thousands of dollars

DISTANCE : distance in miles from the store to the nearest supermarket

The data were analyzed using Minitab and the output follows:

Predictor	Coef	Stdev	t-ratio	p
Constant	175.37	92.62	1.89	0.075
AUTOS	-0.028	0.315	-0.09	0.929
ENTRY	3.775	1.272	2.97	0.008
ANNINC	1.990	4.510	0.44	0.664
DISTANCE	212.41	28.090	7.56	0.000
s = 85.587	R - sq = 95.8%			

- (a) What is the best-fitting regression equation, as given by Minitab?
- (b) What is the standard error of estimate for this equation?
- (c) What fraction of the variation in sales is explained by this regression?
- (d) What sales would you predict for a store located in a neighborhood that had an average annual household income of \$20,000, was 2 miles from the nearest supermarket, was on a road with weekly traffic volume of 100,000 autos, and had an ease of entry of 50?

- 13-20** Rick Blackburn is thinking about selling his house. In order to decide what price to ask, he has collected data for 12 recent closings. He has recorded sales price (in \$1,000s), the number of square feet in the house (in 100s of sq ft.), the number of stories, the number of bathrooms, and the age of the house (in years).

Sales Price	Square Feet	Stories	Bathrooms	Age
49.65	8.9	1	1.0	2
67.95	9.5	1	1.0	6
81.15	12.6	2	1.5	11
81.60	12.9	2	1.5	8
91.50	19.0	2	1.0	22
95.25	17.6	1	1.0	17
100.35	20.0	2	1.5	12
104.25	20.6	2	1.5	11
112.65	20.5	1	2.0	9
149.70	25.1	2	2.0	8
160.65	22.7	2	2.0	18
232.50	40.8	3	4.0	12

- (a) Using whatever computer package is available, determine the best-fitting regression equation for these data.
- (b) What is R^2 for this equation? What does this number measure?
- (c) If Rick's house has 1,800 square feet (=18.0 hundreds of square feet), 1 story, 1.5 bathrooms, and is 6 years old, what sale price can Rick expect?

- 13-21** Allegheny Steel Corporation has been looking into the factors that influence how many millions of tons of steel it is able to sell each year. Management suspects that the following are major factors: the annual national inflation rate, the average price per ton by which imported

steel undercuts Allegheny's prices (in dollars), and the number of cars (in millions) that U.S. automakers are planning to produce in that year. Data for 7 years have been collected:

Year	Y Millions of Tons Sold	X_1 Inflation Rate	X_2 Imported Undercut	X_3 Number of Cars
1993	4.2	3.1	3.10	6.2
1992	3.1	3.9	5.00	5.1
1991	4.0	7.5	2.20	5.7
1990	4.7	10.7	4.50	7.1
1989	4.3	15.5	4.35	6.5
1988	3.7	13.0	2.60	6.1
1987	3.5	11.0	3.05	5.9

- (a) Using whatever computer package is available, determine the best-fitting regression equation for these data.
- (b) What percentage of the total variation in the number of millions of tons of steel sold by Allegheny each year is explained by this equation?
- (c) How many tons of steel should Allegheny expect to sell in a year in which the inflation rate is 7.1, American automakers are planning to produce 6.0 million cars, and the average imported price undercut per ton is \$3.50?

Worked-Out Answer to Self-Check Exercise

SC 13-3 From the computer output we get the following results:

- (a) $\hat{Y} = -1275 + 17.059X_1 + 0.5406X_2 - 0.1743X_3$,
- (b) $R^2 = 87.2\%$; 87.2% of the total variation in Y is explained by the model.
- (c) $\hat{Y} = -1275 + 17.059(169) + 0.5406(10,212) - 0.1743(26,925) = 2,436$ rush returns.

13.4 MAKING INFERENCES ABOUT POPULATION PARAMETERS

In Chapter 12, we noted that the *sample* regression line, $\hat{Y} = a + bX$ (Equation 12-3), estimates the *population* regression line, $Y = A + BX$ (Equation 12-13). The reason we could only estimate the population regression line rather than find it exactly was that the data points didn't fall exactly on the population regression line. Because of random disturbances, the data points satisfied $Y = A + BX + e$ (Equation 12-13a) rather than $Y = A + BX$.

Exactly the same sort of thing happens in multiple regression. **Population regression plane**
Our estimated regression plane

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k \quad [13-5]$$

is an estimate of a true but unknown population regression plane of the form

Population Regression Equation

$$Y = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

[13-7]

Once again, the individual data points usually won't lie exactly on the population regression plane. Consider our IRS problem to see why this is so. Not all payments to informants will be equally effective. Some of the computer hours may be used for collecting and organizing data; others may be used for analyzing those data to seek errors and fraud. The success of the computer in discovering unpaid taxes may depend on how much time is devoted to each of these activities. For these and other reasons, some of the data points will be above the regression plane and some will be below it. Instead of satisfying

Random disturbances moves point off the regression plane

$$Y = A + B_1 X_1 + B_2 X_2 + \dots + B_k X_k \quad [13-7]$$

the individual data points will satisfy

Population Regression Plane Plus Random Disturbance

$$Y = A + B_1 X_1 + B_2 X_2 + \dots + B_k X_k + e \quad [13-7a]$$

The quantity e in Equation 13-7a is a random disturbance, which equals zero on the average. The standard deviation of the individual disturbances is σ_e , and the standard error of estimate, s_e , which we looked at in the last section, is an estimate of σ_e .

Because our *sample* regression plane, $\hat{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$ (Equation 13-5), estimates the unknown population regression plane, $Y = A + B_1 X_1 + B_2 X_2 + \dots + B_k X_k$ (Equation 13-7), we should be able to use it to make inferences about the population regression plane. In this section, we shall make inferences about the slopes (B_1, B_2, \dots, B_k) of the "true" regression equation (the one for the entire population) that are based on the slopes (b_1, b_2, \dots, b_k) of the regression equation estimated from the sample of data points.

Inferences about an Individual Slope B_i

The regression plane is derived from a sample and not from the entire population. As a result, we cannot expect the true regression equation, $Y = A + B_1 X_1 + B_2 X_2 + \dots + B_k X_k$ (the one for the entire population), to be exactly the same as the equation estimated from the sample observations, $\hat{Y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$. Even so, we can use the value of b_i , one of the slopes we calculate from a sample, to test hypotheses about the value of B_i , one of the slopes of the regression plane for the entire population.

Difference between true regression equation and one estimated from sample observations

The procedure for testing a hypothesis about B_i is similar to procedures discussed in Chapters 8 and 9 on hypothesis testing. To understand this process, return to the problem that related unpaid taxes discovered to field-audit labor hours, computer hours, and rewards to informants. On page 690 we pointed out that $b_1 = 0.597$. The first step is to find some value for B_1 to compare with $b_1 = 0.597$.

Testing a hypothesis about B_i

Suppose that over an extended past period of time, the slope of the relationship between Y and X_1 was 0.400. To test if this were still the case, we could define the hypotheses as

$$H_0: B_1 = 0.400 \leftarrow \text{Null hypothesis}$$

$$H_1: B_1 \neq 0.400 \leftarrow \text{Alternative hypothesis}$$

In effect, then, we are testing to learn whether current data indicate that B_1 has changed from its historical value of 0.400.

To find the test statistic for B_1 , it is necessary first to find the *standard error of the regression coefficient*. Here, the regression coefficient we are working with is b_1 , so the standard error of this coefficient is denoted s_{b_1} .

It is too difficult to compute s_{b_1} by hand, but, fortunately, Minitab computes the standard errors of all the regression coefficients for us. For convenience, Figure 13-2 is repeated. The standard errors of the coefficients are given in the column of the output headed “Stdev.”

From the output, we see that s_{b_1} is 0.0811. (Similarly, if we want to test a hypothesis about B_2 , we see that the appropriate standard error to use is $s_{b_2} = 0.0841$.) Once we have found s_{b_1} on the output, we can use Equation 13-8 to standardize the slope of our fitted regression equation:

Standard error of the regression coefficient

Standardizing the regression coefficient

Standardized Regression Coefficient

$$t = \frac{b_i - B_{i_0}}{S_{b_i}} \quad [13-8]$$

where

- b_i = slope of fitted regression
- B_{i_0} = actual slope hypothesized for the population
- s_{b_i} = standard error of the regression coefficient

Why did we use t to denote the standardized statistic? Recall that in simple regression, we used a and b in Equation 12-7 to calculate s_e , and that s_e estimated σ_e , the standard deviation of the disturbances in the data (Equation 12-13a). Then we used s_e in Equation 12-14 to find s_b , the standard error of the regression slope coefficient. We started out with n data points and used them to estimate the **two** coefficients, a and b . Then we based our tests on the t distribution with $n - 2$ degrees of freedom.

Similarly, in multiple regression, we also start out with n data points but we use them to estimate $k + 1$ coefficients: the intercept, a , and k slopes, b_1, b_2, \dots, b_k . These coefficients are then used in Equation 13-6 to calculate s_e , which again estimates σ_e , the standard deviation of the disturbances in the data (Equation 13-7a). Then s_e is used (in an equation that is beyond the scope of this book) to find s_{b_1} .

Regression Analysis

The regression equation is

Predictor	Coeff	Stdev	t-ratio	p
Constant	-45.796	4.878	-9.39	0.000
AUDIT	0.59697	0.08112	7.36	0.000
COMPUTER	1.17684	0.08407	14.00	0.000
REWARDS	0.40511	0.04223	9.59	0.000
s = 0.2861	R - sq = 98.3%			

FIGURE 13-2 MINITAB OUTPUT FOR IRS REGRESSION

Because of this, we base our hypothesis tests on the t distribution with $n - k - 1$ ($=n - (k + 1)$) degrees of freedom.

In our example, the standardized value of the regression coefficient is

$$t = \frac{b_1 - B_{l_0}}{S_{b_1}} \quad [13-8]$$

$$= \frac{0.597 - 0.400}{0.081}$$

$= 2.432 \leftarrow$ Standardized regression coefficient

Suppose we are interested in testing our hypothesis at the 10 percent level of significance. Because we have 10 observations in our sample data, and three independent variables, we know that we have $n - k - 1$ or $10 - 3 - 1 = 6$ degrees of freedom. We look in Appendix Table 2 under the 10 percent column and come down until we find the 6 degrees of freedom row. There, we see that the appropriate t value is 1.943. Because we are concerned whether b_1 (the slope of the sample regression plane) is significantly different from B_1 (the hypothesized slope of the population regression plane), this is a two-tailed test, and the critical values are ± 1.943 . The standardized regression coefficient is 2.432, which is *outside* the acceptance region for our hypothesis test. Therefore, we reject the null hypothesis that B_1 still equals 0.400. In other words, there is enough difference between b_1 and 0.400 for us to conclude that B_1 has changed from its historical value. Because of this, we feel that each additional 100 hours of field-audit labor no longer increases unpaid taxes discovered by \$400,000, as it did in the past.

Conducting the hypothesis test

In addition to hypothesis testing, we can also construct a *confidence interval* for any one of the values of B_i . In the same way that b_i is a point estimate of B_i , such confidence intervals are interval estimates of B_i . To illustrate the process of constructing a confidence interval, let's find a 95 percent confidence interval for B_3 in our IRS problem. The relevant data are

$$\left. \begin{array}{l} b_3 = 0.405 \\ S_{b_3} = 0.0422 \end{array} \right\} \text{from Figure 13-2}$$

$t = 2.447 \leftarrow$ 5 percent level of significance and 6 degrees of freedom

Confidence interval for B_i

With this information, we can calculate confidence intervals like this:

$$\begin{aligned} b_3 + t(s_{b_3}) &= 0.405 + 2.447(0.0422) \\ &= 0.508 \leftarrow \text{Upper limit} \end{aligned}$$

$$\begin{aligned} b_3 - t(s_{b_3}) &= 0.405 - 2.447(0.0422) \\ &= 0.302 \leftarrow \text{Lower limit} \end{aligned}$$

We see that we can be 95 percent confident that each additional \$1,000 paid to informants increases the unpaid taxes discovered by some amount between \$302,000 and \$508,000.

We will often be interested in questions of the form: Does Y really depend on X_i ? For example, we could ask whether unpaid taxes discovered really depend on computer hours. This question

Is an explanatory variable significant?

is often phrased as, “Is X_i a significant explanatory variable for Y ?” A bit of thought should convince you that Y depends on X_i (that is, Y varies when X_i varies) if $B_i \neq 0$, and it doesn’t depend on X_i if $B_i = 0$.

We see that our question leads to hypotheses of the form:

$$H_0: B_i = 0 \leftarrow \text{Null hypothesis: } X_i \text{ is not a significant explanatory variable}$$

$$H_1: B_i \neq 0 \leftarrow \text{Alternative hypothesis: } X_i \text{ is a significant explanatory variable}$$

We can test these hypotheses using Equation 13-8 just as we did when we tested our hypotheses about whether B_1 still equaled 0.400. However, there is an easier way to do this, using the column on the output in Figure 13-2 headed “t-ratio.” Look at Equation 13-8 again:

$$t = \frac{b_i - B_{i_0}}{S_{b_i}} \quad [13-8]$$

Because our hypothesized value for B_i is 0, the standardized value of the regression coefficient, which we shall denote by t_o , becomes

$$t_o = \frac{b_i}{S_{b_i}}$$

The value of t_o is called the “observed” or “computed” t value. This is the number that appears in the column headed “t-ratio” in Figure 13-2. Let’s denote by t_c the “critical” t value that we look up in Appendix Table 2. Then, because the test of whether X_i is a significant explanatory variable is a two-tailed test, we need only check whether $-t_c \leq t_o \leq t_c$

Using computed t values from the Minitab output

Test of Whether a Variable Is Significant

$$-t_c \leq t_o \leq t_c$$

[13-9]

where

- t_c = appropriate t value (with $n - k - 1$ degrees of freedom) for the significance level of the test
- $t_o = b_i / S_{b_i}$ = observed (or computed) t value obtained from computer output

If t_o falls between $-t_c$ and t_c , we accept H_0 and conclude X_i is not a significant explanatory variable. Otherwise, we reject H_0 and conclude that X_i is a significant explanatory variable.

Let’s test, at the 0.01 significance level, whether computer hours is a significant explanatory variable for unpaid taxes discovered. From Appendix Table 2, with $n - k - 1 = 10 - 3 - 1 = 6$ degrees of freedom and $\alpha = 0.01$, we see that $t_c = 3.707$. From Figure 13-2, we see that $t_o = 14.00$. Because $t_o > t_c$, we conclude that computer hours is a significant explanatory variable. In fact, looking at the computed t values for the other two independent variables (field-audit labor hours $t_o = 7.36$ and rewards to informants, $t_o = 9.59$), we see that each of them is also a significant explanatory variable.

Testing the significance of computer hours in the IRS problem

We can also use the column headed “ p ” to test whether X_i is a significant explanatory variable. In fact, using that information, we don’t even need to use Appendix Table 2. The entries in this column are *prob values* for the two-tailed test of the hypotheses:

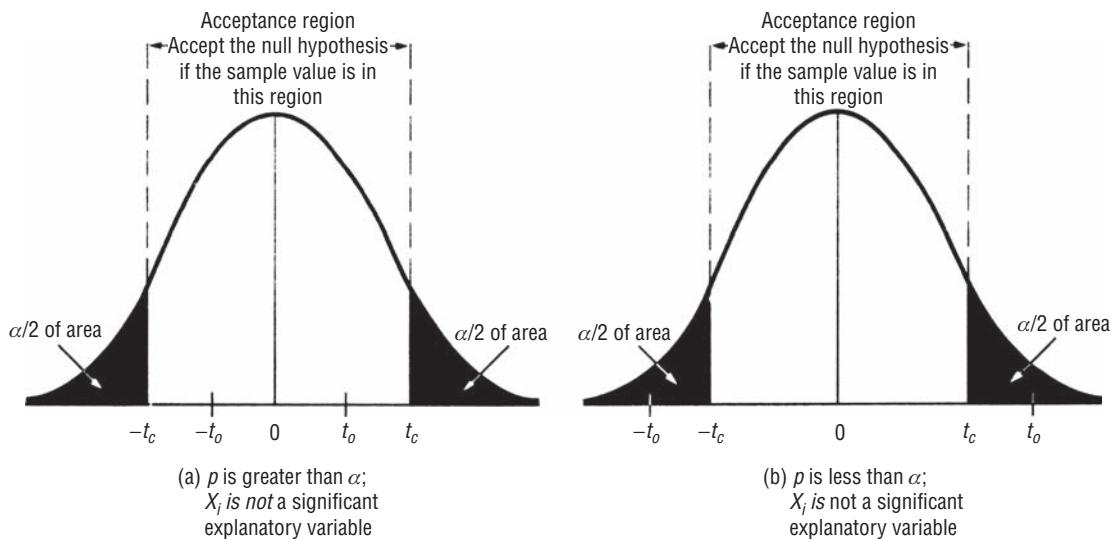


FIGURE 13.3 USING “P” TO SEE WHETHER X_i IS A SIGNIFICANT EXPLANATORY VARIABLE

$$H_0: B_i = 0$$

$$H_r: B_i \neq 0$$

Recall from the discussion in Chapter 9 that these prob values are the probabilities that each b_i would be as far (or farther) away from zero than the observed value obtained from our regression, if H_0 is true. As Figure 13-3 illustrates, we need only compare these prob values with α , the significance level of the test, to determine whether X_i is a significant explanatory variable for Y .

Testing the significance of an explanatory variable is always a two-tailed test. The independent variable X_i is a significant explanatory variable if b_i is significantly *different* from zero, that is, if t_0 is a large positive or a large negative number.

In the IRS example, let's repeat our tests at $\alpha = 0.01$. For each of the three independent variables, p is less than 0.01, so we again conclude that each one is a significant explanatory variable.

Inferences about the Regression as a Whole (Using an F Test)

Suppose you put a piece of graph paper over a dartboard and randomly tossed a bunch of darts at it. After you took out the darts, you would have something that looked very much like a scatter diagram. Suppose you then fit a simple regression line to this set of “observed data points” and calculated r^2 . Because the darts were randomly tossed, you would expect to get a low value of r^2 because in this case, X really doesn't explain Y . However, if you did this many times, occasionally you would observe a high value of r^2 , just by pure chance.

Given any simple (or multiple) regression, **it's natural to ask whether the value of r^2 (or R^2) really indicates that the independent variables explain Y , or might have happened just by chance.** This question is often phrased, “Is the regression as a whole significant?” In the last section, we looked at how to tell whether an individual X_i was a

Significance of the regression as a whole

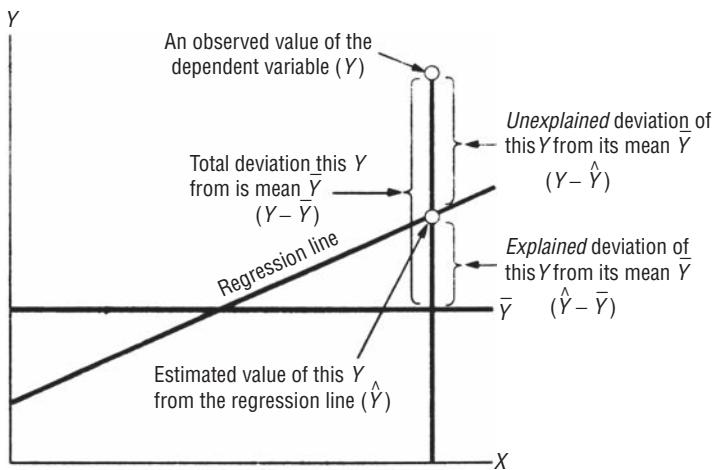


FIGURE 13-4 TOTAL DEVIATION, EXPLAINED DEVIATION, AND UNEXPLAINED DEVIATION FOR ONE OBSERVED VALUE OF Y

significant explanatory variable; now we see how to tell whether all the X_i 's taken together significantly explain the variability observed in Y . Our hypotheses are

$$H_0: B_1 = B_2 = \dots = B_k = 0 \leftarrow \text{Null hypothesis: } Y \text{ doesn't depend on the } X_i\text{'s.}$$

$$H_1: \text{at least one } B_i \neq 0 \leftarrow \text{Alternative hypothesis: } Y \text{ depends on at least one of the } X_i\text{'s.}$$

When we discussed r^2 in Chapter 12, we looked at the total variation in Y , $\sum(Y - \bar{Y})^2$, the part of that variation that is explained by the regression $\sum(\hat{Y} - \bar{Y})^2$, and the unexplained part of the variation, $\sum(Y - \hat{Y})^2$. Figure 13-4 is a duplicate of Figure 12-15. It reviews the relationship between total deviation, explained deviation, and unexplained deviation for a single data point in a simple regression. Although we can not draw a similar picture for a multiple regression, we are doing the same thing conceptually.

In discussing the variation in Y , then, we look at three different terms, each of which is a sum of squares. We denote these by

Analyzing the variation in the Y values

Sums of squares and their degrees of freedom

Three Different Sums of Squares

$$\text{SST} = \text{total sum of squares (i.e., the explained part)} = \sum(Y - \bar{Y})^2$$

$$\text{SSR} = \text{regression sum of squares (i.e., the explained part)} = \sum(\hat{Y} - \bar{Y})^2 \quad [13-10]$$

$$\text{SSE} = \text{error sum of squares (i.e., the unexplained part)} = \sum(Y - \hat{Y})^2$$

These are related by the equation

Decomposing the Total Variation in Y

$$\text{SST} = \text{SSR} + \text{SSE}$$

[13-11]

which says that the total variation in Y can be broken down into two parts, the explained part and the unexplained part.

Each of these sums of squares has an associated number of degrees of freedom. SST has $n - 1$ degrees of freedom (n observations, less 1 degree of freedom because the sample mean is fixed). SSR has k degrees of freedom because there are k independent variables being used to explain Y . Finally, SSE has $n - k - 1$ degrees of freedom because we used our n observations to estimate $k + 1$ constants, a, b_1, b_2, \dots, b_k . If the null hypothesis is true, the ratio below has an F distribution with k numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom.

F Ratio

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n - k - 1)} \quad [13-12]$$

If the null hypothesis is false, then the F ratio tends to be larger than it is when the null hypothesis is true. So if the F ratio is too high (as determined by the significance level of the test and the appropriate value from Appendix Table 6), we reject H_0 and conclude that the regression as a whole is significant.

Figure 13-5 gives Minitab output for the IRS problem. This part of the output includes the computed F ratio for the regression, and is sometimes called the *analysis of variance (ANOVA) for the regression*. You are probably wondering whether this has anything to do with the analysis of variance we discussed in Chapter 11. Yes, it does. Although we did not do so, it is possible to show that the analysis of variance in Chapter 11 also looks at the total variation of all of the observations about the grand mean and breaks it up into two parts: one part explained by the differences among the several groups (corresponding to what we called the between-column variance) and the other part unexplained by those differences (corresponding to what we called the within-column variance). This is precisely analogous to what we just did in Equation 13-11.

Analysis of variance for the regression

For the IRS problem, we see that $\text{SSR} = 29.109$ (with $k = 3$ degrees of freedom), $\text{SSE} = 0.491$ (with $n - k - 1 = 10 - 3 - 1 = 6$ degrees of freedom), and that

$$F = \frac{29.109/3}{0.491/6} = \frac{9.703}{0.082} = 118.33$$

Testing the significance of the IRS regression

The entries in the “MS” column are just the sums of squares divided by their degrees of freedom. For 3 numerator degrees of freedom and 6 denominator degrees of freedom, Appendix Table 6 tells us that

Analysis of Variance					
SOURCE	DF	SS	MS	F	p
Regression	3	29.1088	9.7029	118.52	0.000
Error	6	0.4912	0.0819		
Total	9	29.6000			

FIGURE 13-5 MINITAB OUTPUT: THE ANALYSIS OF VARIANCE

9.78 is the upper limit of the acceptance region for a significance level of $\alpha = 0.01$. Our calculated F value of 118.33 is far above 9.78, so we see that the regression as a whole is highly significant. We can reach the same conclusion by noting that the output tells us that “ p ” is 0.000. Because this prob value is less than our significance level of $\alpha = 0.01$, we conclude that the regression as a whole is significant. In this way, we can use the ANOVA p to do the test without having to use Appendix Table 6 to look up a critical value of F . This is analogous to the way we used the p values in Figure 13-2 for testing the significance of individual explanatory variables.

Multicollinearity in Multiple Regression

In multiple-regression analysis, the regression coefficients often become less reliable as the degree of correlation between the independent variables increases. If there is a high level of correlation between some of the independent variables, we have a problem that statisticians call *multicollinearity*.

Definition and effect of multicollinearity

Multicollinearity might occur if we wished to estimate a firm’s sales revenue and we used both the number of salespeople employed and their total salaries. Because the values associated with these two independent variables are highly correlated, we need to use only one set of them to make our estimate. In fact, adding a second variable that is correlated with the first distorts the values of the regression coefficients. Nevertheless, we can often predict Y well, even when multicollinearity is present.

An example of multicollinearity

Let’s look at an example in which multicollinearity is present to see how it affects the regression. For the past 12 months, the manager of Pizza Shack has been running a series of advertisements in the local newspaper. The ads are scheduled and paid for in the month before they appear. Each of the ads contains a two-for-one coupon, which entitles the bearer to receive two Pizza Shack pizzas

TABLE 13-4 PIZZA SHACK SALES AND ADVERTISING DATA

Month	X_1 Number of Ads Appearing	X_2 Cost of Ads Appearing (00s of dollars)	Y Total Pizza Sales (000s of dollars)
May	12	13.9	43.6
June	11	12.0	38.0
July	9	9.3	30.1
Aug.	7	9.7	35.3
Sept.	12	12.3	46.4
Oct.	8	11.4	34.2
Nov.	6	9.3	30.2
Dec.	13	14.3	40.7
Jan.	8	10.2	38.5
Feb.	6	8.4	22.6
March	8	11.2	37.6
April	10	11.1	35.2

Regression Analysis

The regression equation is

$$\text{SALES} = 16.9 + 2.08 \text{ ADS}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	16.937	4.982	3.40	0.007
ADS	2.0832	0.5271	3.95	0.003

$$s = 4.206 \quad R - \text{sq} = 61.0\%$$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	276.31	276.31	15.62	0.003
Error	10	176.88	17.69		
Total	11	453.19			

FIGURE 13-6 MINITAB REGRESSION OF SALES ON NUMBER OF ADS

while paying for only the more expensive of the two. The manager has collected the data in Table 13-4 and would like to use it to predict pizza sales.

In Figures 13-6 and 13-7, we have given Minitab outputs for the regressions of total sales on number of ads and cost of ads, respectively.

For the regression on number of ads, we see that the observed t value is 3.95. With 10 degrees of freedom and a significance level of $\alpha=0.01$, the critical t value (from Appendix Table 2) is found to be 3.169. Because $t_0 > t_c$ (or, equivalently, because p is less than 0.01), we conclude that the number of ads is a highly significant explanatory variable for total sales. Note also that $r^2 = 61.0$ percent, so that the number of ads explains about 61 percent of the variation in pizza sales.

Regression Analysis

The regression equation is

$$\text{SALES} = 4.17 + 2.87 \text{ COST}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	4.173	7.109	0.59	0.570
COST	2.8725	0.6330	4.54	0.000

$$s = 3.849 \quad R - \text{sq} = 67.3\%$$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	305.04	305.04	20.59	0.000
Error	10	148.15	14.81		
Total	11	453.19			

FIGURE 13-7 MINITAB REGRESSION OF SALES ON THE COST OF ADS

For the regression on the cost of ads, the observed t value is 4.54, so that the cost of ads is even more significant as an explanatory variable for total sales than was the number of ads (for which the observed t value was only 3.95). In this regression, $r^2 = 67.3$ percent, so about 67 percent of the variation in pizza sales is explained by the cost of ads.

Because both explanatory variables are highly significant by themselves, we try to use both of them in a multiple regression. The output is in Figure 13-8.

The multiple regression is highly significant as a whole, because the ANOVA p is 0.006.

The multiple coefficient of determination is $R^2 = 68.4$ percent, so the two variables together explain about 68 percent of the variation in total sales.

However, if we look at the p values for the individual variables in the multiple regression, we see that even at $\alpha = 0.1$, neither variable is a significant explanatory variable.

Using both explanatory variables in a multiple regression

What has happened here? In the simple regression, each variable is highly significant, and in the multiple regression, they are collectively very significant, but individually not significant.

This apparent contradiction is explained once we notice that the number of ads is highly correlated with the cost of ads. In fact, the correlation between these two variables is $r = 0.8949$, so we have a problem with multicollinearity in our data. You might wonder why these two variables are not perfectly correlated. This is because the cost of an ad varies slightly, depending on where it appears in the newspaper. For instance, in the Sunday paper, ads in the TV section cost more than ads in the news section, and the manager of Pizza Shack has placed Sunday ads in each of these sections on different occasions.

Loss of individual significance

Because X_1 and X_2 are closely related to each other, in effect they each explain the same part of the variability in Y . That's why we get $r^2 = 61.0$ percent in the first simple regression, $r^2 = 67.3$ percent in the second simple regression, but an R^2 of only 68.4 percent in the multiple regression. Adding the number of ads as a second explanatory variable to the cost of ads explains only about 1 percent more of the variation in total sales.

Correlation between two explanatory variables

At this point, it is fair to ask, "Which variable is really explaining the variation in total sales in the multiple regression?" The answer is that both are, but **we cannot separate out their individual contributions because they are so highly correlated with each other. As a result of this, their coefficients in the multiple regression have high standard errors, relatively small computed t values, and relatively large prob > | t | values.**

Both variables explain the same thing

How does this multicollinearity affect us? We are still able to make relatively precise predictions when it is present: Note that for the multiple regression (output in Figure 13-8), the standard error of estimate, which determines the width of confidence intervals for predictions, is 3.989, while for the simple regression with the cost of ads as the explanatory variable (output in Figure 13-7), we have $s_e = 3.849$. What we can't do is tell with much precision how sales will change if we increase the number of ads by one. The multiple regression says $b_1 = 0.625$ (that is, each ad increases total pizza sales by about \$625), but the standard error of this coefficient is 1.12 (that is, about \$1,120).

Individual contributions can't be separated out

Variance Inflation Factor (VIF) is a measure to help researcher in identifying the presence multi collinearity between the independent variables. We know that the variance of the OLS estimator for a

Regression Analysis

The regression equation is

$$\text{SALES} = 6.58 + 0.62 \text{ ADS} + 2.14 \text{ COST}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	6.584	8.542	0.77	0.461
ADS	0.625	1.120	0.56	0.591
COST	2.139	1.470	1.45	0.180

$$s = 3.989 \quad R - \text{sq} = 68.4\%$$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	309.99	154.99	9.74	0.006
Error	9	143.20	15.91		
Total	11	453.19			

FIGURE 13-8 MINITAB REGRESSION OF SALES ON THE NUMBER AND COST OF ADS

regression coefficient (say β_i) is given by

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{S_{ii}(1-R_i^2)}$$

Where $S_{ii} = \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$ and R_i^2 is the unadjusted R^2 when we regress X_i against all the other explanatory variables in the model, that is, constant, $X_2, X_3, \dots, X_{i-1}, X_{i+1}, \dots, X_k$. Suppose there is no linear relation between X_i and the other explanatory variables in the model. Then, R_i^2 will be zero and the variance of $\hat{\beta}_i$ will be $\frac{\sigma^2}{S_{ii}}$. Dividing this into the above expression for $\text{Var}(\hat{\beta}_i)$, we obtain the variance inflation factor and tolerance as

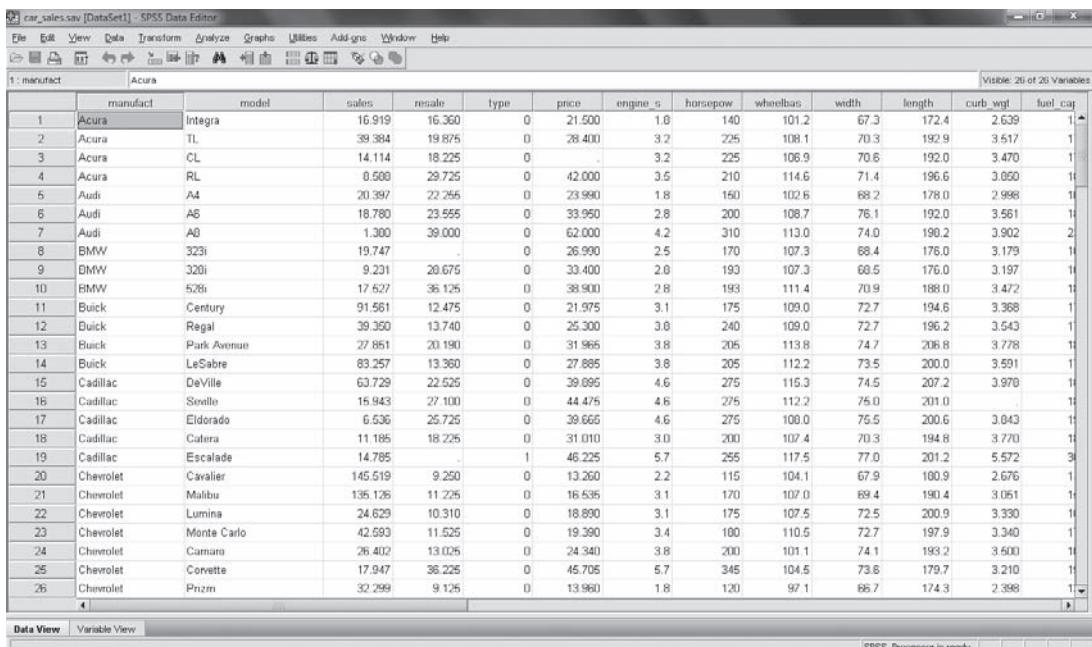
$$\text{VIF}(\hat{\beta}_i) = \frac{1}{1-R_i^2} \quad \text{Tolerance}(\hat{\beta}_i) = \frac{1}{\text{VIF}} = 1-R_i^2$$

It is seen that the higher VIF or the lower the tolerance index, the higher the variance of $\hat{\beta}_i$ and the greater the chance of finding β_i insignificant, which means that severe multi collinearity is present. Thus, these measures can be useful in identifying multi collinearity. We would thus get $k-1$ values for VIF. If any of them is high, then multi collinearity is present. Unfortunately, however, there is no theoretical way to say what the threshold value should be to judge that VIF is "high." Some of the authors, as a thumb rule, use the high value of VIF(>10), as the indicator that the given variable is highly collinear.

HINTS & ASSUMPTIONS

Hint: Making inferences about a multiple regression is conceptually just like what we did in Chapter 12 when we made inferences about a regression line, except here we're dealing with two or more independent variables. Warning: Multicollinearity is a problem you have to deal with in multiple regressions, and developing a common-sense understanding of it is necessary. Remember that you can *still* make fairly precise predictions when it's present. But remember that when it's present, you *can't* tell with much precision how much the dependent variable will change if you "jiggle" one of the independent variables. So our aim should be to minimize multicollinearity. Hint: The best multiple regression is one that explains the relationship among the data by accounting for the largest proportion of the variation in the dependent variable, *with the fewest number of independent variables*. Warning: Throwing in too many independent variables just because you have a computer is not a great idea.

Multiple Linear Regression Using SPSS

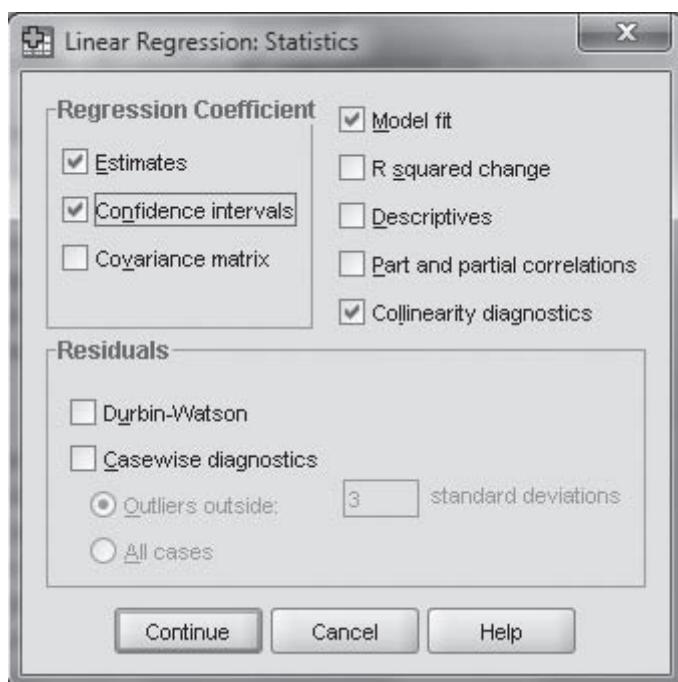
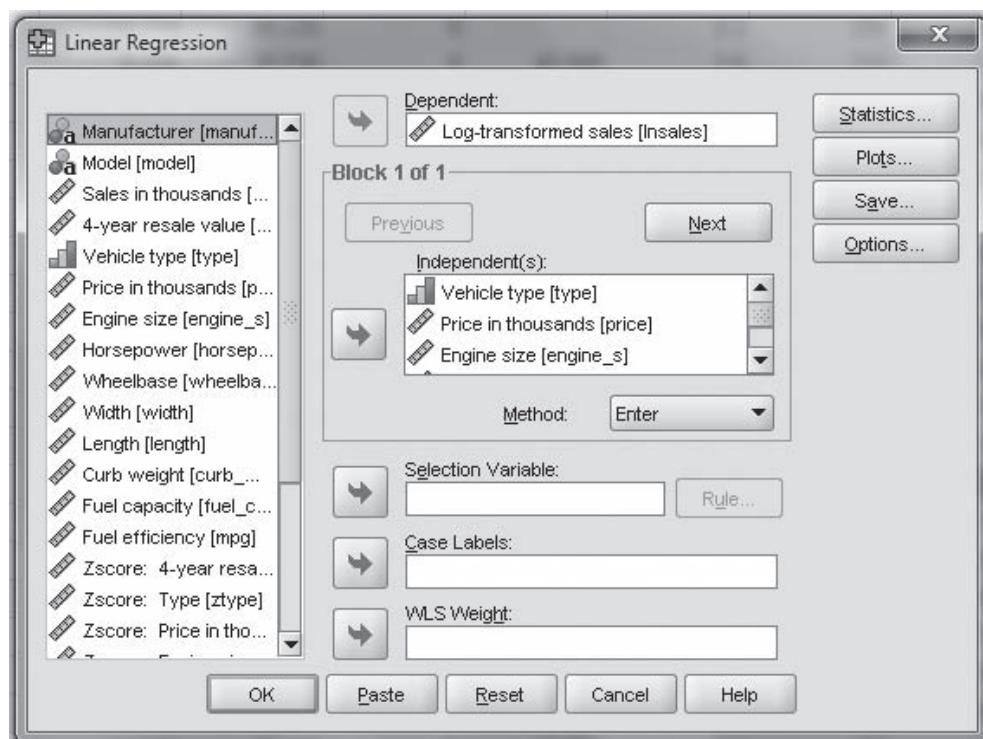


	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg
1	Acura	Integra	16.919	16.360	0	21.500	1.0	140	101.2	67.3	172.4	2.639	1	18.0
2	Acura	TL	39.384	19.875	0	28.400	3.2	225	108.1	70.3	192.9	3.517	1	20.0
3	Acura	CL	14.114	18.225	0	-	3.2	225	106.9	70.6	192.0	3.470	1	18.0
4	Acura	RL	0.588	29.725	0	42.000	3.5	210	114.6	71.4	196.6	3.050	1	20.0
5	Audi	A4	20.397	22.256	0	23.990	1.8	160	102.6	68.2	178.0	2.998	1	18.0
6	Audi	A6	18.780	23.555	0	33.950	2.8	200	108.7	76.1	192.0	3.561	1	20.0
7	Audi	A8	1.300	39.000	0	62.000	4.2	310	113.0	74.0	198.2	3.902	2	20.0
8	BMW	323i	19.747	-	0	26.990	2.5	170	107.3	68.4	176.0	3.179	1	18.0
9	BMW	320i	9.231	20.675	0	33.400	2.8	193	107.3	69.5	176.0	3.197	1	18.0
10	BMW	528i	17.527	36.125	0	38.900	2.8	193	111.4	70.9	188.0	3.472	1	20.0
11	Buick	Century	91.561	12.475	0	21.975	3.1	175	109.0	72.7	194.6	3.388	1	18.0
12	Buick	Regal	39.360	13.740	0	25.300	3.0	240	109.0	72.7	196.2	3.543	1	18.0
13	Buick	Park Avenue	27.861	20.190	0	31.965	3.8	205	113.8	74.7	206.8	3.778	1	20.0
14	Buick	LeSabre	83.257	13.360	0	27.885	3.8	205	112.2	73.5	200.0	3.591	1	18.0
15	Cadillac	DeVille	63.729	22.525	0	39.095	4.6	275	115.3	74.5	207.2	3.978	1	20.0
16	Cadillac	Sentlo	15.943	27.100	0	44.475	4.6	275	112.2	75.0	201.0	-	1	18.0
17	Cadillac	Eldorado	6.536	25.725	0	39.665	4.6	275	100.0	75.5	200.6	3.843	1	18.0
18	Cadillac	Catera	11.185	18.225	0	31.010	3.0	200	107.4	70.3	194.8	3.770	1	18.0
19	Cadillac	Escalade	14.785	-	1	46.225	5.7	295	117.5	77.0	201.2	5.572	3	18.0
20	Chevrolet	Cavalier	145.519	9.250	0	13.260	2.2	115	104.1	67.9	180.9	2.676	1	18.0
21	Chevrolet	Malibu	135.126	11.225	0	16.635	3.1	170	107.0	69.4	190.4	3.061	1	18.0
22	Chevrolet	Lumina	24.629	10.310	0	18.890	3.1	175	107.5	72.5	200.9	3.330	1	18.0
23	Chevrolet	Monte Carlo	42.593	11.525	0	19.390	3.4	160	110.5	72.7	197.9	3.340	1	18.0
24	Chevrolet	Camaro	26.402	13.025	0	24.340	3.8	200	101.1	74.1	193.2	3.600	1	18.0
25	Chevrolet	Corvette	17.947	36.225	0	45.705	5.7	345	104.5	73.6	179.7	3.210	1	18.0
26	Chevrolet	Prizm	32.299	9.125	0	13.960	1.8	120	97.1	66.7	174.3	2.398	1	18.0

Above data is used for regression analysis.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over and underperforming models, you want to establish a relationship between vehicle sales and vehicle characteristics.

For linear regression go to **Analyze > Regression > Linear > Select dependent and independent variables > Go to Statistics > Select Estimates, Confidence Intervals and Collinearity diagnostics > OK**.



The screenshot shows the SPSS Viewer window with the following details:

- Model Summary:**

Mode	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.697*	.486	.449	.98960

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower
- ANOVA:**

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130.300	10	13.030	13.305
	Residual	139.082	141	.979	
	Total	269.383	151		

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower
b. Dependent Variable: Log-transformed sales
- Coefficients:**

Model	Unstandardized Coefficients			Standardized Coefficients		t	Sig.	95% Confidence Interval for B		Tolerance	VIF
	B	Std. Error	Beta	t	Sig.			Lower Bound	Upper Bound		
1	(Constant)	-3.017	2.741	-1.101	.273	-8.435	2.402				
	Vehicle type	.883	.331	.293	2.670	.008	.229	1.537	.304	3.293	
	Price in thousands	-.046	.013	-.502	-3.596	.000	-.072	-.021	.187	5.337	
	Engine size	.356	.180	.281	1.871	.063	-.020	.733	.162	6.159	
	Horsepower	-.002	.004	-.092	-.509	.611	-.011	.008	.112	8.898	
	Wheelbase	.042	.023	.241	1.785	.076	-.004	.068	.200	4.997	
	Width	-.028	.042	-.073	-.676	.500	-.110	.054	.313	3.193	
	Length	.015	.014	.148	1.032	.304	-.013	.043	.178	5.605	
	Curb weight	.156	.350	.075	.447	.655	-.535	.848	.131	7.644	
	Fuel capacity	-.057	.047	-.167	-.203	.231	-.150	.036	.189	5.303	
	Fuel efficiency	.081	.040	.262	2.023	.045	.002	.161	.217	4.604	

a. Dependent Variable: Log-transformed sales

EXERCISES 13.4

Self-Check Exercises

SC 13-4 Edith Pratt is a busy executive in a nationwide trucking company. Edith is late for a meeting because she has been unable to locate the multiple-regression output that an associate produced for her. If the total regression was significant at the 0.05 level, then she wanted to use the computer output as evidence to support some of her ideas at the meeting. The subordinate, however, is sick today and Edith has been unable to locate his work. As a matter of fact, all the information she possesses concerning the multiple regression is a piece of scrap paper with the following on it:

Regression for E. Pratt	
SSR	872.4, with df
SSE	, with 17 df
SST	1023.6, with 24 df

Because the scrap paper doesn't even have a complete set of numbers on it, Edith has concluded that it must be useless. You, however, should know better. Should Edith go directly to the meeting or continue looking for the computer output?

SC13-5 A New England-based commuter airline has taken a survey of its 15 terminals and has obtained the following data for the month of February, where

SALES = total revenue based on number of tickets sold (in thousands of dollars)

PROMOT = amount spent on promoting the airline in the area (in thousands of dollars)

COMP = number of competing airlines at that terminal

FREE = the percentage of passengers who flew free (for various reasons)

Sales (\$)	Promot (\$)	Comp	Free
79.3	2.5	10	3
200.1	5.5	8	6
163.2	6.0	12	9
200.1	7.9	7	16
146.0	5.2	8	15
177.7	7.6	12	9
30.9	2.0	12	8
291.9	9.0	5	10
160.0	4.0	8	4
339.4	9.6	5	16
159.6	5.5	11	7
86.3	3.0	12	6
237.5	6.0	6	10
107.2	5.0	10	4
155.0	3.5	10	4

- (a) Use the following Minitab output to determine the best-fitting regression equation for the airline:

The regression equation is
 $SALES = 172 + 25.9 \text{ PROMOT} - 13.2 \text{ COMP} - 3.04 \text{ FREE}$

Predictor	Coef	Stdev	t-ratio	p
Constant	172.34	51.38	3.35	0.006
PROMOT	25.950	4.877	5.32	0.000
COST	-13.238	3.686	-3.59	0.004
FREE	-3.041	2.342	-1.30	0.221

- (b) Do the passengers who fly free cause sales to decrease significantly? State and test appropriate hypotheses. Use $\alpha = 0.05$.
 (c) Does an increase in promotions by \$1,000 change sales by \$28,000, or is the change significantly different from \$28,000? State and test appropriate hypotheses. Use $\alpha = 0.10$.
 (d) Give a 90 percent confidence interval for the slope coefficient of COMP.

Applications

- 13-22 Mark Lowtown publishes the *Mosquito Junction Enquirer* and is having difficulty predicting the amount of newsprint needed each day. He has randomly selected 27 days over the past year and recorded the following information:

POUNDS = pounds of newsprint for that day's newspaper

CLASIFIED = number of classified advertisements

DISPLAY = number of display advertisements

FULLPAGE = number of full-page advertisements

Using Minitab to regress POUNDS on the other three variables, Mark got the output that follows.

Predictor	Coef	Stdev	t-ratio	p
Constant	1072.95	872.43	1.23	0.232
CLASIFIED	0.251	0.126	1.99	0.060
DISPLAY	1.250	0.884	1.41	0.172
FULLPAGE	250.66	67.92	3.69	0.001

- (a) Mark had always felt that each display advertisement used at least 3 pounds of newsprint. Does the regression give him significant reason to doubt this belief at the 5 percent level?
 - (b) Similarly, Mark had always felt that each classified advertisement used roughly half a pound of newsprint. Does he now have significant reason to doubt this belief at the 5 percent level?
 - (c) Mark sells full-page advertising space to the local merchants for \$30 per page. Should he consider adjusting his rates if newsprint costs him 9¢ per pound? Assume other costs are negligible. State explicit hypotheses and an explicit conclusion. (*Hint:* Holding all else constant, each additional full-page ad uses 250.66 pounds of paper $\times \$0.09$ per pound = \$22.56 cost. Breakeven is at 333.333 pounds. Why? Thus, if the slope coefficient for FULLPAGE is significantly above 333.333, Mark is not making a profit and his rates should be changed.)
- 13-23** Refer to Exercise 13-18. At a significance level of 0.10, which variables are significant explanatory variables for exam scores? (There were 12 students in the sample.)
- 13-24** Refer to Exercise 13-18. The following additional output was provided by Minitab when Bill ran the multiple regression:

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	3134.42	783.60		
Error	7	951.25	135.89		
Total	11	4085.67			

- (a) What is the observed value of F ?
 - (b) At a significance level of 0.05, what is the appropriate critical value of F to use in determining whether the regression as a whole is significant?
 - (c) Based on your answers to (a) and (b), is the regression significant as a whole?
- 13-25** Refer to Exercise 13-19. At a significance level of 0.01, is DISTANCE a significant explanatory variable for SALES?
- 13-26** Refer to Exercise 13-19. The following additional output was provided by Minitab when the multiple regression was run:

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	2861495	715374	102.39	0.000
Error	18	125761	6896.7		
Total	22	2987256			

At the 0.05 level of significance, is the regression significant as a whole?

- 13-27 Henry Lander is director of production for the Alecos Corporation of Caracas, Venezuela. Henry has asked you to help him determine a formula for predicting absenteeism in a meat-packing facility. He hypothesizes that percentage absenteeism can be explained by average daily temperature. Data are gathered for several months, you run the simple regression, and you find that temperature explains 66 percent of the variation in absenteeism. But Henry is not convinced that this is a satisfactory predictor. He suggests that daily rainfall may also have something to do with absenteeism. So you gather data, run a regression of absenteeism on rainfall, and get an r^2 of 0.59. "Eureka!" you cry. "I've got it! With one predictor that explains 66 percent and another that explains 59 percent, all I have to do is run a multiple regression using both predictors, and I'll surely have an almost perfect predictor!" To your dismay, however, the multiple regression has an R^2 of only 68 percent, which is just slightly better than the temperature variable alone. How can you account for this apparent discrepancy?
- 13-28 Juan Armenlegg, manager of Rocky's Diamond and Jewelry Store, is interested in developing a model to estimate consumer demand for his rather expensive merchandise. Because most customers buy diamonds and jewelry on credit, Juan is sure that two factors that must influence consumer demand are the current annual inflation rate and the current prime lending rate at the leading banks in the country. Explain some of the problems that Juan might encounter if he were to set up a regression model based on his two predictor variables.
- 13-29 A new game show, *Check That Model*, asks contestants to specify the minimum number of parameters they need to determine whether a multiple regression model is significant as a whole at $\alpha = 0.01$. You have won the bidding with 4 parameters. Using the information below, determine whether the regression is significant.

$$R^2 = 0.7452$$

$$SSE = 125.4$$

$$n = 18$$

$$\text{Number of independent variables} = 3$$

- 13-30 The Scottish Tourist Agency is interested in the number of tourists who enter the country weekly during the high season (Y). Data have been collected and are presented below:

Tourists (Y) = Number of tourists who entered Scotland in a week (in thousands)

Rate (X_1) = Number of Scottish pounds purchased for \$1 U.S.

Price (X_2) = Number of Scottish pounds charged for round-trip bus fare from London to Edinburgh

Promot (X_3) = Amount spent on promoting the country (in thousands of Scottish pounds)

Temp (X_4) = Mean temperature during the week in Edinburgh (in degrees Celsius)

Tourists (Y)	Rate (X_1)	Price (X_2)	Promot (X_3)	Temp (X_4)
6.9	0.61	40	8.7	15.4
7.1	0.59	40	8.8	15.6
6.8	0.63	40	8.5	15.4
7.9	0.61	35	8.6	15.3
7.6	0.6	35	9.4	15.8
8.2	0.65	35	9.9	16.2
8.0	0.58	35	9.8	16.4
8.4	0.59	35	10.2	16.6
9.7	0.61	30	11.4	17.4
9.8	0.62	30	11.6	17.2
7.2	0.57	40	8.4	17.6
6.7	0.55	40	8.6	16.4

- (a) Using whatever computer package is available, determine the best-fitting regression equation for the tourist agency.
- (b) Is the currency exchange rate a significant explanatory variable? State and test the appropriate hypotheses at a 0.10 significance level.
- (c) Does an increase in promotions by one thousand pounds increase the number of tourists by more than 200? State and test appropriate hypotheses at a 0.05 significance level.
- (d) Give a 95 percent confidence interval for the slope coefficient of Temp.

Worked-Out Answers to Self-Check Exercises

SC 13-4 Because $SST = SSR + SSE$, $SSE = SST - SSR = 1,023.6 - 872.4 = 151.2$.

Because $dfSST = dfSSR + dfSSE$, $dfSSR = dfSST - dfSSE = 24 - 17 = 7$.

$$\text{Thus, } F = \frac{SSR / k}{SSE / (n - k - 1)} = \frac{872.4 / 7}{151.2 / 17} = 14.01.$$

$$F_{CRIT} = F(7, 17, .05) = 2.61.$$

Because $F_{OBS} > F_{CRIT}$, we conclude that the overall regression is significant as a whole; Edith should continue looking for the output so she can use it at the meeting.

SC 13-5 From the computer output, we get the following results:

$$(a) \widehat{SALES} = 172.34 + 25.950PROMOT - 13.238COMP - 3.041FREE$$

$$(b) H_0: B_{FREE} = 0 \quad H_1: B_{FREE} < 0 \quad \alpha = 0.05$$

This is a one-tailed test, and the prob-value on the output is for the two-tailed alternative, $H_1: B_{FREE} \neq 0$. So for our test, the prob-value is $0.221/2 = 0.111 > \alpha = 0.05$, so we cannot reject H_0 ; sales do not decrease significantly as the number of passengers who fly free increases.

$$(c) H_0: B_{PROMOT} = 28 \quad H_1: B_{PROMOT} \neq 28 \quad \alpha = 0.10$$

The observed t value from the regression results is

$$\frac{(b_{PROMOT} - 28)}{s_{b_{PROMOT}}} = \frac{25.950 - 28}{4.877} = -0.420$$

With 11 degrees of freedom and $\alpha = 0.10$ in both tails combined, the critical t values for the test are ± 1.796 , so the observed value is within the acceptance region. We cannot reject H_0 ; the change in SALES for a one-unit (\$1,000) increase in PROMOT is not significantly different from 28, (\$28,000).

- (d) With 11 degrees of freedom, the t value for a 90 percent confidence interval is 1.796, so that interval is

$$\begin{aligned} b_{COMP} \pm 1.796 s_{bCOMP} &= -13.238 \pm 1.796(3.686) \\ &= -13.238 \pm 6.620 = (-19.858, -6.618) \end{aligned}$$

The airline can be 90 percent confident that ticket revenue at an office decreases between approximately \$6,600 and \$19,900 with each additional competing airline.

13.5 MODELING TECHNIQUES

Given a variable we want to explain and a group of potential explanatory variables, there may be several different regression equations we can look at, depending on which explanatory variables we include and how we include them. Each such regression equation is called a *model*. *Modeling techniques* are the various ways in which we can include the explanatory variables and check the appropriateness of our regression' models. There are many different modeling techniques, but we shall look at only two of the most commonly used devices.

Looking at different models

Qualitative Data and Dummy Variables

In all the regression examples we have looked at so far, the data have been numerical, or *quantitative*. But, occasionally, we will be faced with a variable that is categorical, or *qualitative*. In our chapter-opening problem, the director of personnel wanted to see whether the base salary of a salesperson depended on the person's gender. Table 13-5 repeats the data of that problem.

For the moment, ignore the length of employment and use the technique developed in Chapter 9 for testing the difference between means of two populations, to see whether men earn

Reviewing a previous way to approach the problem

TABLE 13-5 DATA FOR GENDER-DISCRIMINATION PROBLEM

Salesmen		Saleswomen	
Months Employed	Base Salary (\$ 1,000s)	Months Employed	Base Salary (\$1,000s)
6	7.5	5	6.2
10	8.6	13	8.7
12	9.1	15	9.4
18	10.3	21	9.8
30	13.0		

more than women. Test this at $\alpha = 0.01$. If we let the men be population 1 and the women be population 2, we are testing

$H_0: \mu_1 = \mu_2 \leftarrow$ Null hypothesis: There is no gender discrimination in base salaries

$H_1: \mu_1 > \mu_2 \leftarrow$ Alternative hypothesis: Women are discriminated against in base salary

$\alpha = 0.01 \leftarrow$ Level of significance

We sketch the analysis below. If you have any trouble following it, you should review briefly pages 728–734.

$$n_1 = 5 \quad n_2 = 4$$

$$\bar{x}_1 = 9.7 \quad \bar{x}_2 = 8.525$$

$$s_1^2 = 4.415 \quad s_2^2 = 2.609$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad [9-3]$$

$$= \frac{4(4.415) + 3(2.609)}{5 + 4 - 2}$$

$$= 3.641$$

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad [9-4]$$

$$= 1.28$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$$= \frac{(9.7 - 8.525) - 0}{1.28}$$

$$= 0.92$$

With 7 degrees of freedom, the critical t value for an upper-tailed test with $\alpha = 0.01$ is 2.998. Because the observed t value of 0.92 is less than 2.998, we cannot reject H_0 .

Our analysis therefore concludes that there does not appear to be any sex discrimination in base salaries. But recall that we have ignored the length-of-employment data thus far in the analysis.

Before we go any farther, look at a scatter diagram of the data. In Figure 13-9, the black points correspond to men and the colored circles correspond to women. The scatter diagram clearly shows that base salary increases with length of service; but if you try to “eyeball” the regression line, you’ll note that the black points tend to be above it and the colored circles tend to be below it.

Figure 13-10 gives the output from a regression of base salary on months employed. From that output, we see that months employed is a very highly significant explanatory variable for base salary.

The old approach doesn't detect any discrimination

“Eyeballing” the data

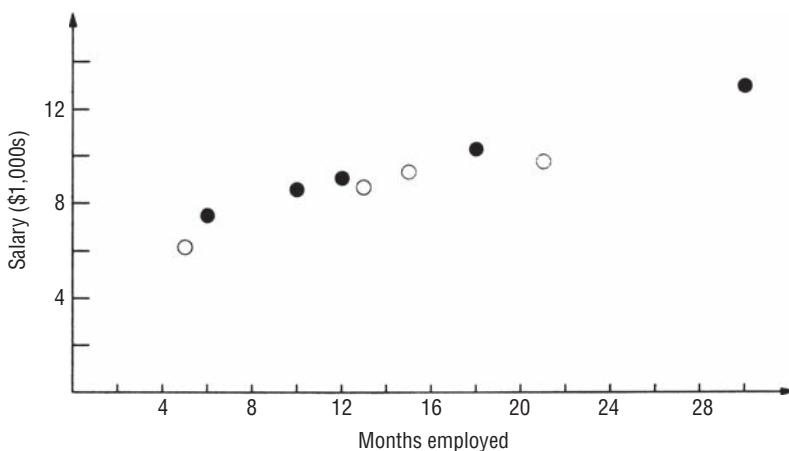


FIGURE 13-9 SCATTER DIAGRAM OF BASE SALARIES PLOTTED AGAINST MONTHS EMPLOYED

Also, $r^2 = 92.6$ percent, indicating that months employed explains about 93 percent of the variation in base salary. Figure 13-11 contains part of the output that we have not seen before, a table of *residuals*. For each data point, the residual is just $Y - \hat{Y}$, which we recognize as the error in the fit of the regression line at that point. In Figure 13-11, FITS1 are the fitted values and RES1 are the residuals.

Perhaps the most important part of analyzing a regression output is looking at the residuals. If the regression includes all the relevant explanatory factors, these residuals ought to be random. Looking at this in another way, if the residuals show any non-random patterns, this indicates that there is something systematic going on that we have failed to take into account.

"Squeezing the residuals"

Regression Analysis

The regression equation is

$$\text{SALARY} = 5.81 + 0.233 \text{ MONTHS}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	5.8093	0.4038	14.39	0.000
MONTHS	0.23320	0.02492	9.36	0.000

$$s = 0.5494 \quad R - \text{sq} = 92.6\%$$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	26.443	26.443	87.61	0.000
Error	7	2.113	0.302		
Total	8	28.556			

FIGURE 13-10 MINITAB REGRESSION OF BASE SALARY ON MONTHS EMPLOYED

ROW	SALARY	FITS1	RES1
1	7.5	7.2085	0.291499
2	8.6	8.1413	0.458684
3	9.1	8.6077	0.492276
4	10.3	10.0069	0.293054
5	13.0	12.8054	0.194607
6	6.2	6.9753	-0.775297
7	8.7	8.8409	-0.140928
8	9.4	9.3073	0.092664
9	9.8	10.7066	-0.906558

FIGURE 13-11 MINITAB TABLE OF RESIDUALS

So we look for patterns in the residuals; or to put it some-what more picturesquely, we “squeeze the residuals until they talk.”

As we look at the residuals in Figure 13-11, we note that the first five residuals are positive. So for the salesmen, we have $Y - \hat{Y} > 0$, or $Y > \hat{Y}$ that is, the regression line falls below these five data points. Three of the last four residuals are negative.

Noticing a pattern in the residuals

And thus for the saleswomen, we have $Y - \hat{Y} < 0$, or $Y < \hat{Y}$, so the regression line lies above three of the four data points. This confirms the observation we made when we looked at the scatter diagram in Figure 13-9. This nonrandom pattern in the residuals suggests that gender *is* a factor in determining base salary.

How can we incorporate the salesperson’s gender *into* the regression model? We do this by using a device called a *dummy variable* (or an *indicator variable*). For the five points that represent salesmen, this variable is given the value 0, and for the four points that represent saleswomen, it is given the value 1. The input data for our regression using dummy variables are given in Table 13-6.

Using dummy variables

To the data in Table 13-6, we fit a regression of the form

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 \quad [13-5]$$

Let’s see what happens if we use this regression to predict the base salary of an individual with X_1 months of service:

$$\text{Salesman: } \hat{Y} = a + b_1 X_1 + b_2(0) = a + b_1 X_1$$

$$\text{Saleswoman: } \hat{Y} = a + b_1 X_1 + b_2(1) = a + b_1 X_1 + b_2$$

For salesmen and saleswomen with the same length of employment, we predict a base salary difference of b_2 thousands of dollars. Now, b_2 is just our estimate of B_2 in the population regression:

$$Y = A + B_1 X_1 + B_2 X_2 \quad [13-7]$$

Interpreting the coefficient of the dummy variable

If there really is discrimination against women, they should earn less than men with the same length of service. In other words, B_2 should be negative. We can test this at the 0.01 level of significance.

Testing for discrimination

TABLE 13-6 INPUT DATA FOR GENDER DISCRIMINATION REGRESSION

	X_1 Months Employed	X_2 Gender	Y Base Salary (\$1,000s)
Men	6	0	7.5
	10	0	8.6
	12	0	9.1
	18	0	10.3
	30	0	13.0
Women	5	1	6.2
	13	1	8.7
	15	1	9.4
	21	1	9.8

$H_0: B_2 = 0 \leftarrow$ Null hypothesis: There is no sex discrimination in base salaries

$H_1: B_2 < 0 \leftarrow$ Alternative hypothesis: Women are discriminated against

$\alpha = 0.01 \leftarrow$ Level of significance

In order to test these hypotheses, we run a regression on the data in Table 13-6. The results of that regression are given in Figure 13-12.

Our hypothesis test is based on the t distribution with $n - k - 1 = 9 - 2 - 1 = 6$ degrees of freedom. For this lower-tailed test, the critical value from Appendix Table 2 is $t_c = -3.143$. From Figure 13-12, we see that the standardized regression coefficient for sex in our test is $t_0 = -3.31$. Figure 13-13 illustrates the critical value, -3.143 , and the standardized coefficient. We see

*Concluding that discrimination
is present*

Regression Analysis

The regression equation is

SALARY = 6.25 + 0.227 MONTHS - 0.789 GENDER

Predictor	Coef	Stdev	t-ratio	p
Constant	6.2485	0.2915	21.44	0.000
MONTHS	0.22707	0.01612	14.09	0.000
GENDER	-0.7890	0.2384	-3.31	0.016

s = 0.3530 R - sq = 97.4%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	27.808	13.904	111.56	0.000
Error	6	0.748	0.125		
Total	8	28.556			

FIGURE 13-12 MINITAB OUTPUT FROM SEX-DISCRIMINATION REGRESSION

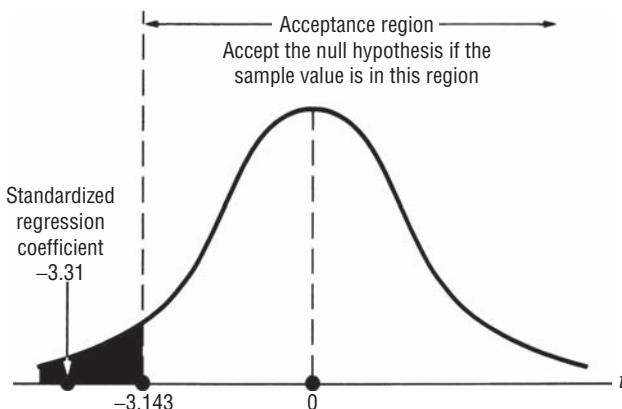


FIGURE 13-13 LEFT-TAILED HYPOTHESIS TEST AT THE 0.01 SIGNIFICANCE LEVEL, SHOWING ACCEPTANCE REGION AND THE STANDARDIZED REGRESSION COEFFICIENT

that the observed b_2 lies outside the acceptance region, so we reject the null hypothesis and conclude that the firm does discriminate against its saleswomen. We also note, in passing, that the computed t value for b_1 in this regression is 14.09, so including gender as an explanatory variable makes months employed even more significant an explanatory variable than it was before. Figure 13-14 gives us Minitab's output of the fitted values and residuals for this regression. Because this was the second regression we ran on these data, Minitab now calls these values FITS2 and RESI2. Note that the residuals for this regression don't seem to show any nonrandom pattern.

Now let's review how we handled the qualitative variable in this problem. We set up a dummy variable, which we gave the value 0 for the men and the value 1 for the women. Then the coefficient of the dummy variable can be interpreted as the difference between a woman's base salary and the base salary for a man. Suppose we had set the dummy variable to 0 for women and 1 for men. Then its coefficient would be the difference between a man's base salary and the base salary for a woman. Can you guess what the regression would have been in this case? It shouldn't surprise you to learn that it would have been

$$\hat{Y} = 5.4595 + 0.22707X_1 + 0.7890X_2$$

The choice of which category is given the value 0 and which the value 1 is totally arbitrary and affects only the sign, not the numerical value of the coefficient of the dummy variable.

Our example had only one qualitative variable (gender), and that variable had only two possible categories (male and female). Although we won't pursue the details here, dummy variable techniques can also be used in problems with several qualitative variables, and those variables can have more than two possible categories.

Interpreting the coefficient of the dummy variable

ROW	SALARY	FITS2	RESI2
1	7.5	7.6109	-0.110921
2	8.6	8.5192	0.080784
3	9.1	8.9734	0.126637
4	10.3	10.3358	-0.035807
5	13.0	13.0607	-0.060692
6	6.2	6.5949	-0.394873
7	8.7	8.4115	0.288537
8	9.4	8.8656	0.534389
9	9.8	10.2281	-0.428053

FIGURE 13-14 MINITAB TABLE OF RESIDUALS

Extensions of dummy variable techniques

Transforming Variables and Fitting Curves

A manufacturer of small electric motors uses an automatic milling machine to produce the slots in the shafts of the motors. A batch of shafts is run and then checked. All shafts in the batch that do not meet required dimensional tolerances are discarded. At the beginning of each new batch, the milling machine is readjusted, because its cutter head wears slightly during the production of the batch. The manufacturer is trying to pick an optimal batch size, but in order to do this, he must know how the size of a batch affects the number of defective shafts in the batch. Table 13-7 gives data for a sample of 30 batches, arranged by ascending size of batch.

Figure 13-15 is a scatter diagram for these data. Because there are two batches of size 250 with 34 defective shafts, two of the points in the scatter diagram coincide (this is indicated by a colored data point in Figure 13-15).

We are going to run a regression of number of defective shafts on the batch size. The output from the regression is in Figures 13-16 and 13-17. What does this output tell us? First of all, we note that batch size

TABLE 13-7 NUMBER OF DEFECTIVE SHAFTS PER BATCH

Batch Size	Number Defective	Batch Size	Number Defective
100	5	250	37
125	10	250	41
125	6	250	34
125	7	275	49
150	6	300	53
150	7	300	54
175	17	325	69
175	15	350	82
200	24	350	81
200	21	350	84
200	22	375	92
225	26	375	96
225	29	375	97
225	25	400	109
250	34	400	112

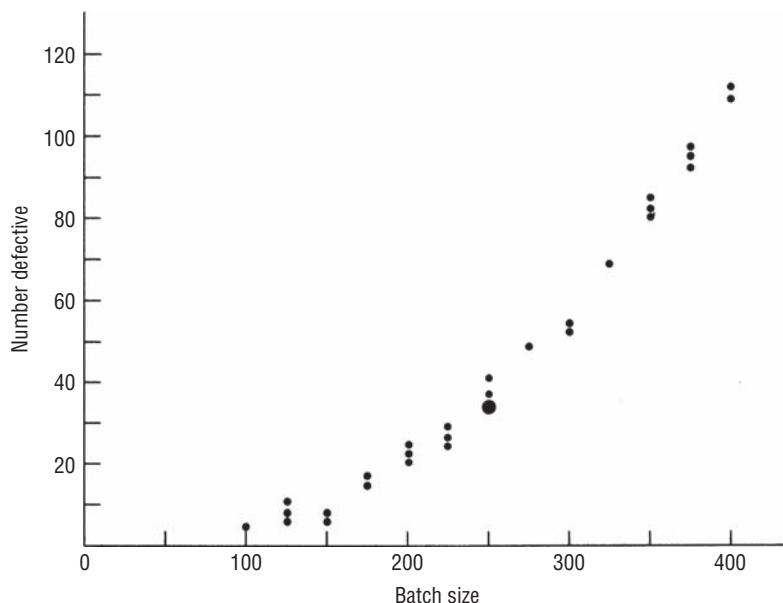


FIGURE 13-15 SCATTER DIAGRAM OF DEFECTIVE SHAFTS PLOTTED AGAINST SIZE OF BATCH

Regression Analysis

The regression equation is

$$\text{DEFECTS} = -47.9 + 0.367 \text{ BATCHSIZ}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	-47.901	4.112	-11.65	0.000
BATCHSIZ	0.36713	0.01534	23.94	0.000
s	7.560	R - sq	= 95.3%	

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	32744	32744	572.90	0.000
Error	28	1600	54		
Total	29	34345			

FIGURE 13-16 MINITAB OUTPUT FROM REGRESSION OF DEFECTS ON BATCH SIZE

ROW	DEFECTS	FITS1	RESI1
1	5	-11.1875	16.1875
2	10	-2.0093	12.0093
3	6	-2.0093	8.0093
4	7	-2.0093	9.0093
5	6	7.1690	-1.1690
6	7	7.1690	-0.1690
7	17	16.3473	0.6527
8	15	16.3473	-1.3473
9	24	25.5256	-1.5256
10	21	25.5256	-4.5256
11	22	25.5256	-3.5256
12	26	34.7039	-8.7039
13	29	34.7039	-5.7039
14	25	34.7039	-9.7039
15	34	43.8822	-9.8822
16	37	43.8822	-6.8822
17	41	43.8822	-2.8822
18	34	43.8822	-9.8822
19	49	53.0605	-4.0605
20	53	62.2387	-9.2387
21	54	62.2387	-8.2387
22	69	71.4170	-2.4170
23	82	80.5953	1.4047
24	81	80.5953	0.4047
25	84	80.5953	3.4047
26	92	89.7736	2.2264
27	96	89.7736	6.2264
28	97	89.7736	7.2264
29	109	98.9519	10.0481
30	112	98.9519	13.0481

FIGURE 13-17 MINITAB OUTPUT OF RESIDUALS

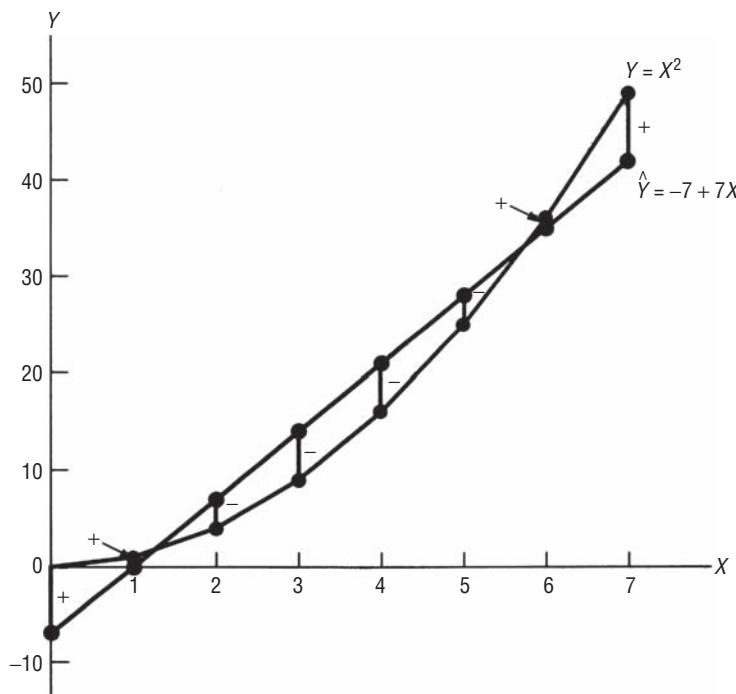


FIGURE 13-18 FITTING A STRAIGHT LINE TO POINTS ON A CURVE

does a fantastic job of explaining the number of defective shafts: The computed t value is 23.94 and $r^2 = 95.3$ percent. However, despite the incredibly high t value, and despite the fact that batch size explains 95 percent of the variation in number of defectives, the residuals in this regression are far from random. Notice how they start out as large positive values, become smaller, then go negative, then become more negative, and then turn around again, finishing up with large positive values.

Noticing a pattern in the residuals

What does this indicate? Look at Figure 13-18, where we have fitted a black regression line ($\hat{Y} = -7 + 7X$) to the eight points $(X, Y) = (0,0), (1,1), (2,4), (3,9), \dots, (7,49)$, all of which lie on the colored curve ($Y = X^2$). The figure also shows the residuals and their signs.

What the pattern suggests

The pattern of residuals that we got in our motor-shaft problem is quite similar to the pattern seen in Figure 13-18. Maybe the shaft data are better approximated by a curve than a straight line. Look back at Figure 13-15. What do you think?

Fitting a curve to the data

But we've fitted only straight lines before. How do we go about fitting a curve? It's simple; all we do is introduce another variable, $X_2 = (\text{batch size})^2$, and then run a multiple regression. The input data are in Table 13-8, and the results are in Figures 13-19 and 13-20.

The curve is much better than the line

Looking at Figure 13-19, we see that batch size and $(\text{batch size})^2$ are *both* significant explanatory variables; their t values are -3.82 and 15.67, respectively. The multiple coefficient of

TABLE 13-8 INPUT FOR FITTING A CURVE TO THE MOTOR-SHAFT DATA

X_1 Batch Size	X_2 (Batch Size) ²	Y Number Defective	X_1 Batch Size	X_2 (Batch Size) ²	Y Number Defective
100	10,000	5	250	62,500	37
125	15,625	10	250	62,500	41
125	15,625	6	250	62,500	34
125	15,625	7	275	75,625	49
150	22,500	6	300	90,000	53
150	22,500	7	300	90,000	54
175	30,625	17	325	105,625	69
175	30,625	15	350	122,500	82
200	40,000	24	350	122,500	81
200	40,000	21	350	122,500	84
200	40,000	22	375	140,625	92
225	50,625	26	375	140,625	96
225	50,625	29	375	140,625	97
225	50,625	25	400	160,000	109
250	62,500	34	400	160,000	112

determination is $R^2 = 99.5$ percent, so together, our two variables explain 99.5 percent of the variation in the number of defective motor shafts. As a final comparison of our two regressions, notice that the standard error of estimate, which measures the dispersion of the sample points around the fitted model,

Regression Analysis

The regression equation is				
DEFECTS = 6.90 - 0.120 BATCHSIZ + 0.000950 SIZESQ				
Predictor	Coef	Stdev	t-ratio	p
Constant	6.898	3.737	1.85	0.076
BATCHSIZ	-0.12010	0.03148	-3.82	0.001
SIZESQ	0.00094954	0.00006059	15.67	0.000
s = 2.423	R - sq =	99.5%		
Analysis of Variance				
SOURCE	DF	SS	MS	F
Regression	2	34186	17093	2911.35
Error	27	159	6	
Total	29	34345		

FIGURE 13-19 MINITAB OUTPUT OF REGRESSION ON BATCH SIZE AND (BATCH SIZE)²

ROW	DEFECTS	FITS2	RESI2
1	.5	4.383	0.61728
2	10	6.721	3.27869
3	6	6.721	-0.72131
4	7	6.721	0.27869
5	6	10.247	-4.24682
6	7	10.247	-3.24682
7	17	14.959	2.04074
8	15	14.959	0.04074
9	24	20.859	3.14138
10	21	20.859	0.14138
11	22	20.859	1.14138
12	26	27.945	-1.94491
13	29	27.945	1.05509
14	25	27.945	-2.94491
15	34	36.218	-2.21811
16	37	36.218	0.78189
17	41	36.218	4.78189
18	34	36.218	-2.21811
19	49	45.678	3.32175
20	53	56.325	-3.32530
21	54	56.325	-2.32530
22	69	68.159	0.84072
23	82	81.180	0.81982
24	81	81.180	-0.18018
25	84	81.180	2.81982
26	92	95.388	-3.38800
27	96	95.388	0.61200
28	97	95.388	1.61200
29	109	110.783	-1.78275
30	112	110.783	1.21725

FIGURE 13-20 MINITAB OUTPUT OF RESIDUALS

is 7.560 for the straight-line model, but only 2.423 for the curved model. **The curved model is far superior to the straight-line model, even though the latter explained 95 percent of the variation!** And remember, it was the pattern we observed in the residuals for the straight-line model that suggested to us that a curved model would be more appropriate. The residuals for the curved model, shown in Figure 13-20, do not exhibit any pattern.

In our curved model, we got our second variable, (batch size)², by doing a *mathematical transformation* of our first variable, batch size. Because we squared a variable, the resulting curved model is known as a *second-degree* (or *quadratic*) regression model. There are many other ways in which we can transform variables to get new variables, and most computer regression packages have these transformations built into them. You do not have to compute the transformed variables by hand, as we did in Table 13-8. Computer packages have the capability to compute all sorts of transformations of one or more variables: sums, differences, products, quotients, roots, powers, logarithms, exponentials, trigonometric functions, and many more.

Transforming variables

HINTS & ASSUMPTIONS

There are many regressions (or models) that can explain the behavior of a dependent variable using a bunch of independent variables. Our job is to include the *right* explanatory variables to find the most effective one. We found that we can even introduce *qualitative* independent variables using dummy variables, and that we can transform variables to fit curves to the data. Warning: Even though the regression output in both of these cases reflects the enormous power of your computer, you still need to rely on your common sense to see whether there are non-random patterns in the residuals. Without that, you cannot tell whether there is something systematic going on in the data that you did not take into account. Hint: The secret of using statistics to make good decisions never changes. It's always an effective combination of data, computers, and common sense.

EXERCISES 13.5**Self-Check Exercises**

SC 13-6 Cindy's, a popular fast-food chain, has recently experienced a marked change in its sales as a result of a very successful advertising campaign. As a result, management is now looking for a new regression model for its sales. The following data have been collected in the 12 weeks since the advertising campaign began.

Time	Sales (in thousands)	Time	Sales (in thousands)
1	4,618	7	19,746
2	3,741	8	34,215
3	5,836	9	50,306
4	4,367	10	65,717
5	5,118	11	86,434
6	8,887	12	105,464

- (a) Use the following Minitab output to determine the best-fitting regression of SALES on TIME:

The regression equation is
 $SALES = -26233 + 9093 \text{ TIME}$

Predictor	Coef	Stdev	t-ratio	p
Constant	-26233	9551	-2.75	0.021
TIME	9093	1298	7.01	0.000
$s = 15518$				R-sq = 83.1%

ROW	SALES	FITS1	RESI1	ROW	SALES	FITS1	RESI1
1	4618	-17140	21758	7	19746	37417	-17671
2	3741	-8047	11788	8	34215	46510	-12295
3	5836	1046	4790	9	50306	55603	-5297
4	4367	10139	-5772	10	65717	64696	1021
5	5118	19231	-14113	11	86434	73789	12645
6	8887	28324	-19437	12	105464	82881	22583

- (b) Are you satisfied with your model as a predictor of SALES? Explain.
 (c) The following output uses TIME and TIMESQR (TIME squared) as explanatory variables. Is this quadratic model better fit to the data? Explain.

The regression equation is

$$\text{SALES} = 13981 - 8142 \text{ TIME} + 1326 \text{ TIMESQR}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	13981	2720	5.14	0.000
TIME	-8141.5	961.9	-8.46	0.000
TIMESQR	1325.72	72.03	18.41	0.000
s = 2631	R-sq = 99.6%			

ROW	SALES	FITS2	RESI2	ROW	SALES	FITS2	RESI2
1	4618	7165	-2547	7	19746	21950	-2204
2	3741	3001	740	8	34215	33695	520
3	5836	1488	4348	9	50306	48090	2216
4	4367	2626	1741	10	65717	65138	579
5	5118	6416	-1298	11	86434	84836	1598
6	8887	12858	-3971	12	105464	107186	-1722

SC 13-7 Below are some data on consumption expenditures, CONSUMP; disposable income, INCOME; and sex of the head of household, SEX, of 12 randomly chosen families. The variable GENDER has been coded:

$$\text{GENDER} = \begin{cases} 1 & \text{if SEX} = \text{'M' (male)} \\ 0 & \text{if SEX} = \text{'F' (female)} \end{cases}$$

Consump	Income (\$)	Sex	Gender
37,070	45,100	M	1
22,700	28,070	M	1
24,260	26,080	F	0
30,420	35,000	M	1
17,360	18,860	F	0
33,520	41,270	M	1
26,960	32,940	M	1
19,360	21,440	F	0
35,680	44,700	M	1
22,360	24,400	F	0
28,640	33,620	F	0
39,720	46,000	M	1

- (a) Use the following Minitab output to determine the best-fitting regression to predict CONSUMP from INCOME and GENDER.

The regression equation is

$$\text{CONSUMP} = 2036 + 0.818 \text{ INCOME} - 1664 \text{ GENDER}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	2036	1310	1.55	0.155
INCOME	0.81831	0.04940	16.56	0.000
GENDER	-1664.2	916.9	-1.82	0.103
$s = 1015$				R-sq = 98.4%

- (b) If disposable income is held constant, is there a significant difference in consumption between households headed by a male versus those where the head of household is female? State explicit hypotheses, test them at the 0.10 level, and state an explicit conclusion.
 (c) Give an approximate 95 percent confidence interval for consumption for a household with disposable income of \$40,000 headed by a male.

Basic Concepts

- 13-31** Describe three situations in everyday life in which dummy variables could be used in regression models.
- 13-32** A restaurant owner with restaurants in two cities believes that revenue can be predicted from traffic flow in front of the restaurant with a quadratic model.
- (a) Describe a quadratic model to predict revenue from traffic flow. State the form of the regression equation.
 - (b) It has been suggested that the city the restaurant is in has an effect on revenue. Extend your model from part (a) by using a dummy variable to incorporate the suggestion. Again, state the form of the regression model.
- 13-33** Suppose you have a set of data points to which you have fitted a linear regression equation. Even though the R^2 for the line is very high, you wonder whether it would be a good idea to fit a second-degree equation to the data. Describe how you would make your decision based on
- (a) A scattergram of the data.
 - (b) A table of residuals from the linear regression.
- 13-34** A statistician collected a set of 20 pairs of data points. He called the independent variable X_1 and the dependent variable Y . He ran a linear regression of Y on X_1 , and he was dissatisfied with the results. Because of some nonrandom patterns he observed in the residuals, he decided to square the values of X_1 ; he called these squared values X_2 . The statistician then ran a multiple regression of Y on both X_1 and X_2 . The resulting equation was

$$\hat{Y} = 200.4 + 2.79X_1 - 3.92X_2$$

The value of s_b_1 was 3.245 and the value of s_b_2 was 1.53. At a 0.05 level of significance, determine whether

- (a) The set of unsquared values of X_1 is a significant explanatory variable for Y .
- (b) The set of squared values of X_1 is a significant explanatory variable for Y .

Applications

- 13-35** Dr. Linda Frazer runs a medical clinic in Philadelphia. She collected data on age, reaction to penicillin, and systolic blood pressure for 30 patients. She established systolic blood pressure as the dependent variable, age as X_1 (independent variable) and reaction to penicillin as X_2 (independent variable). Letting 0 stand for a positive reaction to penicillin and 1 stand for a negative reaction, she ran a multiple regression on her desktop personal computer. The predicting equation was

$$\hat{Y} = 6.7 + 3.5X_1 + 0.489X_2$$

- (a) After the regression had already been run, Dr. Frazer discovered that she had meant to code a positive reaction as 1 and a negative reaction as 0. Does she have to rerun the regression? If so, why? If not, give her the equation she would have gotten if the variable had been coded as she had originally intended.
 - (b) If s_{b_2} has a value of 0.09, does this regression provide evidence at a significance level of 0.05 that the reaction to penicillin is a significant explanatory variable for systolic blood pressure?
- 13-36** Excelsior Notebook computers is reexamining its inventory control policy. They need to accurately predict the number of the EXC-11E computers that will be ordered by suppliers in the weeks to come. The data for the last 15 weeks are presented below

Time	Demand (in 1000's)
1	6.7
2	10.2
3	13.4
4	15.6
5	18.2
6	22.6
7	30.5
8	31.4
9	38.7
10	41.6
11	48.7
12	51.4
13	55.8
14	61.5
15	68.9

- (a) Using any available computer package, fit a linear model with TIME as the independent variable and DEMAND as the dependent variable.
- (b) Fit a quadratic model for the data. Is this model better? Explain.

- 13-37** Below are some data from a local pizza parlor on gross sales (SALES), promotion dollars (PROMO), and type of promotion, including radio, newspaper, or flyers. Assume the pizza parlor used only one type of promotion in any given week. The variables Type1 and Type2 have been coded:

TYPE1 = 1 if radio was used, 0 otherwise

TYPE2 = 1 if flyers were used, 0 otherwise

(when both TYPE1 and TYPE2 are 0, that week's promotion budget was spent on newspaper advertisements).

SALES (in 100s)	PROMO (in 100s)	TYPE1	TYPE2
12.1	3.8	0	1
19.1	6.4	0	1
26.9	7.9	0	0
24.8	8.7	1	0
37.1	12.4	1	0
39.4	15.9	0	1
32.5	11.3	0	0
28.9	9.4	0	0
28.8	8.6	1	0
34.7	12.7	0	1
38.4	14.3	0	0
26.3	6.7	1	0

- (a) Using any available computer package, fit a regression model to predict SALES from PROMO, TYPE1, and TYPE2.
- (b) State the fitted regression function.
- (c) If PROMO is held constant, is there a significant difference between radio and newspaper? State appropriate hypotheses and test at a 0.05 level of significance.
- (d) If PROMO is held constant, is there a significant difference between flyers and newspaper? State appropriate hypotheses and test at a 0.05 level of significance.
- (e) Compute a 90 percent confidence interval for SALES in a week when \$800 is spent using radio advertisements as the only type of promotion.

Worked-Out Answers to Self-Check Exercises

- SC 13-6** From the computer output, we get the following results:

- (a) Predicted SALES = $-26233 + 9093\text{TIME}$.
- (b) Even though R^2 is relatively high (83.1%), this is not a good model because of the pattern in the residuals. They start out large and positive, get smaller, go large and negative, and then grow positive again. Clearly a quadratic model would be better.
- (c) Predicted SALES = $13981 - 8141.5\text{TIME} + 1325.72\text{TIME}^2$.

This model is distinctly better. R^2 has increased to 99.6%, and there is no pattern in the residuals.

SC 13-7 From the computer output, we get the following results:

(a) Predicted CONSUMP = 2036 + 0.818INCOME - 1664GENDER.

(b) $H_0: B_{GENDER} = 0$ $H_1: B_{GENDER} \neq 0$ $\alpha = 0.10$

Since the prob value for our test (0.103) is greater than α (0.10), we cannot reject H_0 ; the gender of the head of the household is not a significant factor in explaining consumption.

(c) Predicted CONSUMP = 2036 + 0.818(40,000) - 1664(1) = \$33,092.

With 9 degrees of freedom, the t value for an approximate 95 percent confidence interval for CONSUMP is 2.262, so that interval is

$$\hat{Y} \pm ts_e = 33,092 \pm 2.262(1,015) = 33,092 \pm 2,296 = (\$30,796, \$35,388).$$

STATISTICS AT WORK

Loveland Computers

Case 13: Multiple Regression and Modeling Lee was pleased to be able to report to Nancy Rainwater that the defects occurring in the keyboard cases were indeed related to the daily recorded low temperature for Loveland. And the warehouse supervisor confirmed the explanation.

“Sure, the components warehouse is heated,” Skip Tremont reported. “But it’s only a couple of gas-fired industrial heaters near the ceiling. When the weather’s just a little bit chilly they work well enough. But on these real cold winter nights, the heaters run all night, but the warehouse gets pretty cold.”

“So we need more heaters?” Nancy queried.

“Not necessarily—the problem is that all the warm air stays up high and it gets pretty cold close to the floor. Then when people start coming in and out during the work shift, the air eventually gets stirred up and the lower level—where everything is stored—comes up to room temperature.”

“So we might be able to cure the problem by installing a couple of ceiling fans,” interjected Tyronza Wilson.

“Just what I was thinking,” said Skip, jumping in his pickup truck and heading for the builders’ supply store. “They’re pretty cheap—I can buy a couple out of my maintenance budget.”

“A great example of quality management!” said Lee. “See, Nancy, the people doing the job already know the answer—you just have to empower them to implement a solution.”

“Well, let me take you to lunch and have you talk to someone who has a more complicated problem.”

Over a plate of tamales, Lee Azko met Sherrel Wright, the advertising manager. Sherrel was a new hire who had been with the company for 6 months. “You’ve met Margot, who’s in charge of marketing. She handles the big picture. My job is to focus on the advertising budget and to target our ads so they result in the highest increase in sales.”

“So how do you decide how much of which media to buy?” asked Lee.

“To tell the truth, before I came, things weren’t very scientific. Your uncle will tell you that when Loveland first started out, the number of ads depended on cash flow. When I came on board, I could see that the ad budget went up or down according to how much money we’d made in the previous quarter. This meant that if we’d had a weak quarter, the company cut back on the ad budget for the next quarter. Margot kept telling them that was the opposite of a good strategy—there are many times when

increasing your advertising will get you out of a sales slump. But I guess they were always in a panic about cash flow. Now it looks as if we're going to get substantial new funding and we have to become more scientific about our advertising plans."

"So how do you decide where to run the ads?" Lee was anxious to learn more about marketing in the real world.

"Well, your uncle says it's an art. He tended to run ads in the magazines he enjoyed reading. But he's the first to admit that he wouldn't be a typical Loveland customer, so he's been pretty receptive to my presentations about cost per thousand, target readership, and so on. The computer monthly magazines are our staple, but there's more of them coming out each month and I have to be choosy about where we spend our dollars. Some of our competitors have been buying four or five-page spreads. We've tried that in a couple of issues but it's hard to know whether they're paying off any more than a single-page advertisement. Sales volume tends to lag behind effective advertising, so it's difficult to measure the success of an individual ad."

"You've already tried monitoring call volume on the 800 numbers, I suppose," Lee commented.

"Well, no. That would be a good idea. Do we keep those statistics?"

"Even if we don't, the phone company can easily give you a daily summary. We'd have to see whether call volume or sales volume was the best indicator." Lee was on a roll.

"Hey, it's not that simple," said Gratia Delaguardia, the company's chief engineer, bringing over a plate of burritos and pulling up a chair. "Mind if I join you?"

"Go right ahead." Sherrel wasn't about to cut off one of the two partners in Loveland Computers.

"No offense to you touchy-feely advertising types, but I think that forces outside of Loveland determine our sales. If the economy is growing, we do well. If there's a recession, we do less well."

"Does that fit with the early years?" asked Lee. "Looks like you had some spectacular growth during tough times in the early 80s."

"And what the competitors do is crucial," Gratia said, ignoring Lee's comment. "You can check that easily. Look at the back numbers of the computer magazines and see how many ad pages they bought 'against' us. And you can also tell their price positions relative to ours for equivalent machines. It's all printed right there in each ad."

Lee made a mental note that this was going to be a lot easier than in many industries, where competitors' prices may be hidden in long-term contracts.

"How do we factor in our newspaper ads?" Sherrel wondered aloud. It costs us a lot to advertise in the *Wall Street Journal*, but it's my hunch that gives us an immediate payoff."

"Let's put our heads together and come up with a plan for how we're going to sort this out," said Lee, signaling the waiter for more picante sauce.

Study Questions: What measure of "advertising success" would you investigate? What factors would you consider in an analysis? How would you handle factors that appear to be irrelevant? In addition to the review of the historical data, are there any "experiments" you would recommend?

CHAPTER REVIEW

Terms Introduced in Chapter 13

Analysis of Variance for Regression The procedure for computing the F ratio used to test the significance of the regression as a whole. It is related to the analysis of variance discussed in Chapter 11.

Coefficient of Multiple Correlation, R The positive square root of R^2

Coefficient of Multiple Determination, R^2 The fraction of the variation of the dependent variable that

is explained by the regression. R^2 measures how well the multiple regression fits the data.

Computed F Ratio A statistic used to test the significance of the regression as a whole.

Computed t A statistic used for testing the significance of an individual explanatory variable.

Dummy Variable A variable taking the value 0 or 1, enabling us to include in a regression model qualitative factors such as sex, marital status, and education level.

Modeling Techniques Methods for deciding which variables to include in a regression model and the different ways in which they can be included.

Multicollinearity A statistical problem sometimes present in multiple-regression analysis in which the reliability of the regression coefficients is reduced, owing to a high level of correlation between the independent variables.

Multiple Regression A statistical process by which several variables are used to predict another variable.

Standard Error of a Regression Coefficient A measure of our uncertainty about the exact value of a regression coefficient.

Transformations Mathematical manipulations for converting one variable into a different form so we can fit curves as well as lines by regression.

Equations Introduced in Chapter 13

13-1

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

p. 680

In multiple regression, this is the formula for the estimating equation that describes the relationship between three variables: Y , X_1 , and X_2 . Picture a two-variable multiple-regression equation as a plane, rather than a line.

13-2

$$\Sigma Y = na + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

p. 681

13-3

$$\Sigma X_1 Y = a \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2$$

p. 681

13-4

$$\Sigma X_2 Y = a \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2$$

p. 683

Solving these three equations determines the values of the numerical constants a , b_1 and b_2 and thus the best-fitting multiple-regression plane in a two-variable multiple regression.

13-5

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

p. 689

This is the formula for the estimating equation describing the relationship between Y and the k independent variables, X_1, X_2, \dots, X_k . Equation 13-1 is the special case of this equation for $k = 2$.

13-6

$$s_e = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - k - 1}}$$

p. 691

To measure the variation around a multiple-regression equation when there are k independent variables, use this equation to find the *standard error of estimate*. The standard error, in this case, has $n - k - 1$ degrees of freedom, owing to the $k + 1$ numerical constants that must be calculated from the data (a, b_1, \dots, b_k) .

13-7

$$Y = A + B_1 X_1 + B_2 X_2 + \dots + B_k X_k$$

p. 698

This is the *population regression equation* for the multiple regression. Its Y intercept is A and it has k slope coefficients, one for each of the independent variables.

13-7a

$$Y = A + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + e$$

p. 699

Because all the individual points in a population do not lie on the population regression equation, the *individual* data points will satisfy this equation, where e is a random disturbance from the population regression equation. On the average, e equals zero because disturbances above the population regression equation are canceled out by disturbances below it.

13-8

$$t = \frac{b_i - B_{i_0}}{s_{b_i}}$$

p. 700

Once we have found s_{b_i} on the computer output, we can use this equation to standardize the observed value of the regression coefficient. Then we test hypotheses about B_i by comparing this standardized value with the critical value(s) of t , with $n - k - 1$ degrees of freedom, from Appendix Table 2.

13-9

$$-t_c \leq t_0 \leq t_c$$

p. 702

To test whether a given independent variable is significant, we use this formula to see whether t_0 , the observed t value (computer output), lies between plus and minus t_c , the critical t value (taken from the t distribution with $n - k - 1$ degrees of freedom). The variable *is* significant when t_0 is *not* in the indicated range. If your computer package gives you prob values, the variable *is* significant when this value is *less than* α , the significance level of the test.

13-10

$$\left. \begin{array}{l} \text{SST} = \text{total sum of squares} \\ \text{SSR} = \text{regression sum of squares} \\ \quad (\text{the explained part of SST}) \\ \text{SSE} = \text{error sum of squares} \\ \quad (\text{the unexplained part of SST}) \end{array} \right\} \begin{array}{l} = \sum (Y - \bar{Y})^2 \\ = \sum (\hat{Y} - \bar{Y})^2 \\ = \sum (Y - \hat{Y})^2 \end{array} \quad \text{p. 704}$$

13-11

$$\text{SST} = \text{SSR} + \text{SSE}$$

p. 704

These two equations enable us to break down the variability of the dependent variable into two parts (one explained by the regression and the other unexplained) so we can test for the significance of the regression as a whole.

13-12

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n - k - 1)}$$

p. 705

This F ratio, which has k numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom, is used to test the significance of the regression as a whole. If F is *bigger* than the critical value, then we conclude that the regression as a whole *is* significant. The same conclusion holds if the ANOVA prob value (from the computer output) is *less than* α , the significance level of the test.

Review and Application Exercises

13-38

Homero Martinez is a judge in Barcelona, Spain. He has recently called you in as a statistical consultant to investigate what purports to be a significant finding. He claims that the number of days a case is in court can be used to estimate the amount of damages that should be awarded. He has gathered data from his court and from the courts of several of his fellow judges. For each of the numbers 1 to 9, he has located a case that took that many days in court, and he has determined the amount (in millions of pesetas) of damages awarded in that case.

The following Minitab results were generated when damages awarded were regressed on days in court.

```

The regression equation is
DAMAGES = - 0.406 + 0.518 DAYS

Predictor      Coef        Stdev      t-ratio      p
Constant     -0.4063      0.2875      -1.41      0.201
DAYS          0.51792     0.0511      10.14      0.000

s = 0.3957    R-sq = 93.6%
Analysis of Variance

SOURCE        DF        SS        MS        F
Regression    1       16.094    16.094    102.77
Error         7       1.096     0.157
Total         8       17.191

ROW        DAMAGES      FITS1      RESI1
1          0.645       0.1117     0.53333
2          0.750       0.6296     0.12042
3          1.000       1.1475     -0.14750
4          1.300       1.6654     -0.36542
5          1.750       2.1833     -0.43333
6          2.205       2.7013     -0.49625
7          3.500       3.2192     0.28083
8          4.000       3.7371     0.26292
9          4.500       4.2550     0.24500

```

Of course, you are quite pleased with these results, because R^2 is very high. But the judge is not convinced that you are right. He says, “This is the worst job I’ve ever seen! I don’t care if this line *does* fit the data I gave you, I can tell by looking at the output that it won’t work for other data. If you can’t do any better, just let me know, and I’ll hire a *smart* statistician!”

(a) Why is the judge dissatisfied with the results?

(b) Suggest a better model that will satisfy the judge.

13-39

Jon Grant, supervisor of the Carven Manufacturing Facility, is examining the relationship among an employee’s score on an aptitude test, prior work experience, and success on the job. An employee’s prior work experience is studied and weighted, yielding a rating between 2 and 12. The measure of on-the-job success is based on a point system involving total output and efficiency, with a maximum possible value of 50. Grant sampled six first-year employees and obtained the following:

	X_1 Aptitude Test Score	X_2 Prior Experience	Y Performance Evaluation
	74	5	28
	87	11	33
	69	4	21
	93	9	40
	81	7	38
	97	10	46

- (a) Develop the estimating equation best describing these data.
 (b) If an employee scored 83 on the aptitude test and had a prior work experience of 7, what performance evaluation would be expected?

13-40 Successful selling is as much an art as a science, but many sales managers believe that personal attributes are important in predicting sales success. Design Alley is a full-service interior design store that sells custom blinds, carpets and wall coverings. The store manager, Dee Dempsey, contracted with a sales-force selection company to conduct pre-hiring tests on four aptitudes. Dee has collected sales growth data for 25 of the salespeople who were hired, along with the scores from the four tests of aptitude: creativity, mechanical ability, abstract thinking, and mathematical calculation. Using a desktop computer, Dee generated the following Minitab output for the best-fitting multiple regression:

```
The regression equation is
GROWTH = 70.1 + 0.422 CREAT + 0.271 MECH + 0.745 ABST = 0.420 MATH

Predictor      Coef          Stdev       t-ratio        p
Constant     70.066        2.130       32.89       0.000
CREAT        0.42160       0.17192       2.45       0.024
MECH         0.27140       0.21840       1.24       0.228
ABST         0.74504       0.28982       2.57       0.018
MATH         0.41955       0.06871       6.11       0.000

s = 2.048 R-sq = 92.6%
Analysis of Variance

SOURCE      DF          SS          MS          F          p
Regression   4          1050.78    262.70     62.64     0.000
Error        20         83.88      4.19
Total        24         1134.66
```

- (a) Write the regression equation for sales growth in terms of the four factors tested?
 (b) How much of the variation in sales growth is explained by the aptitude tests?
 (c) At a significance level of 0.05, which of the aptitude tests are significant explanatory variables for sales growth?
 (d) Is the overall model significant as a whole?
 (e) Jay is a new applicant with scores on the four tests as follows: CREAT = 12, MECH = 14, ABST = 18, and MATH = 30. What sales growth is predicted by the model for this candidate?

13-41 The Money Bank desires to open new checking accounts for customers who will write at least 30 checks per month. To assist in selecting new customers, the bank has studied the relationship between the number of checks written and the age and annual income of eight of their present customers. AGE was recorded to the nearest year, and annual INCOME was recorded in thousands of dollars. The data are as follows:

Checks	Age	Income
29	37	16.2
42	34	25.4

(Continued)

(Contd.)

Checks	Age	Income
9	48	12.4
56	38	25.0
2	43	8.0
10	25	18.3
48	33	24.2
4	45	7.9

- (a) Develop an estimating equation to use age and income to predict the number of checks written per month.
 (b) How many checks per month would be expected from a 35-year-old with annual income of \$22,500?

The proportion of disposable income that consumers spend on different product categories is not the same in all towns—for example, in college towns, sales of pizza are likely to be above average, while the sales of new cars may be below average. In Exercises 13-42 through 13-45, you will construct regressions to try to explain the variability of the EATING variable. (*An important technical note:* Some simple statistical packages have difficulty with large numbers when fitting regressions. If necessary, you can avoid problems by changing the units of the data from thousands of dollars to millions of dollars. For example, for Salem, Oregon, the EATING variable becomes \$216.666 million instead of \$216,666 thousand.)

- 13-42** Develop two simple regression models for EATING, using the population and the median effective buying income per household (EBI) as the independent variables. Which independent variable accounts for more of the variation in the observed sales?
- 13-43** Develop a multiple regression for EATING using both POP and EBI as the explanatory variables. What fraction of the variation in EATING is explained by this model? Is the regression significant as a whole at $\alpha = 0.05$?
- 13-44** Include SINGLE (the number of single-person households in the area) as a third explanatory variable. How much of the variation in EATING is explained now? Is this a significant improvement over the model developed in Exercise 13-43? (Is SINGLE a significant explanatory variable in this regression?)
- 13-45** Because POP was no longer significant in the model developed in Exercise 13-44, run a regression using only EBI and SINGLE as explanatory variables. Use this model to find an approximate 90 percent confidence interval for EATING in a metropolitan area with 20,000 single-person households and a median effective buying income of \$30,000.
- 13-46** Dr. Harden Ricci is a veterinarian in Sacramento, California. Recently, he has been trying to develop a predicting equation for the amount of anesthesia (measured in milliliters) to be used in operations. He feels that the amount used will depend on the weight of the animal (in pounds), length of the operation (in hours), and whether the animal is a cat (coded 0) or a dog (coded 1). He used Minitab to run a regression on his data from 13 recent operations, and got these results:

The regression equation is
 $\text{ANESTHES} = 90.0 + 99.5 \text{ TYPE} + 21.5 \text{ WEIGHT} - 34.5 \text{ HOURS}$

Predictor	Coef	Stdev	t-ratio	p
Constant	90.032	56.842	1.58	0.148
TYPE	99.486	42.374	2.35	0.044
WEIGHT	21.536	2.668	8.07	0.000
HOURS	-34.461	28.607	-1.21	0.259

s = 57.070 R-sq = 95.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	590880	196960	60.47	0.000
Error	9	29312	3256.9		
Total	12	620192			

- (a) What is the predicting equation for amounts of anesthesia, as given by Minitab?
- (b) Give an approximate 95 percent confidence interval for the amount of anesthesia to be used in a 90-minute operation on a 25-pound dog.
- (c) At a significance level of 10 percent, is the amount of anesthesia needed significantly different for dogs and cats?
- (d) At a significance level of 5 percent, is this regression significant as a whole?

13-47

David Ichikawa is a real estate agent who works with developers who build new houses. Although much of his job concerns marketing the finished houses, he also consults with builders on how much they should pay for each lot. In one residential neighborhood, he has collected the following information on closed sales for buildable lots: Recorded sales PRICE (in \$1,000s), SIZE (linear feet of street frontage) and an indicator variable (0 or 1) for whether each lot has a VIEW. From the tax rolls, he can estimate the lot area from the square of an assessment made based on street frontage.

PRICE	SIZE	AREA (= SIZE ²)	VIEW
56.2	175	30625	1
42.5	125	15625	1
67.5	200	40000	1
39.0	115	13225	1
33.3	125	15625	0
29.0	100	10000	0
30.0	108	11664	0
48.0	170	28900	0
44.3	160	25600	0

- (a) Using MINITAB, develop the best-fitting regression line for these data.
- (b) What fraction of the variation in PRICE is accounted for in this equation?

- (c) Find a 90 percent confidence interval for the increase in market value attributable to having a VIEW.

- (d) Was it helpful to use AREA (the square of SIZE) in the regression? Explain.

13-48 Camping-R-Us, a newcomer to the outdoor equipment field, plans to market a two-person, three-season tent for weekend campers. To set a fair price, they look at eight comparable tents currently on the market, in terms of weight and square footage. The data follow:

	Weight (oz)	Sq Ft.	Price
Kelty Nautilus	94	37	\$225
North Face Salamander	90	36	240
REI Mountain Hut	112	35	225
Sierra Designs Meteor Light	92	40	220
Eureka! Cirrus 3	93	48	167
Sierra Designs Clip 3	98	40	212
Eureka! Timberline Deluxe	114	40	217
Diamond Brand Free Spirit	108	35	200

- (a) Calculate the least-squares equation to predict price from weight and square footage.
 (b) If Camping-R-Us' tent weighs 100 ounces and has 46 square feet of space, how much should they charge?

13-49 The Carolina Athletic Association is interested in organizing the First Annual Tarheel Triathlon. To attract top competitors, they wish to establish cash incentives for the top finishers by setting times for both men and women overall winners. Because this course has never been run before, the CAA has chosen 10 races of varying lengths that they consider comparable in weather and course conditions.

Triathlon	Miles			Winning Times (Hr:Min:Sec)	
	Swim	Bike	Run	Men	Women
Bud Light Ironman	2.4	112	26.2	8:09:15	9:00:56
World's Toughest	2.0	100	18.6	8:25:09	9:49:04
Muncie Endurathon	1.2	55.3	13.1	4:05:30	4:40:06
Texas Hill Country	1.5	48	10.0	3:24:24	3:55:02
Leon's Q.E.M.	0.93	24.8	6.2	1:54:32	2:07:10
Sacramento International	0.93	24.8	6.2	1:48:16	2:00:45
Malibu	0.50	18	5.0	1:19:25	1:30:19
Bud Light Endurance	2.4	112	26.2	9:26:30	11:00:29
Wendy's	0.5	20	4.0	1:14:59	1:23:09
Mammoth/Snowcreek	0.6	25	6.2	1:56:07	2:11:49

- (a) Determine the regression equations to predict men's and women's winning times, in terms of the length of each individual race segment. (Convert the times to minutes for use in calculations.)
- (b) Predict the winning times if the Tarheel Triathlon comprises a 1-mile swim, 50-mile bike ride, and a 12.5-mile run.
- (c) If the CAA wants to use the lower limit of an approximate 90 percent confidence interval for the incentive times for men and women, what would these times be?

13-50 Peoria, Illinois, is in the process of modifying its tax structure. Twelve cities of comparable size and economic structure were surveyed as to specific taxes and the associated total tax revenue.

- (a) Use the following data to determine the least-squares equation relating revenue to the three tax rates.

Property	Tax Rates Sales	Tax Revenue (\$ thousands)	Gasoline
1.639%	2.021%	\$28,867.5	3.300 ¢/gal
1.686	1.972	28,850.2	3.300
1.639	2.041	29,011.5	3.300
1.639	2.363	28,806.5	0.131
1.639	2.200	28,821.7	2.540
1.639	2.201	28,774.6	1.560
1.654	2.363	28,803.2	0.000
2.643	1.000	28,685.7	3.300
2.584	1.091	28,671.8	2.998
2.048	1.752	28,671.0	1.826
2.176	1.648	28,627.4	1.555
1.925	1.991	28,670.7	0.757

- (b) Two proposals have been submitted for Peoria. Estimate total tax revenues if the tax rates are

	Property	Sales	Gasoline
Proposal A	2.763%	1.000%	1.0¢/gal
Proposal B	1.639	2.021	3.3

Determine which proposal the city should adopt.

13-51 The National Cranberry Cooperative, an organization formed and owned by growers of cranberries to process and market their berries, is trying to establish a relationship between average price per barrel received in any given year and the total number of barrels sold in the previous year (divided into fresh sales and berries sold for processing).

- (a) Calculate the least-squares equation to predict price from these sales figures.

Sales (in thousands of barrels)			Sales (in thousands of barrels)		
Fresh	Process	Price	Fresh	Process	Price
844	256	15.50	320	460	9.79
965	335	17.15	528	860	10.90
470	672	11.71	340	761	15.88

- 13-52** (b) Predict next year's price per barrel if this year's sales are 980 (fresh) and 360 (process). Cellular phones were introduced in Europe in 1980, and since then, their growth in popularity has been phenomenal. The number of subscribers in subsequent years is contained in the following table.

1981	3,510	1984	143,300	1987	877,850
1982	34,520	1985	288,420	1988	1,471,200
1983	80,180	1986	507,930	1989	2,342,080

Using the number of years since the introduction of cellular phones as the independent variable (i.e., 1981 = 1, etc.), find the least-squares linear equation relating these two variables. Look at the residuals—do they have a noticeable pattern? Find the least-squares quadratic equation. Which appears to be a better fit?

- 13-53** While shopping for a new down sleeping bag, Fred Montana is curious about what features of a bag are most important in determining the bag's price. He picks six Gore-Tex sleeping bags and decides to run a linear regression analysis to find out.

	Down Fill (oz)	Total Weight (lb)	Loft (in.)	Temp. Rating (°F)	Price (\$)
Swallow	14.0	2.00	5.5	20	255
Snow Bunting	18.0	2.25	6.5	10	285
Puffin	24.0	3.13	6.5	10	329
Widgeon	25.5	3.25	7.5	10	395
Tern	32.5	3.63	9.0	-30	459
Snow Goose	41.0	4.25	10.0	-40	509

- (a) Regress price on ounces of down fill, total weight, loft, and temperature rating. Using the prob values, determine which of these variables are significant at the $\alpha = 0.01$ level.
- (b) What about the regression as a whole? Use the ANOVA prob value, again at the $\alpha = 0.01$ level, to determine whether the regression as a whole is significant.
- (c) What problem might there be in using all these variables together? Do the answers to parts (a) and (b) seem to indicate this problem might be present?

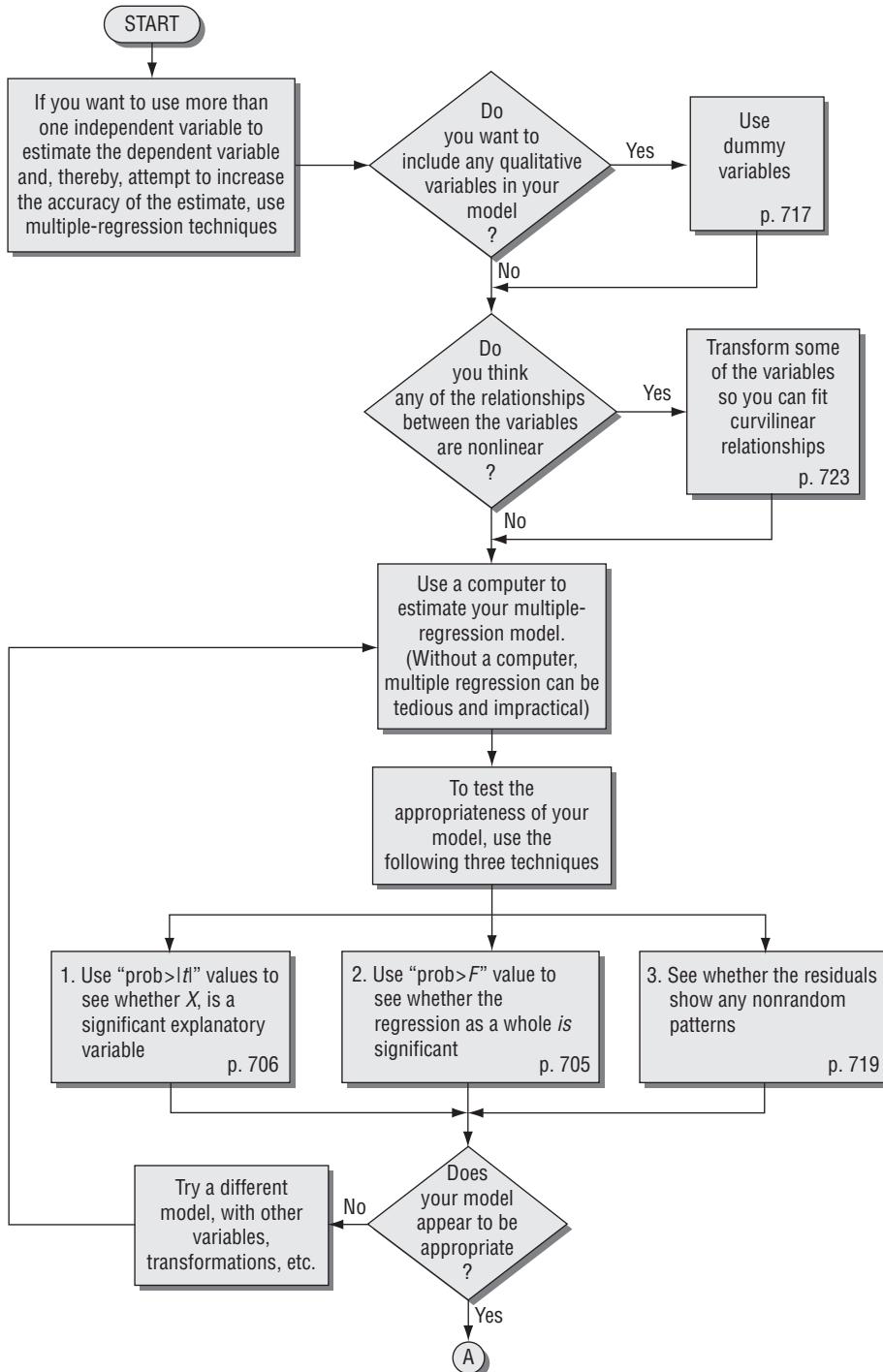


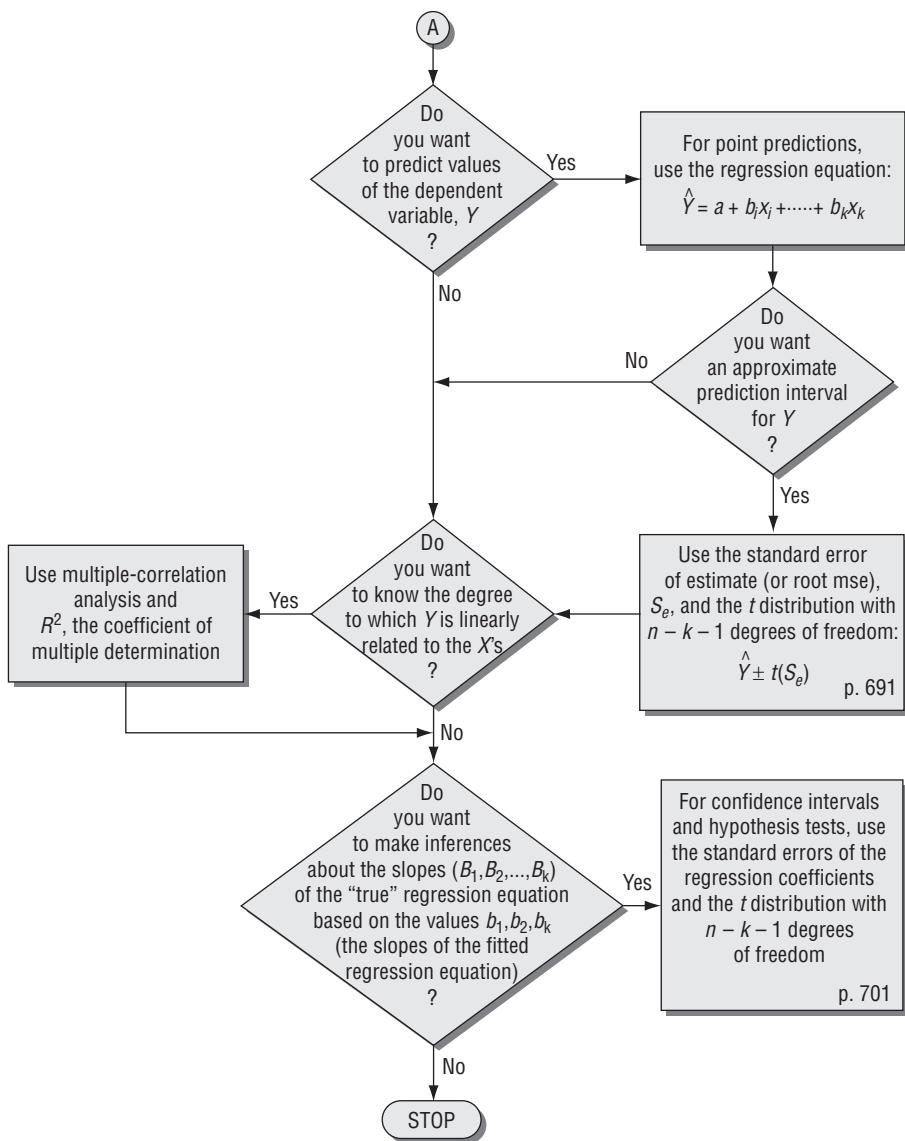
Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Build a regression model to study the level of satisfaction of the customers with e-services provided by their banks on the basis of the importance of e-facilities on bank selection. [Regress Q9 on Q7(a)-(m)]
2. Is there evidence of presence of multicollinearity in the data?
3. If there is multicollinearity problem, what remedial action, if any, would you take?



Flow Chart: Multiple Regression and Modeling





14 Nonparametric Methods

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To test hypotheses when we cannot make any assumptions about the distribution from which we are sampling
 - To know which distribution-free (nonparametric) tests are appropriate for different situations
 - To use and interpret each of six standard nonparametric hypothesis tests
 - To learn the advantages and disadvantages of nonparametric tests
-

CHAPTER CONTENTS

- 14.1 Introduction to Nonparametric Statistics 748
- 14.2 The Sign Test for Paired Data 750
- 14.3 Rank Sum Tests: The Mann–Whitney *U* Test and the Kruskal–Wallis Test 758
- 14.4 The One-Sample Runs Test 772
- 14.5 Rank Correlation 781
- 14.6 The Kolmogorov–Smirnov Test 793

- Statistics at Work 800
- Terms Introduced in Chapter 14 801
- Equations Introduced in Chapter 14 802
- Review and Application Exercises 803
- Flow Chart: Nonparametric Methods 814

Although the effect of air pollution on health is a complex problem, an international organization has decided to make a preliminary investigation of average year-round quality of air and the incidence of pulmonary-related diseases. A preliminary study ranked 11 of the world's major cities from 1 (worst) to 11 (best) on these two variables.

	City										
	A	B	C	D	E	F	G	H	I	J	K
Air-quality rank	4	7	9	1	2	10	3	5	6	8	11
Pulmonary-disease rank	5	4	7	3	1	11	2	10	8	6	9

The health organization's data are different from any we have seen so far in this book: They do not give us the *variable* used to determine these ranks. (We don't know whether the rank of pulmonary disease is a result of pneumonia, emphysema, or other illnesses per 100,000 population.) Nor do we know the values (whether City *D* has twice as much pollution as City *K* or 20 times as much). If we knew the variables and their values, we could use the regression techniques of Chapter 12.

Unfortunately, that is not the case; but even without any knowledge of variables or values, we can use the techniques in this chapter to help the health organization with its problem. ■

14.1 INTRODUCTION TO NONPARAMETRIC STATISTICS

The majority of hypothesis tests discussed so far have made inferences about population *parameters*, such as the mean and the proportion. These parametric tests have used the parametric statistics of samples that came from the population being tested. To formulate these tests, we made restrictive assumptions about the populations from which we drew our samples. In each case in Chapters 8 and 9, for example, we assumed that our samples either were large or came from *normally distributed* populations. But populations are not always normal. And even if a goodness-of-fit test (Chapter 11) indicates that a population *is* approximately normal, we cannot always be sure we're right because the test is not 100 percent reliable. Clearly, there are certain situations in which the use of the normal curve is not appropriate. For these cases, we need alternatives to the parametric statistics and the specific hypothesis tests we've been using so far.

Fortunately, in recent times statisticians have developed useful techniques that do not make restrictive assumptions about the shape of population distributions. **These are known as distribution-free or, more commonly, nonparametric tests.** The hypotheses of a nonparametric test are concerned with something other than the value of a population parameter. A large number of these tests exist, but this chapter will examine only a few of the better known and more widely used ones:

1. The *sign test* for paired data, where positive or negative signs are substituted for quantitative values.
2. A *rank sum test*, often called the *Mann-Whitney U test*, which can be used to determine whether two independent samples have been drawn from the same population. It uses more information than the sign test.
3. Another rank sum test, the *Kruskal-Wallis test*, which generalizes the analysis of variance discussed in Chapter 11 to enable us to dispense with the assumption that the populations are normally distributed.

Parametric statistics

Shortcomings of parametric statistics

Nonparametric statistics

4. The *one-sample runs test*, a method for determining the randomness with which sampled items have been selected.
5. *Rank correlation*, a method for doing correlation analysis when the data are not available to use in numerical form, but when information is sufficient to rank the data first, second, third, and so forth.
6. The *Kolmogorov-Smirnov test*, another method for determining the goodness of fit between an observed sample and a theoretical probability distribution.

Advantages of Nonparametric Methods

Nonparametric methods have a number of clear advantages over parametric methods:

Advantages of nonparametric methods

1. **They do not require us to make the assumption that a population is distributed in the shape of a normal curve or another specific shape.**
2. **Generally, they are easier to do and to understand.** Most nonparametric tests do not demand the kind of laborious computations often required, for example, to calculate a standard deviation. A nonparametric test may ask us to replace numerical values with the order in which those values occur in a list, as has been done in Table 14-1. Obviously, dealing computationally with 1, 2, 3, 4, and 5 takes less effort than working with 13.33, 76.50, 101.79, 113.45, and 189.42.
3. **Sometimes even formal ordering or ranking is not required.** Often, all we can do is describe one outcome as “better” than another. When this is the case, or when our measurements are not as accurate as is necessary for parametric tests, we can use nonparametric methods.

Disadvantages of Nonparametric Methods

Two disadvantages accompany the use of nonparametric tests:

Shortcomings of nonparametric methods

1. **They ignore a certain amount of information.** We have demonstrated how the values 1, 2, 3, 4, and 5 can replace the numbers 13.33, 76.50, 101.79, 113.45, and 189.42. Yet if we represent “189.42” by “5,” we lose information that is contained in the value of 189.42. Notice that in our ordering of the values 13.33, 76.50, 101.79, 113.45, and 189.42, the value 189.42 can become 1, 189.42 and still be the fifth, or largest, value in the list. But if this list is a data set, we can learn more knowing that the highest value is 1,189.42 instead of 189.42 than we can by representing both of these numbers by the value 5.
2. **They are often not as efficient or “sharp” as parametric tests.** The estimate of an interval at the 95 percent confidence level using a nonparametric test may be twice as large as the estimate using a parametric test such as those in Chapters 8 and 9. When we use nonparametric tests, we make a trade-off: We lose sharpness in estimating intervals, but we gain the ability to use less information and to calculate faster.

TABLE 14-1 CONVERTING PARAMETRIC VALUES TO NONPARAMETRIC RANKS

Parametric value	113.45	189.42	76.50	13.33	101.79
Nonparametric value	4	5	2	1	3

EXERCISES 14.1

Basic Concepts

- 14-1** What is the difference between the kinds of questions answered by parametric tests and those answered by nonparametric tests?
- 14-2** The null hypothesis most often examined in nonparametric tests (*choose one answer*)
- Includes specification of a population's parameters.
 - Is used to evaluate some general population aspect.
 - Is very similar to that used in regression analysis.
 - Simultaneously tests more than two population parameters.
- 14-3** What are the major advantages of nonparametric methods over parametric methods?
- 14-4** What are the primary shortcomings of nonparametric tests?
- 14-5** George Shoaf is an interviewer with a large insurance company. George works in the company's home office, and to make the best use of his time, the company requires the receptionist to schedule his interviews according to a precise schedule. There is no 5-minute period unaccounted for, including telephone calls. Unfortunately, the receptionist has been underestimating the amount of time interviews will take, and she has been scheduling too many prospective employees, resulting in long waits in the lobby. Although waiting periods may be short in the morning, as the day progresses and the interviewer gets further behind, the waits become longer. In assessing the problem, should the interviewer assume that the successive waiting times are normally distributed?

Applications

- 14-6** International Communications Corporation is planning to change the benefits package offered to employees. The company is considering different combinations of profit-sharing, health-care, and retirement benefits. Samples of a broad range of benefit combinations were described in a pamphlet and distributed among employees, whose preferences were then recorded. The results follow:

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Profit-sharing—Health-Care—Retirement																			
Combination																			
Number of Preferences																			
15	5	14	4	6	16	7	8	13	3	17	18	12	2	9	1	11	19	10	
52	49	39	38	37	36	32	29	26	25	24	18	15	15	14	10	10	10	9	

Will the company sacrifice any real information by using the ranking test as its decision criterion? (*Hint:* You might graph the data.)

14.2 THE SIGN TEST FOR PAIRED DATA

One of the easiest nonparametric tests to use is the sign test. Its name comes from the fact that it is based on the direction (or signs for pluses or minuses) of a pair of observations, not on their numerical magnitude.

Use the sign test for paired data

Consider the result of a test panel of 40 college juniors evaluating the effectiveness of two types of classes: large lectures by full professors and small sections by graduate assistants. Table 14-2 lists the responses to this request: “Indicate how you rate the effectiveness in transmitting knowledge of these two types of classes by giving them a number from 4 to 1. A rating of 4 is excellent and 1 is poor.” In this case, the sign test can help us determine whether students feel there is a difference between the effectiveness of the two types of classes.

We can begin, as we have in Table 14-2, by converting the evaluations of the two teaching methods into signs. Here a plus sign means the student prefers large lectures, a minus sign indicates a preference for small sections, and a zero represents a tie (no preference). If we count the bottom row of Table 14-2, we get these results:

Number of + signs	19
Number of – signs	11
Number of 0s	10
Total sample size	40

Converting values to signs

Stating the Hypotheses

We are using the sign test to determine whether our panel can discern a real difference between the two types of classes. Because we are testing perceived differences, we shall exclude tie evaluations (0s). We can see that we have 19 plus signs and 11 minus signs, for a total of 30 usable responses. If there is no difference between the two types of classes, p (the probability that the first score exceeds the second score) would be 0.5, and we would expect to get about 15 plus signs and 15 minus signs. We would set up our hypotheses like this:

$H_0: p = 0.5 \leftarrow$ Null hypothesis: There is no difference between the two types of classes

$H_1: p \neq 0.5 \leftarrow$ Alternative hypothesis: There is a difference between the two types of classes

Finding the sample size

If you look carefully at the hypotheses, you will see that the situation is similar to the fair-coin toss we discussed in Chapter 4. If we tossed a fair coin 30 times, p would be 0.5, and we would expect about 15 heads and 15 tails. In that case, we would use the binomial distribution as the appropriate sampling distribution. You may also remember that when np and nq are each at least 5, we can use the normal distribution to approximate the binomial. This is just the case with the results from our panel of college juniors. Thus, we can apply the normal distribution to our test of the two teaching methods.

Choosing the distribution

$p_{H_0} = 0.5 \leftarrow$ Hypothesized proportion of the population that prefers large lectures

$q_{H_0} = 0.5 \leftarrow$ Hypothesized proportion of the population that prefers small sections ($q_{H_0} = 1 - p_{H_0}$)

$n = 30 \leftarrow$ Sample size

$\bar{p} = 0.633 \leftarrow$ Proportion of successes in the sample (19/30)

$\bar{q} = 0.367 \leftarrow$ Proportion of failures in the sample (11/30)

Setting up the problem symbolically

TABLE 14-2 EVALUATION BY 40 STUDENT OF TWO TYPES OF CLASSES

Panel-member number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Score for large lectures (1)	2	1	4	4	3	3	4	2	4	1	3	3	4	4	4	1
Score for small sections (2)	3	2	2	3	4	2	2	1	3	1	2	3	4	4	3	2
Sign of score 1 minus score 2	-	-	+	+	-	+	+	+	+	0	+	0	0	0	+	-

Testing a Hypothesis of No Difference

Suppose the chancellor's office wants to test the hypothesis that there is no difference between student perception of the two types of classes at the 0.05 level of significance. We shall conduct this test using the methods we introduced in Chapter 8. The first step is to calculate the standard error of the proportion:

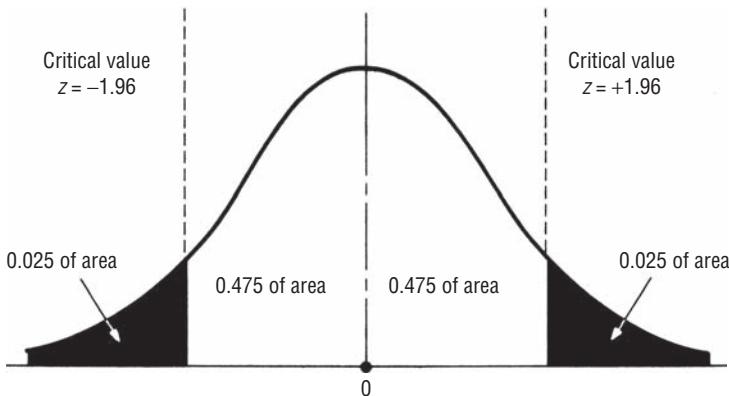
$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{pq}{n}} \\ &= \sqrt{\frac{(0.5)(0.5)}{30}} \\ &= \sqrt{0.00833} \\ &= 0.091 \leftarrow \text{Standard error of the proportion}\end{aligned}\quad [7-4]$$

Because we want to know whether the true proportion is larger or smaller than the hypothesized proportion, this is a two-tailed test. Figure 14-1 illustrates this hypothesis test graphically. The two colored regions represent the 0.05 level of significance.

Next we use Equation 6-2 to *standardize* the sample proportion, \bar{p} , by subtracting p_{H_0} , the hypothesized proportion, and dividing by $\sigma_{\bar{p}}$, the standard error of the proportion.

Calculating the standard error

Illustrating the test graphically

**FIGURE 14-1 TWO-TAILED HYPOTHESIS TEST OF A PROPORTION AT THE 0.05 LEVEL OF SIGNIFICANCE**

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	2	2	4	4	4	4	3	3	2	3	4	3	4	3	1	4	3	2	2	2	1	3	3
3	2	3	3	1	4	3	3	2	2	1	1	1	3	2	2	4	4	3	3	1	1	4	2
-	0	-	+	+	0	+	0	+	0	+	+	+	+	+	-	0	-	-	-	+	0	-	+

$$z = \frac{\bar{p} - p_{H_0}}{\sigma_{\bar{p}}} \quad [6-2]$$

$$= \frac{0.633 - 0.5}{0.091}$$

$$= 1.462$$

Placing this standardized value, 1.462, on the z scale shows that the sample proportion falls well within the acceptance region as shown in Figure 14-2. Therefore, the chancellor should accept the null hypothesis that students perceive no difference between the two types of classes.

A sign test such as this is quite simple to do and applies to both one-tailed and two-tailed tests. It is usually based on the binomial distribution. Remember, however, that we were able to use the normal approximation to the binomial as our sampling distribution because np and nq were both greater than 5. When these conditions are not met, we must use the binomial instead.

Interpreting the results

A final word about the sign test

HINTS & ASSUMPTIONS

Nonparametric tests are very convenient when the real world presents distribution-free data on which a decision must be taken. Hint: Note that the *sign test* is just another application of the familiar *normal approximation to the binomial*, using + and – instead of “success” and “failure.”

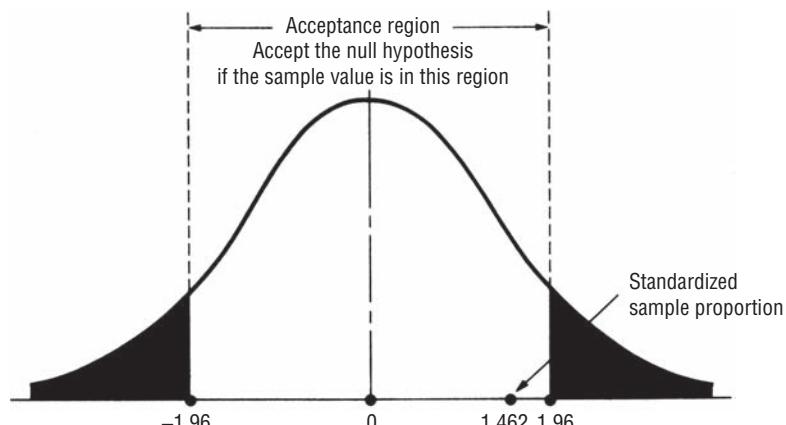


FIGURE 14-2 TWO-TAILED HYPOTHESIS TEST AT THE 0.05 LEVEL OF SIGNIFICANCE, ILLUSTRATING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE PROPORTION

EXERCISES 14.2

Self-Check Exercises

SC 14-1 The following data show employees' rates of defective work before and after a change in the wage incentive plan. Compare the following two sets of data to see whether the change lowered the defective units produced. Use the 0.10 level of significance.

Before	8	7	6	9	7	10	8	6	5	8	10	8
After	6	5	8	6	9	8	10	7	5	6	9	8

SC 14-2 After collecting data on the amount of air pollution in Los Angeles, the Environmental Protection Agency decided to issue strict new rules to govern the amount of hydrocarbons in the air. For the next year, it took monthly measurements of this pollutant and compared them to the preceding year's measurements for corresponding months. Based on the following data, does the EPA have enough evidence to conclude with 95 percent confidence that the new rules were effective in lowering the amount of hydrocarbons in the air? To justify these laws for another year, it must conclude at $\alpha = 0.10$ that they are effective. Will these laws still be in effect next year?

	Last Year*	This Year
Jan.	7.0	5.3
Feb.	6.0	6.1
Mar.	5.4	5.6
April	5.9	5.7
May	3.9	3.7
June	5.7	4.7
July	6.9	6.1
Aug.	7.6	7.2
Sept.	6.3	6.4
Oct.	5.8	5.7
Nov.	5.1	4.9
Dec.	5.9	5.8

*Measured in parts per million.

Applications

14-7 The following data show employees' satisfaction levels (as a percentage) before and after their company was bought by a larger firm. Did the buyout increase employee satisfaction? Use the 0.05 significance level.

Before	98.4	96.6	82.4	96.3	75.4	82.6	81.6	91.4	90.4	92.4
After	82.4	95.4	94.2	97.3	77.5	82.5	81.6	84.5	89.4	90.6

- 14-8** Use the sign test to see whether there is a difference between the number of days required to collect an account receivable before and after a new collection policy. Use the 0.05 significance level.

Before	33	36	41	32	39	47	34	29	32	34	40	42	33	36	29
After	35	29	38	34	37	47	36	32	30	34	41	38	37	35	28

- 14-9** A light-aircraft engine repair shop switched the payment method it used from hourly wage to hourly wage plus a bonus computed on the time required to disassemble, repair, and reassemble an engine. The following are data collected for 25 engines before the change and 25 after the change. At a 0.10 significance level, did the new plan increase productivity?

Hours Required		Hours Required	
Before	After	Before	After
29	32	25	34
34	19	42	27
32	22	20	26
19	21	25	25
31	20	33	31
22	24	34	19
28	25	20	22
31	31	21	32
32	18	22	31
44	22	45	30
41	24	43	29
23	26	31	20
34	41		

- 14-10** Because of the severity of recent winters, there has been talk that the earth is slowly progressing toward another ice age. Some scientists hold different views, however, because the summers have brought extreme temperatures as well. One scientist proposed looking at the mean temperature for each month to see whether it was lower than in the previous year. Another meteorologist at the government weather service argued that perhaps they should look as well at temperatures in the spring and fall months of the last 2 years, so that their conclusions would be based on other than extreme temperatures. In this way, he said, they could detect whether there appeared to be a general warming or cooling trend or just extreme temperatures in the summer and winter months. So 15 dates in the spring and fall were randomly selected, and the temperatures in the last 2 years were noted for a particular location with generally moderate temperatures. Following are the dates and corresponding temperatures for 1994 and 1995.

- (a) Is the meteorologist's reasoning as to the method of evaluation sound? Explain.
- (b) Using a sign test, determine whether the meteorologist can conclude at $\alpha=0.05$ that 1995 was cooler than 1994, based on these data.

Temperature (Fahrenheit)					
Date	1994	1995	Date	1994	1995
Mar. 29	58	57	Oct. 12	54	48
Apr. 4	45	70	May 31	74	79
Apr. 13	56	46	Sept. 28	69	60
May 22	75	67	June 5	80	74
Oct. 1	52	60	June 17	82	79
Mar. 23	49	47	Oct. 5	59	72
Nov. 12	48	45	Nov. 28	50	50
Sept. 30	67	71			

- 14-11** With the concern over radiation exposure and its relationship to the incidence of cancer, city environmental specialists keep a close eye on the types of industry coming into the area and the degree to which they use radiation in their production. An index of exposure to radioactive contamination has been developed and is being used daily to determine whether the levels are increasing or are higher under certain atmospheric conditions.

Environmentalists claim that radioactive contamination has increased in the last year because of new industry in the city. City administrators, however, claim that new, more stringent regulations on industry in the area have made levels lower than last year, even with new industry using radiation. To test their claim, records for 11 randomly selected days of the year have been checked, and the index of exposure to radioactive contamination has been noted. The following results were obtained:

Index of Radiation Exposure												
1994	1.402	1.401	1.400	1.404	1.395	1.402	1.406	1.401	1.404	1.406	1.397	
1995	1.440	1.395	1.398	1.404	1.393	1.400	1.401	1.402	1.400	1.403	1.402	

Can the administrators conclude at $\alpha = 0.15$ that the levels of radioactive contamination have changed or, more specifically, that they have been reduced?

- 14-12** As part of the recent interest in population growth and the sizes of families, a population researcher examined a number of hypotheses concerning the family size that various people look upon as ideal. She suspected that variables of race, sex, age, and background might account for some of the different views. In one pilot sample, the researcher tested the hypothesis that women today think of an ideal family as being smaller than the ideal held by their mothers. She asked each of the participants in the pilot study to state the number of children she would choose to have or that she considered ideal. Responses were anonymous, to guard against the possibility that people would feel obligated to give a socially desirable answer. In addition, people of different backgrounds were included in the sample. The following are the responses of the mother–daughter pairs.

Sample Pair	Ideal Family Size												
	A	B	C	D	E	F	G	H	I	J	K	L	M
Daughter	3	4	2	1	5	4	2	2	3	3	1	4	2
Mother	4	4	4	3	5	3	3	5	3	2	2	3	1

- (a) Can the researcher conclude at $\alpha = 0.03$ that the mothers and daughters do not have essentially the same ideal of family size? Use the binomial distribution.
- (b) Determine whether the researcher could conclude that the mothers do not have essentially the same family-size preferences as their daughters by using the normal approximation to the binomial.
- (c) Assume that for each pair listed, there were 10 more pairs who responded in an identical manner. Calculate the range of the proportion for which the researcher would conclude that there is no difference in the mothers and daughters. Is your conclusion different?
- (d) Explain any differences in conclusions obtained in parts (a), (b), and (c).

14-13 A nationwide used-car company has developed a new instructional video to educate salespeople. Twenty employees' average monthly car sales are presented below for time periods both before and after the video's creation. Does the company have enough evidence to conclude with 95 percent confidence that the video was effective in increasing the average number of cars sold? If we just consider the employees with low sales (less than an average of 12 cars per month before the video), did the video increase their selling performance?

Before	18.4	16.9	17.4	11.6	10.5	12.7	22.3	18.5	17.5	16.4
After	18.6	16.8	17.3	15.6	19.5	12.6	22.3	16.5	18.0	16.4
Before	15.9	18.6	23.5	18.7	9.4	16.3	18.5	17.4	11.3	8.4
After	17.4	18.6	23.5	18.9	15.6	15.4	17.6	17.4	16.5	13.4

Worked-Out Answers to Self-Check Exercises

SC 14-1

Before	8	7	6	9	7	10	8	6	5	8	10	8
After	6	5	8	6	9	8	10	7	5	6	9	8
Sign	-	-	+	-	+	-	+	+	0	-	-	0

12 responses: 4(+), 6(0), 2(0).

For $n = 10$, $p = 0.5$, the probability of 6 or more minuses is 0.3770 (Appendix Table 3). Because $0.3770 > 0.10$, we cannot reject H_0 . The wage incentive plan did not significantly lower the rates of defective work.

SC 14-2

Before	7.0	6.0	5.4	5.9	3.9	5.7	6.9	7.6	6.3	5.8	5.1	5.9
After	5.3	6.1	5.6	5.7	3.7	4.7	6.1	7.2	6.4	5.7	4.9	5.8
Sign	-	+	+	-	-	-	-	-	+	-	-	-

12 responses: 3(+), 9(–).

For $n = 12$, $p = 0.5$, the probability of 9 or more minuses is 0.0729 (Appendix Table 3). Because $0.10 > 0.0729 > 0.05$, they cannot be 95 percent confident that hydrocarbon levels have been lowered, but they will conclude at $\alpha = 0.10$ that the rules are effective. Hence, they will still be in effect next year.

14.3 RANK SUM TESTS: THE MANN–WHITNEY *U* TEST AND THE KRUSKAL–WALLIS TEST

In Chapter 11, we showed how to use analysis of variance to test the hypothesis that several population means are equal. We assumed in such tests that the populations were normally distributed with equal variances. Many times these assumptions cannot be met, and in such cases, we can use two nonparametric tests, neither of which depends on the normality assumptions. Both of these tests are called rank sum tests because the test depends on the ranks of the sample observations.

Use based on the number of populations involved

Rank sum tests are a whole family of tests. We shall concentrate on just two members of this family, the Mann–Whitney *U* test and the Kruskal–Wallis test. We'll use the Mann–Whitney test when only two populations are involved and the Kruskal–Wallis test when more than two populations are involved. Use of these tests will enable us to determine whether independent samples have been drawn from the same population (or from different populations having the same distribution). The use of *ranking* information rather than pluses and minuses is less wasteful of data than the sign test.

Solving a Problem Using the Mann–Whitney *U* Test

Suppose that the board of regents of a large eastern state university wants to test the hypothesis that the mean SAT scores of students at two branches of the state university are equal. The board keeps statistics on all students at all branches of the system. A random sample of 15 students from each branch has produced the data shown in Table 14-3.

Ranking the items to be tested

To apply the Mann–Whitney *U* test to this problem, we begin by ranking all the scores in order from lowest to highest, indicating beside each the symbol of the branch. Table 14-4 accomplishes this.

Next, let's learn the symbols used to conduct a Mann–Whitney *U* test in the context of this problem:

n_1 = number of items in sample 1, that is, the number of students at Branch A

n_2 = number of items in sample 2, that is, the number of students at Branch S

R_1 = sum of the ranks of the items in sample 1: the sum from Table 14-5 of the ranks of all the Branch A scores

Symbols for expressing the problem

R_2 = sum of the ranks of the items in sample 2: the sum from Table 14-5 of the ranks of all the Branch S scores

In this case, both n_1 and n_2 are equal to 15, but it is *not* necessary for both samples to be of the same size. Now in Table 14-5, we can reproduce the data from Table 14-3, adding the ranks from Table 14-4.

TABLE 14-3 SAT SCORES FOR STUDENTS AT TWO STATE UNIVERSITY BRANCHES

Branch A	1,000	1,100	800	750	1,300	950	1,050	1,250
Branch S	920	1,120	830	1,360	650	725	890	1,600
Branch A	1,400	850	1,150	1,200	1,500	600	775	
Branch S	900	1,140	1,550	550	1,240	925	500	

TABLE 14-4 SAT SCORES RANKED FROM LOWEST TO HIGHEST

Rank	Score	Branch	Rank	Score	Branch
1	500	S	16	1,000	A
2	550	S	17	1,050	A
3	600	A	18	1,100	A
4	650	S	19	1,120	S
5	725	S	20	1,140	S
6	750	A	21	1,150	A
7	775	A	22	1,200	A
8	800	A	23	1,240	S
9	830	S	24	1,250	A
10	850	A	25	1,300	A
11	890	S	26	1,360	S
12	900	S	27	1,400	A
13	920	S	28	1,500	A
14	925	S	29	1,550	S
15	950	A	30	1,600	S

Then we can total the ranks for each branch. As a result, we have all the values we need to solve this problem, because we know that

$$n_1 = 15$$

$$n_2 = 15$$

$$R_1 = 247$$

$$R_2 = 218$$

TABLE 14-5 RAW DATA AND RANK FOR SAT SCORES

Branch A	Rank	Branch S	Rank
1,000	16	920	13
1,100	18	1,120	19
800	8	830	9
750	6	1,360	26
1,300	25	650	4
950	15	725	5
1,050	17	890	11
1,250	24	1,600	30
1,400	27	900	12
850	10	1,140	20
1,150	21	1,550	29
1,200	22	550	2
1,500	28	1,240	23
600	3	925	14
775	7	500	1
<hr/> $247 \leftarrow \text{Total ranks}$		<hr/> $218 \leftarrow \text{Total ranks}$	

Calculating the *U* Statistic

Using the values for n_1 and n_2 and the rank sums R_1 and R_2 , we can determine the *U statistic*, a measure of the difference between the ranked observations of the two samples of SAT scores:

***U* Statistic**

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad [14-1]$$

***U* statistic defined**

Computing the *U* statistic

$$\begin{aligned} &= (15)(15) + \frac{(15)(16)}{2} - 247 \\ &= 225 + 120 - 247 \\ &= 98 \leftarrow U \text{ statistic} \end{aligned}$$

If the null hypothesis that the $n_1 + n_2$ observations came from identical populations is true, then this *U* statistic has a sampling distribution with a mean of

Mean of the Sample Distribution of *U*

$$\mu_u = \frac{n_1 n_2}{2}$$

[14-2]

$$\begin{aligned} &= \frac{(15)(15)}{2} \\ &= 112.5 \leftarrow \text{Mean of the } U \text{ statistic} \end{aligned}$$

and a standard error of

Standard Error of the *U* Statistic

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad [14-3]$$

$$\begin{aligned} &= \sqrt{\frac{(15)(15)(15+15+1)}{12}} \\ &= \sqrt{\frac{6,975}{12}} \\ &= \sqrt{581.25} \\ &= 24.1 \leftarrow \text{Standard error of the } U \text{ statistic} \end{aligned}$$

Testing the Hypotheses

The sampling distribution of the U statistic can be approximated by the normal distribution when both n_1 and n_2 are larger than 10. Because our problem meets this condition, we can use the standard normal probability distribution table to make our test. The board of regents wishes to test at the 0.15 level of significance the hypothesis that these samples were drawn from identical populations.

$H_0: \mu_1 = \mu_2$ ← Null hypothesis: There is no difference between the two populations, so they have the same mean

$H_1: \mu_1 \neq \mu_2$ ← Alternative hypothesis: There is a difference between the two populations; in particular, they have different means

$\alpha = 0.15$ ← Level of significance for testing these hypotheses

Stating the hypotheses

Finding the limits of the acceptance region

The board of regents wants to know whether the mean SAT score for students at either of the two schools is better or worse than the other. Therefore, this is a two-tailed hypothesis test. Figure 14-3 illustrates this test graphically. The two colored areas represent the 0.15 level of significance. Because we are using the normal distribution as our sampling distribution in this test, we can determine from Appendix Table 1 that the critical z value for an area of 0.425 is 1.44.

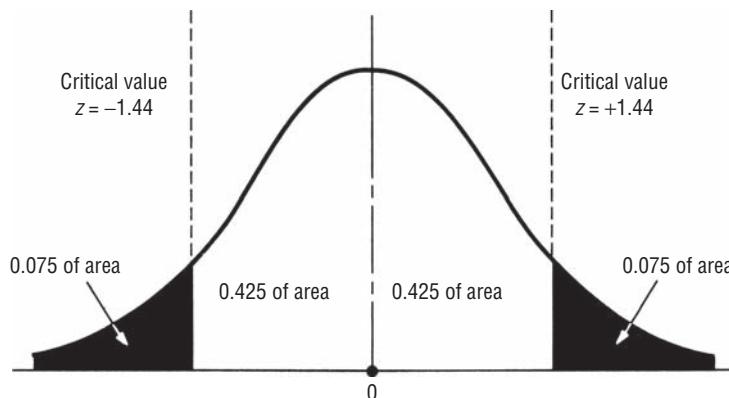
Next, we use Equation 6-2 to *standardize* the sample U statistic, by subtracting μ_U , its mean, and dividing by σ_U , its standard error.

$$z = \frac{U - \mu_u}{\sigma_u} \quad [6-2]$$

$$= \frac{98 - 112.5}{24.1}$$

$$= -0.602$$

Figure 14-4 shows the standardized sample value of U and the critical values of z for the test. The board of regents should notice that the sample statistic does lie within the critical values for the test, and conclude that the distributions, and hence the mean SAT scores at the two schools, are the same.



Illustrating the test graphically

FIGURE 14-3 TWO-TAILED HYPOTHESIS TEST AT THE 0.15 LEVEL OF SIGNIFICANCE

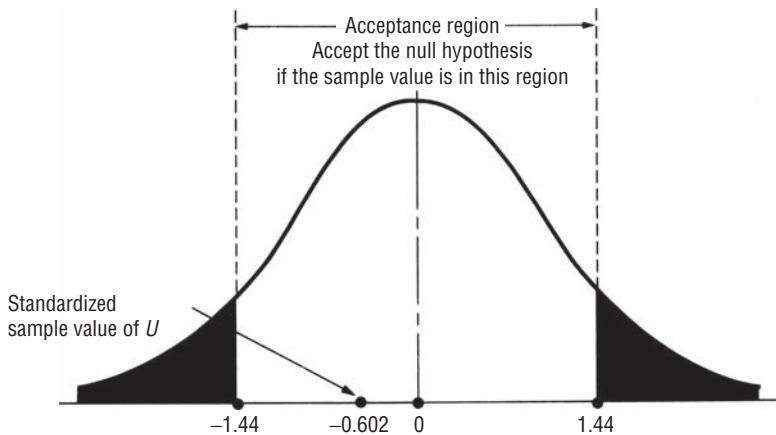


FIGURE 14-4 TWO-TAILED HYPOTHESIS TEST AT THE 0.15 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE U STATISTIC

Special Properties of the *U* Test

The *U* statistic has a feature that enables users to save calculating time when the two samples under observation are of unequal size. We just computed the value of *U* using Equation 14-1:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad [14-1]$$

But just as easily, we could have computed the *U* statistic using the R_2 value, like this:

Alternate Formula for the *U* Statistic

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \quad [14-4]$$

The answer would have been 127 (which is just as far *above* the mean of 112.5 as 98 is *below* it). In this problem, we would have spent the same amount of time calculating the value of the *U* statistic using either Equation 14-1 or Equation 14-4. In other cases, when the number of items is larger in one sample than in the other, choose the equation that will require less work. Regardless of whether you calculate *U* using Equation 14-1 or 14-4, you will come to the same conclusion. Notice that in this example, the answer 127 falls in the acceptance region just as 98 did.

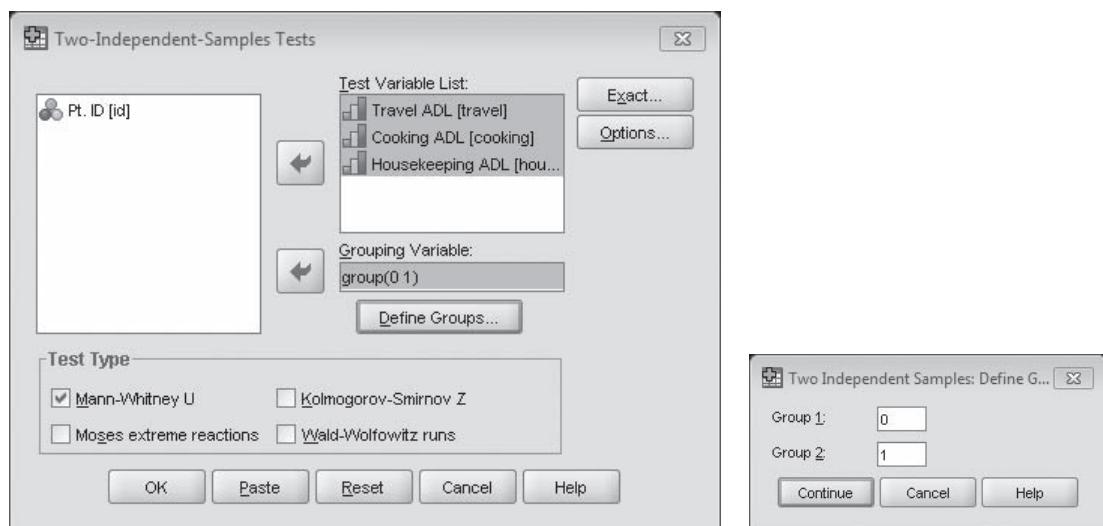
What about *ties* that may happen when we rank the items for this test? For example, what if the two scores ranked 13 and 14 in Table 14-4 both had the value 920? In this case, we would find the average of their ranks $(13 + 14)/2 = 13.5$, and assign the result to both of them. If there were a three-way tie among the scores ranked 13, 14, and 15, we would average these ranks $(13 + 14 + 15)/3 = 14$, and use that value for all three items.

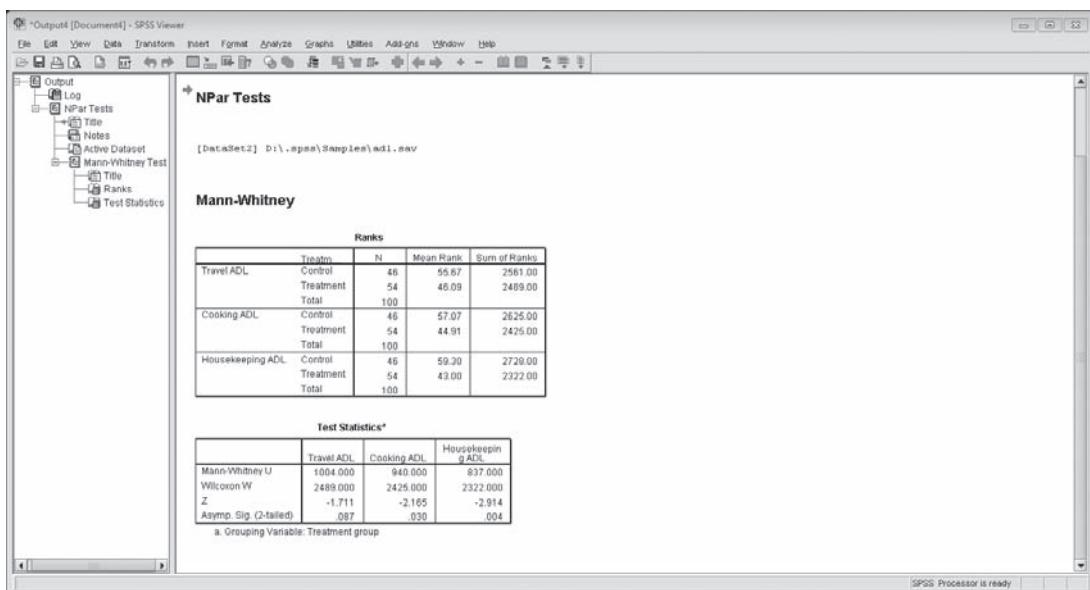
Handling ties in the data

Mann–Whitney U Test Using SPSS

Data of Physicians randomly assigned female stroke patients to receive only physical therapy or physical therapy combined with emotional therapy will be used for it. Three months after the treatments, the Mann-Whitney test is used to compare each group's ability to perform common activities of daily life.

For Mann-Whitney U test go to **Analyze > Non parametric test > Two independent sample test > Define test variables > Define groups > Select Mann Whitney U > OK.**





Solving a Problem Using the Kruskal–Wallis Test

As we noted earlier in this section, the Kruskal–Wallis test is an extension of the Mann–Whitney test to situations where more than two populations are involved. This test, too, depends on the ranks of the sample observations.

In Table 14-6, we have shown the scores of a sample of 20 student pilots on their Federal Aviation Agency written examination, arranged according to which method was used in their training: video cassette, audio cassette, or class room training.

The FAA is interested in evaluating the effectiveness of these three training methods. Specifically, it wants to test at the 0.10 level of significance the hypothesis that the mean written examination scores of student pilots trained by each of these three methods are equal. Because we have more than two populations involved, the Kruskal–Wallis test is appropriate in this instance. To apply the Kruskal–Wallis test to this problem, we begin in Table 14-7 by ranking all the scores in order, from lowest to highest, indicating beside each the symbol of the training method that was used. Ties are handled by averaging ranks, exactly as we did with the Mann–Whitney test.

Testing for differences when more than two populations are involved

Ranking the items to be tested

TABLE 14-6 WRITTEN EXAMINATION SCORES FOR 20 STUDENT PILOTS TRAINED BY THREE DIFFERENT METHODS

Video cassette	74	88	82	93	55	70			
Audio cassette	78	80	65	57	89				
Classroom	68	83	50	91	84	77	94	81	92

TABLE 14-7 WRITTEN EXAMINATION SCORES RANKED FROM LOWEST TO HIGHEST

Rank	Score	Training Method	Rank	Score	Training Method
1	50	C	11	81	C
2	55	VC	12	82	VC
3	57	AC	13	83	C
4	65	AC	14	84	C
5	68	C	15	88	VC
6	70	VC	16	89	AC
7	74	VC	17	91	C
8	77	C	18	92	C
9	78	AC	19	93	VC
10	80	AC	20	94	C

Next, let's learn the symbols used in a Kruskal–Wallis test:

n_j = number of items in sample j

R_j = sum of the ranks of all items in sample j

k = number of samples

$n = n_1 + n_2 + \dots + n_k$, the total number of observations in all samples

Symbols used for a Kruskal–Wallis test

Rearranging data to compute sums of ranks

Computing the k statistic

$$K = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1) \quad [14-5]$$

TABLE 14-8 DATA AND RANK ARRANGED BY TRAINING METHOD

Video Cassette	Rank	Audio Cassette	Rank	Classroom	Rank
74	7	78	9	68	5
88	15	80	10	83	13
82	12	65	4	50	1
93	19	57	3	91	17
55	2	89	16	84	14
70	6			77	8
				94	20
				81	11
				92	18
					$\frac{107}{6} \leftarrow \text{Sum of ranks}$

$$\begin{aligned}
 &= \frac{12}{20(20+1)} \left[\frac{(61)^2}{6} + \frac{(42)^2}{5} + \frac{(107)^2}{9} \right] - 3(20+1) \\
 &= (0.02857) (620.2 + 352.8 + 1,272.1) - 63 \\
 &= 1.143
 \end{aligned}$$

Testing the Hypotheses

The sampling distribution of the K statistic can be approximated by a chi-square distribution *when all the sample sizes are at least 5*. Because our problem meets this condition, we can use the chi-square distribution and Appendix Table 5 for this test. In a Kruskal-Wallis test, the appropriate number of degrees of freedom is $k - 1$, which in this problem is $(3 - 1)$ or 2 because we are dealing with three samples. The hypotheses can be stated as follows:

$H_0: \mu_1 = \mu_2 = \mu_3$ ← Null hypothesis; There are no differences among the three populations, so they have the same mean

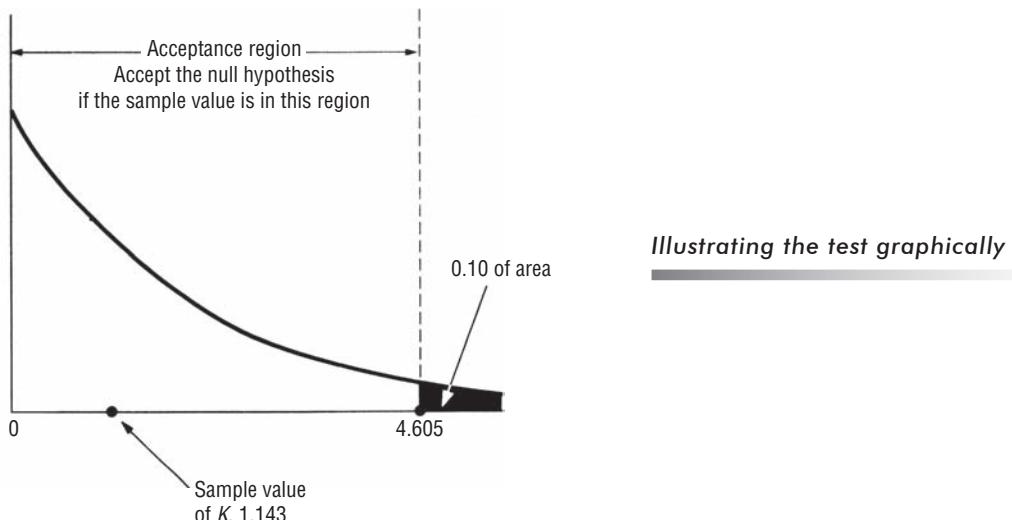
Stating the hypotheses

$H_1: \mu_1, \mu_2$ and μ_3 are not all equal ← Alternative hypothesis: There are differences among the three populations, in particular, they have different means

$\alpha = 0.10$ ← Level of significance for testing these hypotheses

Interpreting the results

Figure 14-5 illustrates a chi-square distribution with 2 degrees of freedom. The colored area represents the 0.10 level of significance. Notice that the acceptance region for the null hypothesis (that there are no differences among the three populations) extends from zero to a chi-square value of 4.605. Obviously, the sample K value of 1.143 is within this acceptance region; therefore, the FAA should accept the null hypothesis and conclude that there are no differences in the results obtained by using the three training methods.



Illustrating the test graphically

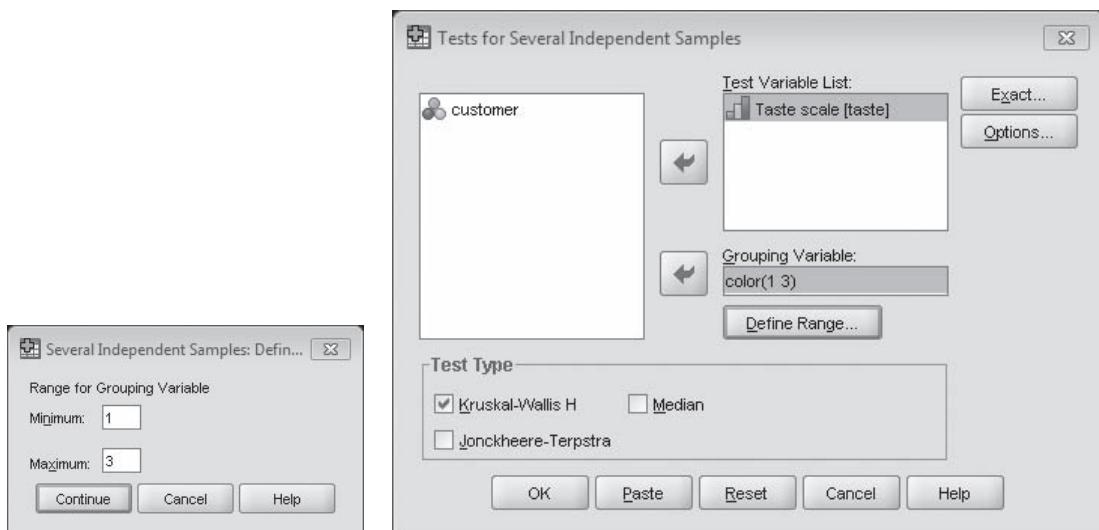
FIGURE 14-5 KRUSKAL-WALLIS TEST AT THE 0.10 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE SAMPLE K STATISTIC

Kruskal-Wallis Test Using SPSS

A screenshot of the SPSS Data Editor window. The title bar says "tasteTest.sav [DataSet5] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. Below the menu is a toolbar with various icons. The main area shows a data grid with 26 rows and 15 columns. The first column is labeled "customer" and contains values from 1 to 26. The second column is labeled "color" and contains values 1, 2, or 3. The third column is labeled "taste" and contains values 1 through 5. The remaining 12 columns are labeled "var" and are mostly empty. A status bar at the bottom right says "SPSS Processor is ready".

Agricultural researchers are studying the effect of mulch color on the taste of crops. Strawberries grown in red, blue, and black mulch were rated by taste-testers on an ordinal scale of one to five (far below to far above average).

For Kruskal-Wallis test go to **Analyze > Non parametric test > k independent samples > Define test variable > Define grouping variable > Select kruskal Wallis H tets > OK.**



The screenshot shows the SPSS interface with the 'Output' tree expanded to show 'NPar Tests'. Under 'NPar Tests', 'Kruskal-Wallis Test' is selected. The results pane displays the following information:

Kruskal-Wallis

	Red	N	Mean Rank
Taste scale	10	9.05	
Blue	10	16.75	
Black	10	20.70	
Total	30		

	Taste scale
Chi-Square	9.751
df	2
Asymp. Sig.	.008

a. Kruskal Wallis Test
b. Grouping Variable: Mulch color

HINTS & ASSUMPTIONS

Rank sum tests such as the Mann-Whitney and the Kruskal-Wallis tests often produce ties in rankings. Hint: When you encounter ties, remember that each tied value gets an *average* rank. If the 10th and 11th items are tied, each of them gets a rank of 10.5. In the case of ties of more than 2 items, they all still get the average rank (a tie in the 3rd, 4th, 5th, and 6th items means all four of them get a rank of $(3 + 4 + 5 + 6)/4 = 4.5$).

EXERCISES 14.3

Self-Check Exercises

SC 14-3 Melisa's Boutique has three mall locations. Melisa keeps a daily record for each location of the number of customers who actually make a purchase. A sample of those data follows. Using the Kruskal-Wallis test, can you say at the 0.05 level of significance that her stores have the same number of customers who buy?

Eastowne Mail	99	64	101	85	79	88	97	95	90	100
Craborchard Mall	83	102	125	61	91	96	94	89	93	75
Fairforest Mall	89	98	56	105	87	90	87	101	76	89

SC 14-4 A large hospital hires most of its nurses from the two major universities in the area. Over the last year, they have been giving a test to the newly graduated nurses entering the hospital to determine which school, if either, seems to educate its nurses better. Based on the following scores (out of 100 possible points), help the personnel office of the hospital

determine whether the schools differ in quality. Use the Mann–Whitney U test with a 10 percent level of significance.

Test Scores												
School A	97	69	73	84	76	92	90	88	84	87	93	
School B	88	99	65	69	97	84	85	89	91	90	87	91 72

Applications

- 14-14** Test the hypothesis of no difference between the ages of male and female employees of a certain company using the Mann–Whitney U test for the sample data. Use the 0.10 level of significance.

Males	31	25	38	33	42	40	44	26	43	35	
Females	44	30	34	47	35	32	35	47	48	34	

- 14-15** The following table shows sample retail prices for three brands of shoes. Use the Kruskal–Wallis test to determine whether there is any difference among the retail prices of the brands throughout the country. Use the 0.01 level of significance.

Brand A	\$89	90	92	81	76	88	85	95	97	86	100
Brand B	\$78	93	81	87	89	71	90	96	82	85	
Brand C	\$80	88	86	85	79	80	84	85	90	92	

- 14-16** A mail-order gift company has the following sample data on dollar sales, separated according to how the order was paid. Test the hypothesis that there is no difference in the dollar amount of orders paid for by cash, by check, or by credit card. Use the Kruskal–Wallis test with a 0.05 level of significance.

Credit-card orders	78	64	75	45	82	69	60
Check orders	110	70	53	51	61	68	
Cash orders	90	68	70	54	74	65	59

- 14-17** The following data show annual hours missed due to illness for the 24 men and women at the Northern Packing Company, Inc. At the 0.10 level of significance, is there any difference attributable to gender? Use the Mann–Whitney U test.

Men	31	44	25	30	70	63	54	42	36	22	25	50
Women	38	34	33	47	58	83	18	36	41	37	24	48

- 14-18** A manufacturer of toys changed the type of plastic molding machines it was using because a new one gave evidence of being more economical. As the Christmas season began, however, productivity seemed somewhat lower than last year. Because production records for the past years were readily available, the production manager decided to compare the monthly output for the 15 months when the old machines were used and the 11 months of production so far this year. Records show these output amounts with the old and new machines.

Monthly Output in Units			
Old Machines		New Machines	
992	966	965	956
945	889	1,054	900
938	972	912	938
1,027	940	850	
892	873	796	
983	1,016	911	
1,014	897	877	
1,258		902	

Can the company conclude at a significance level of 0.10 that the change in machines has reduced output?

- 14-19** Hanks' Hot Dogs has four hot dog stands at Memorial Stadium. Hank knows how many hot dogs are sold at each stand during each football game, and he wants to determine whether the four stands are selling the same number. Using the Kruskal–Wallis test, can you say at the 0.10 significance level that the stands have the same number of hot dog sales?

Game	1	2	3	4	5	6	7	8	9
Visitors north	755	698	725	895	886	794	694	827	814
Visitors south	782	724	754	825	815	826	752	784	789
Home north	714	758	684	816	856	884	774	812	734
Home south	776	824	654	779	898	687	716	889	917

- 14-20** To increase sales during heavy shopping days, a chain of stores selling cheese in shopping malls gives away samples at the stores' entrances. The chain's management defines the heavy shopping days and randomly selects the days for sampling. From a sample of days that were considered heavy shopping days, the following data give one store's sales on days when cheese sampling was done and on days when it was not done.

Sales (in hundreds)												
Promotion days	18	21	23	15	19	26	17	18	22	20	18	21
Regular days	22	17	15	23	25	20	26	24	16	17	23	21

Use the Mann–Whitney U test and a 5 percent level of significance to decide whether the storefront sampling produced greater Sales

- 14-21** A company is interested in knowing whether there is a difference in the output rate for men and women employees in the molding department. Judy Johnson, production manager, was asked to conduct a study in which male and female workers' output was measured for 1 week. Somehow, one of the office clerks misplaced a portion of the data, and Judy was able to locate only the following information from the records of the tests:

$$\sigma_u = 176.4275$$

$$\mu_u = 1,624$$

$$R_1 = 3,255$$

Judy also remembered that the sample size for men, n_2 , had been two units larger than n_1 .

Reconstruct a z value for the test and determine whether the weekly output can be assumed, at a 5 percent level of significance, to be the same for both men and women. Indicate also the values for n_1, n_2 , and R_2 .

- 14-22** A university that accepts students from both rural and urban high schools is interested in whether the different backgrounds lead to a difference in first-year GPA. Data are presented below for 13 randomly selected first-year students of rural background and 16 students of urban background. Use the Mann–Whitney U test with a 5 percent level of significance.

GPA										
Rural	3.19	2.05	2.82	2.16	3.84	4.0	2.91	2.75	3.01	1.98
	2.58	2.76	2.94							
Urban	3.45	3.16	2.84	2.09	2.11	3.08	3.97	3.85	3.72	2.73
	2.81	2.64	1.57	1.87	2.54	2.62				

- 14-23** Twenty salespeople of Henley Paper Company have received sales training during the past year. Some were sent to a national program conducted by Salesmasters. The other received training at the company office conducted by the Henley sales manager. Percentages of selling quotas realized by both groups during last year are shown. Mr. Boyden Henley, president, believes that the backgrounds, sales aptitudes, and motivation of both groups are comparable. At the 0.10 level of significance, has either method of training been better? Use the Mann–Whitney U test.

Percentage of Quota Realized										
Salesmasters	90	95	105	110	100	75	80	90	105	120
Company	80	90	100	120	95	95	90	100	95	105

Worked-Out Answers to Self-Check Exercises

SC 14-3	Eastowne ranks	99	64	101	85	79	88	97	95	90	100
		24	3	26.5	8	6	11	22	20	15.5	25
	Craborchard ranks	83	102	125	61	91	96	94	89	93	75
		7	28	30	2	17	21	19	13	18	4
	Fairforest ranks	89	98	56	105	87	90	87	101	76	89
		13	23	1	29	9.5	15.5	9.5	26.5	5	13

$$n_1 = 10 \quad n_2 = 10 \quad n_3 = 10 \quad \alpha = 0.05$$

$$R_1 = 161 \quad R_2 = 159 \quad R_3 = 145$$

$H_0: \mu_1 = \mu_2 = \mu_3$ $H_1:$ the μ 's are not all the same

$$k = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1)$$

$$\frac{1R_j^2}{nn_j} = \frac{12}{30(31)} \left(\frac{(161)^2}{10} + \frac{(159)^2}{10} + \frac{(145)^2}{10} \right) - 3(31) = 0.196$$

With $3 - 1 = 2$ degrees of freedom and $\alpha = 0.05$, the upper limit of the acceptance region is $x^2 = 5.991$, so we accept H_0 . The average numbers of buyers at the three stores are not significantly different.

SC 14-4	School A	97	69	73	84	76	92	90	88	84	87	93
	ranks	22.5	2.5	5	8	6	20	16.5	13.5	8	11.5	21
	School B	88	99	65	69	97	84	85	89	91	90	87
	ranks	13.5	24	1	2.5	22.5	8	10	15	18.5	16.5	11.5
												72

$$n_1 = 11 \quad n_2 = 13 \quad \alpha = 0.10$$

$$R_1 = 134.5 \quad R_2 = 165.5$$

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = 11(13) + \frac{11(12)}{2} - 134.5 = 74.5$$

$$\mu_u = \frac{n_1 n_2}{2} = \frac{11(13)}{2} = 71.5$$

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{11(13)(25)}{12}} = 17.26$$

The critical values of z are ± 1.645 . The standardized value of U is

$$z = \frac{U - \mu_u}{\sigma_u} = \frac{74.5 - 71.5}{17.26} = 0.174$$

Because the standardized value of U lies within the critical values, we accept H_0 . There is no significant difference between the schools.

14.4 THE ONE-SAMPLE RUNS TEST

So far, we have assumed that the samples in our problems were randomly selected—that is, chosen without preference or bias. What if you were to notice recurrent patterns in a sample chosen by someone else? Suppose that applicants for advanced job training were to be selected without regard to gender from a large population. Using the notation W = woman and M = man, you find that the first group enters in this order:

W, W, W, W, M, M, M, M, W, W, W, W, M, M, M, M

Concept of randomness

By inspection, you would conclude that although the total number of applicants is equally divided between the sexes, the order is not random. A random process would rarely list two items in alternating groups of four. Suppose now that the applicants begin to arrive in this order:

W, M, W, M, W, M, W, M, W, M, W, M, W, M

It is just as unreasonable to think that a random selection process would produce such an orderly pattern of men and women. In this case, too, the *proportion* of women to men is right, but you would be suspicious about the *order* in which they are arriving.

To allow us to test samples for the randomness of their order, statisticians have developed the *theory of runs*. A **run** is a sequence of identical occurrences preceded and followed by different occurrences or by none at all. If men and women enter as follows, the sequence will contain three runs:

$$\underbrace{W,}_{1st} \underbrace{M, M, M, M,}_{2nd} \underbrace{W}_{3rd}$$

And this sequence contains six runs:

$$\underbrace{W, W, W}_{1st} \underbrace{M, M}_{2nd} \underbrace{W}_{3rd} \underbrace{M, M, M, M}_{4th} \underbrace{W, W, W, W}_{5th} \underbrace{M}_{6th}$$

A *test of runs* would use the following symbols if it contained just two kinds of occurrences:

n_1 = number of occurrences of type 1

Symbols used for a runs test

n_2 = number of occurrences of type 2

r = number of runs

Let's apply these symbols to a different pattern for the arrival of applicants:

M, W, W, M, M, M, W, W, W, M, M, W, M, W, W, M

In this case, the values of n_1 , n_2 , and r would be

$n_1 = 8 \leftarrow$ Number of women

$n_2 = 9 \leftarrow$ Number of men

$r = 9 \leftarrow$ Number of runs

A Problem Illustrating a One-Sample Runs Test

A manufacturer of breakfast cereal uses a machine to insert randomly one of two types of toys in each box. The company wants randomness so that every child in the neighborhood does not get the same toy. Testers choose samples of 60 successive boxes to see whether the machine is properly mixing the two types of toys. Using the symbols *A* and *B* to represent the two types of toys, a tester reported that one such batch looked like this:

B, A, B, B, B, A, A, A, B, B, A, B, B, B, A, A, A, A, B,
 A, B, A, A, B, B, B, A, A, B, A, A, A, B, B, A, B, B, A,
 A, A, A, B, B, A, B, B, B, A, A, B, B, A, B, A, A, B, B

The values in our test will be

$n_1 = 29 \leftarrow$ Number of boxes containing toy A

$n_2 = 31 \leftarrow$ Number of boxes containing toy B

$r = 29 \leftarrow$ Number of runs

The Sampling Distribution of the r Statistic

The *number of runs*, or r , is a statistic with its own special sampling distribution and its own test. Obviously, runs may be of differing lengths, and various numbers of runs can occur in one sample. Statisticians can prove that too many or too few runs in a sample indicate that something other than chance was at work when the items were selected. A **one-sample runs test, then, is based on the idea that too few or too many runs show that the items were not chosen randomly.**

The r statistic, the basis of a one-sample runs test

To derive the mean of the sampling distribution of the r statistic, use the following formula:

Mean of the Sampling Distribution of the r Statistic

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad [14-6]$$

Mean and standard error of the r statistic

Applying this to the cereal company, the mean of the r statistic would be

$$\begin{aligned}\mu_r &= \frac{(2)(29)(31)}{29+31} + 1 \\ &= \frac{1,798}{60} + 1 \\ &= 29.97 + 1 \\ &= 30.97 \leftarrow \text{Mean of the } r \text{ statistic}\end{aligned}$$

The standard error of the r statistic can be calculated with this formidable-looking formula:

Standard Error of the r Statistic

$$\sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \quad [14-7]$$

For our problem, the standard error of the r statistic becomes

$$\begin{aligned}\sigma_r &= \sqrt{\frac{(2)(29)(31)(2 \times 29 \times 31 - 29 - 31)}{(29+31)^2(29+31-1)}} \\ &= \sqrt{\frac{(1,798)(1,738)}{(60)^2(59)}} \\ &= \sqrt{14.71} \\ &= 3.84 \leftarrow \text{Standard error of the } r \text{ statistic}\end{aligned}$$

Testing the Hypotheses

In the one-sample runs test, the sampling distribution of r can be closely approximated by the normal distribution if either n_1 or n_2 is larger than 20. Our cereal company has a sample of 60 boxes, so we can use the normal approximation. Management is interested in testing at the 0.20 level the hypothesis that the toys are randomly mixed, so the test becomes

$$\begin{aligned} H_0: & \left\{ \begin{array}{l} \text{In a one-sample runs} \\ \text{test, a symbolic statement} \\ \text{of the hypotheses is} \\ \text{not appropriate} \end{array} \right. & \leftarrow \text{Null hypothesis: The toys are} \\ & \text{randomly mixed} \\ H_1: & \left. \begin{array}{l} \text{not appropriate} \end{array} \right. & \leftarrow \text{Alternative hypothesis: The} \\ & \text{toys are not randomly mixed} \end{aligned}$$

$$\alpha = 0.20 \leftarrow \text{Level of significance for} \\ \text{testing these hypotheses}$$

Because too many *or* too few runs would indicate that the process by which the toys are inserted into the boxes is not random, a two-tailed test is appropriate. Figure 14-6 illustrates this test graphically.

Next we use Equation 6-2 to *standardize* the sample r statistic, 29, by subtracting μ_r , its mean, and dividing by σ_r , its standard error.

$$\begin{aligned} z &= \frac{r - \mu_r}{\sigma_r} & [6-2] \\ &= \frac{29 - 30.97}{3.84} \\ &= -0.513 \end{aligned}$$

Stating the hypotheses

Illustrating the test graphically

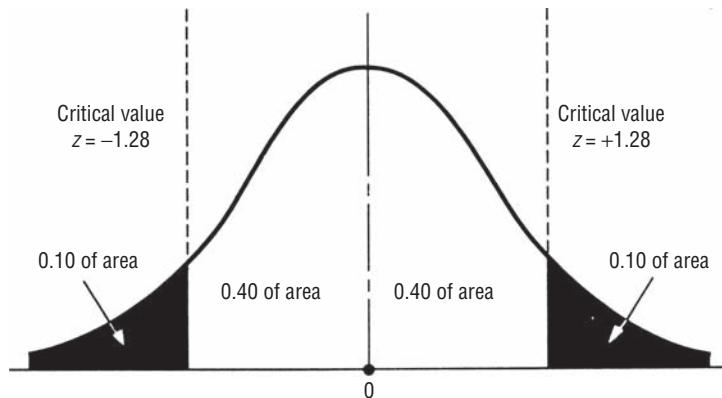


FIGURE 14-6 TWO-TAILED HYPOTHESIS TEST AT THE 0.20 LEVEL OF SIGNIFICANCE

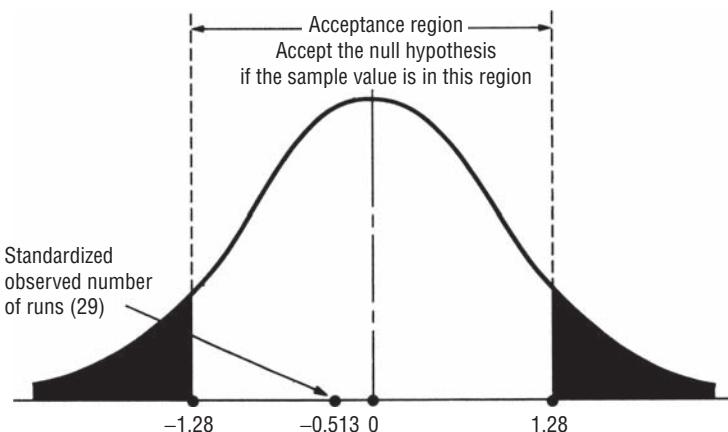


FIGURE 14-7 TWO-TAILED HYPOTHESIS TEST AT THE 0.20 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED OBSERVED NUMBER OF RUNS

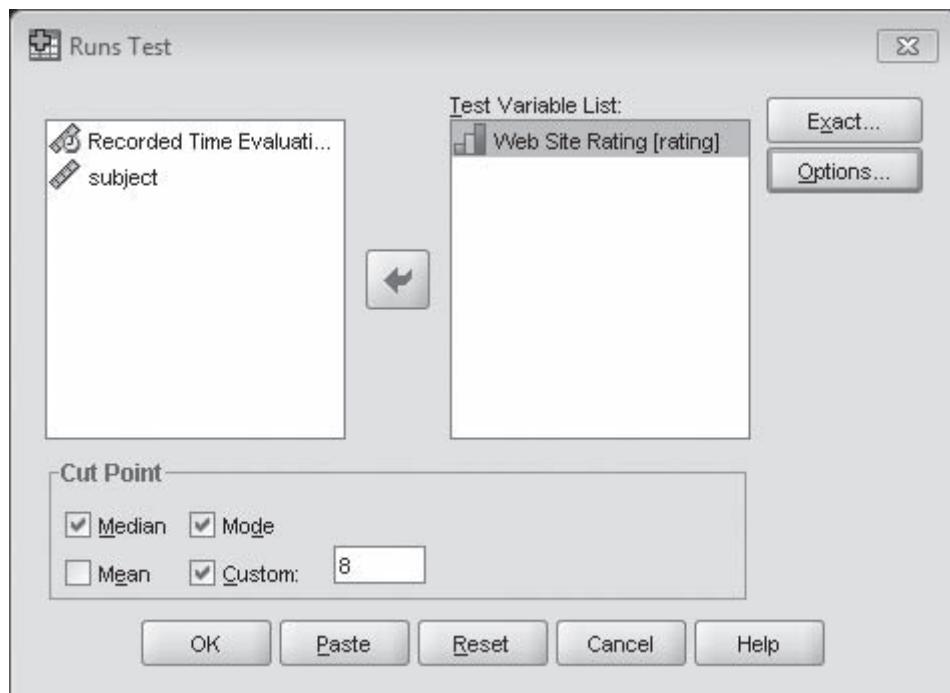
Placing the standardized value on the z scale in Figure 14-7 shows that it falls well within the critical values for this test. Therefore, management should accept the null hypothesis and conclude from this test that toys are being inserted in boxes in random order.

One Sample Run Test Using SPSS

sitteratings.sav [DataSet1] - SPSS Data Editor																																		
	File	Edit	View	Data	Transform	Analyze	Graphs	Utilities	Add-ons	Window	Help																							
15	sitetime	rating	subject	gender	pnoruse	var1	var2	var3	var4	var5	var6	var7	var8	var9	var10	var11	var12	var13	var14	var15	var16	var17	var18	var19	var20	var21	var22	var23	var24	var25				
1	0:08:41:00	8	12 M	Y																														
2	0:09:30:00	7	17 M	Y																														
3	0:10:20:00	8	18 F	Y																														
4	0:10:47:00	6	21 M	Y																														
5	0:11:12:00	10	16 F	Y																														
6	0:12:34:00	8	28 F	N																														
7	0:13:31:00	6	15 F	N																														
8	0:13:50:00	7	11 M	N																														
9	0:15:12:00	8	10 F	N																														
10	0:17:37:00	9	23 F	Y																														
11	0:18:00:00	7	2 M	Y																														
12	0:18:26:00	10	19 F	N																														
13	0:20:04:00	7	31 M	Y																														
14	0:20:41:00	8	25 F	Y																														
15	0:21:18:00	12	24 M	Y																														
16	0:21:51:00	10	26 M	N																														
17	0:23:19:00	12	7 F	Y																														
18	0:25:32:00	9	8 F	Y																														
19	0:25:33:00	11	20 M	N																														
20	0:25:46:00	12	1 M	Y																														
21	0:25:46:00	10	4 F	N																														
22	0:26:30:00	13	13 M	N																														
23	0:27:20:00	13	14 F	Y																														
24	0:28:47:00	12	27 F	N																														
25	0:30:17:00	11	3 M	N																														

For one sample run test we take example of an e-commerce firm enlisted beta tester to browse and then rate their new Web site. Ratings were recorded as soon as each tester finished browsing. The team is concerned that ratings may be related to the amount of time spent browsing.

For one sample run ttest go to **Analyze > Nonparametric tests > Run > Define test variables > Define different cut points > OK.**



The SPSS Viewer window displays the output for the 'Runs Test'. The left pane shows the 'Output' tree with 'NPar Tests' expanded, containing 'Title', 'Notes', 'Active Dataset', 'Descriptive Statistics', 'Runs Test', 'Runs Test 2', and 'Runs Test 3'. The right pane contains two tables: 'Descriptive Statistics' and 'Runs Test'.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles
Web Site Rating	32	11.94	2.365	6	14	25th: 8.00 50th (Median): 10.00 75th: 12.00

Runs Test

Test Value*	Web Site Rating
Test Value*	10
Cases <= Test Value	14
Cases >= Test Value	16
Total Cases	32
Number of Runs	10
Z	-2.263
Asymp. Sig. (2-tailed)	.027

a. Median

Runs Test 2

Test Value*	Web Site Rating
Test Value*	12*
Cases <= Test Value	22
Cases >= Test Value	10
Total Cases	32
Number of Runs	16
Z	.315
Asymp. Sig. (2-tailed)	.752

a. Mode
b. There are multiple modes. The mode with the largest data value is used.

HINTS & ASSUMPTIONS

Runs tests can be used effectively in quality control situations. You will recall from Chapter 10 that variation in quality is either systematic or random, and if it's systematic variation we can correct it. Thus, a runs test can detect the kinds of patterns in output quality that are associated with systematic variation. Hint: Almost all runs tests are two-tailed because the question to be answered is whether there are too many or too few runs. Remember also that runs tests use the r statistic whose distribution can be well described by a normal distribution as long as *either* n_1 or n_2 is larger than 20.

EXERCISES 14.4**Self-Check Exercise**

SC 14-5 Professor Ike Newton is interested in determining whether his brightest students (those making the best grades) tend to turn in their tests earlier (because they can recall the material faster) or later (because they take longer to write down all they know) than the others in the class. For a particular physics test, he observes that the students make the following grades in order of turning their tests in:

Order	Grades									
1–10	94	70	85	89	92	98	63	88	74	85
11–20	69	90	57	86	79	72	80	93	66	74
21–30	50	55	47	59	68	63	89	51	90	88

- (a) If Professor Newton counts those making a grade of 90 and above as his brightest students, then at a 5 percent level of significance, can he conclude the brightest students turned their tests in randomly?
- (b) If 60 and above is passing in Professor Newton's class, then did the students passing versus those not passing turn their tests in randomly? (Also use the 5 percent significance level.)

Basic Concepts

14-24 Test for the randomness of the following sample using the 0.05 significance level:

A, B, A, A, A, B, B, A, B, B, A, A, B, A, B, B, B, B, A, B, B,
A, A, A, B, A, B, A, B, B, A, A, A, B, B, A, A, B, A, A, A, A

Applications

14-25 A sequence of small glass sculptures was inspected for shipping damage. The sequence of acceptable and damaged pieces was as follows:

D, A, A, A, D, D, D, D, A, A, D, D, A, A, A, A, D, A, A, D, D, D, D, D

Test for the randomness of the damage to the shipment using the 0.05 significance level.

- 14-26** The *News and Clarion* kept a record of the gender of people who called the circulation office to complain about delivery problems with the Sunday paper. For a recent Sunday, these data were as follows:

M, F, F, F, M, M, F, M, F, F, F, M, M, M, F, M, F, M, F, F, F, M, M, M, M, M

Using the 0.05 level of significance, test this sequence for randomness. Is there anything about the nature of this problem that would cause you to believe that such a sequence would not be random?

- 14-27** Kerwin County Social Services Agency kept this record of the daily number of applicants for marriage counseling in the order in which they appeared at the agency office in 30 working days.

3, 4, 6, 8, 4, 6, 7, 2, 5, 7, 4, 8, 4, 7, 9, 5, 9, 10,
5, 7, 4, 9, 8, 9, 11, 6, 7, 5, 9, 12

Test the randomness of this sequence by seeing whether the values above and below the mean occur in random order. Use the 0.10 level of significance. Can you think of any characteristic of the environment of this problem that would support the statistical finding you reached?

- 14-28** A restaurant owner has noticed over the years that older couples appear to eat earlier than young couples at his quiet, romantic restaurant. He suspects that perhaps it is because of children having to be left with babysitters and also because the older couples may retire earlier at night. One night, he decided to keep a record of couples' arrivals at the restaurant. He noted whether each couple was over or under 30. His notes are reproduced below. (A = 30 and older; B = younger than 30.)

(5:30 P.M.) A, A, A, A, A, B, A, A, A, A, A, B, B,
B, A, B, B, B, B, A, B, B, A, B, B, (10 P.M.)

At a 5 percent level of significance, was the restaurant owner correct in his thought that the age of his customers at different dining hours is less than random?

- 14-29** Kathy Phillips is in charge of production scheduling for a printing company. The company has six large presses, which frequently break down, and one of Kathy's biggest problems is meeting deadlines when there are unexpected breakdowns in presses. She suspects that the older presses break down earlier in the week than the new presses, because all presses are checked and repaired over the weekend. To test her hypothesis, Kathy recorded the number of all the presses as they broke down during the week. Presses numbered 1, 2, and 3 are the older ones.

Number of Press in Order of Breakdown
1, 2, 3, 1, 4, 5, 3, 1, 2, 5, 1, 3, 6, 2, 3, 6, 2, 2, 3, 5, 4,
6, 4, 2, 1, 3, 4, 5, 5, 1, 4, 5, 2, 3, 5, 6, 4, 3, 2, 5, 4, 3

- At a 5 percent level of significance, does Kathy have a valid hypothesis that the breakdowns of presses are not random?
- Is her hypothesis appropriate for the decision she wishes to make about rescheduling more work earlier in the week on the newer presses?

- 14-30** Martha Bowen, a department manager working in a large marketing-research firm, is in charge of all the research data analyses done in the firm. Accuracy and thoroughness are her responsibility. The department employs a number of research assistants to do some analyses and uses a computer to do other analyses. Typically, each week Martha randomly chooses completed analyses before they are reported and conducts tests to ensure that they have been done correctly and thoroughly. Martha's assistant, Kim Tadlock, randomly chooses 49 analyses per week from those completed and filed each day, and Martha does the reanalyses. Martha wanted to make certain that the selection process was a random one, so she could provide assurances that the computer analyses and those done by hand were both periodically checked. She arranged to have the research assistants place a special mark on the back of the records, so that they could be identified. Kim was unaware of the mark, so the randomness of the test would not be affected. Kim completed her sample with the following data:

Samples of Data Analyses for 1 Week
(1: by computer, 2: by hand)

1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1,
1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

- (a) At a 1 percent significance level, can you conclude that the sample was random?
 - (b) If the sample were distributed as follows, would the sample be random?
- | |
|---|
| 1, |
| 1, 2, 2, 2, 2 |
- (c) Because computer analyses are much faster than those done by hand, and because a number of the analyses can be done by computer, there are about three times as many computer analyses per week as hand analyses. Is there statistical evidence in part (a) to support the belief that somewhere in the sampling process there is something less than randomness occurring? If so, what is the evidence?
 - (d) Does the conclusion you reached in part (c) lead you to any new conclusions about the one-sample runs test, particularly in reference to your answer in part (a)?

- 14-31** Bank of America is curious about the grade level of people who use their ATM at the Student Union. Freshmen and sophomores are classified as type A, juniors and seniors as type B. Data are presented below for 45 people who used the ATM during one Friday afternoon. Test this sequence for randomness at the 0.05 significance level.

BBBAAABAAAAAABBBBABAAAABBAABBBABBBAAAAAABB

- 14-32** The First National Bank of Smithville recorded the gender of the first 40 customers who appeared last Tuesday with this notation:

M, F, M, M, M, F, F, M, M, M, F, M, M, M, M, M, F, F, M,
F, M, M, M, F, M, M, M, M, F, M, M, M, M, F, F, M

At the 0.05 level of significance, test the randomness of this sequence. Is there anything in banking or in the nature of this problem that would lead you to accept intuitively what you have found statistically?

Worked-Out Answer to Self-Check Exercise

SC 14-5 (a) Let G denote those at or above 90, and L denote those below 90:

GLLLGGGLLLLLGLLLLLGLLLLLLGL

$$\begin{aligned} n_1 &= \# \text{ of } G\text{'s} = 6 & r &= 10 \\ n_2 &= \# \text{ of } L\text{'s} = 24 & \alpha &= 0.05 \end{aligned}$$

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(6)(24)}{30} = 10.6$$

$$\begin{aligned} \sigma_r &= \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(6)(24)[2(6)(24) - 6 - 24]}{(30)^2 (29)}} \\ &= 1.69 \end{aligned}$$

The critical values of z are ± 1.96 . The standardized value of r is

$$z = \frac{r - \mu}{\sigma_r} = \frac{10 - 10.6}{1.69} = -0.355$$

so we accept H_0 . The sequence is random.

(b) With P denoting passing (≥ 60) and F denoting failing (< 60), we get

PPPPPPPPPPPPFP PPPPPPFFFFF PPPPFPP

$$\begin{aligned} n_1 &= \# \text{ of } P\text{'s} = 24 & r &= 7 \\ n_2 &= \# \text{ of } F\text{'s} = 6 & \alpha &= 0.05 \end{aligned}$$

$$\mu_r = \frac{2(24)(6)}{30} + 1 = 10.6$$

$$\sigma_r = \sqrt{\frac{2(24)(6)[2(24)(6) - 24 - 6]}{(30)^2 (29)}} = 1.69$$

The critical values of z are ± 1.96 . The standardized value of r is

$$z = \frac{7 - 10.6}{1.69} = -2.13$$

so we reject H_0 because $z < -1.96$. This sequence is not random.

14.5 RANK CORRELATION

Chapters 12 and 13 introduced us to the notion of correlation and to the correlation coefficient, a measure of the closeness of association between two variables. Often in correlation analysis,

Function of the rank-correlation coefficient

information is not available in the form of numerical values such as those we used in the problems of those chapters. But if we can assign rankings to the items in each of the two variables we are studying, a *rank-correlation coefficient* can be calculated. **This is a measure of the correlation that exists between the two sets of ranks, a measure of the degree of association between the variables that we would not have been able to calculate otherwise.**

A second reason for learning the method of rank correlation is to be able to simplify the process of computing a correlation coefficient from a very large set of data for each of two variables. To prove how tedious this can be, try expanding one of the correlation problems in Chapter 12 by a factor of 10 and performing the necessary calculations. Instead of having to do these calculations, we can compute a measure of association that is based on the *ranks* of the observations, *not the numerical values* of the data. This measure is called the Spearman rank-correlation coefficient, in honor of the statistician who developed it in the early 1900s.

Another advantage of using rank correlation

The Coefficient of Rank Correlation

By working a couple of examples, we can learn how to calculate and interpret this measure of the association between two ranked variables. First, consider Table 14-9, which lists five people and compares the academic rank they achieved in college with the level they have attained in a certain company 10 years after graduation. The value of 5 represents the highest rank in the group; the rank of 1, the lowest.

Listing the ranked variables

Using the information in Table 14-9, we can calculate a coefficient of rank correlation between success in college and company level achieved 10 years later. All we need is Equation 14-8 and a few computations.

Calculating the rank-correlation coefficient

Coefficient of Rank Correlation

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad [14-8]$$

where

- r_s = coefficient of rank correlation (notice that the subscript s , from Spearman, distinguishes this r from the one we calculated in Chapter 12)
- n = number of paired observations
- Σ = notation meaning “the sum of”
- d = difference between the ranks for each pair of observation

The computations are easily done in tabular form, as we show in Table 14-10. Therefore, we have all the information we need to find the rank-correlation

TABLE 14-9 COMPARISON OF THE RANKS OF FIVE STUDENTS

Student	College Rank	Company Rank 10 Years Later
John	4	4
Margaret	3	3
Debbie	1	1
Steve	2	2
Lisa	5	5

TABLE 14-10 GENERATING INFORMATION TO COMPUTE THE RANK-CORRELATION COEFFICIENT

Student	College Rank (1)	Company Rank (2)	Difference Between the Two Ranks (1) – (2)	Difference Squared [(1) – (2)] ²
John	4	4	0	0
Margaret	3	3	0	0
Debbie	1	1	0	0
Steve	2	2	0	0
Lisa	5	5	0	0
				$\sum d^2 = 0 \leftarrow \text{Sum of the Squared differences}$

coefficient for this problem:

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} & [14-8] \\
 &= 1 - \frac{6(0)}{5(25 - 1)} \\
 &= 1 - \frac{0}{120} \\
 &= 1 \leftarrow \text{Rank-correlation coefficient}
 \end{aligned}$$

As we learned in Chapter 12, this correlation coefficient of 1 shows that there is a perfect association or *perfect correlation* between the two variables. This verifies what we saw in Table 14-9, the fact that the college and company ranks for each person were identical.

One more example should make us feel comfortable with the coefficient of rank correlation. Table 14-11 illustrates five more

Explaining values of the rank-correlation coefficient

Computing another rank-correlation coefficient

TABLE 14-11 GENERATING DATA TO COMPUTE THE RANK-CORRELATION COEFFICIENT

Student	College Rank (1)	Company Rank (2)	Difference Between the Two Ranks (1) – (2)	Difference Squared [(1) – (2)] ²
Roy	5	1	4	16
David	1	5	-4	16
Jay	3	3	0	0
Charlottee	2	4	-2	4
Kathy	4	2	2	4
				$\sum d^2 = 40 \leftarrow \text{Sum of the Squared differences}$

people, but this time the ranks in college and in a company 10 years later seem to be extreme opposites. We can compute the difference between the ranks for each pair of observations, find d^2 , and then take the sum of all the d^2 s. Substituting these values into Equation 14-8, we find a rank correlation coefficient of -1 :

$$\begin{aligned} r_s &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} & [14-8] \\ &= 1 - \frac{6(40)}{5(25-1)} \\ &= 1 - \frac{240}{120} \\ &= 1 - 2 \\ &= -1 \leftarrow \text{Rank-correlation coefficient} \end{aligned}$$

In Chapter 12, we learned that a correlation coefficient of -1 represents *perfect inverse correlation*. And that is just what happened in our case: The people who did the best in college wound up 10 years later in the lowest ranks of an organization. Now let's apply these ideas.

Interpreting the results

Solving a Problem Using Rank Correlation

Rank correlation is a useful technique for looking at the connection between air quality and the evidence of pulmonary-related diseases that we discussed in our chapter-opening problem. Table 14-12 reproduces the data found by the health organization studying the problem. In the same table, we also do some of the calculations needed to find r_s .

Using the data in Table 14-12 and Equation 14-8, we can find the rank-correlation coefficient for this problem:

Finding the rank-correlation coefficient

$$\begin{aligned} r_s &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} & [14-8] \\ &= 1 - \frac{6(58)}{11(121-1)} \\ &= 1 - \frac{348}{1,320} \\ &= 1 - 0.2636 \\ &= 0.7364 \leftarrow \text{Rank-correlation coefficient} \end{aligned}$$

A correlation coefficient of 0.736 suggests a substantial positive association between average air quality and the occurrence of pulmonary disease, at least in the 11 cities sampled; that is, high levels of pollution go with high incidence of pulmonary disease.

Interpreting the results

How can we test this value of 0.736? We can apply the same methods we used to test hypotheses in Chapter 8 and 9. In performing such tests on r_s , We are trying to avoid the error of concluding that

TABLE 14-12 RANKING OF ELEVEN CITIES

City	Air Quality Rank (1)	Pulmonary-Disease Rank (2)	Difference between the Two Ranks (1) – (2)	Difference Squared $[(1) - (2)]^2$
A	4	5	-1	1
B	7	4	3	9
C	9	7	2	4
D	1	3	-2	4
E	2	1	1	1
F	10	11	-1	1
G	3	2	1	1
H	5	10	-5	25
I	6	8	-2	4
J	8	6	2	4
K	11	9	2	4
Best rank = 11			$\sum d^2 = 58 \leftarrow$ Sum of the squared differences	
Worst rank = 1				

an association exists between two variables if, in fact, no such association exists in the population from which these two samples were drawn, that is, if the *population* rank-correlation coefficient, ρ_s (*rho sub s*), is really equal to zero.

For small values of n , (n less than or equal to 30), the distribution of r_s is not normal, and unlike other small sample statistics we have encountered, it is not appropriate to use the *t*distribution for testing hypotheses about the rank-correlation coefficient. Instead, we use Appendix Table 7, Spearman's Rank Correlation Values, to determine the acceptance and rejection regions for such hypotheses. In our current problem, suppose that the health organization wants to test, at the 0.05 level of significance, the null hypothesis that there is zero correlation in the ranked data of *all* cities in the world. Our problem then becomes:

$$H_0: \rho_s = 0 \leftarrow \text{Null hypothesis: There is no correlation in the ranked data of the population}$$

Testing hypotheses about rank correlation

$$H_1: \rho_s \neq 0 \leftarrow \text{Alternative hypothesis: There is a correlation in the ranked data of the populations}$$

Stating the hypotheses

$$\alpha = 0.05 \leftarrow \text{Level of significance for testing these hypotheses}$$

A two-tailed test is appropriate, so we look at Appendix Table 7 in the row for $n = 11$ (the number of cities) and the column for a significance level of 0.05. There we find that the critical values for r_s are ± 0.6091 , that is, the upper limit of the acceptance region is 0.6091, and the lower limit of the acceptance region is -0.6091.

Figure 14-8 shows the limits of the acceptance region and the rank-correlation coefficient we calculated from the air-quality sample. From this figure, we can see that the rank-correlation coefficient lies outside the acceptance region. Therefore, we would reject the null hypothesis of no correlation and conclude that there is an association between air-quality levels and the incidence of pulmonary disease in the world's cities.

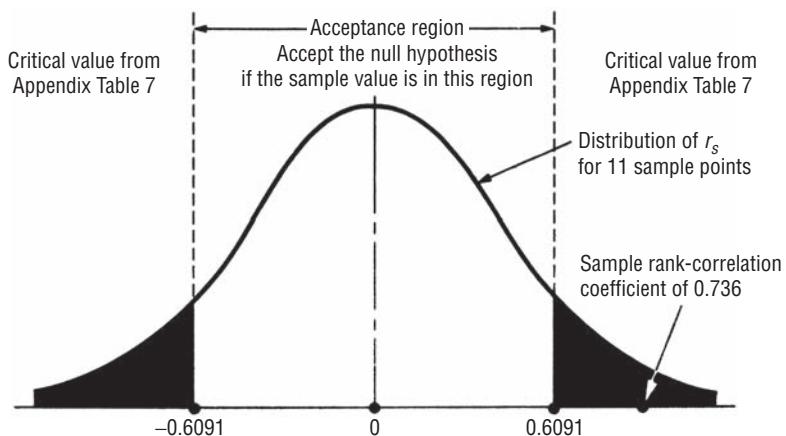


FIGURE 14-8 TWO-TAILED HYPOTHESIS TEST, USING APPENDIX TABLE 7 AT THE 0.05 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND SAMPLE RANK-CORRELATION COEFFICIENT.

If the sample size is greater than 30, we can no longer use Appendix Table 7. However, when n is greater than 30, the sampling distribution of r_s is approximately normal, with a mean of zero and a standard deviation of $1/\sqrt{n-1}$. Thus, the standard error of r_s is,

The appropriate distribution 'for values of n greater than 30

Standard Error of the Coefficient of Rank Correlation

$$\sigma_{r_s} = \frac{1}{\sqrt{n-1}} \quad [14-9]$$

and we can use Appendix Table 1 to find the appropriate z values for testing hypotheses about the population rank correlation.

Example with n greater than 30

As an example of hypothesis testing of rank-correlation coefficients when n is greater than 30, consider the case of a social scientist who tries to determine whether bright people tend to choose spouses who are also bright. He randomly chooses 32 couples and tests to see whether there is a significant rank correlation in the IQs of the couples. His data and computations are given in Table 14-13.

Using the data in Table 14-13 and Equation 14-8, we can find the rank-correlation coefficient for this problem:

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \\
 &= 1 - \frac{6(1,043.5)}{32(1,024 - 1)}
 \end{aligned} \quad [14-8]$$

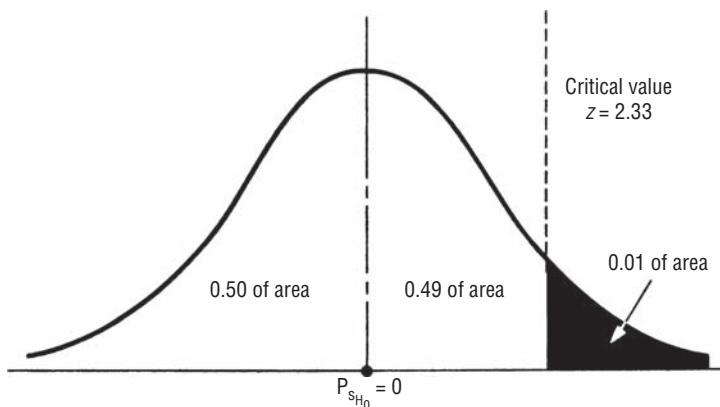
$$\begin{aligned}
 &= 1 - \frac{6,261}{32,736} \\
 &= 1 - 0.1913 \\
 &= 0.8087 \leftarrow \text{Rank-correlation coefficient}
 \end{aligned}$$

If the social scientist wishes to test his hypothesis at the 0.01 level of significance, his problem can be stated:

TABLE 14-13 COMPUTATION OF RANK CORRELATION OF HUSBANDS' AND WIVES' IQS

Couple (1)	Husband's IQ (2)	Wife's IQ (3)	Husband's Rank (4)	Wife's Rank (5)	Difference between Ranks (4) - (5)	Difference Squared $[(4) - (5)]^2$
1	95	95	8	4.5	3.5	12.25
2	103	98	20	8.5	11.5	132.25
3	111	110	26	23	3	9.00
4	92	88	4	2	2	4.00
5	150	106	32	18	14	196.00
6	107	109	24	21.5	2.5	6.25
7	90	96	3	6	-3	9.00
8	108	131	25	32	-7	49.00
9	100	112	17.5	25.5	-8	64.00
10	93	95	5.5	4.5	1	1.00
11	119	112	29	25.5	3.5	12.25
12	115	117	28	30	-2	4.00
13	87	94	1	3	-2	4.00
14	105	109	21	21.5	-0.5	0.25
15	135	114	31	27	4	16.00
16	89	83	2	1	1	1.00
17	99	105	14.5	16.5	-2	4.00
18	106	115	22.5	28	-5.5	30.25
19	126	116	30	29	1	1.00
20	100	107	17.5	19	-1.5	2.25
21	93	111	5.5	24	-18.5	342.25
22	94	98	7	8.5	-1.5	2.25
23	100	105	17.5	16.5	1	1.00
24	96	103	10	15	-5	25.00
25	99	101	14.5	13	1.5	2.25
26	112	123	27	31	-4	16.00
27	106	108	22.5	20	2.5	6.25
28	98	97	12.5	7	5.5	30.25
29	96	100	10	11.5	-1.5	2.25
30	98	99	12.5	10	2.5	6.25
31	100	100	17.5	11.5	6	36.00
32	96	102	10	14	-4	16.00

Sum of the squared differences → $\sum d^2 = 1,043.50$

**FIGURE 14-9 UPPER-TAILED HYPOTHESIS TEST AT THE 0.01 LEVEL OF SIGNIFICANCE**

$H_0: \sigma_s = 0 \leftarrow$ Null hypothesis: There is no rank correlation in the population; that is, husbands' intelligence and wives' intelligence are randomly mixed

$H_1: \sigma_s > 0 \leftarrow$ Alternative hypothesis: The population rank correlation is positive; that is, bright people choose bright spouses

$\alpha = 0.01 \leftarrow$ Level of significance for testing these hypotheses

An upper-tailed test is appropriate. From Appendix Table 1, we find that the critical z value for the 0.01 level of significance is 2.33. Figure 14-9 illustrates this hypothesis test graphically; we show there the colored region in the upper tail of the distribution that corresponds to the 0.01 level of significance.

To compute our test statistic, we first find the standard error of r_s :

$$\begin{aligned} \sigma_{r_s} &= \frac{1}{\sqrt{n-1}} \\ &= \frac{1}{\sqrt{32-1}} = 0.1796 \end{aligned} \quad [14-9]$$

Now we can use Equation 6-2 to *standardize* the rank correlation coefficient, r_s , by subtracting 0, its hypothesized value, and dividing by σ_{r_s} , its standard error.

$$\begin{aligned} z &= \frac{r_s - 0}{\sigma_{r_s}} \\ &= \frac{0.8087}{0.1796} \\ &= 4.503 \end{aligned} \quad [6-2]$$

Figure 14-10 shows the limit of the acceptance region and the standardized rank-correlation coefficient we calculated from the IQ data. In Figure 14-10, we can see that the rank-correlation coefficient lies far outside the acceptance region. Therefore, we would reject the null hypothesis of no correlation and conclude that bright people tend to choose bright spouses.

Interpreting the results

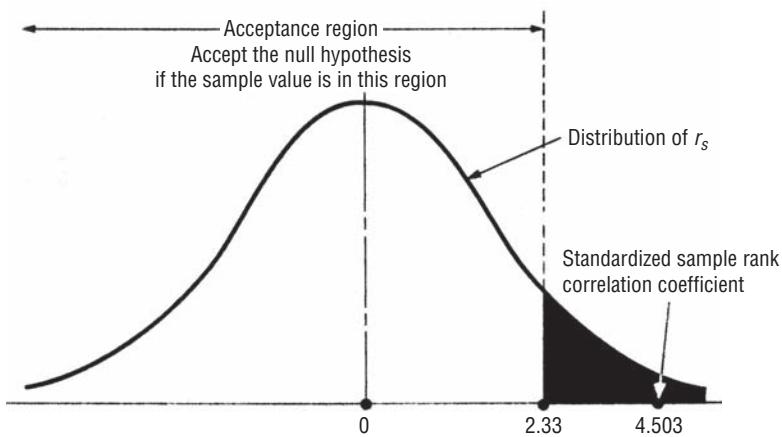


FIGURE 14-10 UPPER-TAILED HYPOTHESIS TEST AT THE 0.01 LEVEL OF SIGNIFICANCE, SHOWING THE ACCEPTANCE REGION AND THE STANDARDIZED SAMPLE RANK-CORRELATION COEFFICIENT

A Special Property of Rank Correlation

Rank correlation has a useful advantage over the correlation method we discussed in Chapter 12. Suppose we have cases in which one or several very extreme observations exist in the original data. By the use of numerical values as was done in Chapter 12, the correlation coefficient may not be a good description of the association that exists between two variables. Yet extreme observations in a *rank*-correlation test will never produce a large rank difference.

Advantage of rank correlation

Consider the following data array of two variables, x and y :

X	10	13	16	19	25
Y	34	40	45	51	117

Because of the large value of the fifth y term, we would get two significantly different answers for r using the conventional and the rank-correlation methods. In this case, the rank-correlation method would be less sensitive to the extreme value. We would assign a rank order of 5 to the numerical value of 117 and avoid the unduly large effect on the value of the correlation coefficient.

HINTS & ASSUMPTIONS

When there are extreme values in the original data, rank correlation can produce more useful results than the correlation method explained in Chapter 12 because extreme observations never produce a large difference in *rank*. Hint: Rank correlation is very useful when data are non-normally distributed. Take the case of university fund-raising where you get a few “big hitter” gifts, lots and lots of gifts below \$100, and a very broad range in between. Using the correlation techniques of Chapter 12 to investigate the relationship between number of appeal mailings and size of gift with this kind of distribution doesn’t make sense because the million-dollar gifts would distort the findings. But using rank correlation in this instance works quite well.

EXERCISES 14.5

Self-Check Exercise

SC 14-6 The following are ratings of aggressiveness (X) and amount of sales in the last year (Y) for eight salespeople. Is there a significant rank correlation between the two measures? Use the 0.10 significance level.

X	30	17	35	28	42	25	19	29
Y	35	31	43	46	50	32	33	42

Applications

14-33 The following are years of experience (X) and average customer satisfaction (Y) for 10 service providers. Is there a significant rank correlation between the two measures? Use the 0.05 significance level.

X	6.3	5.8	6.1	6.9	3.4	1.8	9.4	4.7	7.2	2.4
Y	5.3	8.6	4.7	4.2	4.9	6.1	5.1	6.3	6.8	5.2

14-34 A plant supervisor ranked a sample of eight workers on the number of hours of overtime worked and length of employment. Is the rank correlation between the two measures significant at the 0.01 level?

Amount of overtime	5.0	8.0	2.0	4.0	3.0	7.0	1.0	6.0
Years employed	1.0	6.0	4.5	2.0	7.0	8.0	4.5	3.0

14-35 Most people believe that managerial experience produces better interpersonal relationships between a manager and her employees. The Quail Corporation has the following data matching years of experience on the part of the manager with the number of grievances filed last year by the employees reporting to that manager. At the 0.05 level of significance, does the rank correlation between these two suggest that experience improves relationships?

Years of experience	7	18	17	4	21	27	20	14	15	10
Number of grievances	5	2	4	4	3	2	4	5	4	6

14-36 The Occupational Safety and Health Administration (OSHA) was conducting a study of the relationship between expenditures for plant safety and the accident rate in the plants. OSHA had confined its studies to the synthetic chemical industry. To adjust for the size differential that existed among some of the plants, OSHA had converted its data into expenditures per production employee. The results follow:

**Expenditure by Chemical Companies per Production Employee
in Relation to Accidents per Year**

Company	A	B	C	D	E	F	G	H	I	J	K
Expenditure	\$60	\$37	\$30	\$20	\$24	\$42	\$39	\$54	\$48	\$58	\$26
Accidents	2	7	6	9	7	4	8	2	4	3	8

Is there a significant correlation between expenditures and accidents in the chemical-company plants? Use a rank correlation (with 1 representing highest expenditure and accident rate) to support your conclusion. Test at the 1 percent significance level.

- 14-37** Two business school professors were discussing how difficult it is to predict the success of graduates based on grades alone. One professor thought that the number of years of experience MBAs had before returning for their degrees was probably a better predictor. Using the following data, at the 0.02 level of significance, which rank correlation is a better predictor of career success?

Years experience	4	3	4	3	6	7	1	5	5	2
Grade-point average	3.4	3.2	3.5	3.0	2.9	3.4	2.5	3.9	3.6	3.0
Success rank (10 = top)	4	2	6	5	7	9	1	8	10	3

- 14-38** The Carolina Lighting Company has two trained interviewers to recruit manager trainees for new sales outlets. Although each of the interviewers has a unique style, both are thought to be good preliminary judges of managerial potential. The personnel manager wondered how closely the interviewers would agree, so she had both of them independently evaluate 14 applicants. They ranked the applicants in terms of their degree of potential contribution to the company. The results follow. Use a rank correlation and a 1 percent significance level to determine whether there is a significant positive correlation between the two interviewers' rankings.

Applicant	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Interviewer 1	1	11	13	2	12	10	3	4	14	5	6	9	7	8
Interviewer 2	4	12	11	2	14	10	1	3	13	8	6	7	9	5

- 14-39** Nancy McKenzie, supervisor for a lithographic camera assembly process, feels that the longer a group of employees works together, the higher the daily output rate. She gathered the following data during the first 10 days that one group of employees worked together.

Day	1	2	3	4	5	6	7	8	9	10
Output rate	4.0	7.0	5.0	6.0	8.0	2.0	3.0	0.5	9.0	6.0

Can Nancy conclude at a 5 percent significance level that there is no correlation between the number of days worked together and the daily output?

- 14-40** An electronics firm, which recruits many engineers, wonders whether the cost of extensive recruiting efforts is worth it. If the firm could be confident (using a 1 percent significance level) that the population rank correlation between applicants' résumés scored by the personnel department and interview scores is positive, it would feel justified in discontinuing interviews and relying on résumé scores in hiring. The firm has drawn a sample of 35 engineer applicants in the last 2 years. On the basis of the sample shown, should the firm discontinue interviews and use résumé scores to hire?

Individual	Interview Score	Résumé Score	Individual	Interview Score	Résumé Score
1	81	113	19	81	111
2	88	88	20	84	121
3	55	76	21	82	83
4	83	129	22	90	79
5	78	99	23	63	71
6	93	142	24	78	108
7	65	93	25	73	68
8	87	136	26	79	121
9	95	82	27	72	109
10	76	91	28	95	121
11	60	83	29	81	140
12	85	96	30	87	132
13	93	126	31	93	135
14	66	108	32	85	143
15	90	95	33	91	118
16	69	65	34	94	147
17	87	96	35	94	138
18	68	101			

- 14-41** The following are salary and age data for the 10 Ph.D. candidates graduating this year from the School of Accounting at Northwest University. At the 0.05 level of significance, does the rank correlation of age and salary suggest that older candidates get higher starting salaries?

Salary	Age
\$67,000	29
60,000	25
57,500	30
59,500	35
50,000	27
55,000	31
59,500	32
63,000	38
69,500	28
72,000	34

- 14-42** Dee Boone operates a repair facility for light-aircraft engines. He is interested in improving his estimates of repair time required and believes that the best predictor is the number of operating hours on the engine since its last major repair. Below are data on ten engines Dee worked on recently. At the 0.10 level of significance, does the rank correlation suggest a strong relationship?

Engine	House Since Last Major Repair	House Required to Repair
1	1,000	40
2	1,200	54
3	900	41
4	1,450	60
5	2,000	65
6	1,300	50
7	1,650	42
8	1,700	65
9	500	43
10	2,100	66

Worked-Out Answer to Self-Check Exercise

SC 14-6

X(ranks)	6	1	7	4	8	3	2	5
Y(ranks)	4	1	6	7	8	2	3	5
d	2	0	1	-3	0	1	-1	0
d ²	4	0	1	9	0	1	1	0

$$\sum d^2 = 16 \quad n = 8 \quad \alpha = 0.10$$

$$H_0: \rho_s = 0 \quad H_1: \rho_s \neq 0$$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(16)}{8(63)} = 0.8095$$

From Appendix Table 7, the critical values for r_s are ± 0.6190 . Because $0.8095 > 0.6190$, we reject H_0 . The correlation is significant.

14.6 THE KOLMOGOROV-SMIRNOV TEST

The Kolmogorov-Smirnov test, named for statisticians A. N. Kolmogorov and N. V. Smirnov, is a simple nonparametric method for testing whether there is a significant difference

The K-S test and its advantages

between an observed frequency distribution and a theoretical frequency distribution. The K-S test is therefore another measure of the *goodness-of-fit* of a theoretical frequency distribution, as was the chi-square test we studied in Chapter 11. However, the K-S test has several advantages over the χ^2 test: It is a more powerful test, and it is easier to use because it does not require that data be grouped in any way.

The K-S statistic, D_n , is particularly useful for judging how close the observed frequency distribution is to the expected frequency distribution, because the probability distribution of D_n depends on the sample size n but is independent of the expected frequency distribution (D_n is a distribution-free statistic).

A special advantage

A Problem Illustrating the K-S Test

Suppose that the Orange County Telephone Exchange has been keeping track of the number of “senders” (a type of automatic equipment used in telephone exchanges) that were in use at a given instant. Observations were made on 3,754 different occasions. For capital-investment planning purposes, the budget officer of this company thinks that the pattern of usage follows a Poisson distribution with a mean of 8.5. If he wants to test his hypothesis at the 0.01 level of significance, he can use the K-S test.

We would set up our hypotheses like this:

H_0 : A Poisson distribution with $\lambda = 8.5$ is a good description of the pattern of usage ← Null hypothesis

Stating the hypotheses

H_1 : A Poisson distribution with $\lambda = 8.5$ is not a good description of the pattern of usage ← Alternative hypothesis

$\alpha = 0.01$ ← Level of significance for testing these hypotheses

Computing and comparing expected frequencies

Next, we would list the data that we observed. Table 14-14 lists the observed frequencies and transforms them into observed relative cumulative frequencies.

Now we can use the Poisson formula to compute the expected frequencies.

$$p(x) = \frac{\lambda^x \times e^{-\lambda}}{x!} \quad [5-4]$$

By comparing these expected frequencies with our observed frequencies, we can examine the extent of the difference between them: the absolute deviation. Table 14-15 lists the observed relative cumulative frequencies F_o , the expected relative cumulative frequencies F_e , and the absolute deviations for $x = 0$ to 22.

Calculating the K-S Statistic

To compute the K-S statistic for this problem, you simply pick out D_n , the maximum absolute deviation of F_e from F_o .

Computing the K-S statistic

K-S Statistic

$$D_n = \max |F_e - F_o| \quad [14-10]$$

In this problem, $d_n = 0.2582$ at $x = 9$.

A K-S test must always be a one-tailed test. The critical values for D_n have been tabulated and can be found in Appendix Table 8. By looking in the row for $n = 3,754$ (the sample size) and the column for a significance level of 0.01, we find that the

Computing the critical value

TABLE 14-14 OBSERVED AND RELATIVE CUMULATIVE FREQUENCIES

Number Busy	Observed Frequency	Observed Cumulative Frequency	Observed Relative Cumulative Frequency
0	0	0	0.0000
1	5	5	0.0013
2	14	19	0.0051
3	24	43	0.0115
4	57	100	0.0266
5	111	211	0.0562
6	197	408	0.1087
7	278	686	0.1827
8	378	1,064	0.2834
9	418	1,482	0.3948
10	461	1,943	0.5176
11	433	2,376	0.6329
12	413	2,789	0.7429
13	358	3,147	0.8383
14	219	3,366	0.8966
15	145	3,511	0.9353
16	109	3,620	0.9643
17	57	3,677	0.9795
18	43	3,720	0.9909
19	16	3,736	0.9952
20	7	3,743	0.9971
21	8	3,751	0.9992
22	3	3,754	1.0000

critical value of D_n must be computed using the formula

$$\frac{1.63}{\sqrt{n}} = \frac{1.63}{\sqrt{3,754}} = \frac{1.63}{61.27} = 0.0266$$

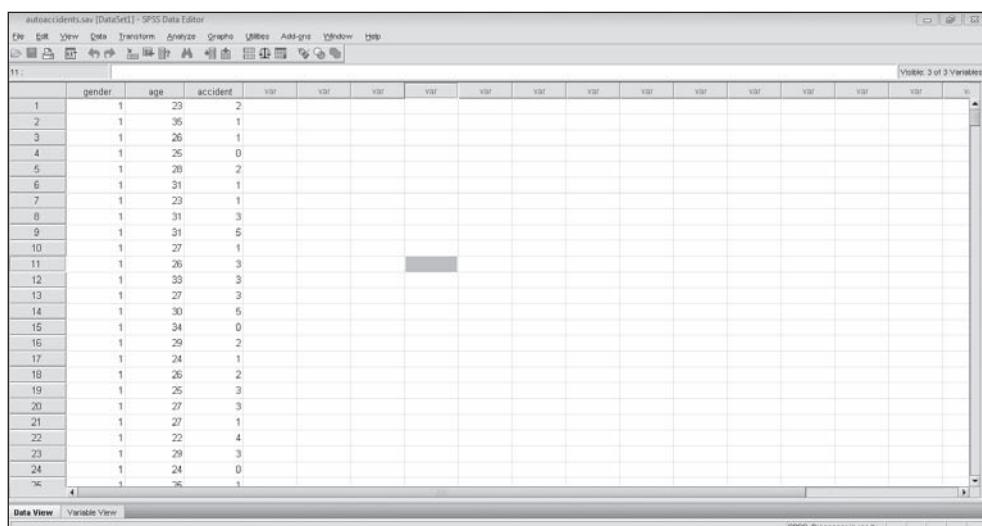
The next step is to compare the calculated value of D_n with the critical value of D_n from the table. If the table value for the chosen significance level is greater than the calculated value of D_n , then we will accept the null hypothesis. Obviously, $0.0266 < 0.2582$, so we reject H_0 and conclude that a Poisson distribution with a mean of 8.5 is *not* a good description of the pattern of sender usage at the Orange County Telephone Exchange.

Our conclusion

TABLE 14-15 RELATIVE OBSERVED CUMULATIVE FREQUENCIES, EXPECTED RELATIVE CUMULATIVE FREQUENCIES, AND ABSOLUTE DEVIATIONS

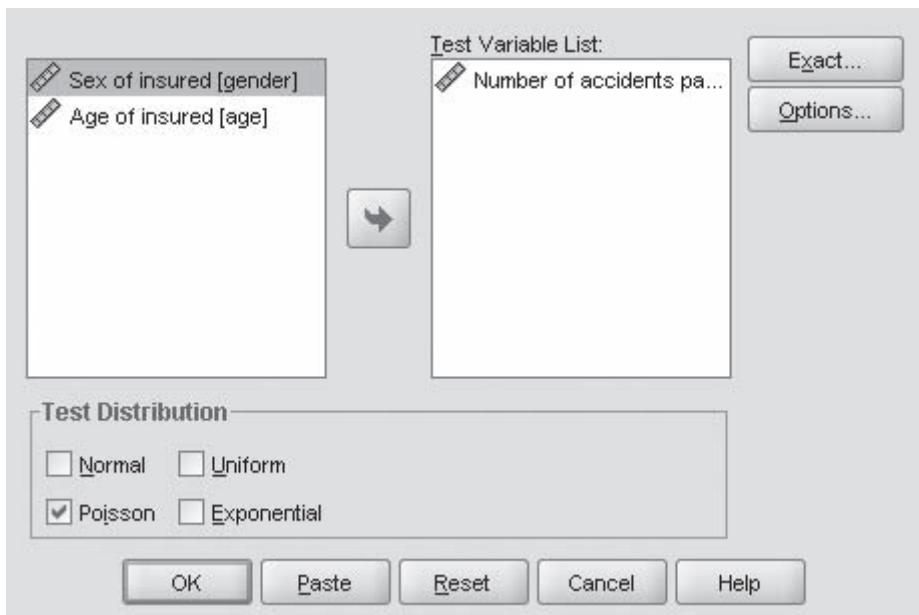
Number Busy	Observed Frequency	Observed Cumulative Frequency	Observed Relative Cumulative Frequency	Expected Relative Cumulative Frequency	$ F_e - F_o $ Absolute Deviation
0	0	0	0.0000	0.0002	0.0002
1	5	5	0.0013	0.0019	0.0006
2	14	19	0.0051	0.0093	0.0042
3	24	43	0.0115	0.0301	0.0186
4	57	100	0.0266	0.0744	0.0478
5	111	211	0.0562	0.1496	0.0934
6	197	408	0.1087	0.2562	0.1475
7	278	686	0.1827	0.3856	0.2029
8	378	1,064	0.2834	0.5231	0.2397
9	418	1,482	0.3948	0.6530	0.2582
10	461	1,943	0.5176	0.7634	0.2458
11	433	2,376	0.6329	0.8487	0.2158
12	413	2,789	0.7429	0.9091	0.1662
13	358	3,147	0.8383	0.9486	0.1103
14	219	3,366	0.8966	0.9726	0.0760
15	145	3,511	0.9353	0.9862	0.0509
16	109	3,620	0.9643	0.9934	0.0291
17	57	3,677	0.9795	0.9970	0.0175
18	43	3,720	0.9909	0.9987	0.0078
19	16	3,736	0.9952	0.9995	0.0043
20	7	3,743	0.9971	0.9998	0.0027
21	8	3,751	0.9992	0.9999	0.0007
22	3	3,754	1.0000	1.0000	0.0000

Kolmogorov-Smirnov Test Using SPSS



For Kolmogorov–Smirnov test we will take data for an insurance analyst who wants to model the number of automobile accidents per driver. She has randomly sampled data on drivers in a certain region and uses the Kolmogorov–Smirnov test to confirm that the number of accidents follows a Poisson distribution.

For Kolmogorov–Smirnov go to **Analyze > Nonparametric tests > One sample Kolmogorov Smirnov test > Select test variable > select test distribution> OK.**



The screenshot shows the SPSS Output window with the following content:

```

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Npar Tests
  /Title
  /Notes
  /Active Dataset
  /One-Sample Kolmogorov-Smirnov Test
  /MISSING ANALYSIS.

Npar Tests
  /One-Sample Kolmogorov-Smirnov Test

[DataSet1] D:\spss\Samples\autoaccidents.sav

One-Sample Kolmogorov-Smirnov Test
  N          Number of accidents past 5 years
  N
  Poisson Parametera Mean      500
  Most Extreme Differences Absolute .065
                                Positive .065
                                Negative -.041
  Kolmogorov-Smirnov Z        1.460
  Asymp. Sig. (2-tailed)     .028

a. Test distribution is Poisson.

```

The output window also displays the menu bar and some other dataset information.

HINTS & ASSUMPTIONS

Think of the Kolmogorov-Smirnov test as another *goodness-of-fit* test, just like the chi-square test in Chapter 11, except that this time it's easier to use because we do not have to do all the arithmetic needed to calculate chi-square. The K-S test just finds the relative cumulative distributions for both observed frequencies and expected frequencies and then tests how far apart they are. If the distance is not significant, then the observed distribution is well described by the theoretical distribution. Hint: K-S tests are *always* one-tailed tests because we are always testing whether differences are greater than a specified level.

EXERCISES 14.6**Self-Check Exercise**

SC 14-7 The following is an observed frequency distribution. Using a normal distribution with $\mu = 6.80$ and $\sigma = 1.24$:

- Find the probability of falling into each class.
- From part (a), compute the expected frequency of each category.
- Calculate D_n .
- At the 0.15 level of significance, does this distribution seem to be well described by the suggested normal distribution?

Value of the variable	≤ 4.009	4.010–5.869	5.870–7.729	7.730–9.589	> 9.590
Observed frequency	13	158	437	122	20

Basic Concepts

14-43 At the 0.05 level of significance, can we conclude that the following data come from a Poisson distribution with $\lambda = 3$?

Number of arrivals per day	0	1	2	3	4	5	6 or more
Number of days	6	18	30	24	11	2	9

14-44 The following is an observed frequency distribution. Using a normal distribution with $\mu = 98.6$ and $\sigma = 3.78$:

- Find the probability of falling into each class.
- From part (a), compute the expected frequency of each category.
- Calculate D_n .
- At the 0.10 significance level, does this distribution seem to be well described by the suggested normal distribution?

Value of the variable	< 92.0	92.0–95.99	96.0–99.99	100–103.99	≥ 104
Observed frequency	69	408	842	621	137

- 14-45** The following is a table of observed frequencies, along with the frequencies to be expected under a normal distribution.
- Calculate the K-S statistic.
 - Can we conclude that these data do in fact come from a normal distribution? Use the 0.10 level of significance.

	Test Score				
	51–60	61–70	71–80	81–90	91–100
Observed frequency	30	100	440	500	130
Expected frequency	40	170	500	390	100

Applications

- 14-46** Kevin Morgan, national sales manager of an electronics firm, has collected the following salary statistics on his field salesforce earnings. He has both observed frequencies and frequencies expected if the distribution of salaries is normal. At the 0.10 level of significance, can Kevin conclude that the distribution of salesforce earnings is normal?

	Earnings (in thousands)						
	25–30	31–36	37–42	43–48	49–54	55–60	61–66
Observed frequency	9	22	25	30	21	12	6
Expected frequency	6	17	32	35	18	13	4

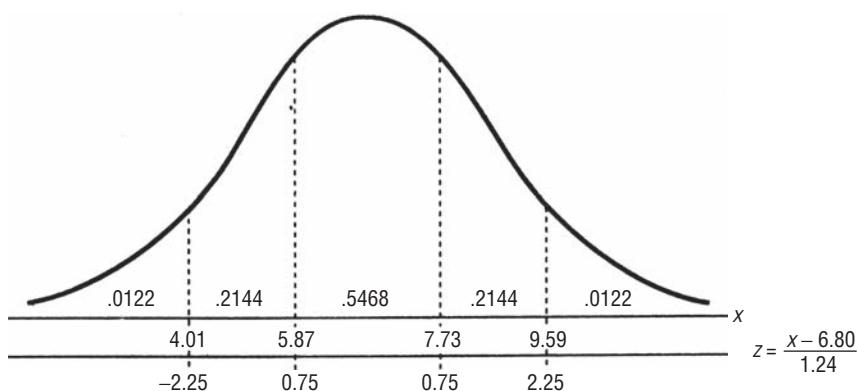
- 14-47** Randall Nelson, salesman for the V-Star company, has seven accounts to visit per week. It is thought that the sales by Mr. Nelson may be described by the binomial distribution, with the probability of selling each account being 0.45. Examining the observed frequency distribution of Mr. Nelson's number of sales per week, determine whether the distribution does in fact correspond to the suggested distribution. Use the 0.05 significance level.

Number of sales per week	0	1	2	3	4	5	6	7
Frequency of the number of sales	25	32	61	47	39	21	18	12

- 14-48** Jackie Denn, an airline food-service administrator, has examined past records from 200 randomly selected cross-country flights to determine the frequency with which low-sodium meals were requested. The number of flights in which 0, 1, 2, 3, or 4 or more low-sodium meals were requested was 25, 45, 67, 43, and 20, respectively. At the 0.05 level of significance, can she reasonably conclude that these requests follow a Poisson distribution with $\lambda = 1$?

Worked-Out Answer to Self-Check Exercise

- SC 14-7** (a) The probabilities of falling into the five classes are the indicated areas under the curve shown in p.800:



- (b) $n = 13 + 158 + 437 + 122 + 20 = 750$. Thus, the expected frequencies are $0.0122(750) = 9.15$, $0.2144(750) = 160.80$, $0.5468(750) = 410.1$, 160.80 , and 9.15 .

(c)	f_o	cum. f_o	F_o	F_e	$ F_e - F_o $
	13	13	0.0173	0.0122	0.0051
	158	171	0.2280	0.2266	0.0014
	437	608	0.8107	0.7734	0.0373 ←
	122	730	0.9733	0.9878	0.0145
	20	750	1.0000	1.0000	0.0000

(d) $D_{table} = \frac{1.14}{\sqrt{n}} = \frac{1.14}{\sqrt{750}} = 0.0416$. $D_n < D_{table}$, so accept H_0 .

The data are well described by the suggested normal distribution.

STATISTICS AT WORK

Loveland Computers

Case 14: Nonparametric Methods “I forgot to tell you,” said Sherrel Wright, the advertising manager, as they headed back to the office, “Margot was looking for you—you better check in with her before you start on this advertising project.”

“I need help!” Margot announced in a voice that could be heard in Cheyenne, Wyoming. “I spent a lot of money to get some data, and now that it’s here, I don’t know what we’ve got.”

“Well I don’t either,” Lee joked, trying to lighten the mood. “Why don’t you tell me what’s going on.”

“For some of the midrange models—basically PCs with fast chips and a reasonable amount of disk storage—we can make them look three different ways. The old AT style machines are the size of a small suitcase. People liked the big box because it had the image of a big, powerful machine. But in the last year or so, some of the very powerful workstations have been made in a pizza box format with a fairly narrow, flat box. So some companies have been offering the midrange in a low-profile format. It’s really just the same innards in a smaller box that does not take up as much desk space. Finally, some competitors have offered a tower configuration. That’s the old AT style tipped on its edge so it can sit on the floor. That eliminates any need for desk space.”

“So which style did Loveland go with?” Lee asked.

"Frankly, we've been all over the place—during different marketing campaigns. Sometimes we've offered two of the three formats, but we've changed back and forth as we've tried to guess what customers want. You'd think that everyone would want the machine on the floor, but it turns out the computer box is 'a useful place to put the monitor, and people who use a lot of floppies don't want to keep reaching under their desks to use the disk drive."

"Okay. So offer all three styles," Lee smiled at this simple-but-elegant solution.

"That just adds to our costs. If we run three styles, we lose the volume discounts that we can get by going with just one. And then we have to advertise three formats while I'm also launching new high-end products and keeping up with demand for our lowest-price machines. I'd like to be able to recommend the single best format to management."

"Well, I don't have a crystal ball," Lee began.

"I don't expect you to. I hired a market-research firm. They ran focus groups in Boulder, New Jersey, and Oregon. There were eight people in each group, and two groups at each site, so altogether I've got 48 response cards—and several hours of videotaped discussions that I'll save you from watching. As you'd expect, we asked the participants to rank the three formats in terms of the style they'd prefer if they were going to buy a personal computer. Then we asked them, if your first choice weren't available, which of the other two formats would you prefer. Tell me how we're going to make some sense out of this so I can make a recommendation to the product-planning group."

Study Questions: How should Lee organize the data and which statistical tests are appropriate? What should Loveland do if the analysis of data from this small group is inconclusive?

CHAPTER REVIEW

Terms Introduced in Chapter 14

Kolmogorov–Smirnov Test A nonparametric test, which does not require that data be grouped in any way, for determining whether there is a significant difference between an observed frequency distribution and a theoretical frequency distribution.

Kruskal–Wallis Test A nonparametric method for testing whether three or more independent samples have been drawn from populations with the same distribution. It is a nonparametric version of ANOVA, which we studied in Chapter 11.

Mann–Whitney *U* Test A nonparametric method used to determine whether two independent samples have been drawn from populations with the same distribution.

Nonparametric Tests Statistical techniques that do not make restrictive assumptions about the shape of a population distribution when performing a hypothesis test.

One-Sample Runs Test A nonparametric method for determining the randomness with which the items in a sample have been selected.

Rank Correlation A method for doing correlation analysis when the data are not available to use in numerical form, but when information is sufficient to rank the data.

Rank-Correlation Coefficient A measure of the degree of association between two variables that is based on the ranks of observations, not their numerical values.

Rank Sum Tests A family of nonparametric tests that use information in a set of data.

Run A sequence of identical occurrences preceded and followed by different occurrences or by none at all.

Sign Test A test for the difference between paired observations where + and – signs are substituted for quantitative values.

Theory of Runs A theory developed to allow us to test samples for the randomness of their order.

Equations Introduced in Chapter 14

$$14-1 \quad U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad \text{p. 760}$$

To apply the Mann–Whitney U test, you need this formula to derive the U statistic, a measure of the difference between the ranked observations of the two variables. R_1 is the sum of the ranks of observations of variable 1; n_1 and n_2 are the numbers of items in samples 1 and 2, respectively. Both samples need not be of the same size.

$$14-2 \quad \mu_U = \frac{n_1 n_2}{2} \quad \text{p. 760}$$

If the null hypothesis of a Mann–Whitney U test is that $n_1 + n_2$ observations came from identical populations, then the U statistic has a sampling distribution with a mean equal to the product of n_1 and n_2 divided by 2.

$$14-3 \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad \text{p. 760}$$

This formula enables us to derive the *standard error of the U statistic* of a Mann–Whitney U test.

$$14-4 \quad U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \quad \text{p. 762}$$

This formula and Equation 14-1 can be used interchangeably to derive the U statistic in the Mann–Whitney U test. To save time, use this formula if the number of observations in sample 2 is significantly smaller than the number of observations in sample 1.

$$14-5 \quad K = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1) \quad \text{p. 765}$$

The formula computes the K statistic used in the Kruskal–Wallis test for different means among three or more populations. The appropriate sampling distribution for K is chi-square with $k - 1$ degrees of freedom, when each sample contains at least five observations.

$$14-6 \quad \mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \text{p. 774}$$

When doing a one-sample runs test, use this formula to derive the mean of the sampling distribution of the r statistic. This r statistic is equal to the *number of runs* in the sample being tested.

$$14-7 \quad \sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \quad \text{p. 774}$$

This formula enables us to derive the *standard error of the r statistic* in a one-sample runs test.

14-8

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$
p. 782

The *coefficient of rank correlation*, r_s , is a measure of the closeness of association between two ranked variables.

14-9

$$\sigma_{r_s} = \frac{1}{\sqrt{n-1}}$$
p. 786

This formula enables us to calculate the *standard error* a hypothesis test on the coefficient of rank correlation.

14-10

$$D_n = \max |F_e - F_o|$$
p. 794

If we compare this computed value to a critical value of D_n in the K-S table, we can test distributional goodness of fit.

Review and Application Exercises

- 14-49** A college football coach has a theory that in athletics, success feeds on itself. In other words, he feels that winning a championship one year increases the team's motivation to win it the next year. He expressed his theory to a student of statistics, who asked him for the records of the team's wins and losses over the last several years. The coach gave him a list, specifying whether the team had won (W) or lost (L) the championship that year. The results of this tally are

W, W, W, W, W, W, L, W, W, W, W, W, L, W, W, W, W, L, L, W, W, W, W, W, W, W

- (a) At a 10 percent significance level, is the occurrence of wins and losses a random one?
- (b) Does your answer to part (a), combined with a sight inspection of the data, tell you anything about the one-sample runs test?

14-50

- A small metropolitan airport recently opened a new runway, creating a new flight path over an upper-income residential area. Complaints of excessive noise had deluged the airport authority to the point that the two major airlines servicing the city had installed special engine baffles on the turbines of the jets to reduce noise and help ease the pressure on the authority. Both airlines wanted to see whether the baffles had helped to reduce the number of complaints that had been brought against the airport. If they had not, the baffles would be removed because they increased fuel consumption. Based on the following random samples of 13 days before the baffles were installed and another 13 days after installation, can it be said at the 0.02 level of significance that installing the baffles had reduced the number of complaints?

Complaints Before and After Baffles Were Installed													
Before	27	15	20	24	13	18	30	46	15	29	17	21	18
After	26	23	19	12	25	9	16	12	28	20	16	14	11

14-51

- The American Broadcasting System (ABS) has invested a sizable amount of money into a new program for television, *High Times*. *High Times* was ABS's entry into the situation-comedy market and featured the happy-go-lucky life in a college dormitory. Unfortunately, the program had not done as well as expected, and the sponsor was considering canceling. To beef up the ratings, ABS introduced co-ed dormitories into the series. The following are the results of

telephone surveys before and after the change in the series. Surveys were conducted in several major metropolitan areas, so the results are a composite from the cities.

- Using a U test, can you infer at the 0.10 significance level that the change in the series format helped the ratings?
- Do the results of your test say anything about the effect of sex on TV program ratings?

Share of Audience Before and After Change to Co-Ed Dormitories																	
Before	22	18	19	20	31	22	25	19	22	24	18	16	14	28	23	15	16
After	25	28	18	30	33	25	29	32	19	16	30	33	17	25			

- 14-52** Overall readiness evaluations for military units are conducted by staff officers, with a maximum score of 100 points. Transport command officers complain that they are rated lower than infantry command officers because most of the staff officers came up through the ranks of the infantry. At the 0.05 level of significance, test the hypothesis of no difference in ratings, based on the readiness evaluations at both units during 10 randomly chosen weeks.

Evaluation Score										
Infantry command	72	80	86	90	95	92	88	96	91	82
Transport command	80	79	90	82	81	84	78	74	85	71

Table RW12-1 on p. 703 presented the results of the 1992 *Business Week* and *U.S. News & World Report* rankings of American business schools. Use that information to answer Exercises 14-53 and 14-54.

- 14-53** Consider the top 10 schools in the overall *Business Week* ranking. Rescaling the student and recruiting-firm rankings for those 10 schools, we get

School	Rankings	
	By Students	By Firms
Northwestern	2	1
Chicago	7	4
Harvard	8	3
Wharton	9	2
Michigan	6	6
Dartmouth	1	10
Stanford	3	7
Indiana	4	8
Columbia	10	5
North Carolina	5	9

At $\alpha = 0.10$, do the firms' rankings differ from the students' rankings?

- 14-54** Considering all 20 of the schools, do the rankings by the two magazines differ significantly, at $\alpha = 0.10$?

- 14-55** The Ways and Means Committee of the U.S. House of Representatives was attempting to evaluate the results of a tax cut given to individuals during the preceding year. The intended purpose had been to stimulate the economy, the theory being that with a tax reduction,

the consumer would spend the tax savings. The committee had employed an independent consumer-research group to select a sample of households and maintain records of consumer spending both before and after the legislation was put into effect. A portion of the data from the research group follows:

Schedule of Consumer Spending					
Household	Before Legislation	After Legislation	Household	Before Legislation	After Legislation
1	\$3,578	\$ 4,296	17	\$11,597	\$12,093
2	10,856	9,000	18	9,612	9,675
3	7,450	8,200	19	3,461	3,740
4	9,200	9,200	20	4,500	4,500
5	8,760	8,840	21	8,341	8,500
6	4,500	4,620	22	7,589	7,609
7	15,000	14,500	23	25,750	24,321
8	22,350	22,500	24	14,673	13,500
9	7,346	7,250	25	5,003	6,072
10	10,345	10,673	26	10,940	11,398
11	5,298	5,349	27	8,000	9,007
12	6,950	7,000	28	14,256	14,500
13	34,782	33,892	29	4,322	4,258
14	12,837	14,297	30	6,828	7,204
15	7,926	8,437	31	7,549	7,678
16	5,789	6,006	32	8,129	8,125

At a significance level of 3 percent, determine whether the tax-reduction policy has achieved its desired goals.

- 14-56** Many entertainment companies have invested in theme parks with tie-ins to hit movies. Attendance depends on many factors, including the weather. Should the weather be considered a random event?
- 14-57** Two television weather forecasters got into a discussion one day about whether years with heavy rainfall tended to occur in spurts. One of them said he thought that there were patterns of annual rainfall amounts, and that several wet years were often followed by a number of drier-than-average years. The other forecaster was skeptical and said she thought that the amount of rainfall for consecutive years was fairly random. To investigate the question, they decided to look at the annual rainfall for several years back. They found the median amount and classified the rainfall as below (B) or above (A) the median annual rainfall. A summary of their results follows:

A, A, A, B, B, B, A, B, A, A, B, B, A, B, A, A, B, B, A, A, A, B, A, A, A,
A, A, A, B, B, B, A, B, B, A, A, A, B, A, A, A, B, A, B, B, A, B, A, B, B

If the forecasters test at a 5 percent significance level, will they conclude that the annual rainfall amounts do not occur in patterns?

- 14-58** Anne J. Montgomery, administrative director of executive education at Southern University, uses two kinds of promotional material to announce seminars: personal letters and brochures. She feels quite strongly that brochures are the more effective method. She has collected data

on numbers of people attending each of the last 10 seminars promoted with each method. At the 0.15 level of significance, is her hunch right?

	Number Attending									
Personal letter	35	85	90	92	88	46	78	57	85	67
Brochure	42	74	82	87	45	73	89	75	60	94

- 14-59** The National Association of Better Advertising for Children (NABAC), a consumer group for improving children's television, was conducting a study on the effect of Saturday morning advertising. Specifically, the group wanted to know whether a significant degree of purchasing was stimulated by advertising directed at children, and if there was a positive correlation between Saturday morning TV advertising time and product sales.

NABAC chose the children's breakfast-cereal market as a sample group. It selected products whose advertising message was aimed entirely at children. The results of the study follow. (The highest-selling cereal has sales rank 1.)

Comparison of TV Advertising Time and Product Sales Advertising Time		
Product	in Minutes	Sales Rank
Captain Grumbles	0.50	10
Obnoxious Berries	3.00	1
Fruity Hoops	1.25	9
OO La Granola	2.00	5
Sweet Tweets	3.50	2
Chocolate Chumps	1.00	11
Sugar Spots	4.00	3
County Cavity	2.50	8
Crunchy Munchies	1.75	6
Karamel Kooks	2.25	4
Flakey Flakes	1.50	7

Can the group conclude that there is a *positive* rank correlation between the amount of Saturday morning advertising time and sales volume of break-fast cereals? Test at the 5 percent significance level.

- 14-60** *American Motoring Magazine* recently tested two brake-disk materials for stopping effectiveness. Data representing stopping distances for both kinds of materials follow. At the 0.05 level of significance, test the hypothesis that there is no difference in the effectiveness of the materials.

	Stopping Distance (feet)									
Graphite bonded	110	120	130	110	100	105	110	130	145	125
Sintered bronze	100	110	135	105	105	100	100	115	135	120

- 14-61** As part of a survey on restaurant quality, a local magazine asked area residents to rank two steak houses. On a scale of 1 to 10, subjects were to rate characteristics such as food quality, atmosphere, service, and price. After data were collected, one of the restaurant owners proposed that various statistical tests be performed. He specifically mentioned that

he would like to see a mean and standard deviation for the responses to each question about each restaurant, in order to see which one had scored better. Several of the magazine workers argued against his suggestions, noting that the quality of input data would not justify a detailed statistical analysis. They argued that what was important was the residents' rankings of the two restaurants. Evaluate the arguments presented by the restaurant owner and the magazine employees.

- 14-62** Senior business students interviewed by the Ohio Insurance Company were asked not to discuss their interviews with others in the school until the recruiter left. The recruiter, however, suspected that the later applicants knew more about what she was looking for. Were her suspicions correct? To find out, rank the interview scores received by subjects given in the table. Then test the significance of the rank correlation coefficient between the scores and interview number. Use the 0.02 significance level.

Interview Number	Score						
1	63	6	57	11	77	16	70
2	59	7	76	12	61	17	75
3	50	8	81	13	53	18	90
4	60	9	58	14	74	19	80
5	66	10	65	15	82	20	89

- 14-63** More than 3 years ago, the Occupational Safety and Health Administration (OSHA) required a number of safety measures to be implemented in the Northbridge Aluminum plant. Now OSHA would like to see whether the changes have resulted in fewer accidents in the plant. It has collected these data:

Accidents at the Northbridge Plant												
	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1992	5	3	4	2	6	4	3	3	2	4	5	3
1993	4	4	3	3	3	4	0	5	4	2	0	1
1994	3	2	1	1	0	2	4	3	2	1	1	2
1995	2	1	0	0	1	2						

- (a) Determine the median number of accidents per month. If the safety measures have been effective, we should find early months falling above the median and later months below the median. Accordingly, there will be a small number of runs above and below the median. Conduct a test at the 0.03 level of significance to see whether the accidents are randomly distributed.
- (b) What can you conclude about the effectiveness of the safety measures?

- 14-64** A large countywide ambulance service calculates that for any given township it serves, during any given 6-hour shift, there is a 35 percent chance of receiving at least one call for assistance. The following is a random sampling of 90 days:

Number of shifts during which calls were received	0	1	2	3	4
Number of days	5	35	30	13	7

At the 0.05 level of significance, do these calls for assistance follow a binomial distribution?

- 14-65** Jim Bailey, owner of Crow's Nest Marina, believes that the number of hours a boat engine has been run in salt water and not the age of the boat is the best predictor of engine failure. His service manager has collected data from his repair records on failed engines. At the 0.05 level of significance, is Jim's hunch right?

Engine	Hours in Salt Water	Age of Engine (years)	Cost of Repair (dollars)
1	300	4	625
2	150	6	350
3	200	3	390
4	250	6	530
5	100	4	200
6	400	5	1,000
7	275	6	550
8	350	6	800
9	325	3	700
10	375	2	600

- 14-66** SavEnergy, an international activist group concerned about the gross domination of Western areas in energy usage, has claimed that population size and energy consumption are negatively correlated. Their opponents claim no correlation is present. Using the following data, test the hypothesis that no rank correlation exists between population and energy consumption, versus SavEnergy's negative correlation claim. Use the 0.10 level of significance.

	1989 Population (000,000 omitted)	Total Energy Consumption (10^{15} joules)
United States	249	68
Latin America	438	16
Africa	646	11
Europe	499	65
Soviet Union	289	54
India	835	9
China	1,100	24

- 14-67** Highway crashes killed more than 75,000 occupants of passenger cars during 1993–1996. Using that grim statistic as a starting point, researchers at the Insurance Institute for Highway Safety computed death rates for the 103 largest-selling vehicle series. Vehicles were categorized as station wagons & vans, four-door cars, two-door cars, or sports & specialty cars. Further stratification in each category labeled vehicles as large, midsize, or small. Looking at the rates (deaths per 10,000 registered vehicles) for four-door cars, the figures are as follows:

Large	1.2	1.3	1.4	1.5	1.5	1.5	1.6	1.8	
Midsize	1.1	1.2	1.2	1.2	1.3	1.3	1.3	1.3	1.4
	1.5	1.6	1.6	1.6	1.7	1.7	1.8	1.9	2.0
	2.3	2.4	2.5	2.6	2.9				2.3
Small	1.1	1.5	1.6	1.7	1.8	2.0	2.0	2.0	2.5
	2.6	2.8	3.2	4.1					

Use the Kruskal–Wallis test to test whether the three population means are equal. Test at the 0.05 level of significance.

- 14-68** The year 1996 was particularly bad for injuries to professional baseball players. From the following data, does a sign test for paired data indicate that American League players suffered significantly more injuries than their National League counterparts? Use a 0.05 level of significance.

Injury Location	AL	NL	Injury Location	AL	NL
Shoulder	46	22	Back	10	7
Neck	3	0	Wrist	10	2
Rib	7	5	Hip	1	1
Elbow	21	19	Hand	6	4
Finger	7	5	Ankle	6	4
Thigh	17	14	Foot	1	4
Groin	7	3	Toe	0	1
Knee	16	18	Other	10	4

- 14-69** Recent research about the kinds of weather patterns that may be correlated with sunspots, has focused on polar temperature (the average temperature in the stratosphere above the North Pole) during periods when certain equatorial winds are blowing. When these winds are from the west, the polar temperature appears to rise and fall with solar activity. When the winds are easterly, the temperature appears to do the opposite of what the sun is doing. From the data, calculate the coefficients of rank correlation between these variables and test, at the 0.05 level of significance, if the hypothesized relationships hold (i.e., positive correlation for westerly winds, negative correlation for easterly winds).

Polar Temperature (°F)		
Solar Activity	East Winds	West Winds
230	-85	-76
160	-97	-86
95	-88	-100
75	-85	-110
100	-90	-108
165	-96	-85
155	-91	-70
120	-76	-100
75	-80	-110
65	-86	-112
125	-90	-99
195	-104	-91
190	-95	-93
125	-99	-99
75	-73	-103

- 14-70** The Model Town Highways has issued a notice in the beginning of February 2012 for the early redemption of some of its infra-bonds. There were a total of 10000 such bonds. The

interest rate for the bonds was 6.5% and scheduled to mature in 2015. The decision for the redemption of the bonds is because of financial reasons. It was decided that the bonds to be selected for redemption should be free from any bias. The bonds selected for redemption were numbered as:

2	10	15	19	35	78	175	549	989	1135	1367	1668
1896	2235	2387	2885	2954	3098	3793	4367	4486	4809	5076	6687
6906	6999	7056	7216	7389	7999	8006	8451	8601	8991	9005	9180
9361	9571	9688	9799								

- (a) Assuming that the infra-bonds were selected randomly for the purpose of redemption, how many you would expect to find with serial numbers between 1 and 2000; 2001 and 4000; 4001 and 6000, 6001 and 8000 and finally 8001 and 10000?
- (b) Is it reasonable to conclude that the infra-bonds called for redemption were selected randomly, using chi-square test of goodness of fit?
- (c) Use Kolmogorov-Smirnov Test to examine the claim that the selection of the bonds is indeed random.
- (d) Compare your results in parts (b) and (c) and give your comments.

- 14-71** Managers in service-operations businesses have to handle peak times, when many customers arrive at once. The manager of the information booth at a suburban mall collected the following data on arrivals per minute between 7:10 and 8:00 on Thursday, the mall's late shopping night:

Number of Arrivals	1	2	3	4	5	6	7	8	9	10	11
Frequency	5	3	2	6	6	2	6	10	4	4	2

Test whether a Poisson distribution with a mean of 6 adequately describes these data. Use the 0.05 level of significance.

- 14-72** The results of the Carolina Athletic Association's first 10K run showed the following order of male and female finishers:

M M M M M M M M M M M M W M M M M M M W M M M
M W M M M M M M M M W M W M W M M M W M M M M W M
M W M M M M M M W M M W M M M W W W W M W M W W M
W M M M W M W W M W W W W M M W M M

Did the women finish randomly throughout? Use the 0.20 level of significance.

- 14-73** Several groups were given a list of 30 activities and technological advances and were asked to rank them, considering the risk of dying as a consequence of each. The results are in the following table. Calculate the rank correlation coefficient of each group relative to the experts' ranking. Which group seemed to have the most accurate perception of the risks involved?

A = Experts

B = League of Women Voters

C = College Students

D = Civic Club Members

Risk	A	B	C	D
Motor vehicles	1	2	5	3
Smoking	2	4	3	4
Alcoholic beverages	3	6	7	5
Handguns	4	3	2	1

(continued)

(contd.)

Risk	A	B	C	D
Surgery	5	10	11	9
Motorcycles	6	5	6	2
X-rays	7	22	17	24
Pesticides	8	9	4	15
Electric power (nonnuclear)	9	18	19	19
Swimming	10	19	30	17
Contraceptives	11	20	9	22
General (private) aviation	12	7	15	11
Large construction	13	12	14	13
Food preservatives	14	25	12	28
Bicycles	15	16	24	14
Commercial aviation	16	17	16	18
Police work	17	8	8	7
Fire fighting	18	11	10	6
Railroads	19	24	23	20
Nuclear power	20	1	1	8
Food coloring	21	26	20	30
Home appliances	22	29	27	27
Hunting	23	13	18	10
Prescription antibiotics	24	28	21	26
Vaccinations	25	30	29	29
Spray cans	26	14	13	23
High school & college football	27	23	26	21
Power mowers	28	27	28	25
Mountain climbing	29	15	22	12
Skiing	30	21	25	16

- 14-74** In testing a new hayfever medication, researchers measured the incidence of adverse side effects of the drug by administering it to a large number of patients and evaluating them against a control group. The percentages of patients reporting 13 types of side effects were recorded. Using a sign test for paired data, can you determine whether, on the whole, either group experienced more adverse side effects? Use the 0.10 significance level.

Side Effect	Drug	Control
A	9.0	18.1
B	6.3	3.8
C	2.9	5.8
D	1.4	1.0
E	0.9	0.6
F	0.9	0.2
G	0.6	0.0
H	4.6	2.7
I	2.3	3.5
J	0.9	0.5
K	0.5	0.5
L	0.0	0.2
M	1.0	1.4

14-75 Commercial banks play an important role in the development of economies by effective and optimum mobilization of resources and their allocation. The banking sector in India has undergone a significant transformation in the past few years because of economic reforms. There is a mix of players in this sector (public sector banks, private banks and foreign banks). This competitive scenario has brought the dimension of customer satisfaction to the forefront. It has become very important for banks to retain their existing customer base besides enlarging the same. The Bhrigus Consultant, a marketing research agency, collects data related to customer satisfaction with the service aspects of the banks. The following table presents the rank of 10 commercial banks as per the customer satisfaction. Analyze the data at 10 percent level of significance and comment whether there is significant change in the satisfaction ranking of the banks?

Name of the Bank	2010 Rank	2011 Rank
Sun Bank	2	4
Lotus Bank	5	5
State Corporation Bank	3	1
Cooperative State Bank	4	3
IBLB Bank	8	7
DBCI Bank	1	2
Vikas bank	6	8
Unnati Bank	7	6
Corporate Bank	10	9
Excel Bank	9	10

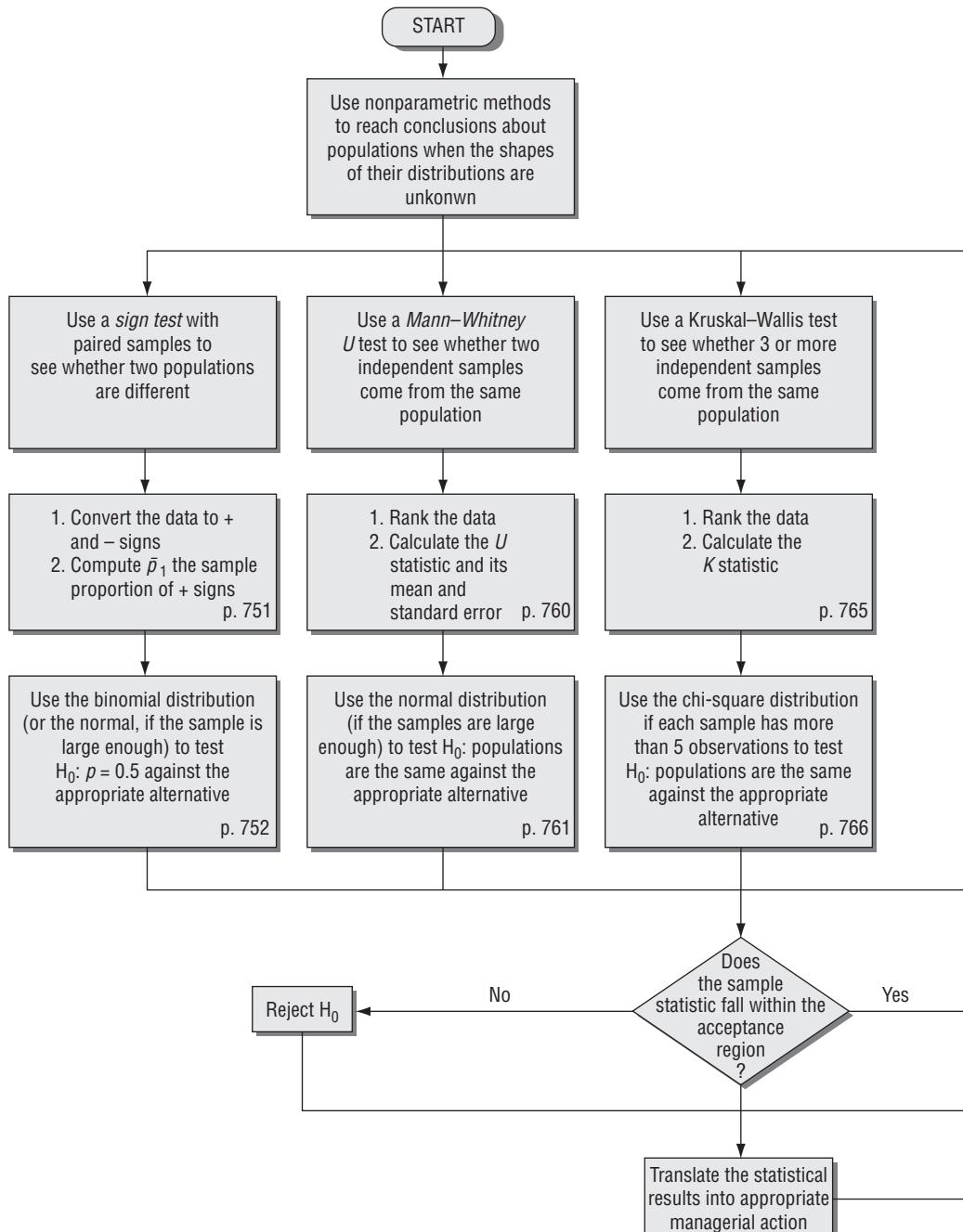


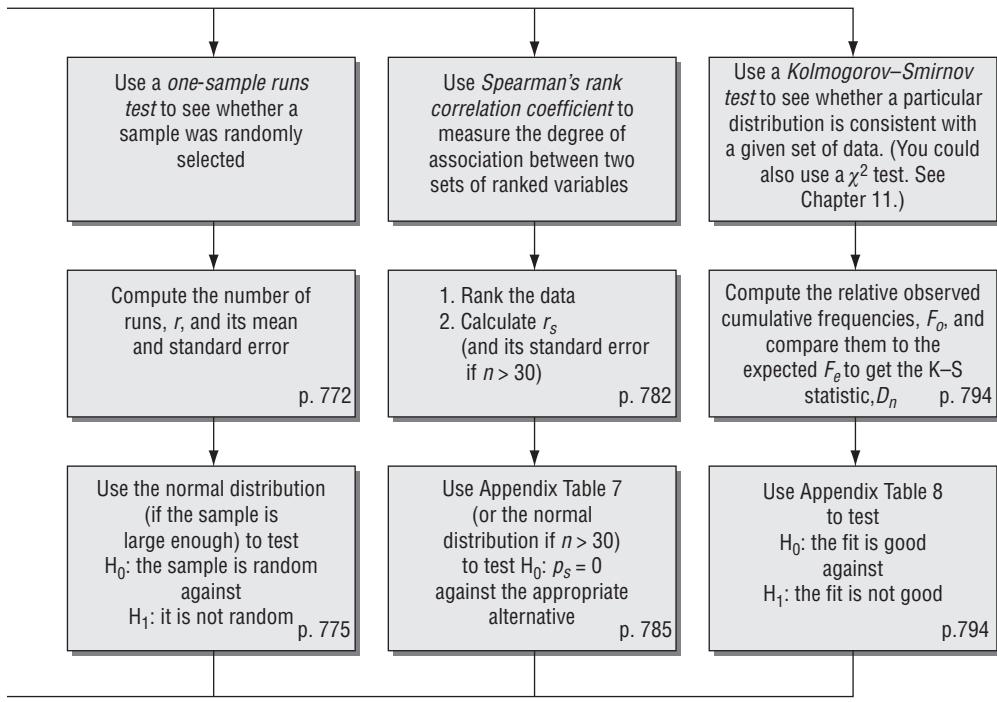
Questions on Running Case: SURYA Bank Pvt. Ltd.

1. Test whether the sample of respondents for this study were selected randomly on the basis of gender. (Runs test on Q15)
2. Test whether the sample of respondents for this study were selected randomly on the basis of marital status. (Runs test on Q16)
3. Test the normality of the variable “Level of satisfaction with e-services”. (Kolmogorov-Smirnov test to Q9)
4. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across the gender. (Mann Whitney U Test to Q9)
5. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different educational groups. (Kruskal_Wallis Test to Q9)
6. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different professions. (Kruskal_Wallis Test to Q9)
7. Test the hypothesis that the level of satisfaction of the customers with regards to the e-services provided by their banks is same across different age groups. (Kruskal_Wallis Test to Q9)



Flow Chart: Nonparametric Methods



Accept H_0

STOP

This page is intentionally left blank.

15

Time Series and Forecasting

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn why forecasting changes that take place over time are an important part of decision making
 - To understand the four components of a time series
 - To use regression-based techniques to estimate and forecast the trend in a time series
 - To learn how to measure the cyclical component of a time series
 - To compute seasonal indices and use them to deseasonalize a time series
 - To be able to recognize irregular variation in a time series
 - To deal simultaneously with all four components of a time series and to use time-series analysis for forecasting
-

CHAPTER CONTENTS

15.1 Introduction	818
15.2 Variations in Time Series	818
15.3 Trend Analysis	820
15.4 Cyclical Variation	832
15.5 Seasonal Variation	838
15.6 Irregular Variation	847
15.7 A Problem Involving All Four Components of a Time Series	848
15.8 Time-Series Analysis in Forecasting	858

■ Statistics at Work	858
■ Terms Introduced in Chapter 15	860
■ Equations Introduced in Chapter 15	860
■ Review and Application Exercises	861
■ Flow Chart: Time Series	867

The management of a ski resort has these quarterly occupancy data over a 5-year period:

Year	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
1991	1,861	2,203	2,415	1,908
1992	1,921	2,343	2,514	1,986
1993	1,834	2,154	2,098	1,799
1994	1,837	2,025	2,304	1,965
1995	2,073	2,414	2,339	1,967

To improve service, management must understand the seasonal pattern of demand for rooms. Using methods covered in this chapter, we shall help the hotel discern such a seasonal pattern, if it exists, and use it to forecast demand for rooms. ■

15.1 INTRODUCTION

Forecasting, or predicting, is an essential tool in any decision-making process. Its uses vary from determining inventory requirements for a local shoe store to estimating the annual sales of video games. The quality of the forecasts management can make is strongly related to the information that can be extracted and used from past data. *Time-series analysis* is one quantitative method we use to determine patterns in data collected over time. Table 15-1 is an example of time-series data.

Time-series analysis is used to detect patterns of change in *Use of time-series analysis* statistical information over regular intervals of time. We *project* these patterns to arrive at an estimate for the future. Thus, time-series analysis helps us cope with uncertainty about the future.

EXERCISES 15.1

Basic Concepts

- 15-1 Of what value are forecasts in the decision-making process?
- 15-2 For what purpose do we apply time-series analysis to data collected over a period of time?
- 15-3 How can one benefit from determining past patterns?
- 15-4 How would errors in forecasts affect a city government?

15.2 VARIATIONS IN TIME SERIES

We use the *term time series* to refer to any group of statistical information accumulated at regular intervals. There are four kinds of change, or variation, involved in time-series analysis:

Four kinds of variation in time-series

1. Secular trend
2. Cyclical fluctuation

TABLE 15.1 TIME SERIES FOR THE NUMBER OF SHIPS LOADED AT MOREHEAD CITY, N.C.

Year	1988	1989	1990	1991	1992	1993	1994	1995
Number	98	105	116	119	135	156	177	208

3. Seasonal variation
4. Irregular variation

With the first type of change, *secular trend*, the value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the Consumer Price Index is an example of secular trend. From year to individual year, the cost of living varies a great deal, but if we examine a long-term period, we see that the trend is toward a steady increase. Figure 15-1(a) shows a secular trend in an increasing but fluctuating time series.

The second type of variation seen in a time series is *cyclical fluctuation*. The most common example of cyclical fluctuation is the business cycle. Over time, there are years when the business cycle hits a peak above the trend line. At other times, business activity is likely to slump, hitting a low point below the trend line. The time between hitting peaks or falling to low points is at least 1 year, and it can be as many as 15 or 20 years. Figure 15-1(b) illustrates a typical pattern of cyclical fluctuation above and below a secular trend line. Note that the cyclical movements do not follow any regular pattern but move in a somewhat un-predictable manner.

The third kind of change in time-series data is *seasonal variation*. As we might expect from the name, seasonal variation involves patterns of change within a year that tend to be repeated from year to year. For example, a physician can expect a substantial increase in the number of flu cases every winter and of poison ivy every summer. Because these are regular patterns, they are useful in forecasting the future. In Figure 15-1(c), we see a seasonal variation. Notice how it peaks in the fourth quarter of each year.

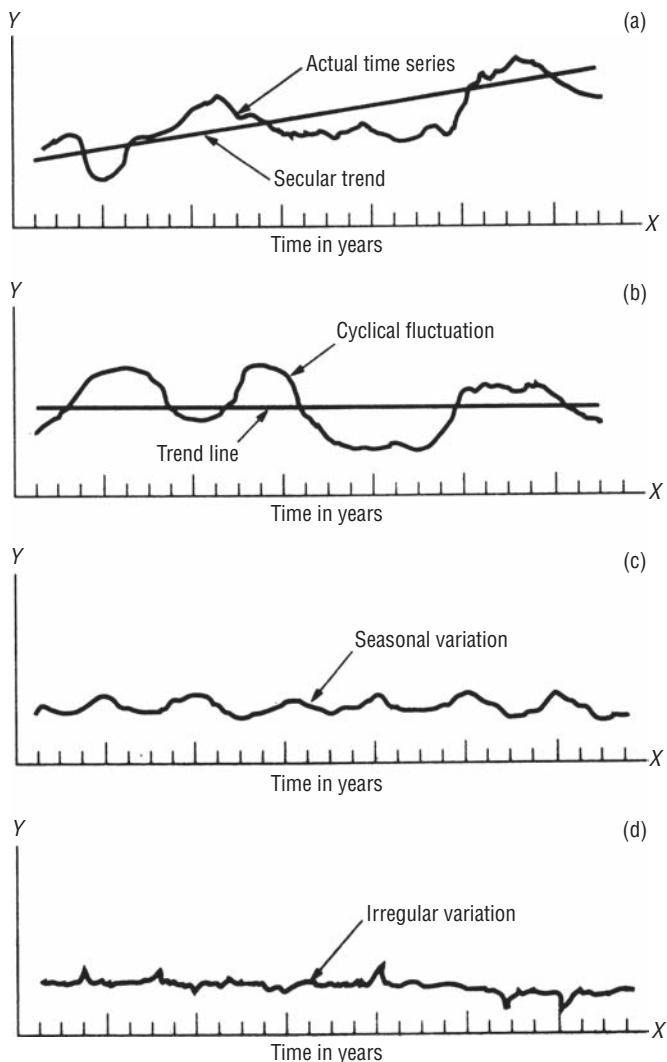
Irregular variation is the fourth type of change in time-series analysis. In many situations, the value of a variable may be completely unpredictable, changing in a random manner. Irregular variations describe such movements. The effects of the Middle East conflict in 1973, the Iranian situation in 1979–1981, the collapse of OPEC in 1986, and the Iraqi situation in 1990 on gasoline prices in the United States are examples of irregular variation. Figure 15-1(d) illustrates irregular variation.

Thus far, we have referred to a time series as exhibiting one or another of these four types of variation. In most instances, however, a time series will contain several of these components. Thus, we can describe the overall variation in a single time series in terms of these four different kinds of variation. In the following sections, we will examine the four components and the ways in which we measure each.

EXERCISES 15.2

Basic Concepts

- 15-5 Identify the four principal components of a time series and explain the kind of change, over time, to which each applies.
- 15-6 Which of the four components of a time series would we use to describe the effect of Christmas sales on a retail department store?
- 15-7 What is the advantage of decomposing a time series into its four components?
- 15-8 Which of the four components of a time series might the U.S. Department of Agriculture use to describe a 7-year weather pattern?
- 15-9 How would a war be accounted for in a time series?
- 15-10 What component of a time series explains the general growth and decline of the steel industry over the last two centuries?
- 15-11 Using the four kinds of variation, describe the behavior of crude oil prices from 1970 to 1987.

**FIGURE 15-1 TIME-SERIES VARIATIONS**

15.3 TREND ANALYSIS

Of the four components of a time series, secular trend represents the long-term direction of the series. One way to describe the trend component is to fit a line visually to a set of points on a graph. Any given graph, however, is subject to slightly different interpretations by different individuals. We can also fit a trend line by the method of least squares, which we examined in Chapter 12. In our discussion, we will concentrate on the method of least squares because visually fitting a line to a time series is not a completely dependable process.

Two methods of fitting a trend line

Reasons for Studying Trends

There are three reasons for why it is useful to study secular trends:

- 1. The study of secular trends allows us to describe a historical pattern.** There are many instances when we can use a past trend to evaluate the success of a previous policy. For example, a university may evaluate the effectiveness of a recruiting program by examining its past enrollment trends.
- 2. Studying secular trends permits us to project past patterns, or trends, into the future.** Knowledge of the past can tell us a great deal about the future. Examining the growth rate of the world's population, for example, can help us estimate the population for some future time.
- 3. In many situations, studying the secular trend of a time series allows us to eliminate the trend component from the series.** This makes it easier for us to study the other three components of the time series. If we want to determine the seasonal variation in ski sales, for example, eliminating the trend component gives us a more accurate idea of the seasonal component.

Three reasons for studying secular trends

Trends can be linear or curvilinear. Before we examine the linear, or straight-line, method of describing trends, we should remember that some relationships do not take that form. The increase of pollutants in the environment follows an upward sloping curve similar to that in Figure 15-2(a). Another common example of a curvilinear relationship is the life cycle of a new business product, illustrated in Figure 15-2(b). When a new product is introduced, its sales volume is low (I). As the product gains recognition and success, unit sales grow at an increasingly rapid rate (II). After the product is firmly established, its unit sales grow at a stable rate (III). Finally, as the product reaches the end of its life cycle, unit sales begin to decrease (IV).

Trend lines take different forms

Fitting the Linear Trend by the Least-Squares Method

Besides trends that can be described by a curved line, there are others that are described by a straight line. These are called linear trends. Before developing the equation for a linear trend, we need to review the general equation for estimating a straight line (Equation 12-3):

$$\text{Equation for estimating a straight line} \rightarrow \hat{Y} = a + bX \quad [12-3]$$

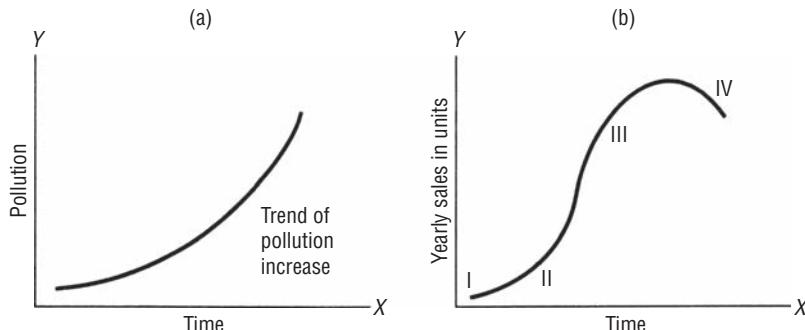


FIGURE 15-2 CURVILINEAR TREND RELATIONSHIPS

where

- \hat{Y} = estimated value of the dependent variable
- X = independent variable (*time* in trend analysis)
- a = *Y*-intercept (the value of *Y* when *X* = 0)
- b = slope of the trend line

We can describe the general trend of many time series using a straight line. But we are faced with the problem of finding the best-fitting line. As we did in Chapter 12, we can use the least-squares method to calculate the best-fitting line, or equation. There we saw that the best-fitting line was determined by Equations 12-4 and 12-5, which are now renumbered as Equations 15-1 and 15-2.

Finding the best-fitting trend line

Slope of the Best-Fitting Regression Line

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \quad [15-1]$$

Y-Intercept of the Best-Fitting Regression Line

$$a = \bar{Y} - b\bar{X} \quad [15-2]$$

where

- Y = values of the dependent variable
- X = values of the independent variable
- \bar{Y} = mean of the values of the dependent variable
- \bar{X} = mean of the values of the independent variable
- n = number of data points in the time series
- a = *Y*-intercept
- b = slope

With Equations 15-1 and 15-2, we can establish the best-fitting line to describe time-series data. However, the regularity of time-series data allows us to simplify the calculations in Equations 15-1 and 15-2 through the process we shall now describe.

Translating, or Coding, Time

Normally, we measure the independent variable *time* in terms such as *weeks*, *months*, and *years*. Fortunately, we can convert these traditional measures of time to a form that simplifies the computation.

Coding the time variable to simplify computation

In Chapter 3, we called this process *coding*. To use coding here, we find the mean time and then subtract that value from each of the sample times. Suppose our time series consists of only three points, 1992, 1993, and 1994. If we had to place these numbers in Equations 15-1 and 15-2, we would find the resultant calculations tedious. Instead, we can transform the values 1992, 1993, and 1994 into corresponding values of -1, 0, and 1, where 0 represents the mean (1993), -1 represents the first year (1992 - 1993 = -1), and 1 represents the last year (1994 - 1993 = 1).

Treating odd and even numbers of elements

We need to consider two cases when we are coding time values. The first is a time series with an *odd number of elements*, as in the

TABLE 15-2 TRANSLATING, OR CODING, TIME VALUES

(a) When there is an <i>odd</i> number of elements in the time series			(b) When there is an <i>even</i> number of elements in the time series		
$X(1)$	$X - \bar{X}$ (2)	Translated, or Coded, Time (3)	$X(1)$	$X - \bar{X}$ (2)	$(X - \bar{X}) \times 2$ (3)
1989	1989–1992 =	-3	1990	1990–1992 $\frac{1}{2}$ =	$-2\frac{1}{2} \times 2 =$
1990	1990–1992 =	-2	1991	1991–1992 $\frac{1}{2}$ =	$-1\frac{1}{2} \times 2 =$
1991	1991–1992 =	-1	1992	1992–1992 $\frac{1}{2}$ =	$-\frac{1}{2} \times 2 =$
1992	1992–1992 =	0	1993	1993–1992 $\frac{1}{2}$ =	$\frac{1}{2} \times 2 =$
1993	1993–1992 =	1	1994	1994–1992 $\frac{1}{2}$ =	$1\frac{1}{2} \times 2 =$
1994	1994–1992 =	2	1995	1995–1992 $\frac{1}{2}$ =	$2\frac{1}{2} \times 2 =$
1995	1995–1992 =	3			
$\Sigma X = 13,944$	\bar{x} (the mean year) = 0		$\Sigma X = 11,955$		\bar{x} (the mean year) = 0
$\bar{X} = \frac{\sum X}{n}$			$\bar{X} = \frac{\sum X}{n}$		
$= \frac{13,944}{7}$			$= \frac{11,955}{6}$		
$= 1992$			$= 1992\frac{1}{2}$		

previous example. The second is a series with an *even number of elements*. Consider Table 15-2. In part (a), on the left, we have an odd number of years. Thus, the process is the same as the one we just described, using the years 1992, 1993, and 1994. In part (b), on the right, we have an even number of elements. In cases like this, when we find the mean and subtract it from each element, the fraction $\frac{1}{2}$ becomes part of the answer. To simplify the coding process and to remove the $\frac{1}{2}$, we multiply each time element by 2. We will denote the “coded,” or translated, time with a lowercase x .

We have two reasons for this translation of time. First, it eliminates the need to square numbers as large as 1992, 1993, 1994, and so on. This method also sets the mean year, \bar{x} , equal to zero and allows us to simplify Equations 15-1 and 15-2.

Now we can return to our calculations of the slope (Equation 15-1) and the Y -intercept (Equation 15-2) to determine the best-fitting line. Because we are using the coded variable x , we replace X and \bar{X} by x and \bar{x} in Equations 15-1 and 15-2. Then, because the mean of our coded time variable \bar{x} is zero, we can substitute 0 for \bar{x} in Equations 15-1 and 15-2, as follows:

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \quad [15-1]$$

$$\begin{aligned} &= \frac{\sum xY - n\bar{x}\bar{Y}}{\sum x^2 - n\bar{x}^2} \leftarrow \begin{cases} \bar{x} \text{ the coded variable, substituted for } \bar{X} \\ \text{and } \bar{x} \text{ substituted for } \bar{X} \end{cases} \\ &= \frac{\sum xY - n0\bar{Y}}{\sum x^2 - n0^2} \leftarrow \bar{x} \text{ replaced by } 0 \end{aligned}$$

Why use coding?

Simplifying the calculation of a and b

Slope of the Trend Line for Coded Time Values

$$b = \frac{\sum xy}{\sum x^2} \quad [15-3]$$

Equation 15-2 changes as follows:

$$\begin{aligned} a &= \bar{Y} - b\bar{x} \\ &= \bar{Y} - b\bar{x} \leftarrow \bar{x} \text{ substituted for } \bar{x} \\ &= \bar{Y} - b0 \leftarrow \bar{x} \text{ replaced by 0} \end{aligned} \quad [15-2]$$

Intercept of the Trend Line for Coded Time Values

$$a = \bar{Y} \quad [15-4]$$

Equations 15-3 and 15-4 represent a substantial improvement over Equations 15-1 and 15-2.

A Problem Using the Least-Squares Method in a Time Series (Even Number of Elements)

Consider the data in Table 15-1, illustrating the number of ships loaded at Morehead City between 1988 and 1995. In this problem, we want to find the equation that will describe the secular trend of loadings. To calculate the necessary values for Equations 15-3 and 15-4, let us look at Table 15-3.

Using the least-squares method

TABLE 15.3 INTERMEDIATE CALCULATIONS FOR COMPUTING THE TREND

X (1)	Y [†] (2)	X - \bar{X} (3)	x (3) × 2 = (4)	xy (4) × (2)	x^2 (4) ²
1988	98	1988 - 1991 1/2 [#] = -3 1/2	-3 1/2 × 2 = -7	-686	49
1989	105	1989 - 1991 1/2 = -2 1/2	-2 1/2 × 2 = -5	-525	25
1990	116	1990 - 1991 1/2 = -1 1/2	-1 1/2 × 2 = -3	-348	9
1991	119	1991 - 1991 1/2 = -1/2	-1/2 × 2 = -1	-119	1
1992	135	1992 - 1991 1/2 = 1/2	1/2 × 2 = 1	135	1
1993	156	1993 - 1991 1/2 = 1 1/2	1 1/2 × 2 = 3	468	9
1994	177	1994 - 1991 1/2 = 2 1/2	2 1/2 × 2 = 5	885	25
1995	208	1995 - 1991 1/2 = 3 1/2	3 1/2 × 2 = 7	1,456	49
$\Sigma X = 15,932$	$\Sigma Y = 1,114$			$\Sigma xy = 1,266$	$\Sigma x^2 = 168$

$$\bar{X} = \frac{\sum X}{n} = \frac{15,932}{8} = 1,991 \frac{1}{2}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{1,114}{8} = 139.25$$

[†]Y is the number of ships.

[#]1991 1/2 corresponds to x = 0.

With these values, we can now substitute into Equations 15-3 and 15-4 to find the slope and the Y -intercept for the line describing the trend in ship loadings:

$$\begin{aligned} b &= \frac{\sum xy}{\sum x^2} \\ &= \frac{1,266}{168} \\ &= 7.536 \end{aligned} \quad [15-3]$$

and

$$\begin{aligned} a &= \bar{Y} \\ &= 139.25 \end{aligned} \quad [15-4]$$

Thus, the general linear equation describing the secular trend in ship loadings is

$$\begin{aligned} \hat{Y} &= a + bx \\ &= 139.25 + 7.536x \end{aligned} \quad [12-3]$$

where

- \hat{Y} = estimated annual number of ships loaded
- x = coded time value representing the number of *half-year* intervals (a minus sign indicates half-year intervals before 1991½; a plus sign indicates half-year intervals after 1991½)

Projecting with the Trend Equation

Once we have developed the trend equation, we can project it to forecast the variable in question. In the problem of finding the secular trend in ship loadings, for instance, we determined that the appropriate secular trend equation was

$$\hat{Y} = 139.25 + 7.536x$$

Now, suppose we want to estimate ship loadings for 1996. First, we must convert 1996 to the value of the coded time (in half-year intervals).

$$\begin{aligned} x &= 1996 - 1991\frac{1}{2} \\ &= 4.5 \text{ years} \\ &= 9 \text{ half-year intervals} \end{aligned}$$

Using our trend line to predict

Substituting this value into the equation for the secular trend, we get $= 139.25 + 67.82$

$$\begin{aligned} \hat{Y} &= 139.25 + 67.82 \\ &= 139.25 + 67.82 \\ &= 207 \text{ ships loaded} \end{aligned}$$

Therefore, we have estimated 207 ships will be loaded in 1996. If the number of elements in our time series had been odd, not even, our procedure would have been the same except that we would have dealt with 1-year intervals, not half-year intervals.

**Finding the slope and
Y-intercept**

Use of a Second-Degree Trend in a Time Series

So far, we have described the method of fitting a straight line to a time series. But many time series are best described by curves, not straight lines. In these instances, the linear trend model does not adequately describe the change in the variable as time changes. To overcome this problem, we often use a parabolic curve, which is described mathematically by a *second-degree equation*. Such a curve is illustrated in Figure 15-3. The general form for an estimated second-degree equation is

Handling time series that are described by curves

General Form for Fitted Second-Degree Curve

$$\hat{Y} = a + bx + cx^2$$

[15-5]

where

- \hat{Y} = estimate of the dependent variable
- a , b , and c = numerical constants
- x = coded values of the time variable.

Again we use the least-squares method to determine the second-degree equation to describe the best fit. The derivation of the second-degree equation is beyond the scope of this text. However, we can determine the value of the numerical constants (a , b , and c) from the following three equations:

Finding the values for a , b , and c

Least-Squares Coefficients for a Second-Degree Trend

Equations to find a ,
 b , and c to fit a parabolic curve

$$\longrightarrow \begin{cases} \sum Y = an + c \sum x^2 & [15-6] \\ \sum x^2 Y = a \sum x^2 + c \sum x^4 & [15-7] \\ b = \frac{\sum xY}{\sum x^2} & [15-3] \end{cases}$$

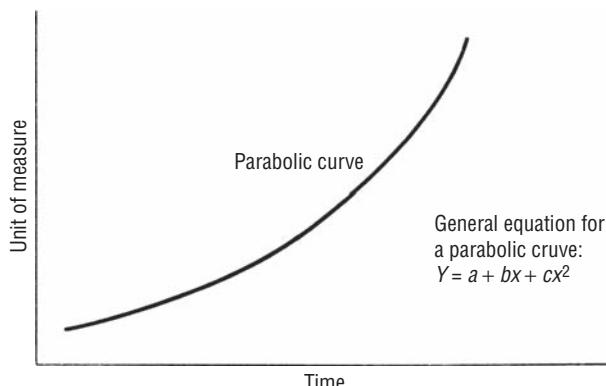


FIGURE 15-3 FORM AND EQUATION FOR A PARABOLIC CURVE

TABLE 15-4 ANNUAL SALES OF ELECTRIC QUARTZ WATCHES

X (year)	1991	1992	1993	1994	1995
Y (unit sales in millions)	13	24	39	65	106

When we find the values of a , b , and c by solving Equations 15-6, 15-7, and 15-3 simultaneously, we substitute these values into the second-degree equation, Equation 15-5.

As in describing a linear relationship, we transform the independent variable, time (X), into a coded form (x) to simplify the calculation. We'll now work through a problem in which we fit a parabolic trend to a time series.

A Problem Involving a Parabolic Trend (Odd Number of Elements in the Time Series)

In recent years, the sale of electric quartz watches has increased at a significant rate. Table 15-4 contains sales information that will help us determine the parabolic trend describing watch sales.

We organize the necessary calculations in Table 15-5. The first step in this process is to translate the independent variable X into a coded time variable x . Note that the coded variable x is listed in 1-year intervals because there is an odd number of elements in our time series. Thus, it is not necessary to multiply the variable by 2.

Substituting the values from Table 15-5 into Equations 15-6, 15-7, and 15-3, we get

$$247 = 5a + 10c \quad \text{①}$$

$$565 = 10a + 34c \quad \text{②} \quad [15-7]$$

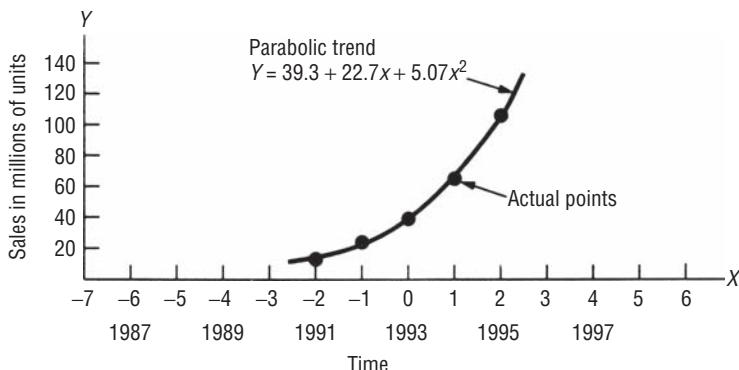
$$b = \frac{227}{10} \quad \text{③} \quad [15-3]$$

From ③, we see that

$$b = 22.7$$

We can find a and c by solving equations ① and ② simultaneously. When we do this, we find that a is 39.3 and c is 5.07.

TABLE 15-5 INTERMEDIATE CALCULATIONS FOR COMPUTING THE TREND

**FIGURE 15-4 PARABOLIC TREND FITTED TO DATA IN TABLE 15-4**

This gives us the appropriate values of a , b , and c to describe the time series presented in Table 15-4 by the following equation

$$\begin{aligned}\hat{Y} &= a + bx + cx^2 \\ &= 39.3 + 22.7x + 5.07x^2\end{aligned}\quad [15-5]$$

Let's graph the watch data to see how well the parabola we just derived fits the time series. We've done this in Figure 15-4.

Does our curve fit the data?

Forecasts Based on a Second-Degree Equation

Suppose we want to forecast watch sales for 2000. To make a prediction, we must first translate 2000 into a coded variable x by subtracting the mean year, 1993.

Making the forecast

$$\begin{aligned}X - \bar{X} &= X \\ 2000 - 1993 &= 7\end{aligned}$$

This coded value ($x = 7$) is then substituted into the second-degree equation describing watch sales:

$$\begin{aligned}\hat{Y} &= 39.3 + 22.7x + 5.07x^2 \\ &= 39.3 + 22.7(7) + 5.07(7)^2 \\ &= 39.3 + 158.9 + 248.4 \\ &= 446.6\end{aligned}$$

We conclude, based on the past secular trend, that watch sales should be approximately 446,600,000 units by 2000. This extraordinarily large forecast suggests, however, that we must be more careful in forecasting with a parabolic trend than we are when using a linear trend. The slope of the second-degree equation in Figure 15-4 is continually increasing. Therefore, the parabolic trend may become a poor estimator as we attempt to predict further into the future. In using the second-degree-equation method, we must also take into consideration factors that may be slowing or reversing the growth rate of the variable.

In our watch example, we can assume that during the time period under consideration, the product is at a very rapid growth stage in its life cycle. But we must realize that as the cycle approaches a

Being careful in interpreting the forecast

mature stage, sales will probably decelerate and no longer be predicted accurately by our parabolic curve. When we calculate predictions for the future, we need to consider the possibility that the trend line may *change*. Such a situation could cause considerable error. It is therefore necessary to exercise particular care when using a second-degree equation as a forecasting tool.

HINTS & ASSUMPTIONS

Warning: “No tree grows to the sky.” that’s a Wall Street proverb meaning that no stock price rises forever. It’s also true here for forecasts made with second-degree equations. Extrapolating the growth rate of a startup company (which starts with zero sales so a dollar of sales is automatically an *infinite* growth rate) is risky. Early growth rates seldom continue.

EXERCISES 15.3

Self-Check Exercises

SC 15-1 Robin Zill and Stewart Griffiths own a small company that manufactures portable massage tables in Hillsborough, North Carolina. Since they started the company, the number of tables they have sold is represented by this time series:

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Tables sold	42	50	61	75	92	111	120	127	140	138

- (a) Find the linear equation that describes the trend in the number of tables sold by Robin and Stewart.
- (b) Estimate their sales of tables in 1998.

SC 15-2 The number of faculty-owned personal computers at the University of Ohio increased dramatically between 1990 and 1995:

Year	1990	1991	1992	1993	1994	1995
Number of PCs	50	110	350	1,020	1,950	3,710

- (a) Develop a linear estimating equation that best describes these data.
- (b) Develop a second-degree estimating equation that best describes these data.
- (c) Estimate the number of PCs that will be in use at the university in 1999, using both equations.
- (d) If there are 8,000 faculty members at the university, which equation is the better predictor? Why?

Applications

15-12 Jeff Richardson invested his life savings and began a part-time carpet-cleaning business in 1986. Since 1986, Jeff’s reputation has spread and business has increased. The average numbers of homes he has cleaned per month each year are:

Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Homes cleaned	6.4	11.3	14.7	18.4	19.6	25.7	32.5	48.7	55.4	75.7	94.3

(a) Find the linear equation that describes the trend in these data.

(b) Estimate the number of homes cleaned per month in 1997, 1998, and 1999.

- 15-13** The owner of Progressive Builders is examining the number of solar homes started in the region in each of the last 7 months:

Month	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number of homes	16	17	25	28	32	43	50

(a) Plot these data.

(b) Develop the linear estimating equation that best describes these data, and plot the line on the graph from part (a) (let x units equal 1 month).

(c) Develop the second-degree estimating equation that best describes these data and plot this curve on the graph from part (a).

(d) Estimate March sales using both curves you have plotted.

- 15-14** Richard Jackson developed an ergonomically superior computer mouse in 1989, and sales have been increasing ever since. Data are presented below in terms of thousands of mice sold per year.

Year	1989	1990	1991	1992	1993	1994	1995	1996
Number sold	82.4	125.7	276.9	342.5	543.6	691.5	782.4	889.5

(a) Develop a linear estimating equation that best describes these data.

(b) Develop a second-degree estimating equation that best describes these data.

(c) Estimate the number of mice that will be sold in 1998, using both equations.

(d) If we assume the rate of increase in mouse sales will decrease soon based on supply and demand, which model would be a better predictor for your answer in part (c)?

- 15-15** Mike Godfrey, the auditor of a state public school system, has reviewed the inventory records to determine whether the current inventory holdings of textbooks are typical. The following inventory amounts are from the previous 5 years:

Year	1991	1992	1993	1994	1995
Inventory (\$1,000)	\$4,620	\$4,910	\$5,490	\$5,730	\$5,990

(a) Find the linear equation that describes the trend in the inventory holdings.

(b) Estimate for him the value of the inventory for the year 1996.

- 15-16** The following table describes first-class postal rates from 1968 to 1996:

Year	1968	1970	1972	1974	1976	1978	1980	1982	1984	1986	1988	1990	1992	1994	1996
Rate (¢)	5	5	8	8	10	13	15	18	20	22	25	25	29	29	32

(a) Develop the linear estimating equation that best describes these data.

(b) Develop the second-degree estimating equation that best describes these data.

(c) Is there anything in the economic or political environment that would suggest that one or the other of these two equations is likely to be the better predictor of postal rates?

- 15-17** Envirotech Engineering, a company that specializes in the construction of antipollution filtration devices, has recorded the following sales record over the last 9 years:

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995
Sales (\$100,000)	13	15	19	21	27	35	47	49	57

- (a) Plot these data.
 (b) Develop the linear estimating equation that best describes these data, and plot this line on the graph from part (a).
 (c) Develop the second-degree estimating equation that best describes these data, and plot this curve on the graph from part (a).
 (d) Does the market to the best of your knowledge favor (b) or (c) as the more accurate estimating method?

15-18 Here are data describing the air pollution rate (in ppm of particles in the air) in a western city:

Year	1980	1985	1990	1995
Pollution rate	220	350	800	2,450

- (a) Would a linear or a second-degree estimating equation provide the better prediction of future pollution in that city?
 (b) Considering the economic, social, and political environment, would you change your answer to part (a)?
 (c) Describe how political and social action could change the effectiveness of either of the estimating equations in part (a).

15-19 The State Department of Motor Vehicles is studying the number of traffic fatalities in the state resulting from drunk driving for each of the last 9 years.

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995
Deaths	175	190	185	195	180	200	185	190	205

- (a) Find the linear equation that describes the trend in the number of traffic fatalities in the state resulting from drunk driving.
 (b) Estimate the number of traffic fatalities resulting from drunk driving that the state can expect in 1996.

Worked-Out Answers to Self-Check Exercises

SC 15-1 (a)	Year	x	Y	xY	x^2
	1987	-9	42	-378	81
	1988	-7	50	-350	49
	1989	-5	61	-305	25
	1990	-3	75	-225	9
	1991	-1	92	-92	1
	1992	1	111	111	1
	1993	3	120	360	9
	1994	5	127	635	25
	1995	7	140	980	49
	1996	9	138	1242	81
	0	956	1,978	330	

$$a = \bar{Y} = \frac{956}{10} = 95.6 \quad b = \frac{\sum xY}{\sum x^2} = \frac{1,978}{330} = 5.9939$$

$$\hat{Y} = 95.6 + 5.9939x \text{ (when } 1991.5 = 0 \text{ and } x \text{ units} = 0.5 \text{ year)}$$

$$(b) \hat{Y} = 95.6 + 5.9939(13) = 173.5 \text{ tables}$$

SC 15-2

Year	x	Y	xY	x^2	x^2Y	x^4
1990	-5	50	-250	25	1,250	625
1991	-3	110	-330	9	990	81
1992	-1	350	-350	1	350	1
1993	1	1,020	1,020	1	1,020	1
1994	3	1,950	5,850	9	17,550	81
1995	5	3,710	18,550	25	92,750	625
0	7,190	24,490	70	113,910	1,414	

$$(a) \quad a = \bar{Y} = \frac{7,190}{6} = 1,198.3333 \quad b = \frac{\sum xY}{\sum x^2} = \frac{24,490}{70} = 349.8571$$

$$\hat{Y} = 1,198.3333 + 349.8571x \text{ (where } 1992.5 = 0 \text{ and } x \text{ units} = 0.5 \text{ year)}$$

- (b) Equations 15.6 and 15.7 become

$$\Sigma Y = na + c \sum x^2 \quad 7,190 = 6a + 70c$$

$$\Sigma x^2 Y = a \sum x^2 + c \sum x^4 \quad 113,910 = 70a + 1,414c$$

Solving these simultaneously, we get

$$a = 611.8750, c = 50.2679$$

$$\hat{Y} = 611.8750 + 349.8571x + 50.2679x^2$$

- (c) Linear forecast: $\hat{Y} = 1,198.3333 + 349.8571(13) = 5,746$ PCs

$$\begin{aligned} \text{Second-degree equation forecast: } \hat{Y} &= 611.8750 + 349.8571(13) + 50.2679(169) \\ &= 13,655 \text{ PCs} \end{aligned}$$

- (d) Neither is very good. The linear trend missed the acceleration in the rate of faculty PC acquisition. The second-degree trend assumed the acceleration would continue, ignoring the fact that there are only 8,000 faculty members.

15.4 CYCLICAL VARIATION

Cyclical variation is the component of a time series that tends to oscillate above and below the secular trend line for periods longer than 1 year. The procedure used to identify cyclical variation is the residual method.

Cyclical variation defined

Residual Method

When we look at a time series consisting of annual data, only the secular-trend, cyclical, and irregular components are considered. (This is true because seasonal variation makes a complete, regular cycle within each year and thus does not affect one year any more than another.) Because we can describe secular trend using a trend line, we can isolate the remaining cyclical and irregular components from the trend. We will assume that the cyclical component explains most of the variation left unexplained by the trend component. (Many real-life time series do not satisfy this assumption. Methods such as Fourier analysis and spectral analysis can analyze the cyclical component for such time series. However, these are beyond the scope of this book.)

If we use a time series composed of annual data, we can find the fraction of the trend by dividing the actual value (Y) by the corresponding trend value (\hat{Y}) for each value in the time series. We then multiply the result of this calculation by 100. This gives us the measure of cyclical variation as a *percent of trend*. We express this process in Equation 15-8:

*Expressing cyclical variation
as a percent of trend*

Percent of Trend	
$\frac{Y}{\hat{Y}} \times 100$	[15-8]

where

- Y = actual time-series value
- \hat{Y} = estimated trend value from the same point in the time series

Now let's apply this procedure.

A farmers' marketing cooperative wants to measure the variations in its members' wheat harvest over an 8-year period. Table 15-6 shows the volume harvested in each of the 8 years. Column Y contains the values of the linear trend for each time period. The trend line has been generated using the methods illustrated in Section 3 of this chapter. Note that when we graph the actual (Y) and the trend (\hat{Y}) values for the 8 years in Figure 15-5, the actual values move above and below the trend line.

Measuring variation

Now we can determine the percent of trend for each of the years in the sample (column 4 in Table 15-7). From this column, we can see the variation in actual harvests around the estimated trend (98.7 to 102.5). We can attribute these cyclical variations to factors such as rainfall and temperature. However, because these factors are relatively unpredictable, we cannot forecast any specific patterns of variation using the method of residuals.

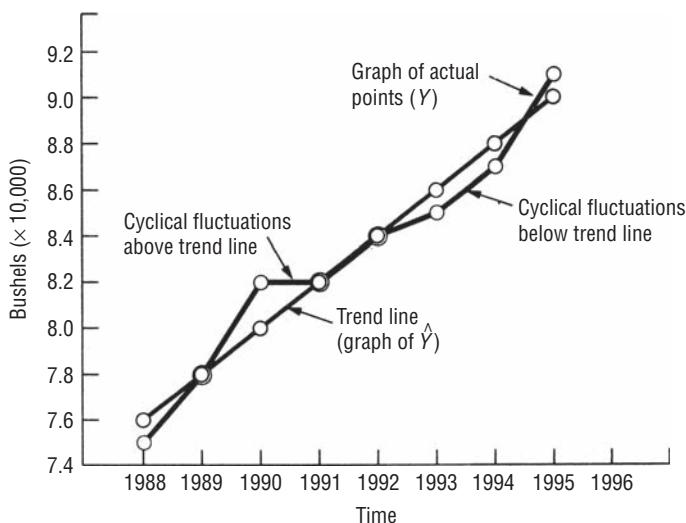
Interpreting cyclical variations

The *relative cyclical residual* is another measure of cyclical variation. In this method, the *percentage deviation* from the trend is found for each value. Equation 15-9 presents the mathematical formula for determining the relative cyclical residuals. As with percent of trend, this measure is also a percentage.

Expressing cyclical variations in terms of relative cyclical residual

TABLE 15.6 GRAIN RECEIVED BY FARMERS' COOPERATIVE OVER 8 YEARS

<i>X</i> Year	<i>Y</i> Actual Bushels ($\times 10,000$)	\hat{Y} Estimated Bushels ($\times 10,000$)
1988	7.5	7.6
1989	7.8	7.8
1990	8.2	8.0
1991	8.2	8.2
1992	8.4	8.4
1993	8.5	8.6
1994	8.7	8.8
1995	9.1	9.0

**FIGURE 15-5 CYCLICAL FLUCTUATIONS AROUND THE TREND LINE****Relative Cyclical Residual**

$$\frac{Y - \hat{Y}}{\hat{Y}} \times 100 \quad [15-9]$$

where

- Y = actual time-series value
- \hat{Y} = estimated trend value from the same point in the time series

Table 15-8 shows the calculation of the relative cyclical residual for the farmers' cooperative problem. Note that the easy way to compute the relative cyclical residual (column 5) is to subtract 100 from the percent of trend (column 4).

TABLE 15.7 CALCULATION OF PERCENT OF TREND

X Year (1)	Y Actual Bushels ($\times 10,000$) (2)	\hat{Y} Estimated Bushels ($\times 10,000$) (3)	$\frac{Y}{\hat{Y}} \times 100$	Percent of Trend (4) = $\frac{(2)}{(3)} \times 100$
1988	7.5	7.6	98.7	
1989	7.8	7.8	100.0	
1990	8.2	8.0	102.5	
1991	8.2	8.2	100.0	
1992	8.4	8.4	100.0	
1993	8.5	8.6	98.8	
1994	8.7	8.8	98.9	
1995	9.1	9.0	101.1	

TABLE 15-8 CALCULATION OF RELATIVE CYCLICAL RESIDUALS

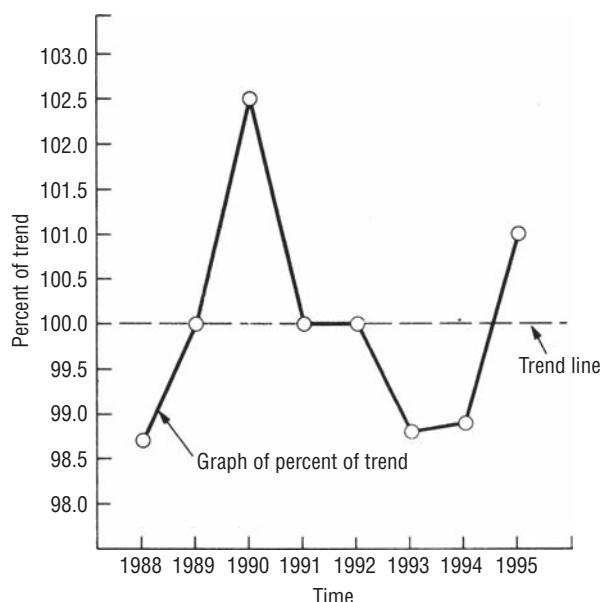
<i>X</i> Year (1)	<i>Y</i> Actual Bushels ($\times 10,000$) (2)	\hat{Y} Estimated Bushels ($\times 10,000$) (3)	$\frac{Y}{\hat{Y}} \times 100$ Percent of Trend (4) = $\frac{(2)}{(3)} \times 100$	$\frac{Y - \hat{Y}}{\hat{Y}} \times 100$ Relative Cyclical Residual (5) = (4) - 100
1988	7.5	7.6	98.7	-1.3
1989	7.8	7.8	100.0	0.0
1990	8.2	8.0	102.5	2.5
1991	8.2	8.2	100.0	0.0
1992	8.4	8.4	100.0	0.0
1993	8.5	8.6	98.8	-1.2
1994	8.7	8.8	98.9	-1.1
1995	9.1	9.0	101.1	1.1

These two measures of cyclical variation, percent of trend and relative cyclical residual, are percentages of the trend. For example, in 1993, the *percent of trend* indicated that the actual harvest was 98.8 percent of the expected harvest for that year. For the same year, the *relative cyclical residual* indicated that the actual harvest was 1.2 percent short of the expected harvest (a relative cyclical residual of -1.2).

We often graph cyclical variation as the percent of trend. Figure 15-6 illustrates how this process eliminates the trend line and isolates

Comparing the two measures of cyclical variation

Graphing cyclical variation

**FIGURE 15-6** GRAPH OF PERCENT OF TREND AROUND THE TREND LINE FOR THE DATA IN TABLE 15-7

the cyclical component of the time series. It must be emphasized that the procedures discussed in this section can be used only for describing past cyclical variations and not for predicting future cyclical variations. Predicting cyclical variation requires the use of techniques which are beyond the scope of this book.

HINTS & ASSUMPTIONS

Remember that *cyclical variation* is the component of a time series that oscillates above and below the trend line for periods *longer* than a year. Warning: *Seasonal variation* makes a complete cycle *within* each year and does not affect one year any more than another. Cyclical variation is measured by two methods. The first method expresses the variation as a percentage *of* the trend, hence its name *percent of trend*. The second method (the *relative cyclical residual*) calculates the variation as a percent deviation *from the trend*.

EXERCISES 15.4

Self-Check Exercise

SC 15-3 The Western Natural Gas Company has supplied 18, 20, 21, 25, and 26 billion cubic feet of gas, respectively, for the years 1991 to 1995.

- Find the linear estimating equation that best describes these data.
- Calculate the percent of trend for these data.
- Calculate the relative cyclical residual for these data.
- In which years does the largest fluctuation from trend occur, and is it the same for both methods?

Applications

15-20 Microprocessing, a computer firm specializing in software engineering, has compiled the following revenue records for the years 1989 to 1995

Year	1989	1990	1991	1992	1993	1994	1995
Revenue ($\times \$100,000$)	1.1	1.5	1.9	2.1	2.4	2.9	3.5

The second-degree equation that best describes the secular trend for these data is

$$\hat{Y} = 2.119 + 0.375x + 0.020x^2, \text{ where } 1992 = 0, \text{ and } x \text{ units} = 1 \text{ year}$$

- Calculate the percent of trend for these data.
- Calculate the relative cyclical residual for these data.
- Plot the percent of trend from part (a).
- In which year does the largest fluctuation from trend occur, and is it the same for both methods?

15-21 The Bulls Eye department store has been expanding market share during the past 7 years, posting the following gross sales in millions of dollars:

Year	1990	1991	1992	1993	1994	1995	1996
Sales	14.8	20.7	24.6	32.9	37.8	47.6	51.7

- (a) Find the linear estimating equation that best describes the data.
 (b) Calculate the percent of trend for these data.
 (c) Calculate the relative cyclical residual for these data.
 (d) In which years does the largest fluctuation from trend occur, and is it the same for both methods?

15-22 Joe Honeg, the sales manager responsible for the appliance division of a large consumer-products company, has collected the following data regarding unit sales for his division during the last 5 years:

Year	1991	1992	1993	1994	1995
Units ($\times 10,000$)	32	46	50	66	68

The equation describing the secular trend for appliance sales is

$$\hat{Y} = 52.4 + 9.2x, \text{ where } 1993 = 0, \text{ and } x \text{ units} = 1 \text{ year}$$

- (a) Calculate the percent of trend for these data.
 (b) Calculate the relative cyclical residual for these data.
 (c) Plot the percent of trend from part (a).
 (d) In which year does the largest fluctuation from trend occur, and is it the same for both methods?

15-23 Suppose you are the capital budgeting officer of a small corporation whose financing requirements over the last few years have been

Year	1989	1990	1991	1992	1993	1994	1995
Millions of dollars required	2.2	2.1	2.4	2.6	2.7	2.9	2.8

The trend equation that best describes these data is

$$\hat{Y} = 2.53 + 0.13x, \text{ where } 1992 = 0, \text{ and } x \text{ units} = 1 \text{ year}$$

- (a) Calculate the percent of trend for these data.
 (b) Calculate the relative cyclical residual for these data.
 (c) In which year does the largest fluctuation from trend occur, and is it the same for both methods?
 (d) As the capital budgeting officer, what would this fluctuation mean for you and the activities you perform?

15-24 Parallel Breakfast Foods has data on the number of boxes of cereal it has sold in each of the last 7 years.

Year	1989	1990	1991	1992	1993	1994	1995
Boxes ($\times 10,000$)	21.0	19.4	22.6	28.2	30.4	24.0	25.0

- (a) Find the linear estimating equation that best describes these data.
 (b) Calculate the percent of trend for these data.
 (c) Calculate the relative cyclical residual for these data.
 (d) In which year does the biggest fluctuation from the trend occur under each measure of cyclical variation? Is this year the same for both measures? Explain.

15-25 Wombat Airlines, an Australian company, has gathered data on the number of passengers who have flown on its planes during each of the last 5 years.

Year	1991	1992	1993	1994	1995
Passengers (in tens of thousands)	3.5	4.2	3.9	3.8	3.6

- Find the linear estimating equation that best describes these data.
- Calculate the percent of trend for these data.
- Calculate the relative cyclical residual for these data.
- Based on the data and your previous calculations, give a one-sentence summary of the position in which Wombat Airlines finds itself.

Worked-Out Answer to Self-Check Exercise

SC 15-3

Year	x	Y	xy	x^2	\hat{Y}	$\frac{Y}{\hat{Y}} \times 100$	$\frac{Y - \hat{Y}}{\hat{Y}} \times 100$
1991	-2	18	-36	4	17.8	101.12	1.12
1992	-1	20	-20	1	19.9	100.50	0.50
1993	0	21	0	0	22.0	95.45	-4.55
1994	1	25	25	1	24.1	103.73	3.73
1995	2	26	52	4	26.2	99.24	-0.76
	0	110	21	10			

$$(a) a = \bar{Y} = \frac{110}{5} = 22 \quad b = \frac{\sum xy}{\sum x^2} = \frac{21}{10} = 2.1$$

$$\hat{Y} = 22 + 2.1x \text{ (where 1993 = 0 and } x \text{ units = 1 year)}$$

- See the next-to-the-last column above for percent of trend.
- See the last column above for relative cyclical residual.
- Largest fluctuation (by both methods) was in 1993.

15.5 SEASONAL VARIATION

Besides secular trend and cyclical variation, a time series also includes seasonal variation. *Seasonal variation* is defined as repetitive and predictable movement around the trend line in *one year or less*. In order to detect seasonal variation, time intervals must be measured in small units, such as days, weeks, months, or quarters.

Seasonal variation defined

We have three main reasons for studying seasonal variation:

- We can establish the pattern of past changes.** This gives us a way to compare two time intervals that would otherwise be too dissimilar. If a flight training school wants to know if a slump in business during December is normal, it can examine the seasonal pattern in previous years and find the information it needs.
- It is useful to project past patterns into the future.** In the case of long-range decisions, secular-trend analysis may be adequate. But for short-run decisions, the ability to predict seasonal fluctuations is often essential. Consider a wholesale food chain that wants to maintain a minimum adequate stock

Three reasons for studying seasonal variation

of all items. The ability to predict short-range patterns, such as the demand for turkeys at Thanksgiving, candy at Christmas, or peaches in the summer, is useful to the management of the chain.

- Once we have established the seasonal pattern that exists, we can eliminate its effects from the time series. This adjustment allows us to calculate the cyclical variation that takes place each year. When we eliminate the effect of seasonal variation from a time series, we have *deseasonalized* the time series.

Ratio-to-Moving-Average Method

In order to measure seasonal variation, we typically use the *ratio-to-moving-average method*. This technique provides an *index* that describes the degree of seasonal variation. The index is based on a mean of 100, with the degree of seasonality measured by variations away from the base. For example, if we examine the seasonality of canoe rentals at a summer resort, we might find that the spring-quarter index is 142. The value 142 indicates that 142 percent of the average quarterly rental occur in the spring. If management recorded 2,000 canoe rentals for all of last year, then the average quarterly rental would be $2,000/4 = 500$. Because the spring-quarter index is 142, we estimate the number of spring rentals as follows:

$$\text{Spring-quarter index} \downarrow \\ \text{Average quarterly rental} \longrightarrow 500 \times \frac{142}{100} = 710 \longleftarrow \text{Seasonalized spring quarter rental}$$

Using the ratio-to-moving-average method of measuring seasonal variation

Our chapter-opening example can illustrate the ratio-to-moving-average method. The resort hotel wanted to establish the seasonal pattern of room demand by its clientele. Hotel management wants to improve customer service and is considering several plans to employ personnel during peak periods to achieve this goal. Table 15-9 contains the quarterly occupancy, that is, the average number of guests during each quarter of the last 5 years.

An example of the ratio-to-moving-average method

We will refer to Table 15-9 to demonstrate the six steps required to compute a seasonal index.

- The first step in computing a seasonal index is to calculate the 4-quarter moving total for the time series. To do this, we total the values for the quarters during the first year, 1991, in

Step 1: Calculate the 4-quarter moving total

TABLE 15-9 TIME SERIES FOR HOTEL OCCUPANCY

Year	Number of Guests per Quarter			
	I	II	III	IV
1991	1,861	2,203	2,415	1,908
1992	1,921	2,343	2,514	1,986
1993	1,834	2,154	2,098	1,799
1994	1,837	2,025	2,304	1,965
1995	2,073	2,414	2,339	1,967

Table 15-9: $1,861 + 2,203 + 2,415 + 1,908 = 8,387$. A moving total is associated with the middle data point in the set of values from which it was calculated. Because our first total of 8,387 was calculated from four data points, we place it opposite the midpoint of those quarters, so it falls in column 4 of Table 15-10, between the rows for the 1991-II and 1991-III quarters.

We find the next moving total by dropping the 1991-1 value, 1,861, and adding the 1992-1 value, 1,921. By dropping the first value and adding the fifth, we keep four quarters in the total. The four values added now are $2,203 + 2,415 + 1,908 + 1,921 = 8,447$. This total is entered in Table 15-10 directly below the first quarterly total of 8,387. We continue the process of “sliding” the 4-quarter total over the time series until we have included the last value in the series. In this example, it is the 1,967 rooms in the fourth quarter of 1995, the last number in column 3 of Table 15-10. The last entry in the moving total column is 8,793. It is between the rows for the 1995-II and 1995-III quarters because it was calculated from the data for the 4 quarters of 1995.

2. In the second step, we compute the 4-quarter moving average by dividing each of the 4-quarter totals by 4.

Step 2: Compute the 4-quarter moving average

In Table 15-10, we divided the values in column 4 by 4, to arrive at the values for column 5.

3. In the third step, we center the 4-quarter moving average.

Step 3: Center the 4-quarter moving average

The moving averages in column 5 all fall halfway between the quarters. We would like to have moving averages associated with each quarter. In order to *center* our moving averages, we associate with each quarter the average of the two 4-quarter moving averages falling just above and just below it. For the 1991-III quarter, the resulting **4-quarter centered moving average** is 2,104.25, that is, $(2,096.75 + 2,111.75)/2$. The other entries in column 6 are calculated the same way. Figure 15-7 illustrates how

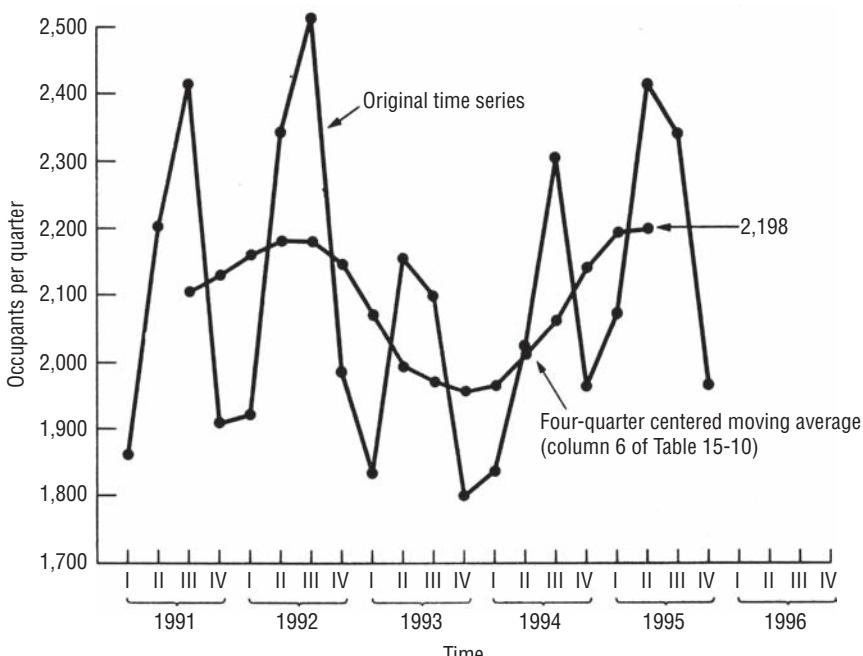


FIGURE 15-7 USING A MOVING AVERAGE TO SMOOTH THE ORIGINAL TIME SERIES

the moving average has smoothed the peaks and troughs of the original time series. The seasonal and irregular components have been smoothed, and the resulting dotted colored line represents the cyclical and trend components of the series.

Suppose we were working with the admissions data for a hospital emergency room, and we wanted to compute *daily* indices. In steps 1 and 2, we would compute 7-day moving totals and moving averages, **and the moving averages would already be centered** (because the middle of a 7-day period is the fourth of those 7 days). In this case, step 3 is unnecessary. Whenever the number of periods for which we want indices is odd (7 days in a week, three shifts in a day), we can skip step 3. However, when the number of periods is even (4 quarters, 12 months, 24 hours), then we must use step 3 to center the moving averages we get with step 2.

Sometimes step 3 can be skipped

4. Next, we calculate the percentage of the actual value to the moving-average value for each quarter in the time series having a 4-quarter moving-average entry. This step allows us to recover the seasonal component for the quarters. We determine this percentage by dividing each of the actual quarter values in column 3 of Table 15-10 by the corresponding 4-quarter centered moving-average values

Step 4: Calculate the percentage of actual value to moving average value

TABLE 15-10 CALCULATING THE 4-QUARTER CENTERED MOVING AVERAGE

Year (1)	Quarter (2)	Occupancy (3)	Step 1: 4-Quarter Moving Total (4)	Step 2: 4-Quarter Moving Total (5) = (4) ÷ 4	Step 3: 4-Quarter Centered Moving Average (6)	Step 4: Percentage of Actual to Moving Average Values (7) = $\frac{(3)}{(6)} \times 100$
1991	I	1,861	8,387	2,096.75	2,104.250	114.8
	II	2,203				
	III	2,415				
	IV	1,908				
1992	I	1,921	8,587	2,146.75	2,159.125	89.0
	II	2,343				
	III	2,514				
	IV	1,986				
1993	I	1,834	8,488	2,122.00	2,070.000	88.6
	II	2,154				
	III	2,098				
	IV	1,799				
1994	I	1,837	7,759	1,939.75	1,965.500	93.5
	II	2,025				
	III	2,304				
	IV	1,965				
1995	I	2,073	8,756	2,189.00	2,193.375	94.5
	II	2,414				
	III	2,339				
	IV	1,967				

in column 6 and then multiplying the result by 100. For example, we find the percentage for 1991-III as follows:

$$\frac{\text{Actual}}{\text{Moving average}} \times 100 = \frac{2,415}{2,104.250} \times 100 \\ = 114.8$$

5. To collect all the percentage of actual to moving-average values in column 7 of Table 15-10, arrange them by quarter.

Step 5: Collect answers from step 4 and calculate the modified mean

Then calculate the *modified mean* for each quarter. The modified mean is calculated by discarding the highest and lowest values for each quarter and averaging the remaining values. In Table 15-11, we present the fifth step and show the process for finding the modified mean.

The seasonal values we recovered for the quarters in column 7 of Table 15-10 still contain the cyclical and irregular components of variation in the time series. By eliminating the highest and lowest values from each quarter, we *reduce* the extreme cyclical and irregular variations. When we average the remaining values, we further smooth the cyclical and irregular components. Cyclical and irregular variations tend to be removed by this process, so the modified mean is an

Reducing extreme cyclical and irregular variations

TABLE 15-11 DEMONSTRATION OF STEP 5 IN COMPUTING A SEASONAL INDEX*

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1991	—	—	114.8	89.6
1992	89.0	107.4	115.3	92.6
1993	88.6	108.0	106.4	92.0
1994	93.5	100.6	111.7	91.8
1995	94.5	109.8	—	—
	<u>182.5</u>	<u>215.4</u>	<u>226.5</u>	<u>183.8</u>

Modified mean:

$$\text{Quarter I: } \frac{182.5}{2} = 91.25$$

$$\text{Quarter II: } \frac{215.4}{2} = 107.70$$

$$\text{Quarter III: } \frac{226.5}{2} = 113.25$$

$$\text{Quarter IV: } \frac{183.8}{2} = 91.90$$

$$\text{Total of indices} = 404.1$$

*Eliminated values are indicated by a colored slash.

TABLE 15-12 DEMONSTRATION OF STEP 6

Quarter	Unadjusted Indices	\times	Adjusting Constant	=	Seasonal Index
I	91.25	\times	0.9899	=	90.3
II	107.70	\times	0.9899	=	106.6
III	113.25	\times	0.9899	=	112.1
IV	91.90	\times	0.9899	=	91.0
			Total of seasonal indices	=	400.0
			Mean of seasonal indices	=	$\frac{400}{4}$ = 100..0

index of the seasonality component. (Some statisticians prefer to use the median value instead of computing the modified mean to achieve the same outcome.)

- 6. The final step, demonstrated in Table 15-12, adjusts the modified mean slightly.** Notice that the four indices in Table 15-11 total 404.1. However, the base for an index is 100. Thus, the four quarterly indices should total 400, and their mean should be 100. To correct for this error, we multiply each of the quarterly indices in Table 15-11 by an adjusting constant. This number is found by dividing the desired sum of the indices (400) by the actual sum (404.1). In this case, the result is 0.9899. Table 15-12 shows that multiplying the indices by the adjusting constant brings the quarterly indices to a total of 400. (Sometimes even after this adjustment, the mean of the seasonal indices is not exactly 100 because of accumulated rounding errors. In this case, however, it is exactly 100.)

Uses of the Seasonal Index

The ratio-to-moving-average method just explained allows us to identify seasonal variation in a time series. The seasonal indices are used to remove the effects of seasonality from a time series. This is called *deseasonalizing* a time series. Before we can identify either the trend or cyclical components of a time series, we must eliminate seasonal variation. To deseasonalize a time series, we divide each of the actual values in the series by the appropriate seasonal index (expressed as a fraction of 100). To demonstrate, we shall deseasonalize the value of the first four quarters in Table 15-9. In Table 15-13, we show the deseasonalizing process using the values for the seasonal indices from Table 15-12. Once the seasonal effect has been eliminated, the deseasonalized values that remain reflect only the trend, cyclical, and irregular components of the time series.

Once we have removed the seasonal variation, we can compute a deseasonalized trend line, which we can then project into the future. Suppose the hotel management in our example estimates from a deseasonalized trend line that the deseasonalized average occupancy for the fourth quarter of the *next* year will be 2,121. When this prediction has been obtained, management must then take the seasonality into account. To do this, it multiplies the deseasonalized predicted average occupancy of 2,121 by the fourth-quarter seasonal index

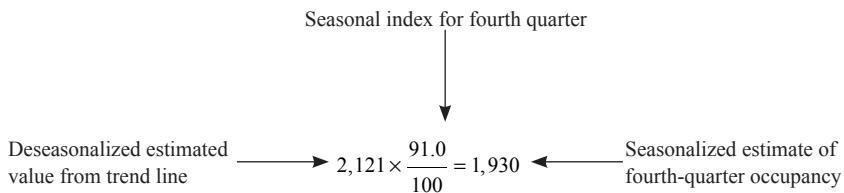
Deseasonalizing a time series

Using seasonality in forecasts

TABLE 15-13 DEMONSTRATION OF DESEASONALIZING DATA

Year (1)	Quarter (2)	Actual Occupancy (3)	$\left(\frac{\text{Seasonal Index}}{100} \right)$ (4)	Deseasonalized Occupancy (5) = (3) ÷ (4)
1991	I	1,861	$\div \left(\frac{90.3}{100} \right)$	= 2,061
1991	II	2,203	$\div \left(\frac{106.6}{100} \right)$	= 2,067
1991	III	2,415	$\div \left(\frac{112.1}{100} \right)$	= 2,154
1991	IV	1,908	$\div \left(\frac{91.0}{100} \right)$	= 2,097

(expressed as a fraction of 100) to obtain a seasonalized estimate of 1,930 rooms for the fourth-quarter average occupancy. Here are the calculations:



HINTS & ASSUMPTIONS

Using seasonal indices to adjust quarterly and monthly data helps us detect the underlying secular trend. Warning: Most reported figures fail to tell us how much seasonal adjustment was used, and in some management decisions this missing information is valuable. For example, if a state motor vehicle department reports last month's new vehicle registrations were 25,000 at a *seasonally adjusted rate*, how would a distributor of an after-market automobile product such as custom floor mats predict demand for *next month* without knowing the *actual* number of new cars? Often, for internal company planning purposes, it is helpful to know both adjusted and unadjusted figures.

EXERCISES 15.5

Self-Check Exercise

- SC 15-4** Using the following percentages of actual to moving average describing the quarterly amount of cash in circulation at the Village Bank in Carrboro, N.C. over a 4-year period, calculate the seasonal index for each quarter.

	Spring	Summer	Fall	Winter
Year				
1992	87	106	86	125
1993	85	110	83	127
1994	84	105	87	128
1995	88	104	88	124

Applications

- 15-26** The owner of The Pleasure-Glide Boat Company has compiled the following quarterly figures regarding the company's level of accounts receivable over the last 5 years ($\times \$1,000$):

	Spring	Summer	Fall	Winter
Year				
1991	102	120	90	78
1992	110	126	95	83
1993	111	128	97	86
1994	115	135	103	91
1995	122	144	110	98

- (a) Calculate a 4-quarter centered moving average.
- (b) Find the percentage of actual to moving average for each period.
- (c) Determine the modified seasonal indices and the seasonal indices.

- 15-27** Marie Wiggs, personnel director for a pharmaceutical company, recorded these percentage absentee rates for each quarter over a 4-year period:

	Spring	Summer	Fall	Winter
Year				
1992	5.6	6.8	6.3	5.2
1993	5.7	6.7	6.4	5.4
1994	5.3	6.6	6.1	5.1
1995	5.4	6.9	6.2	5.3

- (a) Construct a 4-quarter centered moving average and plot it on a graph along with the original data.
 - (b) What can you conclude about absenteeism from part (a)?
- 15-28** Using the following percentages of actual to moving average describing the seasonal sales of sporting goods over a 5-year period, calculate the seasonal index for each season.

Year	Baseball	Football	Basketball	Hockey
1992	96	128	116	77
1993	92	131	125	69
1994	84	113	117	84
1995	97	118	126	89
1996	91	121	124	81

- 15-29** A large manufacturer of automobile springs has determined the following percentages of actual to moving average describing the firm's quarterly cash needs for the last 6 years:

	Spring	Summer	Fall	Winter
1990	108	128	94	70
1991	112	132	88	68
1992	109	134	84	73
1993	110	131	90	69
1994	108	135	89	68
1995	106	129	93	72

Calculate the seasonal index for each quarter. Comment on how it compares to the indices you calculated for Exercise 15-26.

- 15-30** A university's dean of admissions has compiled the following quarterly enrollment figures for the previous 5 years ($\times 100$):

	Fall	Winter	Spring	Summer
1991	220	203	193	84
1992	235	208	206	76
1993	236	206	209	73
1994	241	215	206	92
1995	239	221	213	115

- (a) Calculate a 4-quarter centered moving average.
- (b) Find the percentage of actual to moving average for each period.
- (c) Determine the modified seasonal indices and the seasonal indices.

- 15-31** The Ski and Putt Resort, a combination of ski slopes and golf courses, has just recently tabulated its data on the number of customers (in thousands) it has had during each season of the last 5 years. Calculate the seasonal index for each quarter. If 15 people are employed in the summer, what should winter employment be, assuming both sports have equal labor requirements?

	Spring	Summer	Fall	Winter
1991	200	300	125	325
1992	175	250	150	375
1993	225	300	200	450
1994	200	350	225	375
1995	175	300	200	350

- 15-32** David Curl Builders has collected quarterly data on the number of homes it has started during the last 5 years.

	Spring	Summer	Fall	Winter
1991	8	10	7	5
1992	9	10	7	6
1993	10	11	7	6
1994	10	12	8	7
1995	11	13	9	8

- (a) Calculate the seasonal index for each quarter.
 (b) If David's working capital needs are related directly to the number of starts, by how much should his working capital need decrease between summer and winter?

Worked-Out Answer to Self-Check Exercise

SC 15-4	Year	Spring	Summer	Fall	Winter
	1992	87	106	86	125
	1993	85	110	83	127
	1994	84	105	87	128
	1995	88	104	88	124
	Modified sum	172	211	173	252
	Modified mean	86	105.5	86.5	126
	Seasonal index	85.15	104.46	85.64	124.75

The sum of the modified means was 404, so the adjusting factor was $400/404 = 0.9901$. The seasonal indices were obtained by multiplying the modified means by this factor.

15.6 IRREGULAR VARIATION

The final component of a time series is irregular variation. After we have eliminated trend, cyclical, and seasonal variations from a time series, we still have an unpredictable factor left. Typically, irregular variation occurs over short intervals and follows a random pattern.

Difficulty of dealing with irregular variation

Because of the unpredictability of irregular variation, we do not attempt to explain it mathematically. However, we can often isolate its causes. New York City's financial crisis of 1975, for example, was an irregular factor that severely depressed the municipal bond market. In 1984, the unusually cold temperatures in late December in the southern states were an irregular factor that significantly increased electricity and fuel oil consumption. The Persian Gulf War in 1991 was another irregular factor; it significantly increased airline and ship travel for a number of months as troops and supplies were moved. Not all causes of irregular variation can be identified so easily, however. One factor that allows managers to cope with irregular variation is that over time, these random movements tend to counteract each other.

HINTS & ASSUMPTIONS

Warning: Irregular variation is *very* important but is not explainable mathematically. It's what is "left over" after we eliminate trend, cyclical, and seasonal variation from a time series. In most cases, irregular variation is difficult if not impossible to predict and we never attempt to "fit a line" to account for irregular variation. Hint: Often you will find irregular variation acknowledged with a footnote or a comment on a graph. Examples of this would be "Market closed for Labor Day Holiday" or "Spring break fell in March instead of April last year."

EXERCISES 15.6

Basic Concepts

- 15-33** Why don't we project irregular variations into the future?
- 15-34** Which of the following illustrate irregular variations?
- An extended drought leading to higher food prices.
 - The effect of snow on ski slope business.
 - A one-time federal tax rebate provision for the purchase of new houses.
 - The collapse of crude oil prices in early 1986.
 - The energy use reduction after the 1973 oil embargo.
- 15-35** Make a list of five irregular variations in time series that you deal with as a part of your daily routine.
- 15-36** What allows management to cope with irregular variation in time series?

15.7 A PROBLEM INVOLVING ALL FOUR COMPONENTS OF A TIME SERIES

For a problem that involves all four components of a time series, we turn to a firm that specializes in producing recreational equipment. To forecast future sales based on an analysis of its past pattern of sales, the firm has collected the information in Table 15-14. Our procedure for describing this time series will consist of three stages:

1. Deseasonalizing the time series
2. Developing the trend line
3. Finding the cyclical variation around the trend line

Because the data are available on a quarterly basis, we must first deseasonalize the time series. The steps to do this are shown in Tables 15-15 and 15-16. These steps are the same as those originally introduced in Section 15-5.

In Table 15-15, we have tabulated the first four steps in computing the seasonal index. In Table 15-16, we complete the process.

Once we have computed the quarterly seasonal indices, we can find the deseasonalized values of the time series by dividing the actual sales (in Table 15-14) by the seasonal indices. Table 15-17 (on page 851) shows the calculation of the deseasonalized time-series values.

Step 1: Computing seasonal indices

Finding the deseasonalized values

TABLE 15-14 QUARTERLY SALES

Year	Sales per Quarter ($\times \$10,000$)			
	I	II	III	IV
1991	16	21	9	18
1992	15	20	10	18
1993	17	24	13	22
1994	17	25	11	21
1995	18	26	14	25

TABLE 15-15 CALCULATION OF THE FIRST FOUR STEPS TO COMPUTE THE SEASONAL INDEX

Year (1)	Quarter (2)	Actual Sales (3)	Step 1: 4-Quarter Moving Total (4)	Step 2: 4-Quarter Moving Average (5) = $\frac{(4)}{4}$	Step 3: 4-Quarter Centered Moving Average (6)	Step 4: Percentage of Actual to Moving Average (7) = $\frac{(3)}{(6)} \times 100$
1991	I	16				
	II	21				
	III	9	64	16.00	15.875	56.7
	IV	18	63	15.75	15.625	115.2
1992	I	15	62	15.50	15.625	96.0
	II	20	63	15.75	15.750	127.0
	III	10	63	15.75	16.000	62.5
	IV	18	65	16.25	16.750	107.5
1993	I	17	69	17.25	17.625	96.5
	II	24	72	18.00	18.500	129.7
	III	13	76	19.00	19.000	68.4
	IV	22	76	19.00	19.125	115.0
1994	I	17	77	19.25	19.000	89.5
	II	25	75	18.75	18.625	134.2
	III	11	74	18.50	18.625	59.1
	IV	21	75	18.75	18.875	111.3
1995	I	18	76	19.00		
	II	26	79	19.75	19.375	92.9
	III	14	83	20.75	20.250	128.4
	IV	25				

The second step in describing the components of the time series is to develop the trend line. We accomplish this by applying the least-squares method to the deseasonalized time series (after we have translated the time variable). Table 15-18 presents the calculations to identify the trend component (see page 852).

Step 2: Developing the trend line using the least-squares method

With the values from Table 15-18, we can now find the equation for the trend. From Equations 15-3 and 15-4, we find the slope and Y-intercept for the trend line as follows:

$$\begin{aligned}
 b &= \frac{\sum xY}{\sum x^2} \\
 &= \frac{420.3}{2,660} \\
 &= 0.16
 \end{aligned} \tag{15-3}$$

$$\begin{aligned}
 a &= \bar{Y} \\
 &= 18.0
 \end{aligned} \tag{15-4}$$

TABLE 15-16 STEPS 5 AND 6 IN COMPUTING THE SEASONAL INDEX

Year	Step 5*				
	I	II	III	IV	
1991	—	—	56.7	115.2	
1992	96.0	127.0	62.5	107.5	
1993	96.5	129.7	68.4	115.0	
1994	89.5	134.2	59.1	111.3	
1995	92.9	128.4	—	—	
Modified sum = 188.9		258.1	121.6	226.3	
Modified mean: Quarter I: $\frac{188.9}{2} = 94.45$					
II: $\frac{258.1}{2} = 129.05$					
III: $\frac{121.6}{2} = 60.80$					
IV: $\frac{226.3}{2} = \frac{113.15}{397.45}$					
Step 6†					
Adjusting factor = $\frac{400}{397.45} = 1.0064$					
Quarter	Indices	×	Adjusting Factor	=	Seasonal Indices
I	94.45	×	1.0064	=	95.1
II	129.05	×	1.0064	=	129.9
III	60.80	×	1.0064	=	61.2
IV	113.15	×	1.0064	=	113.9
Sum of seasonal indices				=	400.1

*Arrange percentages-from column 7, Table 15-15, by quarter and find the modified mean.

†Correcting the indices in step 5.

The appropriate trend line is described using the straight-line equation (Equation 12-3), with X replaced by x :

$$\hat{Y} = a + bx \\ = 18 + 0.16x \quad [12-3]$$

We have now identified the seasonal and trend components of the time series. Next, we find the cyclical variation around the trend line. This component is identified by measuring deseasonalized variation around the trend line. In this problem, we will calculate cyclical variation in Table 15-19, using the residual method (see page 853).

Step 3: Finding the cyclical variation

TABLE 15-17 CALCULATION OF DESEASONALIZED TIME-SERIES VALUES

Year (1)	Quarter (2)	Actual Sales (3)	Seasonal Index		Deseasonalized Sales (5) = (3) ÷ (4)
			100 (4)		
1991	I	16	0.951		16.8
	II	21	1.299		16.2
	III	9	0.612		14.7
	IV	18	1.139		15.8
1992	I	15	0.951		15.8
	II	20	1.299		15.4
	III	10	0.612		16.3
	IV	18	1.139		15.8
1993	I	17	0.951		17.9
	II	24	1.299		18.5
	III	13	0.612		21.2
	IV	22	1.139		19.3
1994	I	17	0.951		17.9
	II	25	1.299		19.2
	III	11	0.612		18.0
	IV	21	1.139		18.4
1995	I	18	0.951		18.9
	II	26	1.299		20.0
	III	14	0.612		22.9
	IV	25	1.139		21.9

If we assume that irregular variation is generally short-term and relatively insignificant, we have completely described the time series in this problem using the trend, cyclical, and seasonal components. Figure 15-8 (on page 835) illustrates the original time series, its moving average (containing both the trend and cyclical components), and the trend line.

Now, suppose that the management of the recreation company we have been using as an example wants to estimate the sales volume for the third quarter of 1996. What should management do?

It has to determine the deseasonalized value for sales in the third quarter of 1996 by using the trend equation, $\hat{Y} = 18 + 0.16x$. This requires it to code the time, 1996-III. That quarter (1996-III) is three quarters past 1995-IV, which, we see in Table 15-18, has a coded time value of 19. Adding 2 for each quarter, management finds $x = 19 + 2(3) = 25$. Substituting this value ($x = 25$) into the trend equation produces the following result:

$$\begin{aligned}\hat{Y} &= a + bx \\ &= 18 + 0.16(25) \\ &= 18 + 4 \\ &= 22\end{aligned}$$

Assumptions about irregular variation

Predictions using a time series

Step 1: Determining the deseasonalized value for sales for the period desired

TABLE 15-18 IDENTIFYING THE TREND COMPONENT

Year (1)	Quarter (2)	Desesonalized Sales (Column 5 of Table 15-17) ($\frac{1}{2}x$) Translating or ($\times \$10,000$) (3)	Coding the Time Variable (4)		x (5) = (4) $\times 2$	xY (6) = (5) \times (3)	x^2 (7) = (5) 2
1991	I	16.8		-9 1/2	-19	-319.2	361
	II	16.2		-8 1/2	-17	-275.4	289
	III	14.7		-7 1/2	-15	-220.5	225
	IV	15.8		-6 1/2	-13	-205.4	169
1992	I	15.8		-5 1/2	-11	-173.8	121
	II	15.4		-4 1/2	-9	-138.6	81
	III	16.3		-3 1/2	-7	-114.1	49
	IV	15.8		-2 1/2	-5	-79.0	25
1993	I	17.9		-1 1/2	-3	-53.7	9
	II	18.5		-1/2	-1	-18.5	1
Mean			→ 0*				
	III	21.2		1/2	1	21.2	1
	IV	19.3		1 1/2	3	57.9	9
1994	I	17.9		2 1/2	5	89.5	25
	II	19.2		3 1/2	7	134.4	49
	III	18.0		4 1/2	9	162.0	81
	IV	18.4		5 1/2	11	202.4	121
1995	I	18.9		6 1/2	13	245.7	169
	II	20.0		7 1/2	15	300.0	225
	III	22.9		8 1/2	17	389.3	289
	IV	21.9		9 1/2	19	416.1	361
		$\Sigma Y = 360.9$				$\Sigma xY = 420.3$	$\Sigma x^2 = 2,660$
$\hat{Y} = \frac{\sum Y}{n}$ $= \frac{360.9}{20}$ $= 18.0$							

*We assign the mean of 0 to the middle of the data (1993-II 1/2) and then measure the translated time, x , by 1/2 quarters because we have an even number of periods.

Thus, the deseasonalized sales estimate for 1996-III is \$220,000. This point is shown on the trend line in Figure 15-8.

4. Now management must seasonalize this estimate by multiplying it by the third-quarter seasonal index, expressed as a fraction of 100:

Seasonal index for quarter III from step 6 of Table 15-16

↓

Trend estimate from Equation 12-3 → $22 \times \frac{61.2}{100} = 13.5$ ← Seasonalized estimate

Step 2: Seasonalizing the initial estimate

TABLE 15-19 IDENTIFYING THE CYCLICAL VARIATION

Year (1)	Quarter (2)	<i>Y</i> Deseasonalized Sales (Column 5, Table 15-17)	$a + bx = \hat{Y}^*$ (4)	$\frac{Y}{\hat{Y}} \times 100$ Percent of Trend (15) = $\frac{(3)}{(4)} \times 100$
1991	I	16.8	$18 + 0.16(-19) = 14.96$	112.3
	II	16.2	$18 + 0.16(-17) = 15.28$	106.0
	III	14.7	$18 + 0.16(-15) = 15.60$	94.2
	IV	15.8	$18 + 0.16(-13) = 15.92$	99.2
1992	I	15.8	$18 + 0.16(-11) = 16.24$	97.3
	II	15.4	$18 + 0.16(-9) = 16.56$	93.0
	III	16.3	$18 + 0.16(-7) = 16.88$	96.6
	IV	15.8	$18 + 0.16(-5) = 17.20$	91.9
1993	I	17.9	$18 + 0.16(-3) = 17.52$	102.2
	II	18.5	$18 + 0.16(-1) = 17.84$	103.7
	III	21.2	$18 + 0.16(1) = 18.16$	116.7
	IV	19.3	$18 + 0.16(-3) = 18.48$	104.4
1994	I	17.9	$18 + 0.16(-5) = 18.80$	95.2
	II	19.2	$18 + 0.16(-7) = 19.12$	100.4
	III	18.0	$18 + 0.16(-9) = 19.44$	92.6
	IV	18.4	$18 + 0.16(-11) = 19.76$	93.1
1995	I	18.9	$18 + 0.16(-13) = 20.08$	94.1
	II	20.0	$18 + 0.16(-15) = 20.40$	98.0
	III	22.9	$18 + 0.16(-17) = 20.72$	110.5
	IV	21.9	$18 + 0.16(-19) = 21.04$	104.1

*The appropriate value for x in this equation is obtained from column 5 of Table 15-18.

On the basis of this analysis, the firm estimates that sales for 1996-III will be \$135,000. We must stress, however, that this value is only an estimate and does not take into account the cyclical and irregular components. As we noted earlier, the irregular variation cannot be predicted mathematically. Also, remember that our earlier treatment of cyclical variation was descriptive of past behavior and not predictive of future behavior.

Caution in using the forecast

HINTS & ASSUMPTIONS

A complete analysis of a time series tries to account for *secular trend*, *cyclical variation*, and *seasonal variation*. What's left is *irregular variation*. Warning: Even the best analysis of a time series describes past behavior, and may not be predictive of future behavior. Hint: The correct way to proceed in analyzing all of the components of a time series is to first deseasonalize the time series, then find the trend line, then calculate the cyclical variation around that trend line, and then identify irregular variation from what is left.

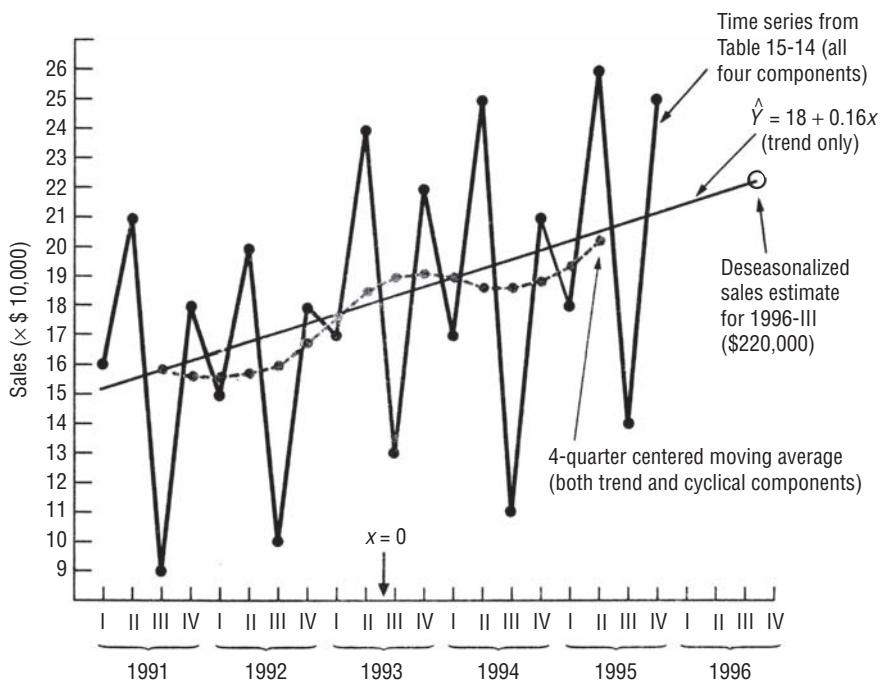


FIGURE 15-8 TIME SERIES, TREND LINE, AND 4-QUARTER CENTERED MOVING AVERAGE FOR QUARTERLY SALES DATA IN TABLE 15-14

EXERCISES 15.7

Self-Check Exercise

SC 15-5 A state commission designed to monitor energy consumption assembled the following seasonal data regarding natural gas consumption, in millions of cubic feet:

	Winter	Spring	Summer	Fall
1992	293	246	231	282
1993	301	252	227	291
1994	304	259	239	296
1995	306	265	240	300

- Determine the seasonal indices and deseasonalize these data (using a 4-quarter centered moving average).
- Calculate the least-squares line that best describes these data.
- Identify the cyclical variation in these data by the relative cyclical residual method.
- Plot the original data, the deseasonalized data, and the trend.

Applications

- 15-37** An environmental agency has been watching New York air quality over a 5-year period and has assembled the following seasonal data regarding amount of contaminants (in parts per million) in the air:

Year	Winter	Spring	Summer	Fall
1992	452	385	330	385
1993	474	397	356	399
1994	494	409	375	415
1995	506	429	398	437
1996	527	454	421	482

- (a) Determine the seasonal indices and deseasonalize these data (using a 4-quarter centered moving average).
- (b) Calculate the least-squares line that best describes these data.
- (c) Identify the cyclic variation in these data by the relative cyclical residual method.
- (d) Plot the original data, the deseasonalized data, and the trend.

- 15-38** The following data describe the marketing performance of a regional beer producer:

Year	Sales by Quarter ($\times \$100,000$)			
	I	II	III	IV
1991	19	24	38	25
1992	21	28	44	23
1993	23	31	41	23
1994	24	35	48	21

- (a) Calculate the seasonal indices for these data. (Use a 4-quarter centered moving average.)
- (b) Deseasonalize these data using the indices from part (a).

- 15-39** For Exercise 15-38:

- (a) Find the least-squares line that best describes the trend in deseasonalized beer sales.
- (b) Identify the cyclical component in this time series by computing the percent of trend.

Worked-Out Answer to Self-Check Exercise

SC 15-5 (a)

Year	Quarter	Actual Gas Usage	4-Quarter Moving Average	Centered Moving Average	Percentage of Actual to Moving Average	Seasonal Index	Deseasonalized Usage
1992	Winter	293				111.66	262.4037
	Spring	246				94.39	260.6208
	Summer	231	263.00	264.000	87.50	86.82	266.0677
	Fall	282	265.00	265.750	106.11	107.13	263.2316
1993	Winter	301	266.50	266.000	113.16	111.66	269.5683
	Spring	252	265.50	266.625	94.51	94.39	266.9774
	Summer	227	267.75	268.125	84.66	86.82	261.4605
	Fall	291	268.50	269.375	108.03	107.13	271.6326
1994	Winter	304	270.25	271.750	111.87	111.66	272.2551
	Spring	259	273.25	273.875	94.57	94.39	274.3935
	Summer	239	274.50	274.750	86.99	86.82	275.2822
	Fall	296	275.00	275.750	107.34	107.13	276.2998
1995	Winter	306	276.50	276.625	110.62	111.66	274.0462
	Spring	265	276.75	277.250	95.58	94.39	280.7501
	Summer	240	277.75			86.82	276.4340
	Fall	300				107.13	280.0336

Year	Winter	Spring	Summer	Fall
1992			87.50	106.11
1993	113.16	94.51	84.66	108.03
1994	111.87	94.57	86.99	107.34
1995	110.62	95.58		
Modified sum	111.87	94.57	86.99	107.34
Seasonal index	111.66	94.39	86.82	107.13

The sum of the modified means was 400.77, so the adjusting factor was $400/400.77 = 0.99808$. The seasonal indices were obtained by multiplying the modified means by this factor.

(b, c)

Year	Quarter	Deseasonalized Usage (\hat{Y})	Deseasonalized trend $\hat{Y} = 270.7161 + 0.6301x$			Relative Cyclical Residual $\frac{\hat{Y} - Y}{\hat{Y}} \times 100$	
			x	xY	x^2		
1992	Winter	262.4037	-15	-3936.0555	225	261.2646	0.44
	Spring	260.6208	-13	-3383.0704	169	262.5248	-0.73
	Summer	266.0677	-11	-2926.7447	121	263.7850	0.87
	Fall	263.2316	-9	-2369.0844	81	265.0452	-0.68
1993	Winter	269.5683	-7	-1886.9781	49	266.3054	1.23
	Spring	266.9774	-5	-1334.8870	25	267.5656	-0.22

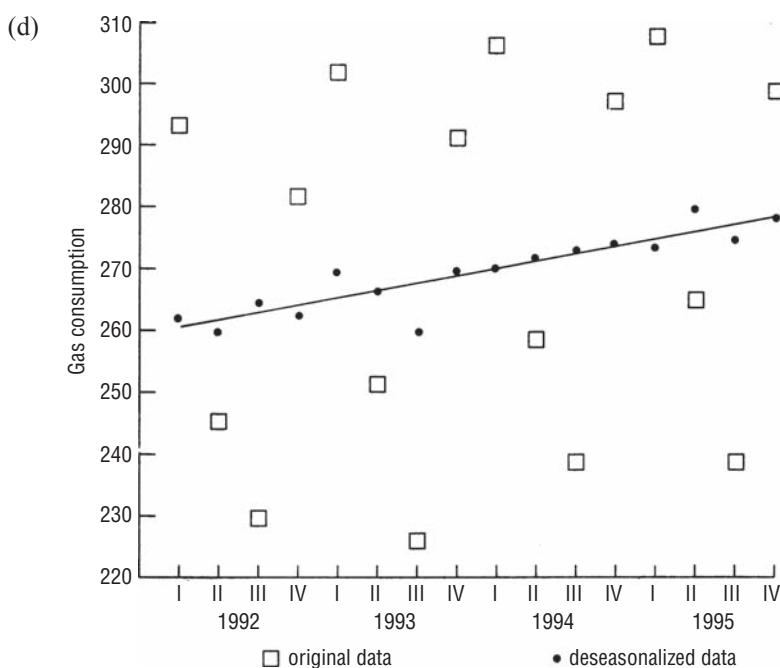
(Continued)

(Contd.)

Year	Quarter	Deseasonalized Usage (\bar{Y})			x	$x\bar{Y}$	x^2	Deseasonalized trend $\hat{Y} = 270.7161 + 0.6301x$	Relative Cyclical Residual $\frac{\bar{Y} - \hat{Y}}{\hat{Y}} \times 100$
1994	Summer	261.4605	-3	-784.3815	9	268.8258	-2.74		
	Fall	271.6326	-1	-271.6326	1	270.0860	0.57		
	Winter	272.2551	1	272.2551	1	271.3462	0.33		
	Spring	274.3935	3	823.1805	9	272.6064	0.66		
	Summer	275.2822	5	1376.4110	25	273.8666	0.52		
	Fall	276.2998	7	1934.0986	49	275.1268	0.43		
1995	Winter	274.0462	9	2466.4158	81	276.3870	-0.85		
	Spring	280.7501	11	3088.2511	121	277.6472	1.12		
	Summer	276.4340	13	3593.6420	169	278.9074	-0.89		
	Fall	280.0336	15	4200.5040	225	280.1676	-0.05		
		4,331.4571	0	856.9239	1,360				

$$a = \bar{Y} = \frac{4,331.4571}{16} = 270.7161 \quad b = \frac{\sum x\bar{Y}}{\sum x^2} = \frac{856.9239}{1,360} = 0.6301$$

$$\hat{Y} = 270.7161 + 0.6301x \text{ (where 1993-IV } \frac{1}{2} = 0 \text{ and } x \text{ units} = \frac{1}{2} \text{ quarter)}$$



15.8 TIME-SERIES ANALYSIS IN FORECASTING

In this chapter, we have examined all four components of a time series. We have described the process of projecting past trend and seasonal variation into the future, while taking into consideration the inherent inaccuracies of this analysis. In addition, we noted that although the irregular and cyclical components do affect the future, they are erratic and difficult to use in forecasting.

We must realize that the mechanical approach of time-series analysis is subject to considerable error and change. It is necessary for management to combine these simple procedures with knowledge of other factors in order to develop workable forecasts. Analysts are constantly revising, updating, and discarding their forecasts. If we wish to cope successfully with the future, we must do the same.

Recognizing limitations of time-series analysis

When using the procedures described in this chapter, we should pay particular attention to two problems:

1. In forecasting, we project past trend and seasonal variation into the future. We must ask, "How regular and lasting were the past trends? What are the chances that these patterns are changing?"
2. How accurate are the historical data we use in time series analysis? If a company has changed from a FIFO (first-in, first-out) to a LIFO (last-in, first-out) inventory accounting system in a period during the time under consideration, the data (such as quarterly profits) before and after the change are not comparable and not very useful for forecasting.

HINTS & ASSUMPTIONS

Warning: Smart managers realize that accounting for most of the variation in a time series of *past* data does *not* mean that this same pattern will continue in the future. Hint: These same smart managers combine the predictions available from time series with intuitive answers to broad what-if questions concern the future business environment (sociological, economic, political) and whether it will change significantly from the environment that existed when the time-series data were gathered.

EXERCISES 15.8

- 15-40** List four errors that can affect forecasting with a time series.
- 15-41** When using a time series to predict the future, what assurances do we need about the historical data on which our forecasts are based?
- 15-42** What problems would you see developing if we used past college enrollments to predict future college enrollments?
- 15-43** How would forecasts using time-series analysis handle things such as
 (a) Changes in the federal tax laws?
 (b) Changes in accounting systems?

STATISTICS AT WORK

Loveland Computers

Case 15: Time Series Lee Azko was resting on well-earned laurels. The complicated regression analysis for the results of advertising expenditures had given Sherrel Wright new confidence in making the

argument for better planning. Even Walter Azko began to accept that some of the firm's success wasn't hit or miss—there really were some rules to this game.

"I never could see the value of running five-or six-page spreads," Uncle Walter said as he rounded the corner of Lee's "office"—a cubicle that was furnished with little except one of the largest and fastest of Loveland's latest personal computers. "Thanks for showing I was right. And you're even making me a believer in that expensive newspaper advertising, too."

"Did Margot say anything about those focus groups?" Lee fished for another compliment.

"We're going to deal with that next week—too early to say. But don't get too comfortable. I have a whole new project for you—go and see Gratia."

Gratia Delaguardia was clearly sharing a joke. The laughter was audible down the corridor—Gratia rated a "real" office, with a door. Lee found Gratia looking at a graph with yet another player on Loveland's team.

"Lee, come on in and meet Roberto Palomar. Bert runs the phone bank—you know, our order department. We were just talking about you."

"Hence the laughter?" Lee was nervous.

"No, no. Take a look at this. Bert's been trying to estimate the number of phone reps we need to have available to take orders. We need to plan for hiring..."

"And to install enough incoming 800 lines," added Roberto, whom everyone called Bert.

"We plotted out the quarterly data," continued Gratia, "and, as an engineer, let me tell you I can recognize a nonlinear trend when I see one." Gratia pointed to a curve that looked like the path of the space shuttle going into orbit. "Of course, we aren't complaining about our growth. It's good to be on a winning team."

"But if we continue this trend," said Bert, sliding a ruler into place on the graph, "within 10 years, we'll have to employ the whole population of Loveland just to staff our phone banks." With that, Gratia and Bert again dissolved with laughter. "Lee, look at these numbers and say it isn't so."

"Well, there's no doubt there's a very strong underlying trend," Lee observed, noting the obvious. "Is there any seasonality—you know, differences from month to month?"

"Good question," Bert replied. "These quarterly totals tend to mask some of the monthly ups and downs. For example, August is always a bust because people are away on vacation. But December is a very heavy month. We're not really in the Christmas gift business, although some home users apparently do ask Santa for a new Loveland Computer. The main effect comes from small businesses that want to book equipment expenditures before the end of the year for tax purposes."

"And I don't suppose the call volume is evenly spaced over the week," Lee ventured.

"Ah, rainy days and Mondays!" Bert answered. "We have a rule of thumb that we do twice as much business on Mondays as on Tuesdays. So we try to avoid training sessions or staff meetings on Mondays. Sometimes the supervisory staff will pitch in—whatever it takes. If we miss a call, a potential customer may buy from one of our competitors."

"But now we're at the point where I really should plan a little better for the number of staff to have available. If I schedule too many people, it's a waste of money and the reps get bored. They'd rather be at home."

"Well I think I can help," Lee offered. "Let me tell you what I'll need."

Study Questions: What data will Lee want to examine? What analysis will be performed? How will Bert make use of the information that Lee develops?

CHAPTER REVIEW

Terms Introduced in Chapter 15

Coding A method of converting traditional measures of time to a form that simplifies computation (often called *translating*).

Cyclical Fluctuation A type of variation in a time series, in which the value of the variable fluctuates above and below a secular trend line.

Deseasonalization A statistical process used to remove the effects of seasonality from a time series.

Irregular Variation A condition in a time series in which the value of a variable is completely unpredictable.

Modified Mean A statistical method used in time-series analysis. Discards the highest and lowest values when computing a mean.

Ratio-to-Moving-Average Method A statistical method used to measure seasonal variation. Uses an index describing the degree of that variation.

Relative Cyclical Residual A measure of cyclical variation, it uses the percentage deviation from the trend for each value in the series.

Residual Method A method of describing the cyclical component of a time series. It assumes that most of the variation in the series not explained by the secular trend is cyclical variation.

Seasonal Variation Patterns of change in a time series within a year; patterns that tend to be repeated from year to year.

Second-Degree Equation A mathematical form used to describe a parabolic curve that may be used in time-series trend analysis.

Secular Trend A type of variation in a time series, the value of the variable tending to increase or decrease over a long period of time.

Time Series Information accumulated at regular intervals and the statistical methods used to determine patterns in such data.

Equations Introduced in Chapter 15

$$15-1 \quad b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \quad \text{p. 822}$$

This formula, originally introduced in Chapter 12 as Equation 12-4, enables us to calculate the *slope of the best-fitting regression line* for any two-variable set of data points. The symbols \bar{X} and \bar{Y} represent the means of the values of the independent variable and dependent variable respectively; n represents the number of data points with which we are fitting the line.

$$15-2 \quad a = \bar{Y} - b\bar{X} \quad \text{p. 822}$$

We met this formula as Equation 12-5. It enables us to compute the *Y-intercept of the best-fitting regression line* for any two-variable set of data points.

$$15-3 \quad b = \frac{\sum xY}{\sum x^2} \quad \text{p. 824}$$

When the individual years (X) are changed to coded time values (x) by subtracting the mean ($x = X - \bar{X}$), Equation 15-1 for the slope of the trend line is simplified and becomes Equation 15-3.

15-4

$$a = \bar{Y}$$

p. 824

In a similar fashion, using coded time values also allows us to simplify Equation 15-2 for the intercept of the trend line.

15-5

$$\hat{Y} = a + bx + cx^2$$

p. 826

Sometimes we wish to fit a trend with a parabolic (or second-degree) curve instead of a straight line ($\hat{Y} = a + bx$). The general form for a fitted second-degree curve is obtained by including the second-degree term (cx^2) in the equation for \hat{Y}

15-6

$$\sum Y = an + c \sum x^2$$

p. 826

15-7

$$\sum x^2 Y = a \sum x^2 + c \sum x^4$$

p. 826

In order to find the least-squares second-degree fitted curve, we must solve Equations 15-6 and 15-7 simultaneously for the values of a and c . The value for b is obtained from Equation 15-3.

15-8

$$\text{Percent of trend} = \frac{Y}{\hat{Y}} \times 100$$

p. 833

We can measure cyclical variation as a *percent of trend* by dividing the actual value (Y) by the trend value (\hat{Y}) and then multiplying by 100.

15-9

$$\text{Relative cyclical residual} = \frac{Y - \hat{Y}}{\hat{Y}} \times 100$$

p. 834

Another measure of cyclical variation is the *relative cyclical residual*, obtained by dividing the deviation from the trend ($Y - \hat{Y}$) by the trend value, and multiplying the result by 100. The relative cyclical residual can easily be obtained by subtracting 100 from the percent of trend.

Review and Application Exercises

15-44

The number of people admitted to Valley Nursing Home per quarter is given in the following table:

	Spring	Summer	Fall	Winter
1992	29	30	41	43
1993	27	34	45	48
1994	33	36	46	51
1995	34	40	47	53

- (a) Calculate the seasonal indices for these data (use a 4-quarter centered moving average).
 (b) Deseasonalize these data using the indices from part (a).

- (c) Find the least-squares line that best describes the trend of the deseasonalized figures.

15-45

Wheeler Airline, a regional carrier, has estimated the number of passengers to be 595,000 (deseasonalized) for the month of December. How many passengers should the company anticipate if the December seasonal index is 128?

- 15-46** An EPA research group has measured the level of mercury contamination in the ocean at a certain point off the East Coast. The following percentages of mercury were found in the water:

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1993	0.3	0.7	0.8	0.8	0.7	0.7	0.6	0.6	0.4	0.7	0.2	0.5
1994	0.4	0.9	0.7	0.9	0.5	0.8	0.7	0.7	0.4	0.6	0.3	0.4
1995	0.2	0.6	0.6	0.9	0.7	0.7	0.8	0.8	0.5	0.6	0.3	0.5

Construct a 4-month centered moving average, and plot it on a graph along with the original data.

- 15-47** A production manager for a Canadian paper mill has accumulated the following information describing the millions of pounds processed quarterly:

	Winter	Spring	Summer	Fall
1992	3.1	5.1	5.6	3.6
1993	3.3	5.1	5.8	3.7
1994	3.4	5.3	6.0	3.8
1995	3.7	5.4	6.1	3.9

- (a) Calculate the seasonal indices for these data (percentage of actual to centered moving average).
- (b) Deseasonalize these data, using the seasonal indices from part (a).
- (c) Find the least-squares line that best describes these data.
- (d) Estimate the number of pounds that will be processed during the spring of 1996.

- 15-48** Describe some of the difficulties in using a linear estimating equation to describe these data:

- (a) Gasoline mileage achieved by U.S. automobiles.
- (b) Fatalities in commercial aviation.
- (c) The grain exports of a single country.
- (d) The price of gasoline.

- 15-49** Magna International is a large Canadian manufacturer of automotive components such as molded door panels. Magna's 1992 annual report listed the company's revenues for the previous ten years (in millions of Canadian dollars):

Year	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Revenue	302.5	493.6	690.4	1,027.8	1,152.5	1,458.6	1,923.7	1,927.2	2,017.2	2,358.8

- (a) Find the least-squares trend line for these data.
- (b) Plot the annual data and the trend line on the same graph. Do the variations from the trend appear random or cyclical?
- (c) Use a computer-based regression package to find the best-fitting parabolic trend for these data. Is c , the coefficient of x^2 , significantly different from zero? Which of the two trend models would you recommend using to forecast Magna's 1993 revenues? Explain.
- (d) Forecast Magna's 1993 revenues.

- 15-50** Comment on the difficulties you would have using a second-degree estimating equation to predict the future behavior of the process that generated these data:

- (a) Sales of personal computers in the United States.

- (b) Use of video games in the United States.
- (c) Premiums for medical malpractice insurance.
- (d) The number of MBAs graduated from U.S. universities.

15-51 The following table shows the number of franchisees of Beauty Bar, Inc., operating at the end of each year:

Year	1990	1991	1992	1993	1994	1995
Number of franchisees	596	688	740	812	857	935

- (a) Find the linear equation that best describes these data.
- (b) Estimate the number of operations manuals (one to a franchisee) that must be printed for 1997.

15-52 An assistant undersecretary in the U.S. Commerce Department has the following data describing the value of grain exported during the last 16 quarters (in billions of dollars):

	I	II	III	IV
1992	1	3	6	4
1993	2	2	7	5
1994	2	4	8	5
1995	1	3	8	6

- (a) Determine the seasonal indices and deseasonalize these data (using a 4-quarter centered moving average).
- (b) Calculate the least-squares line that best describes these data.
- (c) Identify the cyclical variation in these data by the relative cyclical residual method.
- (d) Plot the original data, the deseasonalized data, and the trend.

15-53 Richie Bell's College Bicycle Shop has determined from a previous trend analysis that spring sales should be 165 bicycles (deseasonalized). If the spring seasonal index is 143, how many bicycles should the shop sell this spring?

15-54 With the U.S. Interstate Highway program nearly finished, of what use are old data to the manufacturers of heavy earth-moving equipment as they attempt to forecast sales? What new data would you suggest they use in their forecasting?

15-55 Automobile manufacturing is often cited as an example of a cyclical industry (one subject to changes in demand according to an underlying business cycle). Consider automobile production worldwide (in millions of units) and in the former U.S.S.R. (in hundreds of thousands of units) from 1970 through 1990:

Year	World	U.S.S.R.	Year	World	U.S.S.R.
1970	22.5	3.4	1981	27.5	13.2
1971	26.4	5.3	1982	26.6	13.1
1972	27.9	7.3	1983	30.0	13.2
1973	30.0	9.2	1984	30.5	13.3
1974	25.9	11.2	1985	32.3	13.3
1975	25.0	12.0	1986	32.9	13.3

(Continued)

(Contd.)

Year	World	U.S.S.R.	Year	World	U.S.S.R.
1976	28.8	12.4	1987	33.0	13.3
1977	30.5	12.8	1988	34.3	12.6
1978	31.2	13.1	1989	35.6	12.2
1979	30.8	13.1	1990	35.8	12.6
1980	28.6	13.3			

- (a) Find the least-squares trend line for the worldwide data.
 (b) Plot the worldwide data and the trend line on the same graph. Do the variations from the trend appear random or cyclical?
 (c) Plot the residuals as a percent of trend. Approximately how long is the business cycle shown by these data?
 (d) Consider the output of automobiles in the former U.S.S.R. Discuss its similarities and differences with the patterns you found in parts (a), (b), and (c).

15-56

R. B. Fitch Builders has completed these numbers of homes in the 8 years it has been in business:

Year	1988	1989	1990	1991	1992	1993	1994	1995
Completions	12	11	19	17	19	18	20	23

- (a) Develop a linear estimating equation to describe the trend of completions.
 (b) How many completions should R. B. plan on for 1999?
 (c) Along with the answer to part (b), what advice would you give R. B. about using this forecasting technique?

15-57

As part of an investigation being done by a federal agency into the psychology of criminal activity, a survey of the number of homicides and assaults over the course of a year produced the following results:

Season	Spring	Summer	Fall	Winter
Number of homicides and assaults	31,000	52,000	39,000	29,000

- (a) If the corresponding seasonal indices are 84, 134, 103, and 79, respectively, what are the deseasonalized values for each season?
 (b) What is the meaning of the seasonal index of 79 for the winter season?

15-58

A state's quarterly deseasonalized unemployment percentage figures for years 1991–1995 are as follows:

	I	II	III	IV
1991	7.3	7.2	7.3	8.1
1992	8.7	9.2	9.8	10.5
1993	10.2	9.9	9.2	8.3
1994	7.6	7.4	7.5	7.6
1995	7.4	7.0	6.8	6.5

- (a) Find the linear equation that describes this unemployment trend.
 (b) Calculate the percent of trend for these data.
 (c) Plot the cyclical variation in the unemployment rates from the percent of trend.

- 15-59** The number of confirmed AIDS cases reported at a local health clinic during the 5 years from 1988 to 1992 were 2, 4, 7, 13, and 21, respectively.

- Develop the linear regression line for these data.
- Find the least-squares second-degree curve for these data.
- Construct a table of each year's actual cases, the linear estimates from the regression in part (a), and the second-degree values from the curve in part (b).
- Which regression appears to be the better estimator?

- 15-60** RJ's Grocers has added broiled whole chickens to its line of takeout food for busy professionals who don't have time to cook at home. The number of precooked chickens sold in the first 7 weeks are as follows:

Week	1	2	3	4	5	6	7
Sales	41	52	79	76	72	59	41

- Find the linear regression line that best fits these data.
- Estimate the expected number of sales for week 8.
- Based on the estimate in part (b) and the available data, does the regression accurately describe the sales trend for this item?

- 15-61** The College Town busing system has collected the following count of passengers per season during 1994 and 1995. The deseasonalized data (in thousands of passengers) are

	Spring	Summer	Fall	Winter
1994	593	545	610	575
1995	640	560	600	555

- If the seasonal indices used to deseasonalize these data were 110, 73, 113, and 104, respectively, find the actual passenger counts (in thousands) for these eight seasons.
- Which season in 1995 saw the fewest passengers? The most?
- If the linear estimating equation for these deseasonalized data is $\hat{Y} = 584.75 - 0.45x$ (with x measured in $\frac{1}{4}$ quarters, and $x = 0$ between the winter 1994 and spring 1995 quarters), what is the expected number of actual riders (in thousands) for the fall 1996 season?

- 15-62** Ferris Wheeler, director of the Whirly World amusement park, has provided the following attendance data (in thousands of admissions) for the park's open seasons:

	Spring	Summer	Fall
1992	750	1,150	680
1993	780	1,100	580
1994	800	1,225	610
1995	640	1,050	600

- Calculate the seasonal indices for these data using a 3-period moving average.
- Deseasonalize these data using the seasonal indices from part (a).

- 15-63** A restaurant manager wishes to improve customer service and employee scheduling based on the daily levels of customers in the past 4 weeks. The numbers of customers served in the restaurant during that period were

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week	1 345	310	385	416	597	706	653
	2 418	333	400	515	664	761	702
	3 393	387	311	535	625	711	598
	4 406	412	377	444	650	803	822

Determine the seasonal (daily) indices for these data. (Use a 7-day moving average.)

- 15-64** Suppose television sales by a small appliance chain for the years 1991–1995 were as follows:

Year	1991	1992	1993	1994	1995
Sales	230	250	265	300	310

- (a) Develop the second-degree estimating equation for these data.
- (b) What do the magnitudes of the coefficients a , b , and c tell you about the choice of a second-degree equation for these data?

- 15-65** The Zapit Company has recorded the following numbers (in hundreds of thousands) of total sales of its line of microwave ovens over the last 5 years:

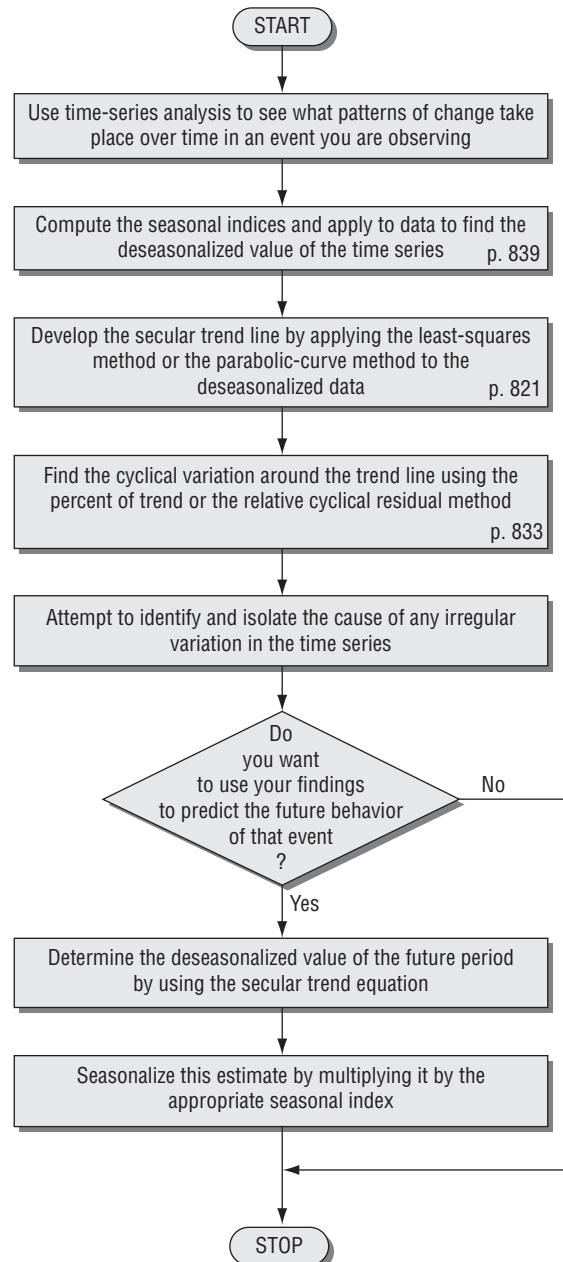
Year	1991	1992	1993	1994	1995
Sales	3.5	3.8	4.0	3.7	3.9

The equation describing the trend for these sales volumes is

$$\hat{Y} = 3.78 + 0.07x, \text{ where } 1993 = 0, \text{ and } x \text{ units} = 1 \text{ year}$$

- (a) Which year had the largest percent of trend?
- (b) Which year was closest to the trend line?

Flow Chart: Time Series



This page is intentionally left blank.

Index Numbers

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To understand that index numbers describe how much economic variables have changed over time
 - To become familiar with the three principal types of indices: price indices, quantity indices, and value indices
 - To understand and avoid problems resulting from the incorrect use of index numbers
 - To learn how to calculate various kinds of index numbers
-

CHAPTER CONTENTS

16.1 Defining an Index Number	870
16.2 Unweighted Aggregates Index	874
16.3 Weighted Aggregates Index	879
16.4 Average of Relatives Methods	888
16.5 Quantity and Value Indices	895
16.6 Issues in Constructing and Using Index Numbers	900

■ Statistics at Work	901
■ Terms Introduced in Chapter 16	902
■ Equations Introduced in Chapter 16	902
■ Review and Application Exercises	904
■ Flow Chart: Index Numbers	910

Precision Metal Products manufactures high-quality fabrications for use in the production of machinery for heavy industry. The company's three principal materials are coal, iron ore, and nickel ore. Management has the following data showing prices of these materials in 1975 and 1995 and quantity data for 1988, a year when purchasing patterns were characteristic of the entire 20-year period.

Raw Material	Qty. Used	Price/Ton	Price/Ton
	1988 (000 tons)	1975	1995
Coal	158	\$7.56	\$19.50
Iron ore	12	9.20	21.40
Nickel ore	5	12.30	36.10

Management would like help in constructing some measure of the change in material prices in the 20-year period. Using the methods in this chapter, we can supply it with such a figure to use in its planning. ■

16.1 DEFINING AN INDEX NUMBER

At some time, everyone faces the question of how much something has changed over a period of time. We may want to know how much the price of groceries has increased so we can adjust our budgets accordingly. A factory manager may wish to compare this month's per-unit production cost with that of the past 6 months. Or a medical research team may wish to compare the number of flu cases reported this year with the number reported in previous years. In each of these situations, the degree of change must be determined and defined. Typically, we use *index numbers* to measure such differences.

Why use an index number?

An index number measures how much a variable changes over time. We calculate an index number by finding the ratio of the current value to a base value. Then we multiply the resulting number by 100 to express the index as a percentage. This final value is the *percentage relative*. Note that the index number for the base point in time is always 100.

What is an index number?

The secretary of state of North Carolina has data indicating the number of new businesses incorporated. The data she collects show that 9,300 were started in 1980, 6,500 in 1985, 9,600 in 1990, and 10,100 in 1995. If 1980 is the base year, she can calculate the index numbers reflecting volume changes using the process presented in Table 16-1.

Computing a simple index

Using these calculations, the secretary of state finds that incorporations in 1985 had an index of 70 relative to 1980. Another way to state this is to say that the number of incorporations in 1985 was 70 percent of the number of incorporations in 1980.

Types of Index Numbers

There are three principal types of indices: the price index, the quantity index, and the value index. A *price index* is the one most frequently used. It compares levels of prices from one period to another. The familiar *Consumer Price Index* (CPI), tabulated by the Bureau of Labor Statistics, measures overall price changes of a variety of consumer goods and services and is used to define the cost of living.

Price index

A *quantity index* measures how much the number or quantity of a variable changes over time. Our example using incorporations determined a quantity index relating the numbers in 1985, 1990, and 1995 to that in 1980.

Quantity index

TABLE 16-1 CALCULATION OF INDEX NUMBERS (BASE YEAR = 1980)

Year (1)	Number of New Incorporations (000) (2)	Ratio (3) = (2) ÷ 9.3	Index or Percentage Relative (4) = (3) × 100
1980	9.3	$\frac{9.3}{9.3} = 1.00$	$1.00 \times 100 = 100$
1985	6.5	$\frac{6.5}{9.3} = 0.70$	$0.70 \times 100 = 70$
1990	9.6	$\frac{9.6}{9.3} = 1.03$	$1.03 \times 100 = 103$
1995	10.1	$\frac{10.1}{9.3} = 1.09$	$1.09 \times 100 = 109$

The last type of index, the *value index*, measures changes in total monetary worth. That is, it measures changes in the dollar value of a variable. In effect, the value index combines price and quantity changes to present a more informative index. In our example, we determined only a quantity index. However, we could have included the dollar effect by computing the total incorporated value for the years under consideration. Table 16-2 presents the corresponding value indices for 1985, 1990, and 1995. From this computation, we can say that the *value index* of incorporations in 1995 was 160. Or we can say that the incorporated value of 1995 increased 60 percent relative to the incorporated value of 1980.

Usually, an index measures change in a variable over a period of time, such as in a time series. However, it can also be used to measure differences in a given variable in different locations. This is done by simultaneously collecting data in different locations and then comparing the data. The

TABLE 16-2 COMPUTING A VALUE INDEX (BASE YEAR = 1980)

Year (1)	Incorporated Value (millions) (2)	Ratio (3) = (2) ÷ 18.4	Index or Percentage Relative (4) = (3) × 100
1980	\$18.4	$\frac{18.4}{18.4} = 1.00$	$1.00 \times 100 = 100$
1985	14.6	$\frac{14.6}{18.4} = 0.79$	$0.79 \times 100 = 79$
1990	26.2	$\frac{26.2}{18.4} = 1.42$	$1.42 \times 100 = 142$
1995	29.4	$\frac{29.4}{18.4} = 1.60$	$1.60 \times 100 = 160$

comparative cost-of-living index, for example, shows that in terms of the cost of goods and services, it is cheaper to live in Austin, Texas, than in New York City.

A single index may reflect a composite, or group, of changing variables. The Consumer Price Index measures the general price level for specific goods and services in the economy. It combines the individual prices of the goods and services to form a composite price index number.

Composite index numbers

Uses of Index Numbers

Index numbers can be used in several ways. It is most common to use them by themselves, as an end result. Index numbers such as the Consumer Price Index are often cited in news reports as general indicators of the nation's economic condition.

Management uses index numbers as part of an intermediate computation to understand other information better. In the chapter on time series, seasonal indices were used to modify and improve estimates of the future. The use of the Consumer Price Index to determine the real buying power of money is another example of how index numbers help increase knowledge of other factors. Table 16-3 shows the weekly salary paid to a secretary over a period of years, the corresponding Consumer Price Index values, and computation of the secretary's real salary. The secretary's dollar salary increased substantially, but the actual buying power of her income increased less rapidly. This can be attributed to the simultaneous rise in the cost-of-living index from 100 to 200.

One use of the Consumer Price Index

Problems Related to Index Numbers

Several things can distort index numbers. The four most common causes of distortion are:

1. Sometimes there is **difficulty in finding suitable data** to compute an index. Suppose the sales manager of Colonial Aircraft is interested in computing an index describing seasonal variation in the sale of the company's small planes. If sales are reported only on an annual basis, he would be unable to determine the seasonal sales pattern.
2. **Incomparability of indices** occurs when attempts are made to compare one index with another after there has been a basic change in what is being measured. If Citizens for Reasonable Transportation compare price indices

Limited data

Incomparability

TABLE 16-3 COMPUTATION OF REAL WAGES

Year (1)	Weekly Salary Paid (2)	Consumer Price Index (3)	(4) = $\frac{(2) \times 100}{(3)}$	Real or Adjusted Salary
1977	\$114.75	100	$114.75 \times \frac{100}{100} =$	\$114.75
1982	145.50	123	$145.50 \times \frac{100}{123} =$	\$118.29
1992	472.98	200	$472.98 \times \frac{100}{200} =$	\$236.49

of automobiles from 1979 to 1989, they find that prices have increased substantially. However, this comparison does not take into consideration technological advances in the quality of automobiles achieved over the time period in consideration.

3. **Inappropriate weighting of factors** can also distort an index. In *Inappropriate weighting* developing a composite index, such as the Consumer Price Index, we must consider changes in some variables to be more important than changes in others. The effect on the economy of a 50-cent-per-gallon increase in the price of gasoline cannot be counterbalanced by a 50-cent decrease in the price of cars. It must be realized that the 50-cent-per-gallon increase in gas cost has a much greater effect on consumers. Thus, greater weight has to be assigned to the increased gas price than to the decrease in the cost of cars.
4. Distortion of index numbers also occurs when **selection of an improper base** occurs. Sometimes a firm selects a base that automatically leads to a result that is in its own interest and proves its initial assumption. If Consumers Against Oil Waste wants to portray oil companies in a bad light, it might measure this year's profits with a recession year as its base for oil profits. This would produce an index that shows oil profits have increased substantially. On the other hand, if Consumers for Unlimited Oil Use wishes to show that this year's profits are minimal, it might select a year with high profits for its base year. Using high profit as a base would probably result in an index indicating a small increase, or maybe even a decline, in oil profits this year. Therefore, we must always consider how and why the base period was selected before we accept a claim based on the result of, comparing index numbers.

Sources of Index Numbers

When managers apply index numbers to everyday problems, they use many sources to obtain the necessary information. The source depends on their information requirements. A firm can use monthly sales reports to determine its seasonal sales pattern. In dealing with broad areas of national economy and the general level of business activity, publications such as the *Federal Reserve Bulletin*, *Moody's*, *Monthly Labor Review*, and the *Consumer Price Index* provide a wealth of data. Many federal and state publications are listed in the U.S. Department of Commerce pamphlet *Measuring Markets*. Almost all government agencies distribute data about their activities, from which index numbers can be computed. Many financial newspapers and magazines provide information from which index numbers can be computed. When you read these sources, you will find that many of them use index numbers themselves

Sources of data for index numbers

EXERCISES 16.1

Basic Concepts

- 16-1 What is the index for a base year?
- 16-2 Explain the differences among the three principal types of indices: *price*, *quantity*, and *value*.
- 16-3 What does the Consumer Price Index measure? Is this based on a single variable or a composite of variables?
- 16-4 What are two basic ways of using index numbers?
- 16-5 What does an index number measure?
- 16-6 How is a percentage relative (index) found?

16.2 UNWEIGHTED AGGREGATES INDEX

The simplest form of a composite index is an *unweighted aggregates index*. *Unweighted* means that all the values considered in calculating the index are of equal importance. *Aggregate* means that we add, or sum, all the values. The principal advantage of an unweighted aggregates index is its simplicity.

An unweighted aggregates index is calculated by adding all the elements in the composite for the given time period and then dividing this result by the sum of the same elements during the base period. Equation 16-1 presents the mathematical formula for computing an unweighted aggregates quantity index.

Computing an unweighted aggregates index

Unweighted Aggregates Quantity Index

$$\frac{\sum Q_i}{\sum Q_0} \times 100 \quad [16-1]$$

where

- Q_i = quantity of each element in the composite for the year in which we want the index
- Q_0 = quantity of each element in the composite for the base year

A word of explanation about the use of the subscript i to indicate the year for which we want to compute the index: Suppose we have quantity data for 1990 (the base year), 1991, and 1992, and we want to compute unweighted aggregates quantity indices for 1991 and 1992. If we use the subscripts 0, 1, and 2 to denote 1990, 1991, and 1992, then the index for 1991 is

$$\frac{\sum Q_1}{\sum Q_0} \times 100$$

and the index for 1992 is

$$\frac{\sum Q_2}{\sum Q_0} \times 100$$

Both of these are captured by the use of the generic subscript i in the numerator of Equation 16-1. We shall use i in the same fashion in the formulas defining all of the index numbers we discuss in this chapter. For sake of brevity, we shall use *current year* to indicate the *year in which we want the index*.

Note that we can substitute either prices or values for quantities in Equation 16-1 to find the general equation for a price index or a value index. Because the ratio is multiplied by 100, the resulting index is technically a percentage. However, it is customary to refer only to the value and to omit the percent sign when discussing index numbers.

The example in Table 16-4 demonstrates how we compute an unweighted index. In this case, we want to measure changes in general price levels on the basis of changes in prices of a few items. The 1990 prices are the base values to which we compare the 1995 prices.

Computing an unweighted Index

From these calculations, we determine that the price index describing the change in these items from 1990 to 1995 is 145. If the elements in this composite are representative of the general price level, we can say that prices rose 45 percent from

Interpreting the index

TABLE 16-4 COMPUTATION OF AN UNWEIGHTED INDEX

Elements in the Composite	Prices	
	1990 P_0	1995 P_1
Milk (1 gallon)	\$1.92	\$3.40
Eggs (1 dozen)	0.81	1.00
Hamburger (1 pound)	1.49	2.00
Gasoline (1 gallon)	$\frac{1.00}{\Sigma P_0 = 5.22}$	$\frac{1.17}{\Sigma P_1 = 7.57}$
Unweighted aggregates price index	$= \frac{\Sigma P_i}{\Sigma P_0} \times 100$	$[16-1]$
	$= \frac{7.57}{5.22} \times 100$	
	$= 1.45 \times 100$	
	$= 145$	

1990 to 1995. However, we cannot expect a sample of four items to reflect accurate price changes for all goods and services. Thus, this calculation provides us with only a very rough estimate.

Suppose we now add the change in price of hand-held electronic calculators from 1990 to 1995 to our composite (Table 16-5). Again, 1990 is the base period against which we compare the 1995 prices.

TABLE 16.5 COMPUTATION OF AN UNWEIGHTED INDEX

Elements in the Composite	Prices	
	1990 P_0	1995 P_1
Milk (1 gallon)	\$1.92	\$3.40
Eggs (1 dozen)	0.81	1.00
Hamburger (1 pound)	1.49	2.00
Gasoline (1 gallon)	1.00	1.17
Hand-held electronic calculator (1)	$\frac{15.00}{\Sigma P_0 = 20.22}$	$\frac{11.00}{\Sigma P_1 = 18.57}$
Unweighted aggregates price index	$= \frac{\Sigma P_i}{\Sigma P_0} \times 100$	$[16-1]$
	$= \frac{18.57}{20.22} \times 100$	
	$= 0.92 \times 100$	
	$= 92$	

Intuitively we know that the previous index of 145 is a more accurate estimate of general price behavior than 92 because more prices rose than fell between 1990 and 1995. Thus, we see the **major disadvantage of an unweighted index. It does not attach greater importance, or weight, to the price change of a high-use item than it does to a low-volume item.** (A family may purchase 50 dozen eggs a year, but it would be unusual for a family to own more than one or two calculators.) A substantial price change for slow-moving items can completely distort an index. For this reason, it is not common to use a simple unweighted index in important analyses.

Limitations of an unweighted index

The deficiencies of an unweighted index suggest that we use a weighted index. There are two ways to calculate more sophisticated indices. Each of these will be discussed in detail in the following sections.

HINTS & ASSUMPTIONS

Warning: An unweighted index can be distorted, and lose its value from changes in a few items in the index that do not fairly represent the situation being studied. Hint: Social Security payments have been “indexed” to the Consumer Price Index, which includes average mortgage costs as a measure of housing costs. But most Social Security recipients are not in the market for a new mortgage. With the exception of those who have an adjustable-rate mortgage, their mortgage payments are fixed and thus their costs are not affected by inflation. Warning: The major disadvantage of an unweighted index is that it does *not* attach greater importance to price changes in a high-use item than it does to a low-use item. Hint: Before you decide which index is appropriate, look carefully at the product/service components of that index to see whether their usage has been constant.

EXERCISES 16.2

Self-Check Exercise

SC 16-1 The VP of sales for Xenon Computer Corporation is examining the commission rate employed for the last 3 years. Below are the commission earnings of the company’s top five sales personnel.

	1993	1994	1995
Guy Howell	\$48,500	\$55,100	\$63,800
Skip Ford	41,900	46,200	60,150
Nelson Price	38,750	43,500	46,700
Nina Williams	36,300	45,400	39,900
Ken Johnson	33,850	38,300	50,200

Using 1993 as the base period, express the commission earnings in 1994 and 1995 in terms of an unweighted aggregates index.

Applications

- 16-7** In an effort to get a measure of economic hardship, the IMF (International Monetary Fund) collected data on the price behavior of five major products imported by a group of less-developed countries. Using 1992 as the base period, express the 1995 prices in terms of an unweighted aggregates index.

Product	A	B	C	D	E
1992 price	\$127	\$532	\$2,290	\$60	\$221
1995 price	\$152	\$651	\$2,314	\$76	\$286

- 16-8** For purposes of bidding on U.S. contracts, the management of a large overseas manufacturing facility are compiling data on wage levels. The following data concern base pay for the different classes of labor in the facility over a 4-year period.

	Wages per Hour			
	1992	1993	1994	1995
Class A	\$8.48	\$9.32	\$10.34	\$11.16
Class B	6.90	7.52	8.19	8.76
Class C	4.50	4.99	5.48	5.86
Class D	3.10	3.47	3.85	4.11

Using 1992 as the base period, calculate the unweighted aggregates wage index for 1993, 1994, and 1995.

- 16-9** A study of college costs has collected data for the amount of tuition a fulltime undergraduate paid during the last 4 years at four schools:

	1993	1994	1995	1996
Eastem U.	\$3,142	\$3,564	\$4,109	\$4,372
State U.	2,816	3,474	3,682	4,019
Western U.	3,582	3,987	4,406	4,819
Central U.	4,014	4,197	4,384	4,671

Using 1993 as a base period, express tuition charges in 1994, 1995, and 1996 in terms of an unweighted aggregates index.

- 16-10** Bill Ivey, the administrator of a small rural hospital, has compiled the information shown regarding food purchased for the hospital kitchen. For the commodities listed, the corresponding price indicates the average price for that year. Using 1994 as the base period, express the prices in 1993 and 1995 in terms of an unweighted aggregates index.

Commodity	1993	1994	1995
Dairy products	\$2.34	\$2.38	\$2.60
Meat products	3.19	3.41	3.36
Vegetable products	0.85	0.89	0.94
Fruit products	1.11	1.19	1.18

- 16-11** A chemical processing plant used five materials in the manufacture of an industrial cleaning agent. The following data indicate the final inventory levels for these materials for the years 1993 and 1995.

Material	A	B	C	D	E
Inventory (tons) 1993	86	395	1,308	430	113
Inventory (tons) 1995	95	380	1,466	469	108

Using 1993 as the base period, express the 1995 inventory levels in terms of an unweighted aggregates index.

- 16-12** John Dykstra, a management trainee in a bank, has collected, information on the bank's transactions for the years 1994 and 1995:

	Withdrawals		Deposits	
	Savings	Checking	Savings	Checking
Number of transactions 1994	169,000	21,843,000	293,000	2,684,000
Number of transactions 1995	158,000	23,241,000	303,000	3,361,000

Using 1994 as the base period, express the number of banking transactions in 1995 in terms of an unweighted aggregates index.

- 16-13** The Bookster Publishing Company began its business of publishing college textbooks in 1993. It is interested in determining how its sales have changed compared to its first year. A summary of the company's records shows how many new books it published in each year in the following areas:

	1993	1994	1995
Biology	48	53	50
Mathematics	32	37	35
History	19	15	22
English	16	20	21
Sociology	24	18	26
Physics	10	26	32
Chemistry	27	26	30
Philosophy	11	8	15

Using 1993 as the base year, calculate the unweighted aggregates quantity index for 1994 and 1995. Interpret the results for the publishing company.

Worked-Out Answer to Self-Check Exercise

SC 16-1	1993 Q_0	1994 Q_1	1995 Q_2
Howell	48,500	55,100	63,800
Ford	41,900	46,200	60,150
Price	38,750	43,500	46,700
Williams	36,300	45,400	39,900
Johnson	33,850	38,300	50,200
	<u>199,300</u>	<u>228,500</u>	<u>260,750</u>
	<u>19,930,000</u>	<u>22,850,000</u>	<u>26,075,000</u>
	99,300	199,300	199,300
Index = $\frac{\sum Q_i}{\sum Q_0} \times 100$:	= 100.0	= 114.7	= 130.8

16.3 WEIGHTED AGGREGATES INDEX

As we have said, often we have to attach greater importance to changes in some variables than to other when we compute an index. This weighting allows us to include more information than just the change in price over time. It also lets us improve the accuracy of the general price level estimate based on our sample. The problem is to decide how much weight to attach to each of the variables in the sample.

The general formula for computing a weighted aggregates price index is

Advantages of weighting in an index

Computing a weighted aggregates index

Weighted Aggregates Price Index

$$\frac{\sum P_i Q}{\sum P_0 Q} \times 100$$

[16-2]

where

- P_i = price of each element in the composite in the current year
- P_0 = price of each element in the composite in the base year
- Q = quantity weighting factor chosen

Consider the sample in Table 16-6. Each of the elements in the composite is taken from Table 16-5 and is weighted according to the volume of sales. The process of weighted aggregates confirms our earlier intuitive impression from page 927 that the general price level had risen (index = 129).

Typically, management uses the quantity of an item consumed as the measure of its importance in computing a weighted aggregates index. This leads to an important question in applying the process: Which quantities are used?

In general, there are three ways to weight an index. The first involves using quantities consumed during the base period in computing each index number. This is called the *Laspeyres method*, after the statistician

Three ways to weight an index

TABLE 16-6 COMPUTATION OF A WEIGHTED AGGREGATES INDEX

Elements in the Composite	Q Volume (billions) (1)	P_0 1990 Prices (2)	P_1 1995 Prices (3)	$P_0 Q$ Weighted Sales (4) = (2) × (1)	$P_1 Q$ Weighted Sales (5) = (3) × (1)
Milk	20.000 (gal)	\$1.92	\$3.40	$1.92 \times 20.000 = 38.40$	$3.40 \times 20.000 = 68.00$
Eggs	3.500 (doz)	0.81	1.00	$0.81 \times 3.500 = 2.84$	$1.00 \times 3.500 = 3.50$
Hamburger	11.000 (lb)	1.49	2.00	$1.49 \times 11.000 = 16.39$	$2.00 \times 11.000 = 22.00$
Gasoline	154.000 (gal)	1.00	1.17	$1.00 \times 154.000 = 154.00$	$1.17 \times 154.000 = 180.18$
Calculators	0.002 (units)	15.00	11.00	$15.00 \times \frac{0.002}{15.00} = 0.03$	$11.00 \times \frac{0.002}{15.00} = 0.02$
				$\sum P_0 Q = 211.66$	$\sum P_1 Q = 273.70$
				Weighted aggregates index = $\frac{\sum P_1 Q}{\sum P_0 Q} \times 100$	
				$= \frac{273.70}{211.66} \times 100$	
				$= 1.29 \times 100$	
				$= 129$	

who developed it. The second uses quantities consumed during the period in question for each index. This is the *Paasche method*, in honor of the person who devised it. The third way is called the *fixed-weight aggregates method*. With this method, one period is chosen, and its quantities are used to find all indices. (Note that if the chosen period is the base period, the fixed-weight aggregates method is the same as the Laspeyres method.)

Laspeyres Method

The Laspeyres method, which uses quantities consumed during the base period, is the method most commonly used because it requires quantity measures for only one period. Because each index number depends on the same base price and quantity, management can compare the index of one period directly with the index of another. Suppose a steel manufacturer's price index is 103 in 1992 and 125 in 1995, using 1990 base prices and quantities. The company concludes that the general price level has increased 22 percent from 1992 to 1995. To calculate the Laspeyres index, the company first multiplies the current-period price by the base-period quantity for each element in the composite, and then it sums each of the resulting values. Next it multiplies the base-period price by the base-period quantity for each element, and again it sums the resulting values. By dividing the first sum by the second and multiplying the result by 100, management can convert this value to a percentage relative. Equation 16-3 presents the formula used to determine the Laspeyres index.

The Laspeyres method

Computing a Laspeyres index

Laspeyres Price Index

$$\frac{\sum P_i Q_0}{\sum P_0 Q_0} \times 100$$

[16-3]

TABLE 16-7 CALCULATION OF A LASPEYRES INDEX

Elements in the Composite (1)	P_0 Base Price 1991 (2)	P_1 Current Price 1995 (3)	Average Quantity Consumed in 1991 by a Family (4)	$P_0 Q_0$ (5) = (2) × (4)	$P_1 Q_0$ (6) = (3) × (4)
Bread (1 loaf)	\$0.91	\$1.19	200 loaves	\$182	\$238
Potatoes (1 lb)	0.79	0.99	300 lb	237	297
Chicken (3-lb fryer)	3.92	4.50	100 chickens	392	450
					$\sum P_0 Q_0 = 811$
					$\sum P_1 Q_0 = 985$
$\text{Laspeyres price index} = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100$ $= \frac{985}{811} \times 100$ $= 1.21 \times 100$ $= 121$					

where

- P_i = prices in the current year
- P_0 = prices in the base year
- Q_0 = quantities sold in the base year

Let's work an example to demonstrate how the Laspeyres method is used. Suppose we want to determine changes in price level between 1991 and 1995. Table 16-7 contains the pertinent data for 1991 and 1995.

If we have selected a representative sample of goods, we can conclude that the general price index for 1995 is 121 based on the 1991 index of 100. Alternatively, we can say that prices have increased by 21 percent. Notice that we have used the average quantity consumed in 1991 rather than the total quantity consumed. Actually, it does not matter which is used, as long as we apply the same quantity measure throughout the problem. Typically, we select the quantity measure that is easiest to find.

One advantage of the Laspeyres method is the comparability of one index with another. If we had the 1992 prices for the previous example, we would be able to find a value for the 1992 general price index. This index could be compared directly with the 1995 index. Using the same base-period quantity allows us to make a direct comparison.

Another advantage is that many commonly used quantity measures are not tabulated every year. A firm might be interested in some variable whose quantity measure is computed once every 10 years. The Laspeyres method uses only one quantity measure, that of the base year, so the firm does not need yearly tabulations to measure quantities consumed.

The primary disadvantage of the Laspeyres method is that it does not take into consideration changes in consumption patterns. Items purchased in large quantities just a few years ago may be relatively

Example using the Laspeyres method

Drawing conclusions from the calculated index

Advantages of the Laspeyres method

Disadvantage of the Laspeyres method

[16-3]

unimportant today. Suppose the base quantity of an item differs greatly from the quantity for the period in question. Then the change in that item's price indicates very little about the change in the general price level.

Paasche Method

The second way to compute a weighted aggregates price index is the Paasche method. Finding a Paasche index is similar to finding a Laspeyres index. The difference is that the weights used in the Paasche method are the quantity measures for the *current* period rather than for the *base* period.

The Paasche index is calculated by multiplying the current-period price by the current-period quantity for each item in the composite and summing these products. Then the baseperiod price is multiplied by the current-period quantity for each item, and the results are summed. The first sum is divided by the second sum, and the resulting value is multiplied by 100 to convert the value into a percentage relative. Equation 16-4 defines the method for calculating a Paasche index.

Difference between Paasche and Laspeyres methods

Computing a Paasche index

Paasche Price Index

$$\frac{\sum P_i Q_i}{\sum P_0 Q_i} \times 100 \quad [16-4]$$

where

- P_i = current-period prices
- P_0 = base-period prices
- Q_i = current-period quantities

With this equation, we can rework the problem in Table 16-7. Notice that we have discarded the quantities consumed in 1991. They have been replaced by the quantities consumed in 1995. Table 16-8 presents the information necessary for this modified problem.

TABLE 16-8 CALCULATION OF A PAASCHE INDEX

Elements in the Composite (1)	P_1 Current Price 1995 (2)	P_0 Base Price 1995 (3)	Average Quantity Consumed in 1995 by a Family (4)	$P_1 Q_1$ (5) = (2) × (4)	$P_0 Q_1$ (6) = (3) × (4)
Bread (1 loaf)	\$1.19	\$0.91	200 loaves	\$238	\$182
Potatoes (1 lb)	0.99	0.79	100 lb	99	79
Chicken (3-lb fryer)	4.50	3.92	300 chickens	1,350	1,176
				$\sum P_1 Q_1 = 1,687$	$\sum P_0 Q_1 = 1,437$
				Paasche price index = $\frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100$	
				$= \frac{1,687}{1,437} \times 100$	
				$= 1.17 \times 100$	
				$= 117$	

In this analysis, we find that the price index for 1995 is 117. As you see from Table 16-7, the price index calculated by the Laspeyres method is 121. The difference between these indices reflects the change in consumption patterns of the three variables in the composite.

The Paasche method is particularly helpful because it combines the effects of changes in price and consumption patterns. Thus, it is a better indicator of general changes in the economy than the Laspeyres method. In our examples, the Paasche index shows a trend toward less-expensive goods and services because it indicates a price level increase of 17 percent instead of the 21 percent increase calculated using the Laspeyres method.

One of the principal disadvantages of the Paasche method is the need to tabulate quantity measures for each period examined. Often, quantity information for each period is either expensive to gather or unavailable. It would be hard, for example, to find reliable sources of data to determine quantity measures of 100 food products consumed in different countries for each of several years.

Each value for a Paasche price index is the result of both price and quantity changes from the base period. **Because the quantity measures used for one index period are usually different from the quantity measures for another index period, it is impossible to attribute the difference between the two indices to price changes only.** Thus, it is difficult to compare indices from different periods as calculated by the Paasche method.

Interpreting the difference between the two methods

Advantage of the Paasche method

Disadvantages of the Paasche method

Fixed-Weight Aggregates Method

The third technique used to assign weights to elements in a composite is the fixed-weight aggregates method. It is similar to both the Laspeyres and Paasche methods. However, instead of using base-period or current-period weights (quantities), it uses weights from a representative period. The representative weights are referred to as fixed weights. The fixed weights and the base prices do not have to come from the same period.

We calculate a fixed-weight aggregates price index by multiplying the current-period prices by the fixed weights and summing the results. Then we multiply the base-period prices by the fixed weights and sum them. Finally, we divide the first sum by the second and multiply by 100 to convert the ratio to a percentage relative. The formula used to calculate a fixed-weight aggregates price index is presented in Equation 16-5.

Fixed-weight aggregates index

Computing a fixed-weight aggregates index

Fixed-Weight Aggregates Price Index

$$\frac{\sum P_i Q_2}{\sum P_0 Q_2} \times 100 \quad [16-5]$$

where

- P_i = current-period prices
- P_0 = base-period prices
- Q_2 = fixed weights

We can demonstrate the process used to calculate a fixed-weight aggregates price index by solving our chapter-opening example. Recall that management wants to determine the price-level changes of raw

Example of a fixed-weight aggregates index

TABLE 16-9 COMPUTATION OF A FIXED-WEIGHT AGGREGATES INDEX

Raw Materials	Q_2 Quantity Consumed 1988 (thousands of tons)	P_0 Average Price 1975 (\$ per ton)	P_1 Average Price 1995 (\$ per ton)	$P_0 Q_2$ Weighted Aggregate 1975 (5) = (3) × (2)	$P_1 Q_2$ Weighted Aggregate 1995 (6) = (4) × (2)
(1)	(2)	(3)	(4)		
Coal	158	\$7.56	\$19.50	\$1,194.48	\$3,081.00
Iron	12	9.20	21.40	110.40	256.80
Nickel ore	5	12.30	36.10	61.50	180.50
				$\sum P_0 Q_2 = 1,366.38$	$\sum P_0 Q_2 = 3,518.30$
$\text{Fixed-weight aggregates price index} = \frac{\sum P_1 Q_2}{\sum P_0 Q_2} \times 100$ $= \frac{3,518.30}{1,366.38} \times 100$ $= 2.57 \times 100$ $= 257$ [16-5]					

materials consumed by the company between 1975 and 1995. It has accumulated the information in Table 16-9. From examination of past purchasing records, management has decided that the quantities purchased in 1988 were characteristic of the purchasing patterns during the 20-year period. The 1975 price level is the base price in this analysis. Calculation of the fixed-weight aggregates index is shown in Table 16-9. The company management concludes from this analysis that the general price level has increased 157 percent over the 20-year period.

The primary advantage of a fixed-weight aggregates price index is the flexibility in selecting the base price and the fixed weight (quantity). In many cases, the period that a company wishes to use as the base-price level may have an uncharacteristic consumption level. Therefore, by being able to select a different period for the fixed weight, the company can improve the accuracy of the index. This index also allows a company to change the price base without changing the fixed weight. This is useful because quantity measures are often expensive or impossible to obtain for certain periods.

Advantage of a fixed-weight aggregates index

HINTS & ASSUMPTIONS

The three methods covered in this section all produce a *weighted aggregates index* by using the *quantities consumed* as a basis for the weighting. Hint: The only real difference among them is the period each uses to select these quantities. The *Laspeyres* method uses quantities from the base period. The *Paasche* method uses quantities from the period in question. The *fixed-weight aggregates* method uses quantities from a chosen period. Hint: If the chosen period in the *fixed-weight aggregates* method is the base period, this method becomes the *Laspeyres* method. Warning: Choosing the period to use for weighting requires careful observation and common sense. The decision maker is looking for a period that has *characteristic consumption*, which means a period that most nearly reflects the reality of the situation. There is no mathematical formula that will give you the right answer to this.

EXERCISES 16.3

Self-Check Exercises

SC 16-2 Bill Simpson, owner of a California vineyard, has collected the following information describing the prices and quantities of harvested crops for the years 1992–1995.

Type of Grape	Price (per ton)				Quantity Harvested (tons)			
	1992	1993	1994	1995	1992	1993	1994	1995
Ruby Cabernet	\$108	\$109	\$113	\$111	1,280	1,150	1,330	1,360
Barbera	93	96	96	101	830	860	850	890
Chenin Blanc	97	99	106	107	1,640	1,760	1,630	1,660

Construct a Laspeyres index for each of these 4 years using 1992 as the base period.

SC 16-3 Use the data from Exercise SC 16-2 to calculate a fixed-weight index for each year using 1992 prices as the base and the 1995 quantities as the fixed weight.

SC 16-4 Use the data from Exercise SC 16-2 to calculate a Paasche index for each year using 1993 as the base period.

Applications

16-14 Eastern Digital has developed a substantial market share in the PC computer industry. The prices and number of units sold for their top four computer products from 1993 to 1996 were:

Model	Selling Price (\$)				Number Sold (thousands)			
	1993	1994	1995	1996	1993	1994	1995	1996
ED 107	1,894	1,906	1,938	1,957	84.6	86.9	98.4	107.5
ED Electra	2,506	2,560	2,609	2,680	38.4	42.5	55.6	67.5
ED Optima	1,403	1,440	1,462	1,499	87.4	99.4	109.7	134.6
ED 821	1,639	1,650	1,674	1,694	75.8	78.9	82.4	86.4

Construct a Laspeyres index for each of these 4 years using 1993 as the base period.

16-15 Use the data from Exercise 16-14 to calculate a fixed-weight index for each year using 1993 prices as the base and the 1996 quantities as the fixed weights.

16-16 Use the data from Exercise 16-14 to calculate a Paasche index for each year using 1994 as the base period.

16-17 Julie Pristash, the marketing manager of Mod-Stereo, a manufacturer of blank cassette tapes, has compiled the following information regarding unit sales for 1993–1995. Using the average quantities sold from 1993 to 1995 as the fixed weights, calculate the fixed weight index for each of the years 1993 to 1995 based on 1993.

Length of Tape (minutes)	Retail Price			Average Quantity (× 100,000) 1993–1995
	1993	1994	1995	
30	\$2.20	\$2.60	\$2.85	32
60	2.60	2.90	3.15	119
90	3.10	3.20	3.25	75
120	3.30	3.35	3.40	16

- 16-18** Gray P. Saeurs owns the corner fruitstand in a small town. After hearing many complaints that his prices constantly change during the summer, he has decided to see whether this is true. Based on the following data, help Mr. Saeurs calculate the appropriate weighted aggregate price indices for each month. Use June as the base period. Is your result a Laspeyres index or a Paasche index?

Fruit	Price per Pound			No of Pounds Sold
	June	July	Aug.	
Apples	\$0.59	\$0.64	\$0.69	150
Oranges	0.75	0.65	0.70	200
Peaches	0.87	0.90	0.85	125
Watermelons	1.00	1.10	0.95	350
Cantaloupes	0.95	0.89	0.90	150

- 16-19** Charles Widget is in charge of keeping in stock certain items that his company needs in repairing its machines. Since he started this job 3 years ago, he has been observing the changes in the prices for the items he keeps in stock. He arranged the data in the following table in order to calculate a fixed-weight aggregates price index. Perform the calculations Mr. Widget would do using 1993 as the base year.

Item	Price per Item			Average No Used During 3-Year Period
	1993	1994	1995	
W-gadget	\$1.25	\$1.50	\$2.00	900
X-gadget	6.50	7.00	6.25	50
Y-gadget	5.25	5.90	6.40	175
Z-gadget	0.50	0.80	1.00	200

Worked-Out Answers to Self-Check Exercises

SC 16-2

Type of Grape	1992	1992	1993	1994	1995
	Q_0	P_0	P_1	P_2	P_3
Ruby Cabernet	1,280	108	109	113	111
Barbera	830	93	96	96	101
Chenin Blanc	1,640	97	99	106	107
	1992	1993	1994	1995	
	$P_0 Q_0$	$P_1 Q_0$	$P_2 Q_0$	$P_3 Q_0$	
	138,240	139,520	144,640	142,080	
	77,190	79,680	79,680	83,830	
	159,080	162,360	173,840	175,480	
	374,510	381,560	398,160	401,390	

$$\text{Laspeyres Index} = \frac{\sum P_i Q_0}{\sum P_0 Q_0} \times 100 : \quad \begin{array}{c} \frac{37,451,000}{374,510} \\ \quad = 100.0 \end{array} \quad \begin{array}{c} \frac{38,156,000}{374,510} \\ \quad = 101.9 \end{array} \quad \begin{array}{c} \frac{39,816,000}{374,510} \\ \quad = 106.3 \end{array} \quad \begin{array}{c} \frac{40,139,000}{374,510} \\ \quad = 107.2 \end{array}$$

SC 16-3

Type of Grape	1995	1992	1993	1994	1995
	Q_3	P_0	P_1	P_2	P_3
Ruby Cabernet	1,360	108	109	113	111
Barbera	890	93	96	96	101
Chenin Blanc	1,660	97	99	106	107
	1992	1993	1994	1995	
	$P_0 Q_3$	$P_1 Q_3$	$P_2 Q_3$	$P_3 Q_3$	
	146,880	148,240	153,680	150,960	
	82,770	85,440	85,440	89,890	
	161,020	164,340	175,960	177,620	
	<u>390,670</u>	<u>398,020</u>	<u>415,080</u>	<u>418,470</u>	

$$\text{Fixed-Weight Index} = \frac{\sum P_i Q_3}{\sum P_0 Q_3} \times 100 : \quad \begin{array}{c} \frac{39,067,000}{390,670} \\ \quad = 100.0 \end{array} \quad \begin{array}{c} \frac{39,802,000}{390,670} \\ \quad = 101.9 \end{array} \quad \begin{array}{c} \frac{41,508,000}{390,670} \\ \quad = 106.2 \end{array} \quad \begin{array}{c} \frac{41,847,000}{390,670} \\ \quad = 107.1 \end{array}$$

SC 16-4

Type of Grape	1992	1993	1994	1995	1992	1993	1994	1995
	P_1	P_0	P_2	P_3	Q_1	Q_0	Q_2	Q_3
Ruby Cabernet	108	109	113	111	1,280	1,150	1,330	1,360
Barbera	93	96	96	101	830	860	850	890
Chenin Blanc	97	99	106	107	1,640	1,760	1,630	1,660

1992		1994		1995	
$P_1 Q_1$	$P_0 Q_1$	$P_2 Q_2$	$P_0 Q_2$	$P_3 Q_3$	$P_0 Q_3$
138,240	139,520	150,290	144,970	150,960	148,240
77,190	79,680	81,600	81,600	89,890	85,440
159,080	162,360	172,780	161,370	177,620	164,340
<u>374,510</u>	<u>381,560</u>	<u>404,670</u>	<u>387,940</u>	<u>418,470</u>	<u>398,020</u>

$$\text{Paasche Index} = \frac{\sum P_i Q_i}{\sum P_0 Q_i} \times 100 : \quad \begin{array}{c} \frac{37,451,000}{381,560} \\ \quad = 98.2 \end{array} \quad \begin{array}{c} \frac{40,467,000}{387,940} \\ \quad = 104.3 \end{array} \quad \begin{array}{c} \frac{41,847,000}{398,020} \\ \quad = 105.1 \end{array}$$

16.4 AVERAGE OF RELATIVES METHODS

Unweighted Average of Relatives Method

As an alternative to the aggregates methods, we can use the average of relatives method to construct an index. Once again, we will use a price index to introduce the process.

Actually, we used a form of the average of relatives method in calculating the simple index in Table 16-1 on page 877. In that one-product example, we calculated the percentage relative by dividing the number of incorporations in the current year, Q_1 , by the number in the base year, Q_0 , and multiplying the result by 100.

With more than one product (or activity), we first find the ratio of the current price to the base price for each product and multiply each ratio by 100. We then add the resulting percentage relatives and divide by the number of products. (Notice that the aggregates methods discussed in Section 16-3 differ from this method. They sum all the prices *before* finding the ratio.) Equation 16-6 presents the general form for the *unweighted average of relatives* method.

Computing an unweighted average of relatives index

Unweighted Average of Relatives Price Index

$$\frac{\sum \left(\frac{P_i}{P_0} \times 100 \right)}{n} \quad [16-6]$$

where

- P_i = current-period prices
- P_0 = base-period prices
- n = number of elements (or products) in the composite

In Table 16-10, we rework the problem in Table 16-4 on page 881 using the unweighted average of relatives method rather than the unweighted aggregates method.

Comparing the unweighted aggregates index and the unweighted average of relatives index

Based on this analysis, the general price-level index for 1995 is 138. In Table 16-4, the unweighted aggregates index for the same problem is 145. Obviously, there is a difference between these two indices. With the unweighted average of relatives method, we compute the average of the ratios of the prices for each product. With the unweighted aggregates method, we compute the ratio of the sums of the prices of each product. Notice that this is not the same as assigning some items more weight than others. Rather, the average of relatives method converts each element to a relative scale where each element is represented as a *percentage* rather than an *amount*. Because of this, each of the elements in the composite is measured against a base of 100.

Weighted Average of Relatives Method

Most problems management has to deal with require weighting by *importance*. Thus, it is more common to use the *weighted average of relatives method* than the unweighted method. When we

TABLE 16-10 COMPUTATION OF AN UNWEIGHTED AVERAGE OF RELATIVES INDEX

Product (1)	P_0 1990 Prices (2)	P_1 1995 Prices (3)	$\text{Ratio} \times 100$ (4) = $\frac{(3)}{(2)} \times 100$
Milk (1 gal)	\$1.92	\$3.40	$\frac{3.40}{1.92} \times 100 = 1.77 \times 100 = 177$
Eggs (1 doz)	0.81	1.00	$\frac{1.00}{0.81} \times 100 = 1.23 \times 100 = 123$
Hamburger (1 lb)	1.49	2.00	$\frac{2.00}{1.49} \times 100 = 1.34 \times 100 = 134$
Gasoline (1 gal)	1.00	1.17	$\frac{1.17}{1.00} \times 100 = 1.17 \times 100 = 117$
$\text{Unweighted average of relatives index} = \frac{\sum \left(\frac{P_1}{P_0} \times 100 \right)}{n}$ [16.6]			
$= \frac{551}{4}$ = 138			

computed a weighted aggregates price index in Section 16-3, we used the quantity consumed to weight the elements in the composite. To assign weights using the weighted average of relatives, we use the value of each element in the composite. (The value is the total dollar volume obtained by multiplying price by quantity.)

With the weighted average of relatives method, there are several ways to determine weighted value. As in the Laspeyres method, we can use the base value found by multiplying the base quantity by the base price. Using the base value will produce exactly the same result as calculating the index using the Laspeyres method. Because the result is the same, the decision to use the Laspeyres method or the weighted average of relatives method often depends on the availability of data. If value data are more readily available, the weighted average of relatives method is used. We use the Laspeyres method when quantity data are more readily obtained.

Equation 16-7 is used to compute a weighted average of relatives price index. This is a general equation into which we can substitute values from the base period, the current period, or any fixed period.

Different ways to determine weights

Computing a weighted average of relatives index

Weighted Average of Relatives Price Index

$$\frac{\sum \left[\left(\frac{P_i}{P_0} \times 100 \right) (P_n Q_n) \right]}{\sum P_n Q_n} \quad [16-7]$$

where

- $P_n Q_n$ = value
- P_0 = prices in the base period
- P_i = prices in the current period
- P_n and Q_n = quantities and prices that determine values we use for weights. In particular, $n = 0$ for the base period, $n = i$ for the current period, and $n = 2$ for a fixed period that is not a base or current period.

If we wish to compute a weighted average of relatives index using base values, $P_0 Q_0$, the equation would be

Weighted Average of Relatives Price Index with Base Year Values as Weights

$$\frac{\sum \left[\left(\frac{P_i}{P_0} \times 100 \right) (P_0 Q_0) \right]}{\sum P_0 Q_0} \quad [16-8]$$

Equation 16-8 is equivalent to the Laspeyres method for any given problem.

In addition to the specific cases of the general form of the weighted average of relatives method, we can use values determined by multiplying the price from one period by the quantity from a different period. Usually, however, we find Equations 16-7 and 16-8 adequate.

Here is an example. The information in Table 16-11 comes from Table 16-7 on page 887. We have base quantities and base prices, so we will use Equation 16-8. The price index of 122 differs slightly from the 121 calculated in Table 16-7 using the Laspeyres method, but only because of intermediate rounding.

As was the case for weighted aggregates, when we use base values, $P_0 Q_0$, or fixed values, $P_2 Q_2$, for weighted averages, we can readily compare the price level of one period with that of another. However, when we use current values, $P_1 Q_1$, in computing a weighted average of relatives price index, we *cannot* directly compare values from different periods because both the prices and the quantities may have changed. Thus, we usually use either base values or fixed values when computing a weighted average of relatives index.

Relation of weighted average of relatives to the Laspeyres method

Example of a weighted average of relatives index

Using base values, fixed values, or current values

TABLE 16-11 COMPUTING A WEIGHTED AVERAGE OF RELATIVES INDEX

Elements in the Composite (1)	Price		Quantity 1991 P_0 (4)	Percentage Price Relative $\frac{P_1}{P_0} \times 100$ (5) = $\frac{(3)}{(2)} \times 100$	Base Value $P_0 Q_0$ (6) = (2) × (4)	Weighted Percentage Relative (7) = (5) × (6)
	1991 P_0 (2)	1995 P_1 (3)				
Bread (1 loaf)	\$0.91	\$1.19	200 loaves	$\frac{1.19}{0.91} \times 100 = 131$	182	23,842
Potatoes (1 lb)	0.79	0.99	300 lb	$\frac{0.99}{0.79} \times 100 = 125$	237	29,625
Chicken (3-lb fryer)	3.92	4.50	100 fryers	$\frac{4.50}{3.92} \times 100 = 115$	392	45,080
$\sum P_0 Q_0 = 811$ $\sum \left[\left(\frac{P_1}{P_0} \times 100 \right) (P_0 Q_0) \right] = 98,547$ $\sum \left[\left(\frac{P_i}{P_0} \times 100 \right) (P_0 Q_0) \right] = \frac{98,547}{\sum P_0 Q_0}$ $= \frac{98,547}{811} \rightarrow 122$						
Weighted average of relatives index = [16-8]						

HINTS & ASSUMPTIONS

Hint: The *average of relatives* methods described in this section differ from those in the last section because they use the *total dollar volume consumed* as a basis for the weighting instead of just the quantities consumed. That's why each of them involves a price \times quantity calculation. These kinds of indices are used by gasoline refineries and coffee blenders that must use different amounts of raw materials to produce a blended product that is pretty much the same month after month.

EXERCISES 16.4**Self-Check Exercise**

SC 16-5 As a part of the evaluation of a possible acquisition, a New York City conglomerate has collected this sales information:

Product	Average Annual Price		Total Dollar Value (Thousands)
	1993	1995	1993
Calculators	\$27	\$20	\$150
Radios	30	42	900
Portable TVs	157	145	1,370

- (a) Calculate the unweighted average of relatives price index using 1993 as the base period.
 (b) Calculate the weighted average of relatives price index using the dollar value for each product in 1993 as the appropriate set of weights and 1995 as the base year.

Applications

- 16-20** F. C. Linley, owner of the San Mateo Seals, collected information regarding the ticket prices and volume for his franchise over the last 4 years.

	Average Annual Price				Tickets Sold ($\times 10,000$)			
	1992	1993	1994	1995	1992	1993	1994	1995
Box seats	\$6.50	\$7.25	\$7.50	\$8.10	26	27	31	28
General admission	3.50	3.85	4.30	4.35	71	80	89	90

Calculate a weighted average of relatives price index for each of the years 1992 through 1995 using 1993 as the base year and for weighting.

- 16-21** The following table contains information from the raw-material purchase records of a tire manufacturer for the years 1993–1995.

Material	Average Annual Purchase Price/Ton			Value of Purchase (thousands)
	1993	1994	1995	1995
Butadiene	\$ 17	\$ 15	\$ 11	\$ 50
Styrene	85	89	95	210
Rayon cord	348	358	331	1,640
Carbon black	62	58	67	630
Sodium pyrophosphate	49	56	67	90

Calculate a weighted average of relatives price index for each of those 3 years using 1995 for weighting and for the base year.

- 16-22** A Tennessee public interest group has surveyed the labor cost of automobile repairs in three major Tennessee cities (Knoxville, Memphis, and Nashville). With the following information, construct an unweighted average of relatives price index using the 1991 prices as a base.

Type of Repair	1991	1993	1995
Replacement of water pump	\$ 35	\$ 37	\$ 41
Replacement of engine valves (6 cyl.)	189	205	216
Wheel balancing	26	29	30
Tune-up (minor)	16	16	18

- 16-23** Garret Cage, the president of a local bank, is interested in the average levels of total savings and checking accounts for each of the last 3 years. He sampled days from each of these years; using the levels on those days, he determined the following yearly averages:

	1993	1994	1995
Saving accounts	\$1,845,000	\$2,320,000	\$2,089,000
Checking accounts	385,000	447,000	491,000

Calculate an unweighted average of relatives index for each year using 1993 as the base period.

- 16-24** InfoTech has researched the unit price and total value of memory chips imported into the United States in 1994 and 1996

Product	Price		Total Dollar Value (Thousands)
	1994	1996	1994
1-megabyte chips	\$ 42	\$ 65	957
4-megabyte chips	\$180	\$247	487
16-megabyte chips	\$447	\$612	349

- (a) Calculate the unweighted average of relatives price index for 1996 using 1994 as the base period.
 (b) Calculate the weighted average of relatives price index for 1996 using the dollar value for each product in 1994 as the appropriate set of weights and 1994 as the base year.

- 16-25** A survey of transatlantic passenger rates for roundtrip flights from New York to various European cities produced these results:

Destination	Average Annual Passenger Rates					Passengers (x 1,000)
	1991	1992	1993	1994	1995	
Paris	\$690	\$714	\$732	\$777	\$783	2,835
London	648	654	675	696	744	5,175
Munich	702	723	753	768	798	2,505
Rome	840	867	903	939	975	2,145

Calculate the weighted average of relatives index for each of the years 1991 through 1994 using 1995 as the base year and for weighting.

- 16-26** In a study of group health insurance policies commissioned by the Rhode Island Medical Care Association, the following sample of average individual rates was collected. Using 1994 as the base period, calculate an unweighted average of relatives price index for each year.

Insurance Group	1992	1993	1994	1995
Physicians	\$54	\$65	\$86	\$103
Students	39	41	55	76
Government employees	48	61	76	93
Teachers	46	58	75	96

- 16-27** A new motel chain hopes to place its first motel in Boomingville, but before it makes a commitment to start construction, it wants to check the room prices charged nightly by the other motels and hotels. After sending an employee to investigate the prices, the motel chain received data in the following form:

Hotel	Price per Room per Night			No. Rooms Rented
	1993	1994	1995	
Happy Hotel	\$35	\$37	\$42	1,250
Room Service Rooms	25	26	28	950
Executive Motel	45	45	51	1,000
Country Inn	37	38	44	600
Family Fun Motel	26	30	31	2,075

Help the company determine the relative prices using 1993 as the base year and using an unweighted average of relatives index.

- 16-28** The Quick-Stop Gas Station has been selling road maps to its customers for the past 3 years. The maps that are sold are of the nearest city, the county the gas station is in, the state it is in, and the entire United States. From the following table, calculate the weighted average of relatives price indices for 1994 and 1995 using 1993 as the base year.

Maps	Quantity Sold			
	1993	1994	1995	1993
City	\$0.75	\$0.90	\$1.10	1,000
County	0.75	0.90	1.00	400
State	1.00	1.50	1.50	1,000
United States	2.50	2.75	2.75	220

Worked-Out Answer to Self-Check Exercise

SC 16-5

Product	<u>1993</u> $\frac{P_1}{P_0}$	<u>1995</u> $\frac{P_1}{P_0}$	$\frac{P_1}{P_0}$	$P_0 Q_0$	$\left(\frac{P_1}{P_0}\right)(P_0 Q_0)$
Calculators	\$27	\$20	0.7407	150	111.11
Radios	30	42	1.4000	900	1,260.00
TVs	157	145	0.9236	1,370	1,265.33
			<u>3.0643</u>	<u>2,420</u>	<u>2,636.44</u>

$$(a) \text{ Index} = \frac{\sum \left(\frac{P_i}{P_0} \times 100 \right)}{n} = \frac{306.43}{3} = 102.1$$

$$(b) \text{ Index} = \frac{\sum \left(\frac{P_i}{P_0} \times 100 \right) (P_0 Q_0)}{\sum P_0 Q_0} = \frac{263,644}{2,420} = 108.9$$

16.5 QUANTITY AND VALUE INDICES

Quantity Indices

Our discussion of index numbers up to now has concentrated on price indices so that it would be easier to understand the general concepts. However, we can also use index numbers to describe quantity and value changes. Of these two, we use quantity indices more often. The Federal Reserve Board calculates quarterly indices in its monthly publication *The Index of Industrial Production* (IIP). The IIP measures the quantity of production in the areas of manufacturing, mining, and utilities. It is computed using a weighted average of relatives quantity index in which the fixed weights (prices) and the base quantities are measured from 1977.

In times of inflation, a quantity index provides a more reliable measure of actual output of raw materials and finished goods than a corresponding value index does. Similarly, agricultural production is best measured using a quantity index because it eliminates misleading effects due to fluctuating prices. We often use a quantity index to measure commodities that are subject to considerable price variation.

Any of the methods discussed in previous sections of this chapter to determine price indices can be used to calculate quantity indices. When we computed price indices, we used quantities or values as weights. Now that we want to compute quantity indices, we use prices or values as weights. Let's consider the construction of a weighted average of relatives quantity index.

The general process for computing a weighted average of relatives quantity index is the same as that used to compute a price index. Equation 16-9 describes the formula for this type of quantity index. In this equation, value is determined by multiplying quantity by price. The value associated with each quantity is used to weight the elements in the composite.

Using a quantity index

Advantages of a quantity index

Computing a weighted average of relatives quantity index

Weighted Average of Relatives Quantity Index

$$\frac{\sum \left[\left(\frac{Q_i}{Q_0} \times 100 \right) (Q_n P_n) \right]}{\sum Q_n P_n} \quad [16-9]$$

where

- Q_i = quantities for the current period
- Q_0 = quantities for the base period
- P_n and Q_n = quantities and prices that determine values we use for weights. In particular, $n = 0$ for the base period, $n = 1$ for the current period, and $n = 2$ for a fixed period that is not a base or current period.

Consider the problem in Table 16-12. We use Equation 16-9 to compute a weighted average of relatives quantity index. The value $Q_n P_n$ is determined from the base period and is therefore symbolized $Q_0 P_0$.

TABLE 16-12 COMPUTATION OF A WEIGHTED AVERAGE OF RELATIVES QUANTITY INDEX

Elements in the Composite (1)	Quantities (billions of bushels)		Price (per bushel)	$\frac{Q_1}{Q_0} \times 100$ Percentage Relatives (5) = $\frac{(3)}{(2)} \times 100$	Base Value (6) = (2) \times (4)	$\frac{Q_1}{Q_0} \times 100 \times Q_0 P_0$ Weighted Relatives (7) = (5) \times (6)
	Q_0 (2)	Q_1 (3)				
	P_0 (4)					
Wheat	29	24	\$3.80	$\frac{24.0}{29.0} \times 100 = 83$	$29 \times 3.80 = 110.20$	9,146.60
Corn	3	2.5	2.91	$\frac{2.5}{3} \times 100 = 83$	$3 \times 2.91 = 8.73$	724.59
Soybeans	12	14	6.50	$\frac{14.0}{12.0} \times 100 = 117$	$\frac{12 \times 6.50 = 78.00}{\sum Q_0 P_0 = 196.93}$	9,126.00
$\text{Weighted average of relatives quantity index} = \frac{\sum \left[\left(\frac{Q_1}{Q_0} \times 100 \right) (Q_0 P_0) \right]}{\sum Q_0 P_0}$ $= \frac{18,997.19}{196.93}$ $= 96$ [16-9]						

Value Indices

A value index measures general changes in the total value of some variable. Because value is determined both by price and quantity, a value index actually measures the combined effects of price and quantity changes. The principal disadvantage of a value index is that it does not distinguish between the effects of these two components.

Nevertheless, a value index is useful in measuring overall changes. Medical insurance companies, for example, often cite the sharp increase in the *value* of payments awarded in medical malpractice suits as the primary reason for discontinuing malpractice insurance. In this situation, value involves both a greater number of payments and larger cash amounts awarded.

A disadvantage of a value index

Advantages of a value index

HINTS & ASSUMPTIONS

A quantity index is often used in production decisions because it avoids the effects of inflation and price fluctuations due to market dynamics. Hint: Think about your pizza delivery service, whose total dollar revenue may decrease during periods of high use of discount coupons. Because the company expects the *quantity* of pizzas to increase as a result of discounting, a quantity index is more useful in making decisions about reordering cheese, toppings, and dough and scheduling delivery people.

EXERCISES 16.5

Self-Check Exercise

- SC 16-6** William Olsen, owner of a real estate office, has collected the following sales information for each of the firm's sales personnel

Salesperson	Value of Sales ($\times \$ 1,000$)			
	1992	1993	1994	1995
Thompson	490	560	530	590
Alfred	630	590	540	680
Jackson	760	790	810	840
Blockard	230	250	240	360

Calculate an unweighted average of relatives value index for each year using 1992 as the base period.

Basic Concepts

- 16-29** Explain the principal disadvantage in using value indices.
16-30 What is the major difference between a weighted aggregates index and a weighted average of relatives index?

Applications

- 16-31** The financial VP of the American division of Banshee Camera Company is examining the company's cash and credit sales over the last 5 years.

	Value of Sales ($\times \$ 100,000$)				
	1991	1992	1993	1994	1995
Credit	5.66	6.32	6.53	6.98	7.62
Cash	2.18	2.51	2.48	2.41	2.33

Calculate an unweighted average of relatives value index for each year using 1991 as the base period.

- 16-32** A Georgia firm manufacturing heavy equipment has collected the following production information about the company's principal products. Calculate a weighted aggregates quantity index using the quantities and prices from 1995 as the bases and the weights.

Product	Quantities Produced			Cost of Production/ Unit (thousands)
	1993	1994	1995	
River barges	92	118	85	\$ 33
Railroad gondola cars	456	475	480	56
Off-the-road trucks	52	56	59	116

- 16-33** Arkansas Electronics has marketed three basic types of calculators: for the business sector, the scientific sector, and a simple model capable of basic computational functions. The following information describes unit sales for the past 3 years:

Model	Number Sold ($\times 100,000$)			Price 1995
	1993	1994	1995	
Business	11.85	13.32	15.75	\$34.00
Scientific	10.32	11.09	10.18	69.00
Basic	7.12	7.48	7.89	13.00

Calculate the weighted average of relatives quantity indices using the prices and quantities from 1995 to compute the value weights with 1993 as the base year.

- 16-34** In preparation for an appropriations hearing, the police commissioner of a Maryland town has collected the following information:

Type of Crime	1992	1993	1994	1995
Assault and rape	110	128	134	129
Murder	30	45	40	48
Robbery	610	720	770	830
Larceny	2,450	2,630	2,910	2,890

Calculate the unweighted average of relatives quantity index for each of these years using 1995 as the base period.

- 16-35** Recycled Sounds has collected the following sales information for five different styles of music. Data are presented in hundreds of compact discs sold per year.

Type	Number of Sales					
	1991	1992	1993	1994	1995	1996
Soft rock	642.4	721.5	842.6	895.3	905.6	951.2
Hard rock	325.8	347.8	398.5	406.3	418.7	426.4
Classical	118.3	123.6	174.3	176.2	174.9	185.3
Jazz	125.6	122.4	137.8	149.6	172.9	205.4
Alternative	208.7	252.7	405.9	608.9	942.7	987.4

Calculate an unweighted average of relatives quantity index for each year using 1991 as the base year.

- 16-36** After encouraging a chemical company to make its employees handle certain dangerous chemicals with protective gloves, the Public Health Agency is now interested in seeing whether this ruling has had its effect in curbing the number of cancer deaths in that area. Before this rule went into effect, cancer was widespread not only among the workers at the company, but also among their families, close friends, and neighbors. The following data show what these numbers were in 1973 before the ruling and what they were after the ruling in 1993.

Age Group	No. in Population for 1973	Deaths in 1973	Deaths in 1993
<4 yr	5,000	400	125
4–15 yr	4,000	295	200
16–35 yr	24,000	1,230	1,000
36–60 yr	19,000	700	450
>60 yr	7,000	1,100	935

Use a weighted aggregates index of the number of deaths using the 1973 population size as the weights to help the Public Health Agency understand what has happened to the cancer rate.

- 16-37** A veterinarian has noticed she has treated a large number of pets this past winter. She wonders whether this number was spread across the 3 winter months evenly or whether she treated more pets in any certain month. Using December as the base period, calculate the weighted average of relatives quantity indices for January and February.

	Number Treated			Price per Visit	
	Dec.	Jan.	Feb.	Average for 3 Months	
Cats	100	200	95	\$ 55	
Dogs	125	75	200	65	
Parrots	15	20	15	85	
Snakes	10	5	5	100	

Worked-Out Answer to Self-Check Exercise

SC 16-6	1992	1993	1994	1995	1992	1993	1994	1995
Salesperson	V_0	V_1	V_2	V_3	V_0/V_0	V_1/V_0	V_2/V_0	V_3/V_0
Thompson	490	560	530	590	1.000	1.143	1.082	1.204
Alfred	630	590	540	680	1.000	0.937	0.857	1.079
Jackson	760	790	810	840	1.000	1.039	1.066	1.105
Blockard	230	250	240	360	1.000	1.087	1.043	1.565
					4.000	4.206	4.048	4.953

$$\text{Index} = \frac{\sum \left(\frac{V_i}{V_0} \times 100 \right)}{4} : \frac{400.0}{4} \quad \frac{420.6}{4} \quad \frac{404.8}{4} \quad \frac{495.3}{4}$$

$$= 100.0 \quad = 105.2 \quad = 101.2 \quad = 123.8$$

16.6 ISSUES IN CONSTRUCTING AND USING INDEX NUMBERS

In this chapter, we have used examples with small samples and short time spans. Actually, index numbers are computed for composites with many elements, and they cover long periods of time. This produces relatively accurate measures of changes. However, even the best index numbers are imperfect.

Imperfections in index numbers

Problems in Construction

Although there are many problems in constructing index numbers, there are three principal areas of difficulty:

1. **Selecting an item to be included in a composite.** Almost all indices are constructed to answer a particular question. Thus, the items included in the composite depend on the question. The Consumer Price Index asks, “How much has the price of a certain group of items purchased by moderate-income urban Americans changed from one period to another?” From this question, we know that only the items that reflect the purchases of moderate-income urban families should be included in the composite. We must realize that the Consumer Price Index will less accurately reflect price changes of goods purchased by low or high-income rural families than by moderate-income urban families.

Which items should be included in a composite?

2. **Selecting the appropriate weights.** In the previous sections of this chapter, we emphasized that the weights selected should represent the relative importance of the various elements. Unfortunately, what is appropriate in one period may become inappropriate in a short period of time. This must be kept in mind when comparing values of indices computed at different times.

Need for selection of appropriate weights

3. **Selecting the base period.** Typically, the base period selected should be a normal period, preferably a fairly recent period. “Normal” means that the period should not be at either the peak or the trough of a fluctuation. One technique to avoid using an irregular period is to average the values of several consecutive periods to determine a normal value. The U.S. Bureau of Labor Statistics uses the average of 1982, 1983, and 1984 consumption patterns to compute the Consumer Price Index. Management often tries to select a base period that coincides with the base period for one or more of the major indices, such as the Index of Industrial Production. Use of a common base allows management to relate its index to the major indices.

What is a normal base period?

Caveats in Interpreting an Index

In addition to these problems in constructing indices, there are several common errors made in interpreting indices:

1. **Generalization from a specific index.** One of the most common misinterpretations of an index is generalization of the results. The Consumer Price Index measures how prices of a particular combination of goods purchased by moderate-income urban Americans have changed. Despite its specific definition, the Consumer Price Index is often described as reflecting the cost of living for

Problems with generalizing from an index

all Americans. Although it is related to the cost of living to some degree, to say that it measures the change in the cost of living is not correct.

- 2. Lack of general knowledge regarding published indices.** Part of the problem leading to the first error is lack of knowledge of what the various published indices measure. All the well-known indices are accompanied by detailed statements concerning measurement. Management should become familiar with exactly what each index measures
- 3. Effect of time span on an index.** Factors related to an index tend to change with time. In particular, the appropriate weights tend to change. Thus, unless the weights are changed accordingly, the index becomes less reliable.
- 4. Quality changes.** One common criticism of index numbers is that they do not reflect changes in the quality of the items they measure. If the quality has indeed changed, then the index either understates or overstates the price-level changes. For example, if we construct an index number to describe price changes in pocket calculators over the last decade, the resulting index would underestimate the actual change that is due to rapid technological improvements in calculators.

Additional knowledge needed

Time affects an index

Lack of measurement of quality

EXERCISES 16.6

Basic Concepts

- 16-38** What is the effect of time on the weighting of a composite index?
- 16-39** List several preferences for the choice of base period.
- 16-40** Describe a technique used to avoid the use of an irregular period for a base.
- 16-41** Is it correct to say that the Consumer Price Index measures the cost of living?
- 16-42** What problems exist with index numbers if the quality of an item changes?

STATISTICS AT WORK

Loveland Computers

Case 16: Index Numbers “Lee, help me figure out these shipping charges.” Walter Azko was looking at a contract about half an inch thick. “The way we do our buying, the manufacturers are responsible for delivering an order to the airport and then an international shipping agent arranges for all the paperwork and loading. Sometimes it feels as if I’m paying the agents more for shipping the goods than I pay the manufacturer for making them. This contract right here is a good example. They want more than 10 percent more than I paid them for a similar shipment last quarter. When I called them, they gave some excuse about the cost of living going up.”

“But not by 10 percent,” Lee interjected.

“No, and the price of jet fuel went down, so the air-freight bill should be less.”

“Well, at least you don’t have to worry about exchange rates,” Lee said, glancing over the contract. “This says you’re to make payment in U.S. dollars.”

“That’s true—we do send them a check in dollars and they clear it through the local branch of an American bank. Even though the dollar isn’t quite the universal currency it once was, people still think it’s less risky than many other currencies. But once the agent has cashed the check, they still have to exchange dollars for local currency. They can’t pay their warehouse workers in dollars. So, even though the price is stated in dollars, I can tell that I get a better deal when the dollar is ‘strong’ against other currencies.”

"The cost of living is one factor, the cost of aviation fuel is another, and the exchange rate is the third. Does that cover everything?"

"I suppose so," Walter replied. "But with three things going up and down, it's hard to bargain with the agent and tell him I think they're too high."

"I think I have a way I can help," Lee offered cheerfully. "Can I take the afternoon to go down to Denver and talk with the international department of our bank?"

Study Questions: What solution is Lee going to propose as a way to evaluate the proposed price for the shipping agent's contract? What information will Lee be looking for in the bank's international department?

CHAPTER REVIEW

Terms Introduced in Chapter 16

Consumer Price Index The U.S. government prepares this index, which measures changes in the prices of a representative set of consumer items.

Fixed-Weight Aggregates Method To weight an aggregates index, this method uses as weights quantities consumed during some representative period.

Index of Industrial Production Prepared monthly by the Federal Reserve Board, the IIP measures the quantity of production in the areas of manufacturing, mining, and utilities.

Index Number A ratio that measures how much a variable changes over time.

Laspeyres Method To weight an aggregates index, this method uses as weights the quantities consumed during the base period.

Paasche Method In weighting an aggregates index, the Paasche method uses as weights the quantities consumed during the current period.

Percentage Relative Ratio of a current value to a base value with the result multiplied by 100.

Price Index Compares levels of prices from one period to another.

Quantity Index A measure of how much the number or quantity of a variable changes over time.

Unweighted Aggregates Index Uses all the values considered and assigns equal importance to each of these values.

Unweighted Average of Relatives Method To construct an index number, this method finds the ratio of the current price to the base price for each product, adds the resulting percentage relatives, and then divides by the number of products.

Weighted Aggregates Index Using all the values considered, this index assigns weights to these values.

Weighted Average of Relatives Method To construct an index number, this method weights by importance the value of each element in the composite.

Equations Introduced in Chapter 16

$$16-1 \quad \text{Unweighted aggregates quantity index} = \frac{\sum Q_L}{\sum Q_0} \times 100 \qquad \text{p. 874}$$

To compute an unweighted aggregates index, divide the sum of the current-year quantities of the elements in the index by the sum of the base-year quantities and multiply the result by 100.

16-2 Weighted aggregates price index = $\frac{\sum P_i Q_i}{\sum P_0 Q_i} \times 100$ p. 879

For a weighted aggregates price index using quantities as weights, obtain the weighted sum of the current-year prices by multiplying each price in the index by its associated quantity and summing the results. Then divide this weighted sum by the weighted sum of the base-year prices and multiply the result by 100.

16-3 Laspeyres index = $\frac{\sum P_i Q_0}{\sum P_0 Q_0} \times 100$ p. 880

The Laspeyres price index is a weighted aggregates price index using the base-year quantities as weights.

16-4 Paasche index = $\frac{\sum P_i Q_i}{\sum P_0 Q_i} \times 100$ p. 882

To get the Paasche price index, we compute a weighted aggregates price index using the current-year quantities for weights.

16-5 Fixed-weight aggregates price index = $\frac{\sum P_i Q_2}{\sum P_0 Q_2} \times 100$ p. 883

The fixed-weight aggregates price index is a weighted aggregates price index whose weights are the quantities from a representative year, not necessarily either the base year or the current year.

16-6 Unweighted average of relatives price index = $\frac{\sum \left(\frac{P_i}{P_0} \times 100 \right)}{n}$ p. 888

We compute an unweighted average of relatives price index by multiplying the ratios of current prices to base prices by 100, summing the results, and then dividing by the number of elements used in the index.

16-7 Weighted average of relatives price index = $\frac{\sum \left[\left(\frac{P_i}{P_0} \times 100 \right) (P_n Q_n) \right]}{\sum P_n Q_n}$ p. 890

With this index, we weight the relative prices by the values for a fixed reference period and divide the weighted sum of relative prices by the sum of the weights. If we use the base year values as weights, we get

16-8
$$\frac{\sum \left[\left(\frac{P_i}{P_0} \times 100 \right) (P_0 Q_0) \right]}{\sum P_0 Q_0}$$
 p. 890

which is the same as the Laspeyres price index.

16-9

$$\text{Weighted average of relatives quantity index} = \frac{\sum \left[\left(\frac{Q_i}{Q_0} \times 100 \right) (Q_n P_n) \right]}{\sum Q_n P_n} \quad \text{p. 895}$$

In this quantity index, we weight the relative quantities by the values for a fixed reference period and divide the weighted sum by the sum of the weights.

Review and Application Exercises

- 16-43 Kamischika Motorcycles began producing three models of mopeds in 1993. For the 3 years 1993 through 1995, sales were as follows:

Model	Average Annual Price			Units Sold ($\times 10,000$)		
	1993	1994	1995	1993	1994	1995
I	\$139	\$155	\$149	3.7	4.1	7.6
II	169	189	189	2.3	4.6	8.1
III	199	205	219	1.6	2.1	3.4

- (a) Calculate the weighted average of relatives price indices using the prices and quantities from 1995 as the bases and weights.
- (b) Calculate the weighted average of relatives price indices using the total dollar values for each year as the weights and 1995 as the base year.

- 16-44 These data indicate the value (in millions of dollars) of the principal products exported by a developing country. Determine unweighted aggregate value indices for 1993 and 1995 based on 1991.

Commodity	1991	1993	1995
Coffee	\$834	\$1,436	\$1,321
Sugar	96	118	122
Copper	241	258	269
Zinc	142	125	106

- 16-45 In a survey of U.S. coal production for 4 years, the following information was collected. Using the value of the 1992 production for weighting and 1992 as the base year, calculate the weighted average of relatives quantity index for each of the 4 years.

Type of Coal	Production (millions of tons)				Value (\$ millions)
	1989	1990	1991	1992	
Anthracite	7.4	6.8	7.1	7.2	90
Bituminous	595	580	601	625	5,050

- 16-46** A survey by the National Dairy Products Association produced the following information. Construct a Laspeyres index with 1991 as the base period.

Product	Average Price per Unit		Total Quantity (billions)
	1991	1995	
Cheese (lb)	\$1.45	\$1.49	2.6
Milk (gal)	1.60	1.61	47.6
Butter (lb)	0.70	0.80	3.1

- 16-47** Robert Barry, Ltd., a garment consulting firm, has examined the pricing trends of clothing items for a client. This table contains the results of the survey (shown in unit prices):

Products	1992	1993	1994	1995
Jeans	\$13.00	\$13.00	\$15.00	\$15.00
Jackets	19.00	19.50	22.00	24.00
Shirts	12.00	11.00	12.00	13.00

Calculate an unweighted average of relatives index for each year using 1992 as the base period.

- 16-48** What problem would exist in comparing price indices describing computer sales over the past few decades?

- 16-49** The VP of sales for the National Hospital Supply Company conducted a survey of travel expenses incurred by selected salespeople. Of particular interest were the following data regarding expenditures for taxis and the price paid per mile.

Salespeople	Expenditures on Taxis			Average Price/Mile 1991
	1991	1992	1993	
A	\$704	\$985	\$1,391	0.52
B	635	875	1306	0.55
C	752	1,023	1,523	0.59
D	503	696	1,106	0.56
E	593	781	1,215	0.55

Calculate an unweighted average of relatives index for each year using 1993 as the base period.

- 16-50** This information describes the unit sales of a bicycle shop for 3 years:

Model	Number Sold			Price 1993
	1993	1994	1995	
Sport	45	48	56	\$ 89
Touring	64	67	71	104
Cross-country	28	35	27	138
Sprint	21	16	28	245

Calculate the weighted average of relatives quantity indices using the prices and quantities from 1993 to compute the value weights, with 1993 as the base year.

- 16-51** The Dow Jones Industrial Average (DJIA) is an index number used by many people as a proxy for describing the overall strength of prices on the New York Stock Exchange. It is based on the sum of the prices of single shares of the common stock of 30 large companies traded on the exchange. This sum is then adjusted to account for splits and changes in the companies whose shares make up the index.

- Two of the stocks in the index are Coca Cola, which was trading around \$44 per share in late July 1993, and Westinghouse, which was then trading around \$17 per share. What information does the DJIA ignore by simply adding single-share prices? Does a 10-percent rise in the price of Westinghouse stock have the same effect as a 10-percent rise in the share price of Coca Cola?
- The total annual return of U.S. common stocks has been about 11 percent, as an average, over long time periods. But stockbrokers sometimes choose low points in the market (selected with hindsight) to express gains over time. At the end of 1992, the DJIA stood at 3301. Calculate an index number for how well stocks have done recently, based on the bottom of the market after the October 1987 crash, when the DJIA stood at 1739. Compare this with an index number based on the August 1987 high point of the market, when the DJIA was 2722.

- 16-52** Pem Jenkins runs a lumberyard and has the following information on costs for 3 years:

Costs	1991	1993	1995
Wages	\$24,378	\$36,421	\$37,613
Lumber	1,816	2,019	2,136
Utilities	638	681	701

Construct an unweighted aggregates index for production costs in 1991 and 1995 using 1993 as the base year.

- 16-53** An Ohio consumer protection agency has surveyed the price changes of a meatpacking company. The following table contains the average annual per-pound prices for a sample of the firm's products. Construct an unweighted average of relatives price index using the prices from 1993 as the base period.

Products	1993	1994	1995
Sirloin	\$1.69	\$1.81	\$1.85
Chuck	0.91	1.15	1.24
Bologna	1.45	1.58	1.53
Hot dogs	0.99	1.03	1.01
Rib eyes	2.39	2.61	2.56

- 16-54** Why must one exercise caution in selecting a base period?

- 16-55** Tameka Robinson, a purchasing agent, has compiled the following price information. Using 1992 as the base period, calculate the unweighted aggregates price index for 1993, 1994, and 1995.

Material	1992	1993	1994	1995
Aluminum	\$0.96	\$0.99	\$1.03	\$1.06
Steel	1.48	1.54	1.55	1.59
Brass tubing	0.21	0.25	0.26	0.31
Copper wire	0.06	0.08	0.07	0.09

- 16-56** A USDA survey of grain production for selected areas in the United States yielded this information:

Product	Quantities Produced (millions of bushels)					Price per Bushel 1991
	1991	1992	1993	1994	1995	
Wheat	610	620	640	630	650	\$ 4.40
Corn	390	390	410	440	440	3.60
Oats	100	90	120	130	150	1.20
Rye	10	20	10	10	20	24.00
Barley	160	150	120	190	180	2.10
Soybeans	130	140	160	120	130	5.60

Using the prices from 1991 for weights, calculate the weighted aggregates quantity indices for each year.

- 16-57** John Pringle, an international mineral trader, has collected the following information on prices and quantities of minerals exported by an African country for the years 1994 and 1995. Calculate a Paasche index for 1995 using 1994 as the base period.

Mineral	Quantity (million tons)		Price (per lb)	
	1995	1994	1994	1995
Copper	38.1		\$0.59	\$0.63
Lead	53.5		0.17	0.16
Zinc	86.4		0.21	0.23

- 16-58** A European automobile manufacturer has compiled the following information on car sales of one U.S. manufacturer:

Size	Average Annual Price (hundreds)			Units Sold ($\times 1,000$)		
	1991	1993	1995	1991	1993	1995
Subcompact	\$62	\$68	\$70	32	65	86
Compact	76	78	80	45	68	73
Sedan	90	98	106	462	325	386

- Calculate the weighted average of relatives price indices using the prices and quantities from 1993 as the bases and weights.
- Calculate the weighted average of relatives price indices using the total dollar values for each year as the weights and 1993 as the base year.

- 16-59** Sylvia Jensen, cost analyst for a major appliance firm, has compiled price data for four of the company's products. The figures (given in unit prices) for 1993 through 1996 are shown in the table.

Products	1993	1994	1995	1996
Dishwasher	\$219	\$241	\$272	\$306
Washing machine	362	385	397	413
Dryer	229	241	261	275
Refrigerator	562	580	598	625

Using 1993 as the base period, express the prices in 1994, 1995, and 1996 in terms of an unweighted aggregates index.

- 16-60** The budget director for a New England college wants to keep track of the budget that each engineering department requires to recruit new graduate students. He has received the following data from four departments.

Department	Total Expenditures		
	1994	1995	1996
Mechanical	\$3,642	\$3,891	\$4,253
Chemical	3,888	4,052	4,425
Biomedical	4,251	4,537	4,724
Electrical	3,764	4,305	4,297

Calculate an unweighted average of relatives index for each year using 1994 as the base period.

- 16-61** In 1991, the average weekly wage for a certain group of households was \$422.60. In 1996, the average weekly wage for the same group was \$521.35. The Consumer Price Index in 1996 using 1991 as a base, was 152. Calculate the "real" average weekly wage for this group in 1996.

- 16-62** A national shopping survey was conducted to study the average weekly buying habits of a typical family in 1992 and 1996. The data collected are as follows:

Items	1992		1996	
	Unit Price	Quantity	Unit Price	Quantity
Cheese (8 oz)	\$1.19	2	\$2.09	1
Bread (1 loaf)	0.79	3	1.09	3
Eggs (1 doz)	0.84	2	1.35	1
Milk (1 gal)	1.36	2	2.39	2

Calculate a Paasche index for 1996 using 1992 as the base period.

- 16-63** Snow Mountain has several different ticket prices, including discounts for people who own property in the area, handicapped skiers, and snowboarders. The average number of tickets sold per ski-day was as follows:

	1993	1994	1995	1996
Local	73	76	112	107
Snowboard	101	129	163	162
Disabled	163	189	271	268
Regular price	183	210	303	298

Calculate the unweighted average of relatives quantity index for each of these years using 1996 as the base period.

- 16-64** Francis Hill, president of an agricultural trade consulting company, has obtained the following information on grain (prices and sales) exported by the United States.

Product	Amount Exported (in millions of tons)				Price per Ton 1994
	1992	1993	1994	1995	
Wheat	4.6	6.7	4.0	5.2	\$2,680
Feed grains	4.9	6.2	1.8	1.2	2,270
Soybeans	4.7	5.7	1.2	1.8	3,430

Compute the weighted aggregates quantity indices for each year using the prices for 1994 as weights and 1994 as the base year.

- 16-65** Andrea Graham, a budget analyst for a long-distance phone company, has collected price and sales volume data for phone calls from New York to Boston. The data for each of the three rate schedules are as follows:

Rate (times)	Price per Call (per minute)		Total # Calls (millions) 1991
	1991	1996	
Day (8 A.M. – 5 P.M.)	\$0.17	\$0.19	5.2
Evening (5 P.M. – 11 P.M.)	0.13	0.16	8.7
Night (11 P.M. – 8 A.M.)	0.09	0.12	10.3

Construct a Laspeyres price index using 1991 as the base period.

- 16-66** The Reliable Bus Company provides transportation for its own town, and in addition, it sells buses to neighboring towns. The company has collected the following data in order to analyze its sales for years 1992, 1994, and 1996.

Town	Average Selling Price per Bus			Number of Buses Sold 1994
	1992	1994	1996	
Greenville	\$21,206	\$24,210	\$26,235	17
Hampton	17,129	19,722	22,109	14
Middletown	25,723	28,657	32,481	21

Construct a Laspeyres index using 1994 as the base period.

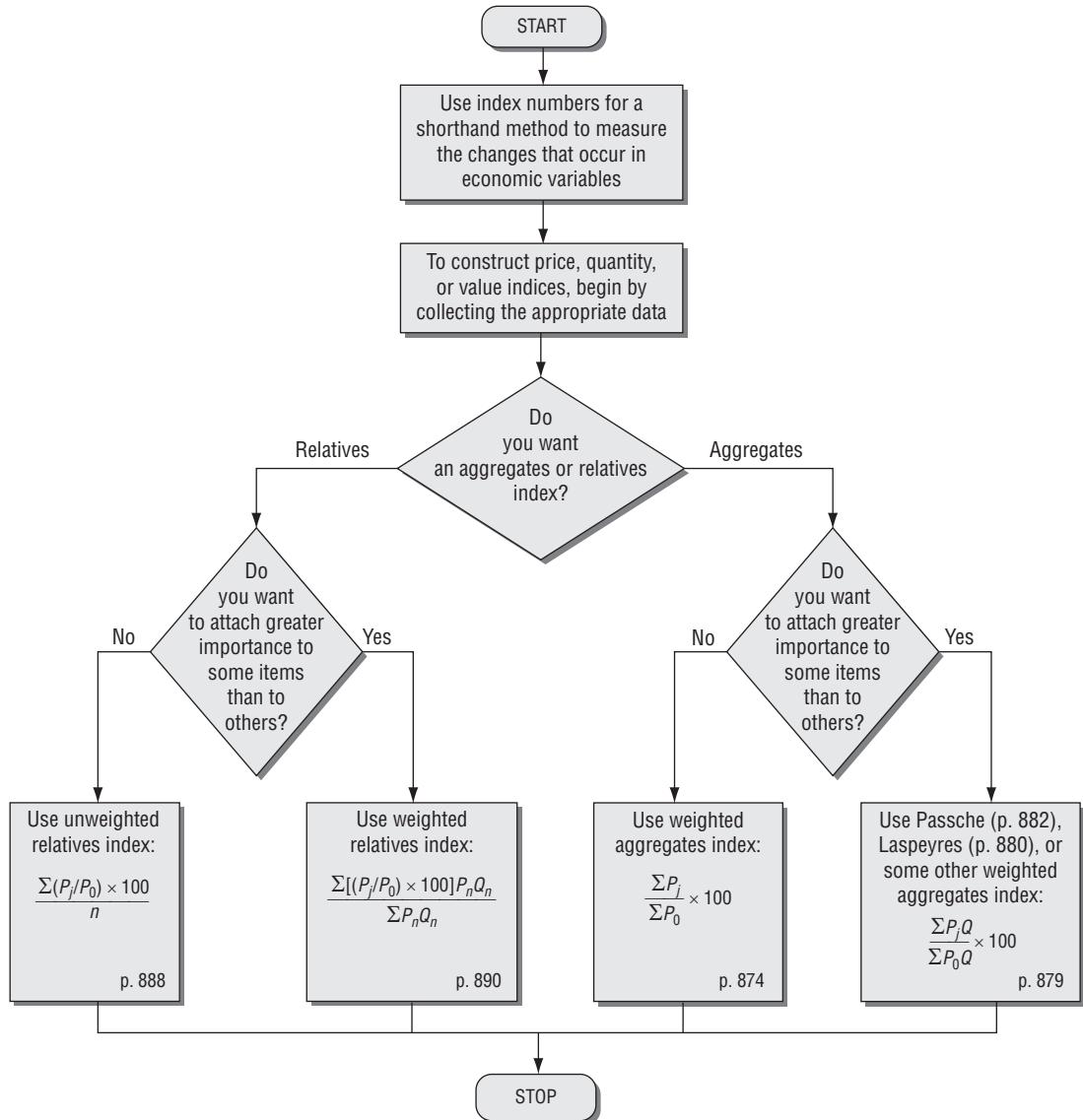
- 16-67** A local fast-food restaurant wants to examine how sales are changing for each of its four most popular menu items. The data for the years 1993 through 1996 follow.

Menu Item	Unit Price				Quantity Sold (millions)			
	1993	1994	1995	1996	1993	1994	1995	1996
Hamburger	\$0.58	\$0.62	\$0.69	\$0.79	2.1	2.5	2.0	1.8
Chicken sandwich	1.89	2.09	2.18	2.25	1.5	1.2	1.8	2.1
French fries	0.84	0.89	0.99	0.99	2.9	2.7	2.3	2.4
Onion rings	0.91	0.99	1.14	1.19	3.1	2.4	2.0	1.6

Calculate a fixed-weight aggregates index for each year using 1993 prices as the base and the 1996 quantities as the fixed weights.

- 16-68** Use the data from Exercise 16-67 to calculate a Paasche index for each year using 1995 as the base period.

Flow Chart: Index Numbers



Decision Theory

LEARNING OBJECTIVES

After reading this chapter, you can understand:

- To learn methods for making decisions under uncertainty
 - To use expected value and utility as decision criteria
 - To understand why additional information is useful and to calculate its value
 - To help decision makers supply needed probability values even when they do not understand probability theory
 - To learn how to use decision trees to structure and analyze complex decision-making problems
-

CHAPTER CONTENTS

17.1 The Decision Environment	912
17.2 Expected Profit under Uncertainty: Assigning Probability Values	913
17.3 Using Continuous Distributions: Marginal Analysis	922
17.4 Utility as a Decision Criterion	931
17.5 Helping Decision Makers Supply the Right Probabilities	935
17.6 Decision-Tree Analysis	939

■ Statistics at Work	952
■ Terms Introduced in Chapter 17	953
■ Equations Introduced in Chapter 17	953
■ Review and Application Exercises	954

Acme Fruit and Produce Wholesalers buys tomatoes, then sells them to retailers. Acme currently pays \$20 a box. Tomatoes sold on the same day bring \$32 a box. Extremely perishable, tomatoes not sold on the first day are worth only \$2 a box. Acme has calculated that the mean past daily sales is 60 boxes and that the standard deviation of past daily sales is 10 boxes. Using the techniques introduced in this chapter, we can tell Acme how many boxes to order each day to maximize profits.

In Section 5-3 (beginning on page 220), we introduced you to the idea of using expected value in decision making. There we worked through a simple problem involving the purchase of strawberries for resale. That kind of problem is part of a set of problems that can be solved using the techniques developed in this chapter.

In the last 35 years, managers have used newly developed statistical techniques to solve problems for which information was incomplete, uncertain, or in some cases almost completely lacking. This new area of statistics has a variety of names: *statistical decision theory*, *Bayesian decision theory* (after the Reverend Thomas Bayes, whom we introduced in Chapter 4), or simply *decision theory*. These names are used interchangeably.

When we did hypothesis testing, we had to decide whether to accept or to reject the stated hypothesis. In decision theory, we must decide among alternatives by taking into account the *monetary* repercussions of our actions. A manager who must select from among a number of available investments should consider the profit or loss that might result from each alternative. Applying decision theory involves selecting an alternative and having a reasonable idea of the economic consequences of choosing that action. ■

What is decision theory?

17.1 THE DECISION ENVIRONMENT

Decision theory can be applied to problems whether the time span is 5 years or 1 day, whether they involve financial management or a plant assembly line, and whether they are in the public or private sector. Regardless of the environment, most of these problems have common characteristics. As a result, decision makers approach their solutions in fairly consistent ways. The elements common to most decision-theory problems are these:

- An objective the decision maker is trying to reach.** If the objective is to minimize downtime of expensive machinery, the manager may try to find the optimal number of spare motors to be kept on hand for quick repairs. Success in finding that number can be measured by counting downtime each month.
- Several courses of action.** The decision should involve a choice among alternatives (called *acts*). In our example involving spare motors, the various acts open to the decision maker include stocking one, two, three, four, or five spare motors or choosing not to stock any spare motors.
- A calculable measure of the benefit or worth of the various alternatives.** In general, these costs can be negative or positive and are called *payoffs*. Cost accountants should be able to determine the cost of lost production time resulting from a motor burnout both when a spare is on hand and when one is not available. But sometimes the payoffs involve consequences that are more than solely financial. Imagine trying to decide the optimal number of spare generators a hospital might require in the event of a power failure. Not having enough could cost lives as well as money.
- Events beyond the control of the decision maker.** These uncontrollable occurrences are often called *outcomes* or *states of nature*, and their existence creates difficulties as well as interest in decision making under uncertainty. Such events could be the number of motors in our expensive

Elements common to decision-theory problems

production machinery that will burn out in a given month. Preventive maintenance will reduce motor burnouts, but they will still happen.

- 5. Uncertainty concerning which outcome or state of nature will actually happen.** In our example, we are uncertain about how many motors will burn out. This uncertainty is generally handled by the use of probabilities assigned to the various events that might take place, say, a 0.1 chance of losing five motors a month.

EXERCISES 17.1

Applications

- 17-1** Wholesale Lamps has been in contact with Leerie's, a local retail lamp shop, about supplying it with a special chrome tree lamp, which the shop wants to use as a drawing card in an upcoming sale. Wholesale Lamps must order the lamps in 2 days to deliver them by the sale date. Wholesale's cost is \$49 for the lamps; it will sell them to Leerie's for \$54. Wholesale is uncertain about the number Leerie's desires but guesses that it will be between 15 and 20. One of the managers has assigned probabilities to the various numbers that Leerie's might order. The manager of Wholesale Lamps does not foresee a market for the lamps it does not sell to Leerie's. Leerie's is expected to submit the order tomorrow. Should the manager of Wholesale Lamps use decision theory to order the lamps for Leerie's?
- 17-2** Adventures, Inc., is a source of capital for entrepreneurs starting new firms in the field of genetic engineering. Lisa Levin, a partner in Adventures, has been examining several business proposals that have recently been made to her. Each proposal describes a new venture, outlines its potential market, and solicits investment by Adventures. Lisa has just finished reading the chapter on decision theory in her father's statistics text. She thinks decision theory provides a methodology that can help her decide which ventures to support and at what level. Is Lisa correct? If so, what information does she need in order to apply decision theory to her problem? If not, why not?
- 17-3** The 8th Avenue Book Store relied on Grambler News Service to supply it with several well-known magazines. Each week, Grambler would deliver a predetermined number of *Today's Romances*, among others, and pick up any unsold copies of the previous week's magazines. The number of copies that the bookstore would sell was never known for sure, but the manager did have past sales data. Grambler charged its bookstands \$1.60 for magazines that sold for \$2.95. Management of the bookstore wanted to get maximum profitability from the sale of its magazines and was considering the optimal number of *Today's Romances* to order. Should the manager of the bookstore use decision theory to decide the number of magazines to stock?

17.2 EXPECTED PROFIT UNDER UNCERTAINTY: ASSIGNING PROBABILITY VALUES

Buying and selling strawberries, as in our example in Chapter 5, is only one case in which decisions have to be made under uncertainty. Another involves a newspaper dealer who buys newspapers for 30¢ each and sells them for 50¢ each. Any papers not sold by the end of the day are completely worthless to him. The dealer's problem is to determine the optimal number he should order each day.

Buying decision under conditions of uncertainty

On days when he stocks more than he sells, his profits are reduced by the cost of the unsold papers. On days when buyers request more copies than he has in stock, he loses sales and makes smaller profits than he could have.

The dealer has kept a record of his sales for the past 100 days (Table 17-1). This information is a distribution of the dealer's past sales. Because sales volume can take on only a limited number of values, the distribution is discrete. We will assume, for purposes of discussion, that the dealer will sell only the numbers of papers listed—not, say, 412, 525, or 637. Furthermore, the dealer has no reason to believe that sales volume will take on any other value in the future.

This information tells the dealer something about the historical pattern of his sales. Although it does not tell him what quantity the buyers will request tomorrow, it does tell him that there are 45 chances in 100 that the quantity will be 500 papers. Therefore, a probability of 0.45 is assigned to the sales figure of 500 papers. The probability column in Table 17-1 shows the relationship between the total observations of sales (100 days) and the number of times each possible value of daily sales appeared in the 100 observations. The probability of each sales level occurring is thus derived by dividing the total number of times each value has appeared in the 100 observations by the total number of observations, that is, $15/100$, $20/100$, $45/100$, $15/100$, and $5/100$.

Computing probabilities of sales levels

TABLE 17-1 DISTRIBUTION OF NEWSPAPER SALES

Daily Sales	Number of Days Sold	Probability of Each Number Being Sold
300	15	0.15
400	20	0.20
500	45	0.45
600	15	0.15
700	5	0.05
	100	1.00

Maximizing Profits Instead of Minimizing Losses

Back in Section 5-3, when we first introduced you to using expected value in decision making, we used an approach that minimized losses and led us to an optimal stocking pattern for our strawberry dealer. It is just as easy to find the optimal stocking pattern by *maximizing profits*, and that's just what we'll do at this point.

Recall that our fruit and vegetable wholesaler in Chapter 5 bought strawberries at \$20 a case and resold them at \$50 a case. There we assumed that the product had no value if not sold on the first day (a restriction we shall soon lift). If buyers call for more cases tomorrow than the wholesaler has in stock, profits suffer by \$30 (selling price minus cost) for each case he cannot sell. On the other hand, costs also result from stocking *too many* units on a given day. If the wholesaler has 13 cases in stock but sells only 10, he makes a profit of \$300, or \$30 a case on 10 cases. But this profit must be reduced by \$60, the cost of the three cases not sold and of no value.

A Chapter 5 problem worked another way

TABLE 17-2 CASES SOLD DURING 100 DAYS

Daily Sales	Number of Days Sold	Probability of Each Number Being Sold
10	15	0.15
11	20	0.20
12	40	0.40
13	25	0.25
	100	1.00

A 100-day observation of past sales gives the information shown in Table 17-2. The probability values there are obtained just as they were in Table 5-6.

TABLE 17-3 CONDITIONAL PROFIT TABLE

Possible Demand (Sales) in Cases	Possible Stock Action			
	10 Cases	11 Cases	12 Cases	13 Cases
10	\$300	\$280	\$260	\$240
11	300	330	310	290
12	300	330	360	340
13	300	330	360	390

Notice that there are only four discrete values for sales volume, and as far as we know, there is no discernible pattern in the sequence in which these four values occur. We assume that the retailer has no reason to believe sales volume will behave differently in the future.

Calculating Conditional Profits

To illustrate this retailer's problem, we can construct a table showing the results in dollars of all possible combinations of purchases and sales. The only values for purchases and for sales that have meaning to us are 10, 11, 12, and 13 cases, because the retailer has no reason to consider buying fewer than 10 or more than 13 cases.

Table 17-3, called a *conditional profit table*, shows the profit resulting from any possible combination of supply and demand. **Conditional profit table** The profits could be either positive or negative (although they are all positive in this example) and are conditional in that a certain profit results from taking a specific stocking action (ordering 10, 11, 12, or 13 cases) and selling a specific number of cases (10, 11, 12, or 13 cases).

Table 17-3 reflects the losses that occur when stock remains unsold at the end of a day. Notice, too, that the retailer forgoes potential additional profit when customers demand more cases than he has stocked.

Observe that the stocking of 10 cases each day will always result in a profit of \$300. Even on days when buyers want 13 cases, the retailer can sell only 10. When the retailer stocks 11 cases, his profit will be \$330 on days when buyers request 11, 12, or 13 cases. But on days when he has 11 cases in stock and buyers buy only 10 cases, profit drops to \$280. The \$300 profit on the 10 cases sold must be reduced by \$20, the cost of the unsold case. A stock of 12 cases will increase daily profits to \$360, but only on days when buyers want 12 or 13 cases. Should buyers want only 10 cases, profit is reduced to \$260; the \$300 profit on the sale of 10 cases is reduced by \$40, the cost of two unsold cases. Stocking 13 cases will result in a profit of \$390 (a \$30 profit on each case sold, with no unsold cases) when there is a market for 13 cases. When buyers purchase fewer than 13 cases, such a stock action results in profits of less than \$390. For example, with a stock of 13 cases and sale of only 11 cases, the profit is \$290; the profit on 11 cases, \$330, is reduced by the cost of two unsold cases (\$40).

Such a conditional profit table does *not* show the retailer how many cases he *should* stock each day in order to maximize profits. It reveals the outcome only if a specific number of cases is stocked and a specific number of cases is sold. Under conditions of uncertainty, the retailer does not know in advance the size of any day's market. However, he must still decide which number of cases, stocked consistently, will maximize profits over a long period of time.

Explaining elements in the conditional profit table

Function of the conditional profit table

TABLE 17-4 EXPECTED PROFIT FROM STOCKING 10 CASES

Market Size in Cases (1)	Conditional Profit (2)		Probability of Market Size (3)		Expected Profit (4)
10	\$300	×	0.15	=	\$ 45.00
11	300	×	0.20	=	60.00
12	300	×	0.40	=	120.00
13	300	×	0.25	=	75.00
			1.00		\$300.00

Calculating Expected Profits

The next step in determining the best number of cases to stock is assigning probabilities to the possible outcomes or profits. We saw in Table 17-2 that the probabilities of the possible values for the retailer's sales are as follows:

Cases	10	11	12	13
Probability	0.15	0.20	0.40	0.25

Using these probabilities and the information contained in Table 17-3, we can now compute the expected profit of each possible stock action.

We stated in Chapter 5 that we can compute the expected value of a random variable by weighting each possible value the variable can take by the probability of its taking on that value. Using this procedure, we can compute the expected daily profit from stocking 10 cases each day. See Table 17-4. The figures in column 4 of Table 17-4 are obtained by weighting the conditional profit of each possible sales volume (column 2) by the probability of that conditional profit occurring (column 3). The sum in the last column is the expected daily profit resulting from stocking 10 cases each day. It is not surprising that this expected profit is \$300 because we saw in Table 17-3 that stocking 10 cases each day would always result in a daily profit of \$300, regardless of whether buyers wanted 10, 11, 12, or 13 cases.

The same computation for a daily stock of 11 units can be made, as we have done in Table 17-5. This tells us that if the retailer stocks 11 cases each day, his expected daily profit over time will be \$322.50. Eighty-five percent of the time

TABLE 17-5 EXPECTED PROFIT FROM STOCKING 11 CASES

Market Size in Cases	Conditional Profit		Probability of Market Size		Expected Profit
10	\$280	×	0.15	=	\$42.00
11	330	×	0.20	=	66.00
12	330	×	0.40	=	132.00
13	330	×	0.25	=	82.50
			1.00		\$322.50

TABLE 17-6 EXPECTED PROFIT FROM STOCKING 12 CASES

Market Size in Cases	Conditional Profit		Probability of Market Size		Expected Profit
10	\$260	×	0.15	=	\$39.00
11	310	×	0.20	=	62.00
12	360	×	0.40	=	144.00
13	360	×	0.25	=	90.00
			<u>1.00</u>		<u>\$335.00</u>
					← action

the daily profit will be \$330; on these days, buyers ask for 11, 12, or 13 cases. However, column 3 tells us that 15 percent of the time the market will take only 10 cases, resulting in a profit of only \$280. It is this fact that reduces the daily expected profit to \$322.50.

For 12 and 13 units, the expected daily profit is computed as **For 12 and 13 units** shown in Tables 17-6 and 17-7, respectively.

We have now computed the expected profit of each of the four stock actions open to the retailer. These expected profits are

- If 10 cases are stocked each day, the expected daily profit is \$300.00.
- If 11 cases are stocked each day, the expected daily profit is \$322.50.
- If 12 cases are stocked each day, the expected daily profit is \$335.00.
- If 13 cases are stocked each day, the expected daily profit is \$327.50.

The *optimal stock action* is the one that results in the greatest **Optimal solution** expected profit—the largest daily average profits and thus the maximum total profits over a period of time. In this illustration, the proper number to stock each day is 12 cases, because that quantity will give the highest possible average daily profits under the conditions given.

We have *not* reduced uncertainty in the problem facing the **What the solution means** retailer. Rather, we have used his past experience to determine the best stock action open to him. He still does not know how many cases will be requested on any given day. There is no guarantee that he will make a profit of \$335.00 tomorrow. However, if he stocks 12 cases each day under the conditions given, he will have *average* profits of \$335.00 per day. This is the *best* he can do, because the choice of any one of the other three possible stock actions will result in a lower expected daily profit.

TABLE 17-7 EXPECTED PROFIT FROM STOCKING 13 CASES

Market Size in Cases	Conditional Profit		Probability of Market Size		Expected Profit
10	\$240	×	0.15	=	\$36.00
11	290	×	0.20	=	58.00
12	340	×	0.40	=	136.00
13	390	×	0.25	=	97.00
			<u>1.00</u>		<u>\$327.50</u>

Expected Profit with Perfect Information

Now suppose that the retailer in our illustration could remove all uncertainty from his problem by obtaining complete and accurate information about the future, referred to as *perfect* information. This does not mean that sales would not vary from 10 to 13 cases per day. Sales would still be 10 cases per day 15 percent of the time, 11 cases 20 percent of the time, 12 cases 40 percent of the time, and 13 cases 25 percent of the time. However, with perfect information, the retailer would know in advance how many cases were going to be called for each day.

Definition of perfect information

Under these circumstances, the retailer would stock today the exact number of cases buyers will want tomorrow. For sales of 10 cases, the retailer would stock 10 cases and realize a profit of \$300. When sales were going to be 11 cases, he would stock exactly 11 cases, thus realizing a profit of \$330.00.

Use of perfect information

Table 17-8 shows the conditional profit values that are applicable to the retailer's problem if he has perfect information. Knowing the size of the market in advance for a particular day, the retailer chooses the stock action that will maximize his profits. This means he buys and stocks quantities that avoid *all* losses from obsolete stock as well as *all* losses that reflect lost profits on unfilled requests for strawberries.

Expected profit with perfect information

We can now compute the expected profit with perfect information. This is shown in Table 17-9. The procedure is the same as that already used, but you will notice that the conditional profit figures in column 2 of Table 17-9 are the maximum profits possible for each sales volume. When buyers buy 12 cases, for example, the retailer will always make a profit of \$360 with perfect information because he will have stocked exactly 12 cases. With perfect information, then, our retailer could count on making an average profit of \$352.50 a day. This is a significant figure because it is the *maximum expected profit* possible.

TABLE 17-8 CONDITIONAL PROFIT TABLE WITH PERFECT INFORMATION

Possible Sales in Cases	Possible Stock Action			
	10 Cases	11 Cases	12 Cases	13 Cases
10	\$300	—	—	—
11	—	\$330	—	—
12	—	—	\$360	—
13	—	—	—	\$390

TABLE 17-9 EXPECTED PROFIT WITH PERFECT INFORMATION

Market Size in Cases	Conditional Profit with Perfect Information		Probability of Market Size		Expected Profit with Perfect Information
10	\$300	×	0.15	=	\$45.00
11	330	×	0.20	=	66.00
12	360	×	0.40	=	144.00
13	390	×	0.25	=	97.50
			1.00		\$352.50

Expected Value of Perfect Information

Assuming that a retailer could obtain a perfect predictor about the future, what would be its value to him? He must compare the cost of that information with the additional profit he would realize as a result of having the information.

The retailer in our example can earn average daily profits of \$352.50 if he has perfect information about the future (see Table 17-9). His best expected daily profit without the predictor is only \$335.00 (see Tables 17-4 to 17-7). The difference of \$17.50 is the maximum amount the retailer would be willing to pay, per day, for a perfect predictor, because that is the maximum amount by which he can increase his expected daily profit. The difference is the *expected value of perfect information* and is referred to as EVPI. There is no sense in paying more than \$17.50 for the predictor; to do so would cost more than the knowledge is worth.

Value of perfect information

Why do we need the value of perfect information?

Calculating the value of additional information in the decision-making process is a serious problem for managers. In our illustration, we found that our retailer would pay \$17.50 a day for a perfect predictor. Only infrequently, however, can we secure a perfect predictor. In most decision-making situations, managers are really attempting to evaluate the worth of information that will enable them to make better, rather than perfect, decisions.

HINTS & ASSUMPTIONS

Warning: All of the examples used in this section have involved discrete distributions; that is, we've allowed the random variable to take on only a few values. This is not reflective of most real-world situations, but makes it easy for us to do the calculations necessary to introduce this idea. With discrete outcomes, the expected profit is *not* necessarily one of the outcomes. Hint: A 50 percent chance of making a \$10 profit coupled with a 50 percent chance of making no profit gives an expected profit of \$5. But with a discrete distribution the outcome will be *either* \$10 or zero! Some real-world situations also turn out like this. A parcel of undeveloped land can be worth either \$5 million or \$250,000, depending on where a new airport is finally located. The land may also be sold for \$500,000 to a speculator who hopes for the final \$5 million sale.

EXERCISES 17.2

Self-Check Exercise

SC 17-1 The Writer's Workbench operates a chain of word-processing franchises in college towns. For an hourly fee of \$8.00, Writer's Workbench provides access to a personal computer, word-processing software, and a printer to students who need to prepare papers for their classes. Paper is provided at no additional cost. The firm estimates that its hourly variable cost per machine (principally due to paper, ribbons, electricity, and wear and tear on the computers and printers) is about 85¢. Deborah Rubin is considering opening a Writer's Workbench franchise in Ames, Iowa. A preliminary market survey has resulted in the following probability distribution of the number of machines demanded per hour during the hours she plans to operate:

Number of machines	22	23	24	25	26	27
Probability	0.12	0.16	0.22	0.27	0.18	0.05

If she wishes to maximize her profit contribution, how many machines should Deborah plan to have? What is the hourly expected value of perfect information in this situation? Even if Deborah could obtain a perfectly accurate forecast of the demand for each and every hour, why wouldn't she be willing to pay up to the EVPI for that information in this situation?

Applications

- 17-4** Center City Motor Sales has recently incorporated. Its chief asset is a franchise to sell automobiles of a major American manufacturer. CCMS's general manager is planning the staffing of the dealership's garage facilities. From information provided by the manufacturer and from other nearby dealerships, he has estimated the number of annual mechanic hours that the garage will be likely to need.

Hours	10,000	12,000	14,000	16,000
Probability	0.2	0.3	0.4	0.1

The manager plans to pay each mechanic \$9.00 per hour and to charge customers \$16.00. Mechanics will work a 40-hour week and get an annual 2-week vacation.

- (a) Determine how many mechanics Center City should hire.
- (b) How much should Center City pay to get perfect information about the number of mechanics it needs?

- 17-5** Airport Rent-A-Car is a locally operated business in competition with several major firms. ARC is planning a new deal for customers who want to rent a car for only one day and return it to the airport. For \$24.95, the company will rent a small economy car to a customer, whose only other expense is to fill the car with gas at the day's end. ARC is planning to buy a number of small cars from the manufacturer at a reduced price of \$6,750. The big question is how many to buy. Company executives have decided on the following estimated probability distribution of the number of cars rented per day:

Number of cars rented	10	11	12	13	14	15
Probability	0.18	0.19	0.21	0.15	0.14	0.13

The company intends to offer the plan 6 days a week (312 days per year) and anticipates that its variable cost per car per day will be \$2.25. After using the cars for 1 year, ARC expects to sell them and recapture 45 percent of the original cost. Ignoring the time value of money and any noncash expenses, determine the optimal number of cars for ARC to buy.

- 17-6** For several years, the Madison Rhodes Department Store had featured personalized pencils as a Christmas special. Madison Rhodes purchased the pencils from its supplier, who provided the embossing machine. The personalizing was done on the department store premises. Despite the success of the pencil sales, Madison Rhodes had received comments that the quality of the lead in the pencils was poor, and the store had found a different supplier. The new supplier, however, would be unable to begin servicing the department store until after the first of January. Madison Rhodes was forced to purchase its pencils one final time from its original supplier to meet Christmas demand. It was important, therefore, that pencils not be overstocked, and yet the manager was adamant about not losing too many customers because of stockouts. The pencils came packed 15 to the box, 72 boxes to the case. Madison Rhodes

paid \$60 per case and sold the pencils for \$1.50 per box. Labor costs are 37.5¢ per box sold. Based on previous years' sales, management constructed the following schedule:

Expected sales (cases)	15	16	17	18	19	20
Probability	0.05	0.20	0.30	0.25	0.10	0.10

- (a) How many cases should Madison Rhodes order?
 (b) What's the expected profit?

17-7

Emily Scott, head of a small business consulting firm, must decide how many M.B.A.s to hire as full-time consultants for the next year. (Emily has decided that she will not bother with any part-time employees.) Emily knows from experience that the probability distribution on the number of consulting jobs her firm will get each year is as follows:

Consulting jobs	24	27	30	33
Probability	0.3	0.2	0.4	0.1

Emily also knows that each M.B.A. hired will be able to handle exactly three consulting jobs per year. The salary of each M.B.A. is \$60,000. Each consulting job is worth \$30,000 to Emily's firm. Each consulting job that the firm is awarded but cannot complete costs the firm \$10,000 in future business lost.

- (a) How many M.B.A.s should Emily hire?
 (b) What is the expected value of perfect information to Emily?

17-8

As a fund-raiser for a student organization, some students have decided to sell individual pizzas outside the Union on Friday. Each pizza will sell for \$1.75 and costs the organization \$.77. Historical sales indicated that between 55 and 60 dozen pizzas will be sold with the probability distribution given below:

Dozens of pizzas	55	56	57	58	59	60
Probability	0.15	0.20	0.10	0.35	0.15	0.05

To maximize the profit contribution, how many pizzas should be ordered? Assume pizzas must be ordered by the dozen. What is the expected value of perfect information in this problem? What is the maximum amount the organization would be willing to pay for perfect information?

17-9

Manfred Baum, merchandise manager for the Grant Shoe Company, is planning production decisions for the coming year's summer line of shoes. His chief concern is estimating the sales of a new design of fashion sandals. These sandals have posed problems in the past for two reasons: (1) the limited selling season does not provide enough time for the company to produce a second run of a popular item, and (2) the styles change dramatically from year to year, and unsold sandals become worthless. Manfred has discussed the newest design with salespeople and has formulated the following estimates of how the item will sell:

Pairs (thousands)	45	50	55	60	65
Probability	0.25	0.30	0.20	0.15	0.10

Information from the production department reveals that the sandal will cost \$15.25 per pair to manufacture, and marketing has informed Manfred that the wholesale price will be \$31.35 a pair. Using the expected-value decision criterion, calculate the number of pairs that Manfred should recommend that the company produce.

Worked-Out Answer to Self-Check Exercise

SC 17-1 The payoff table below gives both conditional and expected profits.

Machines needed	22	23	24	25	26	27	Expected Profit
Probability	0.12	0.16	0.22	0.27	0.18	0.05	
22	157.30	157.30	157.30	157.30	157.30	157.30	157.30
23	156.45	164.45	164.45	164.45	164.45	164.45	163.49
24	155.60	163.60	171.60	171.60	171.60	171.60	168.40
supplied	25	154.75	162.75	170.75	178.75	178.75	171.55
	26	153.90	161.90	169.90	177.90	185.90	172.54 ←
	27	153.05	161.05	169.05	177.05	185.05	193.25
							172.09

She should have 26 machines.

$$\text{EVPI} = 157.30(0.12) + 164.45(0.16) + 171.60(0.22) + 178.75(0.27) \\ + 185.90(0.18) + 193.25(0.05) - 172.54 = \$1.787$$

Because she cannot every hour adjust the number of machines she will have available, an hour-by-hour forecast of demand is of little value to her in this situation.

17.3 USING CONTINUOUS DISTRIBUTIONS: MARGINAL ANALYSIS

In many inventory problems, the number of computations required makes the use of conditional-profit and expected-profit tables difficult. Our previous illustration contained only four possible stock actions and four possible sales levels, resulting in a conditional-profit table containing 16 possibilities for conditional profits. If we had 300 possible values for sales volume and an equal number of calculations for determining conditional and expected profit, we would have to do a great many computations. The marginal approach avoids this problem.

Limitations of the tabular approach

Marginal analysis is based on the fact that when an additional unit of an item is bought, two fates are possible: the unit will be sold or it will not be sold. The sum of the probabilities of these two events must be 1. (For example, if the probability of selling the additional unit is 0.6, then the probability of not selling it must be 0.4.)

If we let p represent the probability of selling one additional unit, then $1 - p$ must be the probability of not selling it. If the additional unit is sold, we shall realize an increase in our conditional profits as a result of the profit from the additional unit. We refer to this as *marginal profit*, or *MP*. In our previous illustration about the retailer, the marginal profit resulting from the sale of an additional unit is \$30, the selling price (\$50) minus the cost (\$20).

Derivation of marginal profit

Table 17-10 illustrates this point. If we stock 10 units each day and daily demand is for 10 or more units, our conditional profit is \$300 per day. Now we decide to stock 11 units each day. If the eleventh unit is sold (and this is the case when demand is for 11, 12, or 13 units), our conditional profit is increased to \$330 per day. Notice that the increase in conditional profit does not follow merely from stocking the eleventh unit. Under the conditions assumed in the problem, this increase in profit will result only when demand is for 11 or more units. This will be the case 85 percent of the time.

We must also consider how profits would be affected by stocking an additional unit and not selling it. This reduces our conditional

Marginal loss

TABLE 17-10 CONDITIONAL PROFIT TABLE

Possible Demand (sales) in Cases	Probability of Market Size	Possible Stock Actions			
		10 Cases	11 Cases	12 Cases	13 Cases
10	0.15	\$300	\$280	\$260	\$240
11	0.20	300	330	310	290
12	0.40	300	330	360	340
13	0.25	300	330	360	390

profit. The amount of the reduction is referred to as the *marginal loss (ML)* resulting from the stocking of an item that is not sold. In our previous example, the marginal loss was \$20 per unit, the cost of the item.

Table 17-10 also illustrates marginal loss. Once more we decide to stock 11 units. If the eleventh unit (the marginal unit) is not sold, the conditional profit is \$280. The \$300 conditional profit when 10 units were stocked and 10 were sold is reduced by \$20, the cost of the unsold unit.

Additional units should be stocked as long as the expected marginal profit from stocking each of them is greater than the expected marginal loss from stocking each. **Derivation of stocking rule** **The size of each day's order should be increased up to the point where the expected marginal profit from stocking one more unit if it sells is just equal to the expected marginal loss from stocking that unit if it remains unsold.**

In our illustration, the probability distribution of demand is

Market Size	Probability of Market Size
10	0.15
11	0.20
12	0.40
13	0.25
	1.00

This distribution tells us that as we increase our stock, the probability of selling one additional unit (this is p) decreases. If we increase our stock from 10 to 11 units, the probability of selling all eleven is 0.85. This is the probability that demand will be for 11 units or more. Here is the computation:

Probability that demand will be for 11	0.20
Probability that demand will be for 12	0.40
Probability that demand will be for 13	0.25
Probability that demand will be for 11 or more units	0.85

If we add a twelfth unit, the probability of selling all 12 units is reduced to 0.65 (the sum of the probabilities of demand for 12 or 13 units). Finally, the addition of a thirteenth unit carries with it only a 0.25 probability of our selling all 13 units, because demand will be for 13 units only 25 percent of the time.

Deriving the Minimum Probability Equation

The *expected marginal profit* from stocking and selling an additional unit is the marginal profit of the unit multiplied by the probability that the unit will be sold; this is $p(MP)$. The *expected marginal loss* from stocking and not selling an additional unit is the marginal loss incurred if the unit is unsold multiplied by the probability that the unit will not be sold; this is $(1 - p)(ML)$. We can generalize that the retailer in this situation would stock up to the point at which

$$p(MP) = (1 - p)(ML) \quad [17-1]$$

This equation describes the point at which the expected marginal profit from stocking and selling an additional unit, $p(MP)$, is equal to the expected marginal loss from stocking and not selling the unit, $(1 - p)(ML)$. As long as $p(MP)$ is larger than $(1 - p)(ML)$, additional units should be stocked, because the expected profit from such a decision is greater than the expected loss.

In any given inventory problem, there will be only *one* value of p for which the maximizing equation will be true. We must determine that value in order to know the optimal stock action to take. We can do this by taking our maximizing equation and solving it for p in the following manner:

$$p(MP) = (1 - p)(ML) \quad [17-1]$$

Multiplying the two terms on the right side of the equation, we get

$$p(MP) = ML - p(ML)$$

Collecting terms containing p , we have

$$p(MP) + p(ML) = ML$$

or

$$p(MP + ML) = ML$$

Dividing both sides of the equation by $MP + ML$ gives

Minimum Probability Required to Stock Another Unit

$$p^* = \frac{ML}{MP + ML}$$

[17-2]

Minimum-probability equation

The symbol p^* represents the minimum required probability of selling at least one additional unit to justify the stocking of that additional unit. The retailer should stock additional units as long as the probability of selling at least an additional unit is greater than p^* .

We can now compute p^* for our illustration. The marginal profit per unit is \$30 (the selling price minus the cost); the marginal loss per unit is \$20 (the cost of each unit); thus

$$p^* = \frac{ML}{MP + ML} = \frac{\$20}{\$30 + \$20} = \frac{\$20}{\$50} = 0.40 \quad [17-2]$$

This value of 0.40 for p^* means that in order to make the stocking of an additional unit justifiable, we must have at least a 0.40 *cumulative* probability of selling that unit or more. In order to determine the

TABLE 17-11 CUMULATIVE PROBABILITIES OF SALES

Sales Units	Probability of This Sales Level	Cumulative Probability That Sales Will Be at This Level or Greater
10	0.15	1.00
11	0.20	0.85
12	0.40	0.65
13	0.25	0.25

probability of selling each additional unit we consider stocking, we must compute a series of cumulative probabilities, as we have done in Table 17-11.

The cumulative probabilities in the right-hand column of Table 17-11 represent the probabilities that sales will reach or exceed each of the four sales levels. For example, the 1.00 that appears beside the 10-unit sales level means that we are 100 percent of selling 10 or more units. This must be true because our assumes that one of the four sales levels will *always* occur.

Calculation of cumulative probabilities

The 0.85 probability value beside the 11-unit sales figure means that we are only 85 percent sure of selling 11 or more units. This can be calculated in two ways. First, we could add the chances of selling 11, 12, or 13 units:

$$\begin{array}{r}
 11 \text{ Units} & 0.20 \\
 12 \text{ Units} & 0.40 \\
 13 \text{ Units} & + 0.25 \\
 \hline
 \textbf{0.85} = \text{probability of selling 11 or more}
 \end{array}$$

Or we could reason that sales of 11 or more units include all possible outcomes except sales of 10 units, which has a probability of 0.15.

$$\begin{array}{r}
 \text{All possible outcomes} & 1.00 \\
 \text{Probability of selling 10} & -0.15 \\
 \hline
 \textbf{0.85} = \text{Probability of selling 11 or more}
 \end{array}$$

The cumulative probability value of 0.65 assigned to sales of 12 units or more can be established in similar fashion. Sales of 12 or more must mean sales of 12 or 13 units.

$$\begin{array}{r}
 \text{Probability of selling 12} & 0.40 \\
 \text{Probability of selling 13} & + 0.25 \\
 \hline
 \textbf{0.65} = \text{probability of selling 12 or more}
 \end{array}$$

And, of course, the cumulative probability of selling 13 units is still 0.25, because we have assumed that sales will never exceed 13.

As we mentioned previously, the value of p decreases as the level of stock increases. This causes the expected marginal profit to decrease and the expected marginal loss to increase until, at some point, stocking an additional unit would not be profitable.

We have said that additional units should be stocked as long as the probability of selling at least an additional unit is greater than p^* . We can now apply this rule to our probability distribution of sales and determine how many units should be stocked.

Stocking rule

In this case, the probability of selling 11 or more units is 0.85, a figure clearly greater than our p^* of 0.40; thus, we should stock an eleventh unit. The expected marginal profit from stocking this unit is greater than the expected marginal loss from stocking it. We can verify this as follows:

$$\begin{aligned} p(MP) &= 0.85(\$30) = \$25.50 \text{ expected marginal profit} \\ (1-p)(ML) &= 0.15(\$20) = \$3.00 \text{ expected marginal loss} \end{aligned}$$

A twelfth unit should be stocked because the probability of selling 12 or more units (0.65) is greater than the required p^* of 0.40. Such action will result in the following expected marginal profit and expected marginal loss:

$$\begin{aligned} p(MP) &= 0.65(\$30) = \$19.50 \text{ expected marginal profit} \\ (1-p)(ML) &= 0.35(\$20) = \$7.00 \text{ expected marginal loss} \end{aligned}$$

Twelve is the *optimal* number of units to stock, because the addition of a thirteenth unit carried with it only a 0.25 probability that it will be sold, and that is less than our required p^* of 0.40. The following figures reveal why the thirteenth unit should not be stocked:

$$\begin{aligned} p(MP) &= 0.25(\$30) = \$7.50 \text{ expected marginal profit} \\ (1-p)(ML) &= 0.75(\$20) = \$15.00 \text{ expected marginal loss} \end{aligned}$$

Optimal stocking level for this problem

If we stock a thirteenth unit, we add more to expected loss than we add to expected profit.

Notice that the use of marginal analysis leads us to the same conclusion that we reached with the use of conditional profit and expected profit tables. Both methods of analysis suggest that the retailer should stock 12 units each period.

Our strategy, to stock 12 cases every day, assumes that daily sales is a random variable. In actual practice, however, daily sales often take on recognizable patterns depending on the particular day of the week. In retail sales, Saturday is generally recognized as being a higher-volume day than, say, Tuesday. Similarly, Monday retail sales are typically less than those on Friday. In situations with recognizable patterns in daily sales, we can apply the techniques we have learned by computing an optimal stocking level for *each* day of the week. For Saturday, we would use as our input data past sales experience for Saturdays only. Each of the other 6 days could be treated in the same fashion. Essentially, this approach represents nothing more than recognition of, and reaction to, discernible patterns in what may at first appear to be a completely random environment.

Adjusting the optimal stocking level

Using the Standard Normal Probability Distribution

We first learned the concept of the standard normal probability distribution in Chapter 5. We can now use this idea to help us solve a decision-theory problem using a continuous distribution.

Assume that a manager sells an article having normally distributed sales with a mean of 50 units daily and a standard deviation in daily sales of 15 units. The manager purchases this article for \$4 per unit and sells it for \$9 per unit. If the article is not sold on the selling day, it is worth nothing. Using the marginal method of calculating optimal inventory purchase levels, we can calculate our required p^* :

$$\begin{aligned} p^* &= \frac{ML}{MP + ML} \quad [17-2] \\ &= \frac{\$4}{\$5 + \$4} = 0.44 \end{aligned}$$

Solving a problem using marginal analysis

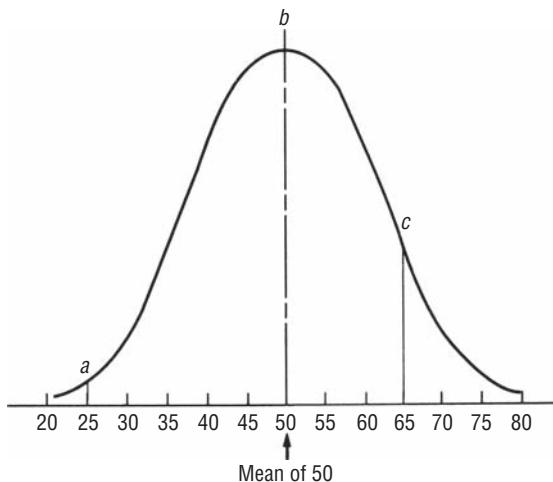


FIGURE 17-1 NORMAL DISTRIBUTION OF PAST DAILY SALES

Suppose the manager considers stocking 25 units, line *a*. Most of the entire area under the curve lies to the right of the vertical line drawn at 25; thus, the probability is great that the manager will sell 25 units or more. If he considers stocking 50 units (the mean), one-half the entire area under the curve lies to the right of vertical line *b*; thus, he is 0.5 sure of selling the 50 units or more. Now, say he considers stocking 65 units. Only a small portion of the entire area under the curve lies to the right of line *c*; thus, the probability of selling 65 or more units is quite small.

Figure 17-2 illustrates the 0.44 probability that must exist before it pays our manager to stock another unit. He will stock additional units until he reaches point *Q*. If he stocks a larger quantity, the shaded area under the curve drops below 0.44 and the probability of selling another unit or more falls below the required 0.44. How can we locate point *Q*? As we saw in Chapter 5, we can use Appendix Table 1 to determine how many standard deviations it takes to include any portion of the area under the curve measuring from the mean to any point such as *Q*. In this particular case, because we know that the shaded area must be 0.44 of the total area, the area from the mean to point *Q* must be 0.06 (the total area from the mean to the right tail is 0.50). Looking in the body of the table, we find that 0.06 of the area under the curve is located between the mean and a point 0.15 standard deviation to the right of the mean. Thus we know that point *Q* is 0.15 standard deviation to the right of the mean (50).

We have been given the information that 1 standard deviation for this distribution is 15 units; so 0.15 times this would be 2.25 units. Because point *Q* is 2.25 units to the right of the mean (50), it must be at about 52 units. This is the optimal order for the manager to place: 52 units per day.

This means that the manager must be 0.44 sure of selling at least an additional unit before it would pay to stock that unit. Let us reproduce the curve of past sales and determine how to incorporate the marginal method with continuous distributions of past daily sales.

Now refer to Figure 17-1. If we erect a vertical line *b* at 50 units, the area under the curve to the right of this line is one-half the total area. This tells us that the probability of selling 50 or more units is 0.5. *The area to the right of any such vertical line represents the probability of selling that quantity or more.* As the area to the right of any vertical line decreases, so does the probability that we will sell that quantity or more.

Using the standard normal probability distribution in marginal analysis

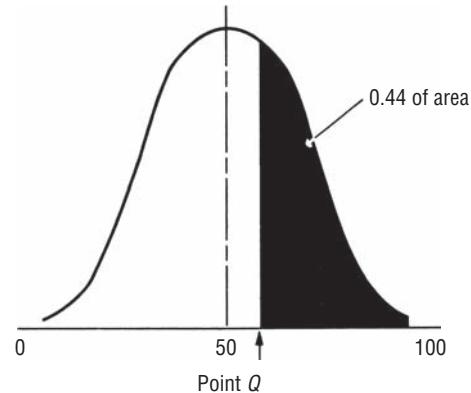


FIGURE 17-2 NORMAL PROBABILITY DISTRIBUTION, WITH 0.44 OF THE AREA UNDER THE CURVE SHADeD

Optimal solution for this problem

Now that we have been through one problem using a continuous probability distribution, we can work our chapter-opening problem involving these data for a normally distributed daily sales record:

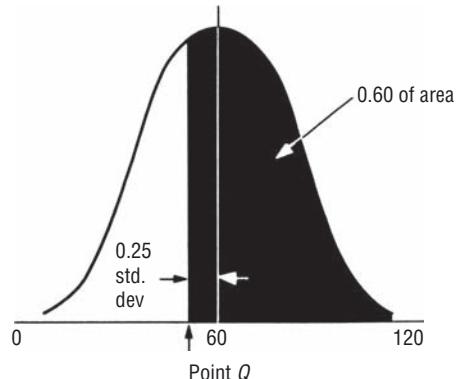
Chapter-opening problem

Mean of past daily sales	60 boxes
Standard deviation of past daily sales distribution	10 boxes
Cost per box	\$20
Selling price per box	\$32
Value if not sold on first day	\$ 2

As we did in the previous problem, we first calculate the p^* that is required to justify the stocking of an additional box. In this instance,

Minimum required probability

$$\begin{aligned}
 p^* &= \frac{ML}{MP + ML} & [17-2] \\
 &= \frac{\$20 - \$2}{\$12 + (\$20 - \$2)} & \text{Notice that a salvage value of } \$2 \\
 && \text{is deducted from the cost of } \$20 \\
 && \text{to obtain the } ML \\
 &= \frac{\$18}{\$12 + \$18} \\
 &= \frac{\$18}{\$30} = 0.60
 \end{aligned}$$



We can now illustrate the probability on a normal curve by marking off 0.60 of the area under the curve, starting from the right-hand end of the curve, as in Figure 17-3.

The manager wants to increase his order size until it reaches point Q . Now point Q lies to the *left* of the mean, whereas in the preceding problem it lay to the *right*. How can we locate point Q ? Because 0.50 of the area under the curve is located between the mean and the right-hand tail, 0.10 of the shaded area must be to the left of the mean, $(0.60 - 0.50 = 0.10)$. In the body of Appendix Table 1, the nearest value to 0.10 is 0.0987, so we want to find a point Q with 0.0987 of the area under the curve contained between the mean and point Q . The table indicates point Q is 0.25 standard deviation from the mean. We now solve for point Q as follows:

$$0.25 \times \text{standard deviation} = 0.25 \times 10 \text{ boxes} = 2.5 \text{ boxes}$$

$$\begin{aligned}
 \text{Point } Q &= \text{mean less } 2.5 \text{ boxes} \\
 &= 60 - 2.5 \text{ boxes} = 57.5 \text{ boxes}
 \end{aligned}$$

Optimal solution for chapter-opening problem

HINTS & ASSUMPTIONS

Warning: Use of the maximum *expected profit* calculated from a single sales distribution as your decision rule assumes that the sales distribution you are dealing with represents *all* of the information you have about demand. If you have information that sales on Saturday, for example, are better represented by a different distribution, then you must treat Saturday as a separate decision and calculate a stocking level for Saturday that will probably be different from that for the other 6 days. Hint: This is how good managers make decisions anyhow. Instead of accepting that every day of the week has identical market characteristics, it's long been known that strong, discernable daily differences exist. These daily differences are themselves quite different in certain countries. Hint: Whereas Saturday is a prime shopping day in the United States, Saturday sales would be near zero in Israel because it is their sabbath.

EXERCISES 17.3

Self-Check Exercise

SC 17-2 Floyd Guild operates a newsstand near the 53rd Street station of the IC South Shore and Suburban line. The *City Herald* is the most popular of the newspapers that Floyd stocks. Over many years, he has observed that daily demand for the *Herald* is well described by a normal distribution with mean $\mu = 165$ and standard deviation $\sigma = 40$. Copies of the *Herald* sell for 30¢, but the publisher charges Floyd only 20¢ for each copy he orders. If any *Heralds* are left over at the end of the evening commuting hours, Floyd sells them to Jesselman's Fish Market down the street for a dime each. If Floyd wishes to maximize his expected daily profit, how many copies of the *Herald* should he order?

Applications

- 17-10** Highway construction in North Dakota is concentrated in the months from May through September. To provide some protection to the crews at work on the highways, the Department of Transportation (DOT) requires that large, orange MEN WORKING signs be placed in advance of any construction. Because of vandalism, wear and tear, and theft, the DOT purchases new signs each year. Although the signs are made under the auspices of the Department of Correction, the DOT is charged a price equivalent to one it would pay were it to buy the signs from an outside source. The interdepartmental charge for the signs is \$21 if more than 35 of the same kind are ordered. Otherwise, the cost per sign is \$29. Because of budget pressures, the DOT attempts to minimize its costs both by not buying too many signs and by attempting to buy in sufficiently large quantity to get the \$21 price. In recent years, the department has averaged purchases of 78 signs per year, with a standard deviation of 15. Determine the number of signs the DOT should purchase.
- 17-11** The town of Green Lake, Wisconsin, is preparing for the celebration of the seventy-ninth Annual Milk and Dairy Day. As a fund-raising device, the city council once again plans to sell souvenir T-shirts. The T-shirts, printed in six colors, will have a picture of a cow and the words "79th Annual Milk and Dairy Day" on the front. The city council purchases heat-transfer patches from a supplier for \$0.75 and plain white cotton T-shirts for \$1.50. A local merchant supplies the appropriate heating device and also purchases all unsold white cotton

T-shirts. The council plans to set up a booth on Main Street and sell the shirts for \$3.25. The transfer of the color to the shirt will be completed when the sale is made. In the past year, similar shirt sales have averaged 200 with a standard deviation of 34. The council knows that there will be no market for the patches after the celebration. How many patches should the city council buy?

- 17-12** Jack buys hot dogs each morning for his stand in the city. Jack prides himself on slow-roasted, always-fresh hot dogs. As a result, he will sell only hot dogs purchased that morning. Each hot dog plus bun and condiments sells for \$1.50 and costs Jack \$0.67. Assume Jack can purchase any number of hot dogs. Because tomorrow is Friday, Jack knows tomorrow's hot dog demand will be normally distributed with mean 375 hot dogs and variance 400. If Jack has any hot dogs left over, he either eats them or gives them away to the less fortunate, earning no additional revenue. If Jack wants to maximize his profits, how many hot dogs should he purchase? How many hot dogs should he buy if leftover hot dogs could always be sold for \$0.50?
- 17-13** Bike Wholesale Parts was established in the early 1980s in response to demands of several small and newly established bicycle shops that needed access to a wide variety of inventory but were not able to finance it themselves. The company carries a wide variety of replacement parts and accessories but does not maintain any stock of completed bicycles. Management is preparing to order $27'' \times 1\frac{1}{4}''$ rims from the Flexspin Company in anticipation of a business upturn expected in about 2 months. Flexspin makes a superior product, but the lead time required necessitates that wholesalers make only one order, which must last through the critical summer months. In the past, Bike Wholesale Parts has sold an average of 120 rims per summer with a standard deviation of 28. The company expects that its stock of rims will be depleted by the time the new order arrives. Bike Wholesale Parts has been quite successful and plans to move its operations to a larger plant during the winter. Management feels that the combined cost of moving some items such as rims and the existing cost of financing them is at least equal to the firm's purchase cost of \$7.30. Accepting management's hypothesis that any unsold rims at the end of the summer season are permanently unsold, determine the number of rims the company should order if the selling price is \$8.10.
- 17-14** The B&G Cafeteria features barbecued chicken each Thursday, and Priscilla Alden, the cafeteria manager, wants to ensure that the cafeteria will make money on this dish. Including labor and other costs of preparation, each portion of chicken costs \$1.35. The \$2.15 selling price per portion is such a bargain that the barbecued chicken special has become a very popular item. Data taken from the last year indicate that demand for the special is normally distributed with mean $\mu = 190$ portions and standard deviation $\sigma = 32$ portions. If B&G Cafeteria prepares two portions of barbecued chicken from each whole chicken it cooks, how many chickens should Priscilla order each Thursday?
- 17-15** Paige's Tire Service stocks two types of radial tires: polyester-belted and steel-belted. The polyester-belted radials cost the company \$30 each and sell for \$35. The steel-belted radials cost the company \$45 and sell for \$60. For various reasons, Paige's Tire Service will not be able to reorder any radials from the factory this year so it must order just once to satisfy customers' demand for the entire year. At the end of the year, owing to new tire models, Paige will have to sell all its inventory of radials for scrap rubber at \$5 each. The annual sales of both types of radial tires are normally distributed with the following means and standard deviations:

Radial Tire Type	Annual Mean Sales	Standard Deviation
Polyester-belted	300	50
Steel-belted	200	20

- (a) How many polyester-belted radials should be ordered?
 (b) How many steel-belted radials should be ordered?

Worked-Out Answer to Self-Check Exercise

$$\text{SC 17-2 } MP = 50 - 20 = 30 \quad ML = 20 - 10 = 10$$

$$p^* = \frac{ML}{MP + ML} = \frac{10}{40} = 0.25, \text{ which corresponds to } 0.67\sigma, \text{ so}$$

he should order $\mu + 0.67\sigma = 165 + 0.67(40) = 191.8$, or 192 copies.

17.4 UTILITY AS A DECISION CRITERION

So far in this chapter, we have used expected value (expected profit, for example) as our decision criterion. We assumed that if the expected profit of alternative A was better than that of alternative B, then the decision maker would certainly choose alternative A. Conversely, if the expected loss of alternative C was greater than the expected loss of alternative D, then we assumed that the decision maker would surely choose D as the better course of action.

Different decision criteria

Shortcomings of Expected Value as a Decision Criterion

There are situations, however, in which the use of expected value as the decision criterion would get a manager into serious trouble. Suppose an entrepreneur owns a new factory worth \$2 million. Suppose further that there is only one chance in 1,000, (0.001) that it will burn down this year. From these two figures, we can compute the expected loss:

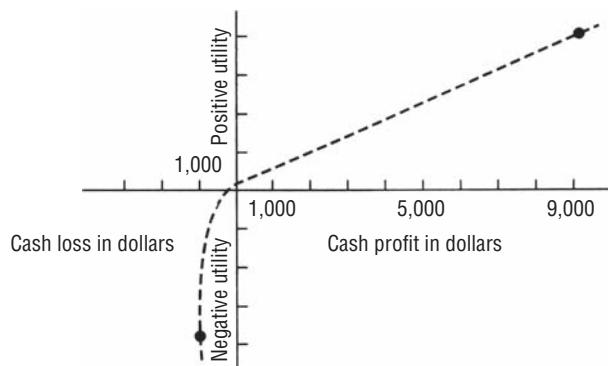
Expected value is sometimes inappropriate

$$0.001 \times \$2,000,000 = \$2,000 = \text{expected loss by fire}$$

An insurance representative offers to insure the building for \$2,250 this year. If the entrepreneur applies the notion of minimizing expected losses, he will refuse to insure the building. The expected loss of insuring (\$2,250) is higher than the expected loss by fire. However, if the businessman feels that a \$2 million uninsured loss would wipe him out, he will probably discard expected Value as his decision criterion and buy the insurance at the extra cost of \$250 per year per policy (\$2,250 – \$2,000). He would choose *not* to minimize expected loss in this case.

Take an example closer, perhaps, to student life. You are a student with just enough money to get through the semester. A friend offers to sell you a 0.9 chance of winning \$10 for just \$1. You would most likely think of the problem in terms of expected values and reason as follows: “Is 0.9 – \$10 greater than \$1?” Because \$9 (the expected value of the bet) is nine times greater than the cost of the bet (\$1), you might feel inclined to take your friend up on this offer. Even if you lose, the loss of \$1 will not affect your situation materially.

A personal example

**FIGURE 17-4 UTILITY OF VARIOUS PROFITS AND LOSSES**

Now your friend offers to sell you a 0.9 chance of winning \$1,000 for \$100. The question you would now ponder is, “Is $0.9 \times \$1,000$ greater than \$100?” Of course, \$900 (the expected value of the bet) is still nine times the cost of the bet (\$100), but you would more than likely think twice before putting up your money. Why? Because even though the pleasure of winning \$1,000 would be high, the pain of losing your hard-earned \$100 might be more than you care to experience.

Say, finally, that your friend offers to sell you a 0.9 chance at winning \$10,000 for your total assets, which happen to be \$1,000. If you use expected value as your decision criterion, you would ask the question, “Is $0.9 \times \$10,000$ greater than \$1,000?” You would get the same answer as before: yes. The expected value of the bet (\$9,000) is still nine times greater than the cost of the bet (\$1,000), but now you would probably refuse your friend, not because the expected value of the bet is unattractive, but because the thought of losing all your assets is completely unacceptable as an outcome.

In this example, you changed the decision criterion away from **Function of utility** expected value when the thought of losing \$1,000 was too painful, despite the pleasure to be gained from \$10,000. At this point, you no longer considered the expected value; you thought solely of *utility*. In this sense, utility is the pleasure or displeasure one would derive from certain outcomes. Your utility curve in Figure 17-4 is linear around the origin (\$1 of gain is as pleasurable as \$1 of loss is painful in this region), but it turns down rapidly when the potential loss rises to levels near \$1,000. Specifically, this utility curve shows us that from your point of view, the displeasure from losing \$1,000 is about equal to the pleasure from winning nine times that amount. The shape of one’s utility curve is a product of one’s psychological makeup, one’s expectations about the future, and the particular decision or act being evaluated. A person can have one utility curve for one situation and quite a different one for the next situation.

Different Utilities

The utility curves of three different managers’ decisions are shown on the graph in Figure 17-5. We have arbitrarily named these managers David, Ann, and Jim. Their attitudes are readily apparent from analysis of their utility curves. David is a cautious and conservative businessman. A move to the right of the zero-profit point increases his utility only very slightly, whereas a move to the left of the zero-profit point decreases his utility rapidly. In terms of numerical values, David’s utility curve indicates that going from \$0 to \$100,000 profit increases his utility by a value of 1 on the vertical scale, while moving into the loss range by only \$40,000 decreases his utility by the same value of 1 on the vertical scale. David will avoid situations in which high losses might occur; he is said to be averse to risk. **Attitudes toward risk**

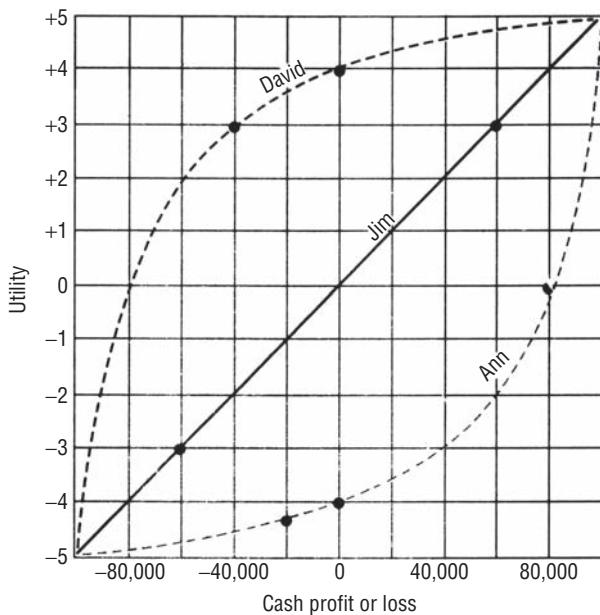


FIGURE 17-5 THREE UTILITY CURVES

Ann is quite another story. We see from her utility curve that a profit increases her utility by much more than a loss of the same amount decreases it. Specifically, increasing her profits \$20,000 (from \$80,000 to \$100,000) raises her utility from 0 to +5 on the vertical scale, but lowering her profits \$20,000 (from \$0 to -\$20,000) decreases her utility by only 0.33, from 0 to -4.33. Ann is a player of long shots; she feels strongly that a large loss would not make things much worse than they are now, but that a big profit would be quite rewarding. She will take large risks to earn even larger gains.

Jim, fairly well-off financially, is the kind of businessman who would not suffer greatly from a \$60,000 loss nor increase his wealth significantly with a \$60,000 gain. Pleasure from making an additional \$60,000 or pain from losing it would be of about equal intensity. **Because his utility curve is linear, he can effectively use expected value as his decision criterion, whereas David and Ann must use utility. Jim will act when the expected value is positive, David will demand a high expected value for the outcome, and Ann may act when the expected value is negative.**

Who would use expected value?

HINTS & ASSUMPTIONS

An important prerequisite to understanding the behavior of investors is realizing that their utility curves are not all the same. Specifically, some “high rollers” are attracted to high-risk investments that can result in losing the entire investment or making a fortune. Presumably, such people with significant net worths can afford the loss. On the other hand, people with moderate net worths and heavy family obligations tend to be risk averse and invest only when the expected outcome is positive. An interesting question for you to discuss with your classmates is why the elderly are victims of “get rich quick” investment schemes far out of proportion to their number in the population.

EXERCISES 17.4**Applications**

- 17-16** Bill Johnson's income places him in the 50 percent bracket for federal income tax purposes. Johnson often supplies venture capital to small startup firms in return for some type of equity position in the firm. Recently, Bill has been approached by Circutronics, a small firm entering the microcircuitry industry. Circutronics has requested \$1.6 million backing. Because of his tax position, Bill invests in tax-exempt municipal securities when he cannot find any attractive ventures to back. Currently, he has a large position in North Carolina Eastern Municipal Power Agency bonds, which are yielding a return of 9.43 percent. Bill considers this 9.43 percent after-tax return to be his utility breakeven point. Above that point, his utility rises very rapidly; below, it drops slightly because he can well afford to lose the money.
- What dollar return must Circutronics promise before Bill will consider financing it?
 - Graph Bill's utility curve.
- 17-17** The Enduro Manufacturing Company is a partnership producing structural-steel building components. Financial manager and partner William Flaherty is examining potential projects that the firm might undertake in the coming fiscal year. The company has a target rate of return of 10 percent on its investment, but because there is no outside financing and interference, the partners have accepted projects with rates of return between 0 and 100 percent. Above 10 percent, the partners' utility rises very rapidly; between 0 and 10 percent, it rises only slightly above 0; below 0, it falls very rapidly; Flaherty is considering several projects that will cause Enduro to invest \$250,000. Plot the firm's utility curve.
- 17-18** An investor is convinced that the price of a share of PDQ stock will rise in the near future. PDQ stock is currently selling for \$57 a share. Upon inspecting the latest quotes on the options market, the investor finds that she can purchase an option at a cost of \$5 per share, allowing her to buy PDQ for \$55 per share within the next 2 months. She can also purchase an option to buy the stock within a 4-month period; this option, which costs \$10 per share, also has an exercise price of \$55 per share. She has estimated the following probability distributions for the stock price on the days the options expire:

Price	50	55	60	65	70	75
Probability at 2 months	0.05	0.15	0.15	0.25	0.35	0.05
Probability at 4 months	0	0.05	0.05	0.20	0.30	0.40

The investor plans to exercise her option just before its expiration if PDQ stock is selling for more than \$55 and immediately sell the stock at that market price. Of course, if the stock is selling for \$55 or less when the option expires, she will lose the entire purchase cost of the option. The investor is relatively conservative, with the following utility values for changes in her dollar assets:

Change	+1,500	+1,000	+500	0	-500	-1,000
Utility	1.0	0.9	0.8	0.7	0.1	0.0

She is considering one of three alternatives:

- To buy a 2-month option on 100 shares.
- To buy a 4-month option on 100 shares.
- Not to buy at all.

Which of these alternatives will maximize her expected utility?

17.5 HELPING DECISION MAKERS SUPPLY THE RIGHT PROBABILITIES

The two problems we worked using the normal probability distribution (pp. 926–928) required us to know both the mean (μ) and the standard deviation (σ). But how can we make use of a probability distribution when past data are missing or incomplete? By working through a problem, we shall see how we can often generate the required values by using an *intuitive* approach.

Missing information

An Intuitive Approach to Estimating the Mean and Standard Deviation

Assume that you are thinking about purchasing a machine to replace hand labor on an operation. The machine will cost \$10,000 per year to operate and will save \$8 for each hour it operates. To break even, then, it must operate at least $\$10,000/\$8 = 1,250$ hours annually. If you are interested in the probability that it will run more than 1,250 hours, you must know something about the distribution of running times, specifically, the mean and standard deviation of this distribution. But because you do not have a history of the machine's operation, where would you find these figures?

We could ask the foreman of this operation, who has been closely involved with the process, to guess the mean running time of the machine. Let us say that his best estimate is 1,400 hours. But how would he react if you asked him to give you the standard deviation of the distribution? This term may not be meaningful to him, and yet he probably has some intuitive notion of the dispersion of the distribution of running times. Most people understand betting odds, so let us approach him on that basis.

Estimating the mean

We begin by counting off an equal distance on each side of his mean, say, 200 hours. This gives us an interval from 1,200 to 1,600 hours. Then we ask, "What are the odds the number of hours will lie between 1,200 and 1,600 hours?" If he has had any experience with betting, he should be able to reply. Suppose he says, "I think the odds it will run between 1,200 and 1,600 hours are 4 to 3." We show his answer on a probability distribution in Figure 17-6.

Estimating the standard deviation

Figure 17-6 illustrates the foreman's reply that the odds are 4 to 3 the machine will run between 1,200 and 1,600 hours rather than outside those limits. What should we do next? First, we label the 1,600-hour

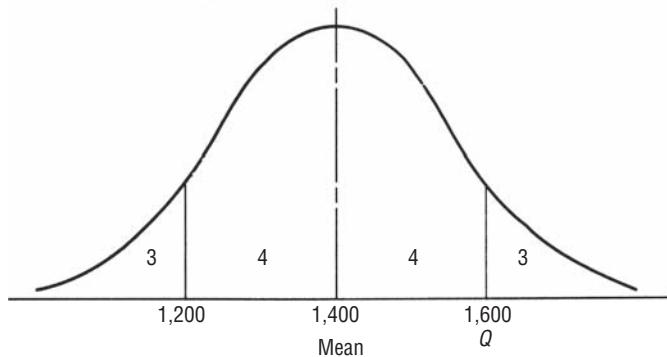
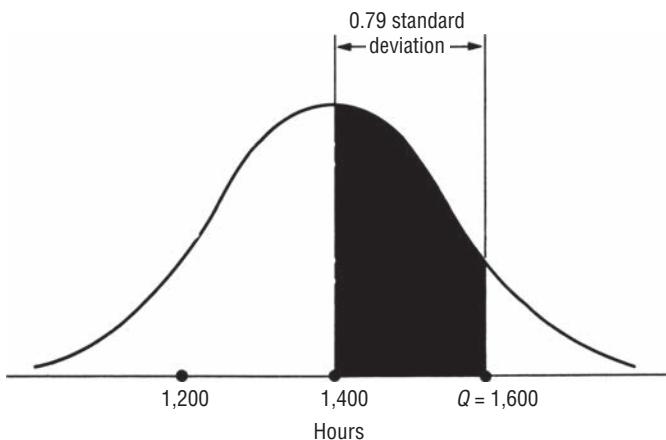


FIGURE 17-6 FOREMAN'S ODDS INTERVALS FOR OPERATING TIMES OF PROPOSED MACHINES

**FIGURE 17-7 DETERMINATION OF STANDARD DEVIATION FROM FOREMAN'S ODDS**

point on the distribution in Figure 17-6 point Q . Then we can see that the area under the curve between the mean and point Q according to the foreman's estimates is $4/7$ of *half* the area under the entire curve, or $4/14 = (0.2857)$ of the *total* area under the curve.

Look at Figure 17-7. If we turn to Appendix Table 1 for the value 0.2857, we find that point Q is 0.79 standard deviation to the right of the mean. Because we know that the distance from the mean to Q is 200 hours, we see that

$$0.79 \text{ standard deviation} = 200 \text{ hours}$$

and thus

$$\begin{aligned} 1 \text{ standard deviation} &= 200/0.79 \\ &= 253 \text{ hours} \end{aligned}$$

Now that we know the mean and standard deviation of the distribution of running times, we can calculate the probability of the machine's running fewer than its break-even point of 1,250 hours:

$$\begin{aligned} \frac{1,250 - 1,400}{253} &= \frac{-150}{253} \\ &= -0.59 \text{ standard deviation} \end{aligned}$$

Calculating the break-even probability

Figure 17-8 illustrates this situation. In Appendix Table 1, we find that the area between the mean of the distribution and a point 0.59 standard deviation below the mean (1,250 hours) is 0.2224 of the total area under the curve. To 0.2224, we add 0.5, the area from the mean to the right-hand tail. This gives us 0.7224. Because 0.7224 is the probability that the machine will operate *more* than 1,250 hours, the chances that it will operate fewer than 1,250 hours (its break-even point) are $1 - 0.7224$, or 0.2776. Apparently, this is not too risky a situation.

This problem illustrates how we can make use of other people's knowledge about a situation without requiring them to understand the intricacies of various statistical techniques.

Securing information for models

Had we expected the foreman to comprehend the theory behind our calculations, or had we even attempted to explain the theory to him, we might never have been able to benefit from his practical

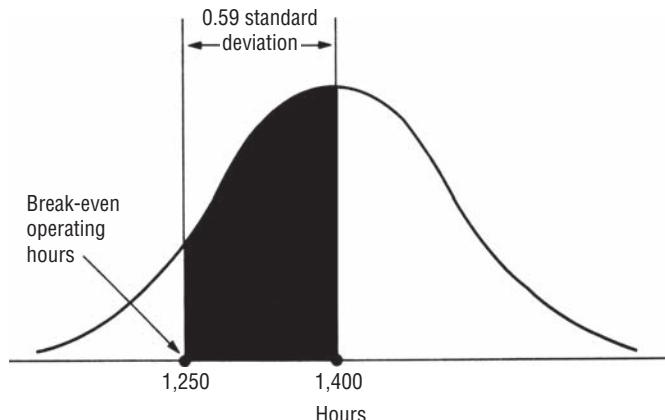


FIGURE 17-8 PROBABILITY THE MACHINE WILL OPERATE BETWEEN 1,250 AND 1,400 HOURS

wisdom concerning the situation. By using language and terms of reference that he understood, we were able to get the foreman to give us workable estimates of the mean and standard deviation of the distribution of operating times for the machine we contemplated purchasing. In this case (and for that matter, in most others, too), it is wiser to accommodate the ideas and knowledge of other people in your models than to search until you find a situation that will fit a model that has already been developed.

HINTS & ASSUMPTIONS

If you used only the methods in this chapter to make decisions, you would not be very likely to wind up successful. And if all you used to make decisions was your intuition, there would be lots of situations where you would miss out on opportunities. But when you combine high intelligence, strong intuition, and sound quantitative models, the chances of winning increase dramatically. Hint: The people with the strongest intuitive ideas about how things work and what is possible and what is more likely to happen are not the “quant jocks,” but ordinary people who have a lot of experience and probably little knowledge of expected value models. The real challenge is to capture the industry wisdom of these veterans and focus it on making sensible decisions when the future is unknown.

EXERCISES 17.5

Self-Check Exercise

- SC 17-3** John Stein is the scheduling director of SATPlus Services, a firm that guarantees that its preparatory course for the college board exams will increase a student's combined score on the verbal and quantitative parts of those exams by at least 120 points. Each student taking the course is charged \$275 in tuition, and it costs SATPlus about \$3,300 in salaries, supplies, and facility rental costs to teach the course. John will not schedule the course in any location where he cannot be at least 90 percent certain that SATPlus will earn no less than \$2,200 in profit. Reviewing a marketing survey that he just received from Charlottesville, Virginia, he has decided that if the course is offered there, he can expect about 30 students to enroll.

He also feels that the odds are about 8 to 5 that actual enrollment will be between 25 and 35 students and that it is appropriate to use the normal distribution to describe course enrollment. Should John schedule the course in Charlottesville?

Applications

- 17-19** Northwestern Industrial Pipe Company is considering the purchase of a new electric arc welder for \$2,100. The welder is expected to save the firm \$5 an hour when it can be used in place of the present, less efficient welder. Before making the decision, Northwestern's production manager noted there were only about 185 hours a year of welding on which the new arc welder could be substituted for the present one. He gave 7 to 3 odds that the actual outcome would be within 25 hours of this estimate. In addition, he felt secure in assuming that the number of hours was well described by a normal distribution. Can Northwestern be 98 percent sure that the new electric arc welder will pay for itself over a 3-year period?
- 17-20** Relman Electric Battery Company has felt the effects of a recovering economy as demand for its products has risen in recent months. The company is considering hiring six new people for its assembly operation. Plant production manager Mike Casey, whose performance is evaluated in part by cost efficiency, does not want to hire additional employees unless they can be expected to have jobs for at least 6 months. If the employees are terminated involuntarily before that time, the company is forced by union rules to pay a substantial termination bonus. Additionally, if employees are laid off within 6 months after hiring, the company's unemployment insurance rate is raised. Relman's corporate economist expects that the upswing in the economy will last at least 8 months and gives 7 to 2 odds that the length of the upswing will be within a 1-month range of that figure. Casey wants to be 95 percent sure that he will not have to lay off any newly hired employees. Should he hire six new people at this time?
- 17-21** Speedy Rabbit courier service operates a fleet of 30 cars covering many miles each day. Currently, the cars use regular fuel at a cost of \$1.059 per gallon, and the fleet fuel efficiency is about 36 miles per gallon (mpg). A recent report indicates that if they switch to premium at a cost of \$1.229 per gallon, each car would see an increase of 6.4 mpg. The company will switch fuel provided they can be 95 percent certain they will save money, which they will do if the fleet fuel efficiency is less than 40 mpg. They believe that the odds are about 6 to 4 that current fuel efficiency is between 33 and 39 mpg and that it is appropriate to use the normal distribution to describe fuel efficiency. Should they switch fuel?
- 17-22** Natalie Larsen, a traveling sales representative for Nova Products, is considering the purchase of a new car for business use. The car she has in mind has a sticker price of \$13,497, but she thinks she can bargain the dealer down to \$12,250. Because her car is used solely for business purposes, Natalie can deduct 31¢ a mile for operating expenses. She will buy the car only if the resulting tax savings will pay for the car over its lifetime. Natalie has been in a combined federal and state 34 percent tax bracket for some years, and it appears she will remain there for the foreseeable future. A reputable automotive magazine states that the average life of the car she is considering is 120,000 miles. The article further states that the odds are 4 to 3 that the actual life of the car will be within 12,000 miles of 120,000. What is the probability that the car will run long enough for Natalie to break even?
- 17-23** The Newton Pines Police Force is considering purchasing a VASCAR radar unit to be installed on the town's single police cruiser. The town council has balked at the idea, because it is not certain that the unit is worth its price of \$2,000. Police Chief Buren Hubbs has stated that he

is sure that the unit will pay for itself through the increased number of \$20 citations that he and his deputy will give. Buren has been overheard to say that he will give 9 to 1 odds that the increase in citations in the first year will be between 95 and 135 if the unit is purchased. He expects that there will be 115 more tickets given if the cruiser is equipped with VASCAR. Can the town council be 99 percent sure that the unit will be paid for by the increase in revenue from citations in the first year?

- 17-24** You are planning to invest \$15,000 in Infometrics common stock if you can be reasonably certain that its price will rise to \$60 a share within 6 months. You ask two knowledgeable brokers the following questions:

- What is your best estimate of the highest price at which Infometrics will sell in the next 6 months?
- What odds will you give that your estimate will be off by no more than \$5?

Their responses are as follow:

Broker	Best Estimate	Odds
A	68	2 to 1
B	65	5 to 1

If you had decided that you would buy the stock only if each broker was at least 80 percent certain that it would be selling for at least \$60 sometime within the next 6 months, what should you do?

Worked-Out Answer to Self-Check Exercise

SC 17-3 $8/26 = 0.3077$, corresponding to 0.87σ , so $\sigma = 5/0.87 = 5.75$ students. To earn a profit of \$2,200, they will have to enroll at least $\frac{3,330 + 2,200}{275} = 20$ students, corresponding to $z = \frac{20 - 30}{5.75} = -1.74$. $P(z > -1.74) = 0.9591$. Because this exceeds the necessary 0.90, he should schedule the course in Charlottesville.

17.6 DECISION-TREE ANALYSIS

A *decision tree* is a graphic model of a decision process. With it, we can introduce probabilities into the analysis of complex decisions involving many alternatives and future conditions that are not known but that can be specified in terms of a set of discrete probabilities or a continuous probability distribution. Decision-tree analysis is a useful tool in making decisions concerning investments, the acquisition or disposal of physical property, project management, personnel, and new-product strategies.

Decision-tree fundamentals

The term *decision tree* is derived from the physical appearance of the usual graphic representation of this technique. A decision tree is like the probability trees we introduced in Chapter 4. But a decision tree contains *not only*, the probabilities of outcomes, *but also* the conditional monetary (or utility) values attached to those outcomes. Because of this, we can use these trees to indicate the expected values of different actions we can take. Decision trees have standard symbols:

- Squares symbolize *decision points*, where the decision maker must choose among several possible actions. From these decision *nodes*, we draw one *branch* for each of the possible actions.

TABLE 17-12 DISTRIBUTION OF SNOWFALL AND PROFIT FOR SNOW FUN SKI RESORT

Amount of Snow	Profit	Probability of Occurrence
More than 40 inches	\$120,000	0.4
20 to 40 inches	40,000	0.2
Less than 20 inches	-40,000	0.4

- Circles represent *chance events*, where some state of nature is realized. These chance events are not under the decision maker's control. From these chance nodes, we draw one branch for each possible outcome.

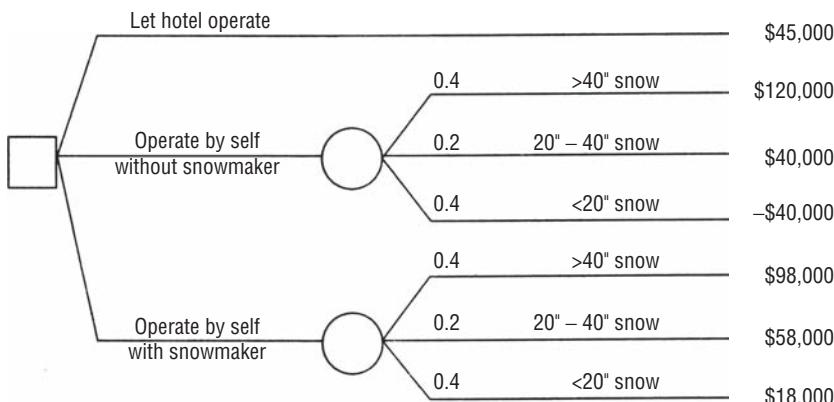
Let's use a decision tree to help Christie Stem, the owner and general manager of the Snow Fun Ski Resort, decide how the hotel should be run in the coming season. Christie's profits for this year's skiing season will depend on how much snowfall occurs during the winter. On the basis of previous experience, she believes the probability distribution of snowfall and the resulting profit can be summarized by Table 17-12.

Decision-tree example: Running a ski resort

Christie has recently received an offer from a large hotel chain to operate the resort for the winter, guaranteeing her a \$45,000 profit for the season. She has also been considering leasing snowmaking equipment for the season. If the equipment is leased, the resort will be able to operate full time, regardless of the amount of natural snowfall. If she decides to use snowmakers to supplement the natural snowfall, her profit for the season will be \$120,000 minus the cost of leasing and operating the snow-making equipment. The leasing cost will be about \$12,000 per season, regardless of how much it is used. The operating cost will be \$10,000 if the natural snowfall is more than 40 inches, \$50,000 if it is between 20 and 40 inches, and \$90,000 if it is less than 20 inches.

Christie's decision tree

Figure 17-9 illustrates Christie's problem as a decision tree. The three branches emanating from the decision node represent her three possible ways to operate the resort this winter: hiring the hotel chain, running it herself without snowmaking equipment, and running it by herself with the snowmakers. Each of the last two branches terminates in a chance node representing the amount of snow that will fall during the season. Each

**FIGURE 17-9 CHRISTIE STEM'S DECISION TREE**

of these nodes has three branches emanating from it, one for each possible value of snowfall, and the probabilities of that much snow are indicated on each branch. Notice that time flows from left to right in the tree; that is, nodes at the left represent actions or chance events that occur before nodes that fall farther to the right. It is very important to maintain the proper time sequence when constructing decision trees.

At the end of each rightmost branch is the net profit that Christie will earn if a path is followed from the root of the tree (at the decision node) to the top of the tree. For example, if she operates the resort herself with the snowmaker and the snowfall is between 20 and 40 inches, her profit will be \$58,000, (\$120,000 less \$12,000 to lease the snowmaker and \$50,000 to operate it). The other net profits are calculated similarly.

We can now begin to analyze Christie's decision tree. (The **Rules for analyzing a decision tree** process starts from the right (at the top of the tree) and works back to the left (to the root of the tree). In this *rollback* process, by working from right to left, we make the future decisions first and then roll them back to become part of earlier decisions.) We have two rules directing this process:

1. If we are analyzing a *chance node* (circle), we calculate the expected value at that node by multiplying the probability on each branch emanating from the node by the profit at the end of that branch and then summing the products for all of the branches emanating from the node.
2. If we are analyzing a *decision node* (square), we let the expected value at that node be the maximum of the expected values for all of the branches emanating from the node. In this way, we choose the action with the largest expected value and we *prune* the branches corresponding to the less profitable actions. We mark those branches with a double slash to indicate that they have been pruned.

For Christie's decision, which is illustrated in Figure 17-10, the **Christie's optimal decision** expected value of hiring the hotel chain to manage the resort is \$45,000. If she operates the resort herself and doesn't use the snowmaking equipment, her expected profit is

$$\$40,000 = \$120,000(0.4) + \$40,000(0.2) - \$40,000(0.4)$$

If she uses the snowmakers, her expected profit is

$$\$58,000 = \$98,000(0.4) + \$58,000(0.2) + \$18,000(0.4)$$

Thus, her optimal decision is to operate Snow Fun herself with snowmaking equipment.

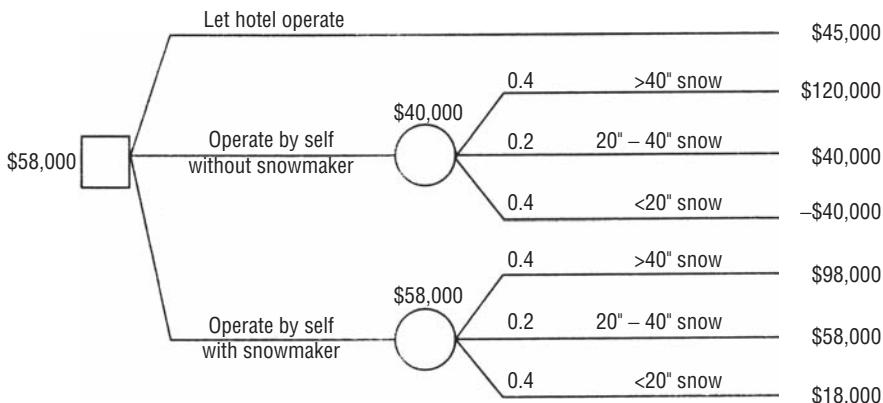


FIGURE 17-10 CHRISTIE STEM'S ANALYZED DECISION TREE

Decision Trees and New Information: Using Bayes' Theorem to Revise Probabilities

Just as Christie is getting ready to decide whether to let the hotel chain operate Snow Fun or to operate it herself, she receives a call from Meteorological Associates offering to sell her a forecast of snowfall in the coming season. The price of the forecast will be \$2,000. The forecast will indicate either that the snowfall will be above normal or else that it will be below normal. After doing a bit of research, Christie learns that Meteorological Associates is a reputable firm whose forecasts have been quite good in the past, although, of course, they haven't been perfectly reliable. In the past, the firm has forecast above-normal snowfall in 90 percent of all years when the natural snowfall has been above 40 inches, in 60 percent of all years when it has been between 20 and 40 inches, and in 30 percent of the years in which it has been below 20 inches.

In order to incorporate this new information and decide whether she should purchase the snowfall forecast, Christie has to use Bayes' Theorem (which we discussed in Chapter 4) to see how the results of the forecast will cause her to revise the snowfall probabilities that she is using to make her decision. The forecast will have some value to her if it will cause her to change her decision and avoid taking a less-than-optimal action. However, before doing the calculations necessary to apply Bayes' Theorem, she decides to see first how much a perfectly reliable forecast of the snowfall would be worth. The calculation of this EVPI can be done with the tree given in Figure 17-11. In this figure, we have reversed the time sequence of Christie's decision and when she learns the season's level of snowfall. In Figure 17-9, she had to decide how to operate the resort, and she then learned the amount of snowfall by actually experiencing it. If a perfectly reliable forecast were available, she would learn how much snow would fall *before she had to decide how to operate the resort*.

Let's examine Figure 17-11 carefully. Even though Christie is trying to determine the worth of a perfectly reliable forecast, she can't know beforehand what the result of the forecast will be. Forty

Cost and value of new information

Incorporating new information

Expected value of perfect information

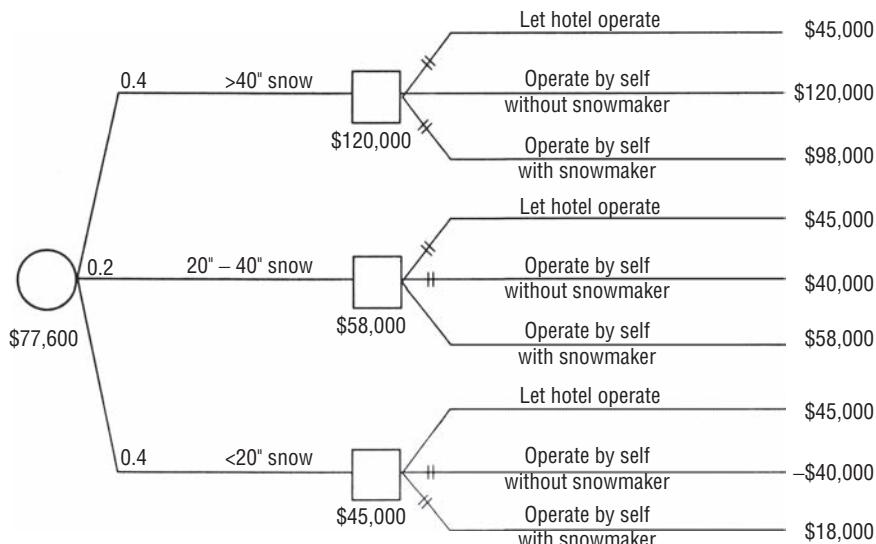


FIGURE 17-11 CHRISTIE'S TREE WITH A PERFECTLY RELIABLE FORECAST

percent of the time, there will be over 40 inches of snow in a skiing season. So, the probability is 0.4 that the forecast will be for over 40 inches of snow. When the snowfall is at that level, Christie's best course of action is to operate the resort herself, without using snow-making equipment, and her profit will be \$120,000. In another 20 percent of all seasons, when snowfall is between 20 and 40 inches, Christie will earn \$58,000 by operating the resort herself and using snowmakers to supplement the meager natural snowfall. Finally, in years with less than 20 inches of natural snowfall (and this happens 40 percent of the time), she should take the \$45,000 profit available by letting the hotel chain operate Snow Fun. With a perfectly reliable forecast, we thus see that Christie's expected profit would be

$$\$77,600 = \$120,000(0.4) + \$58,000(0.2) + \$45,000(0.4)$$

Because her best course of action without the forecast (operating Snow Fun herself with the snowmaking equipment) has an expected profit of only \$58,000, her EVPI is \$19,600 ($\$77,600 - \$58,000$).

Updating probabilities with Bayes' Theorem

Because the forecast from Meteorological Associates is not perfectly reliable, it will be worth less than \$19,600. Nevertheless, Christie sees that additional information about the amount of snowfall can be quite valuable. Will the Meteorological Associates forecast be worth its \$2,000 cost? The answer to this question can be found in Table 17-13 and Figure 17-12. Table 17-13 uses the same format we used in Chapter 4 to do the calculations for using Bayes' theorem to update the snowfall probabilities, given the results of the forecast.

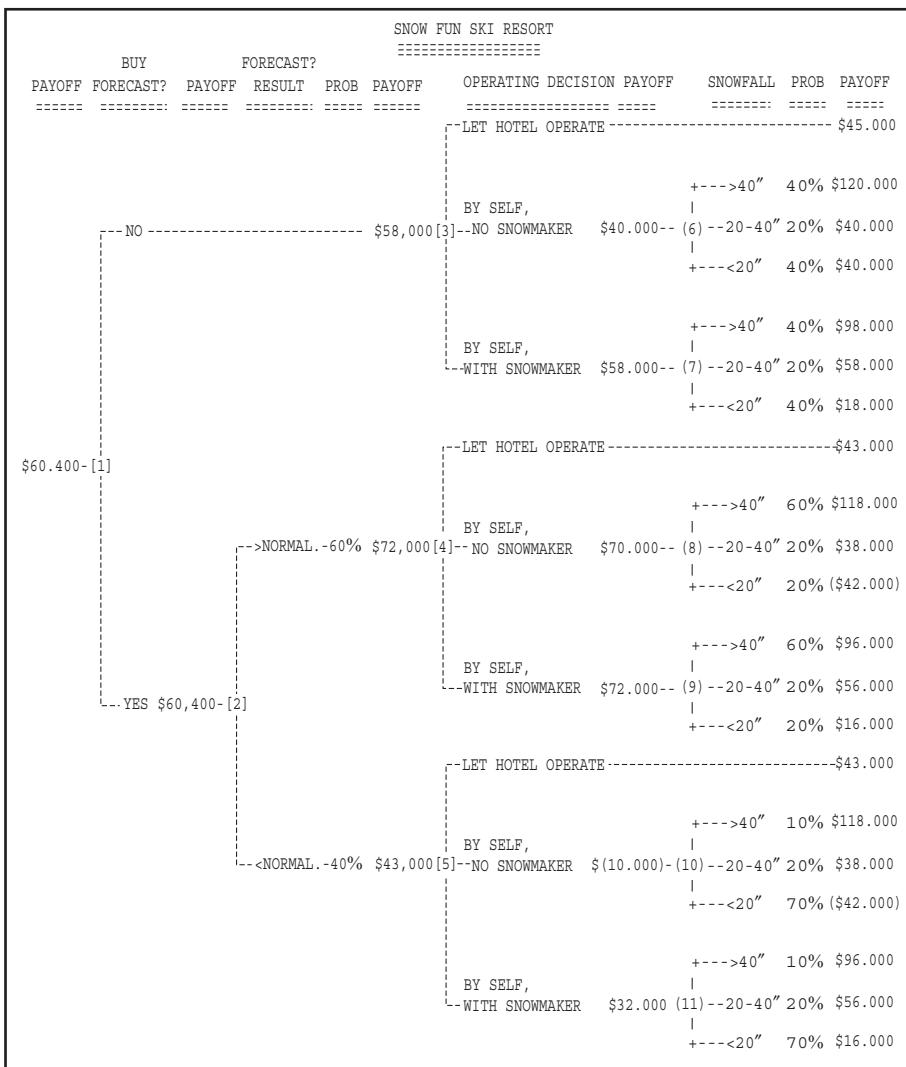
Notice how the probabilities change. If the forecast is for above-normal snowfall, Christie's probability that there will be more than 40 inches of snow increases to 0.6 from its initial value of 0.4. With a forecast for below-normal snowfall, she revises her probability downward to 0.1.

Figure 17-12 gives the entire tree, including the option to buy the forecast from Meteorological Associates. Let's review the rollback procedure for this tree. The top of the tree (from node 3 on) is the same as Figure 17-10. The bottom of the tree (from node 2 on) analyzes Christie's options if she buys the forecast. At the chance nodes 8, 9, 10, and 11, she has calculated expected values using rule 1 on p. 999. Using rule 2, she decides at node 4 that she will run the resort by herself (but hedges her bets by using the snowmaking equipment) if the forecast is for above-normal snowfall. She decides at node 5, on the other hand, that she will accept the hotel chain's offer to operate Snow Fun if the forecast is for below-normal snowfall.

Analyzing the entire tree

TABLE 17-13 CHRISTIE'S POSTERIOR PROBABILITIES

Forecast	Event (snowfall)	P(event)	P(forecast event)	P(forecast & event)	P(event forecast)
Above normal	Over 40"	0.4	0.9	$0.4 \times 0.9 = 0.36$	$0.36/0.60 = 0.6$
	20"-40"	0.2	0.6	$0.2 \times 0.6 = 0.12$	$0.12/0.60 = 0.2$
	Under 20"	0.4	0.3	$0.4 \times 0.3 = 0.12$	$0.12/0.60 = 0.2$
P(above normal) = 0.60					
Below normal	Over 40"	0.4	0.1	$0.4 \times 0.1 = 0.04$	$0.04/0.40 = 0.1$
	20"-40"	0.2	0.4	$0.2 \times 0.4 = 0.08$	$0.08/0.40 = 0.2$
	Under 20"	0.4	0.7	$0.4 \times 0.7 = 0.28$	$0.28/0.40 = 0.7$
P(below normal) = 0.40					

**FIGURE 17-12 CHRISTIE STEM'S COMPLETE DECISION TREE**

Continuing to work her way back through the tree, at node 2 she finds that the expected value of buying the forecast is \$60,400. Finally, at node 1, Christie decides that she should pay Meteorological Associates the \$2,000 that it is charging for its forecast because the resulting expected profit of \$60,400 is more than the \$58,000 she expects to earn without buying the forecast.

In summary, we see that Christie's optimal decision is to buy the ***Christie's optimal decision*** forecast. Then, if the forecast is for above-normal snowfall, she should operate the resort by herself, but hedge her bets by using the snowmaking equipment. However, if the forecast is for below-normal snowfall, she should accept the hotel chain's offer to operate Snow Fun for her. If she follows this course of action, she expects her profit for the season to be \$60,400. Even

INPUT DATA AND BAYES' REVISIONS FOR CHRISIE STEM AND THE SNOW FUN SKI RESORT											
SNOWFALL STATE	PRIOR PROB	PROFIT WITHOUT SNOWMAKER		PROFIT WITH SNOWMAKER		FORECAST RESULT		JOINT PROBABILITIES		REVISED PROBABILITIES	
		SNOWMAKER	OPERATING COST	SNOWMAKER	>NORMAL	<NORMAL	>NORMAL	<NORMAL	>NORMAL	<NORMAL	
>40"	40%	\$120,000	\$10,000	\$98,000	90%	10%	36%	4%	60%	10%	
20-40"	20%	\$40,000	\$50,000	\$58,000	60%	40%	12%	8%	20%	20%	
>20"	40%	(\$40,000)	\$90,000	\$18,000	30%	70%	12%	28%	20%	70%	
PROFIT FROM=>		\$45,000	\$120,000 <=REVENUE WITH SNOWMAKER		60%		40% <=PROBABILITY OF FORECAST RESULTS		\$2,000 <=COST OF FORECAST		
HOTEL LEASE		\$ 12,000 <=COST OF SNOWMAKER LEASE									

FIGURE 17-13 SPREADSHEET WITH CHRISTIE'S INPUT AND BAYES' THEOREM CALCULATIONS

after paying \$2,000 for the forecast, she is \$2,400 better off than she would be if she didn't use it. What is the maximum amount she would be willing to pay for the forecast? She would pay up to an additional \$2,400 for it and still expect to earn at least as much as she could earn without buying it. Thus, the expected value of the forecast (sometimes called the *expected value of sample information*, or EVSI) is \$4,400, and this is the maximum amount that Christie would be willing to pay for it.

You probably noticed that Figure 17-12 (Christie's expanded decision tree) was output from a computer. In fact, we constructed the tree and did the Bayes' Theorem calculations and the rollback procedure using the Lotus 1-2-3 spreadsheet program on a personal computer. (Figure 17-13 gives the input data and the Bayes' computations from our spread-sheet.) Similar analysis can be done with many other spreadsheet programs. A discussion of how to do this kind of analysis is given by J. Morgan Jones in "Decision Analysis Using Spreadsheets," *The European Journal of Operations Research* 26(3) (1986): 385-400. There is also some special-purpose software designed specifically for analyzing decision trees. See the survey article by Dennis Buede, "Aiding Insight, 11," *OR/MS Today* 21(3) (June 1994): 62-68.

Decision trees on the personal computer

Christie is pleased with the results of this analysis, but she still isn't sure that she should go ahead and implement the optimal policy. Her uncertainty stems from the fact that she doesn't know for sure that leasing the snowmaking equipment will cost \$12,000 for the season. That was the amount her friend, Betsy Anderson, had paid last year for snowmakers at her place, The Quaking Aspen Lodge. But there are many differences, among them the fact that Snow Fun's slopes are longer than Quaking Aspen's and that there are several more firms renting snowmakers this year. Christie is reasonably certain that the cost of leasing the equipment will be somewhere between \$5,000 and \$20,000.

Changing some input data

She has realized that there are only three reasonable courses of action (*strategies*) to take:

1. Don't buy the forecast and operate Snow Fun herself using the snowmakers.
2. Buy the forecast and operate Snow Fun herself without snowmakers if the predicted snowfall is above normal, but accept the hotel chain's offer if below-normal snowfall is predicted.
3. Buy the forecast and operate Snow Fun herself with snowmakers if the predicted snowfall is above normal, but accept the hotel chain's offer if below-normal snowfall is predicted.

Reasonable strategies

With her original \$12,000 "guesstimate" of the leasing cost, Christie's optimal decision is to follow the third strategy. She wonders how other possible leasing costs between \$5,000 and \$20,000 will affect her optimal

Sensitivity analysis

SENSITIVITY ANALYSIS ON SNOWMAKER LEASE COST						
=====						
STRATEGY 1: OPERATE BY SELF WITH SNOWMAKERS						
STRATEGY 2: BUY FORECAST AND						
OPERATE BY SELF W/O SNOWMAKERS IF >NORMAL						
LET HOTEL CHAIN OPERATE IF <NORMAL						
STRATEGY 3: BUY FORECAST AND						
OPERATE BY SELF WITH SNOWMAKERS IF >NORMAL						
LET HOTEL CHAIN OPERATE IF <NORMAL						
COST OF	STRATEGY /	EXPECTED	PROFIT	OPTIMAL STRATEGY	EXPECTED VALUE	MAXIMUM TO PAY FOR FORECAST
SNOW-MAKERS	1	2	3			
\$5,000	\$65,000	\$59,200	\$64,600	1	\$65,000	\$1,600
\$6,000	\$64,000	\$59,200	\$64,000	1 OR 3	\$64,000	\$2,000
\$7,000	\$63,000	\$59,200	\$63,400	3	\$63,400	\$2,400
\$8,000	\$62,000	\$59,200	\$62,800	3	\$62,800	\$2,800
\$9,000	\$61,000	\$59,200	\$62,200	3	\$62,200	\$3,200
\$10,000	\$60,000	\$59,200	\$61,600	3	\$61,600	\$3,600
\$11,000	\$59,000	\$59,200	\$61,000	3	\$61,000	\$4,000
\$12,000	\$58,000	\$59,200	\$60,400	3	\$60,400	\$4,400
\$13,000	\$57,000	\$59,200	\$59,800	3	\$59,800	\$4,800
\$14,000	\$56,000	\$59,200	\$59,200	2 OR 3	\$59,200	\$5,200
\$15,000	\$55,000	\$59,200	\$58,600	2	\$59,200	\$6,200
\$16,000	\$54,000	\$59,200	\$58,000	2	\$59,200	\$7,200
\$17,000	\$53,000	\$59,200	\$57,400	2	\$59,200	\$8,200
\$18,000	\$52,000	\$59,200	\$56,800	2	\$59,200	\$9,200
\$19,000	\$51,000	\$59,200	\$56,200	2	\$59,200	\$10,200
\$20,000	\$50,000	\$59,200	\$55,600	2	\$59,200	\$11,200

FIGURE 17-14 SENSITIVITY ANALYSIS ON THE COST OF LEASING THE SNOWMAKING EQUIPMENT

strategy and expected profit, if at all. Although such a *sensitivity analysis* is tedious to do by hand, it is quite easy to do in Lotus 1-2-3, and Figure 17-14 shows Christie what to do as the cost of leasing the snowmaking equipment varies from \$5,000 to \$20,000. If the cost is between \$5,000 and \$6,000, she should adopt the first strategy. (At exactly \$6,000, she is indifferent between the first and third strategies.) For costs between \$6,000 and \$14,000, strategy 3 is optimal. (At exactly \$14,000, she is indifferent between the second and third strategies.) Finally, if the cost is above \$14,000, she should adopt strategy 2.

The last column in Figure 17-14 gives the maximum amount that Christie will be willing to pay for the snowfall forecast. She is including this calculation in her analysis because she has heard a rumor that Meteorological Associates has gotten so much business that they are considering increasing their fees. These figures will be useful to her if she has to negotiate the fee for the forecast.

We have just seen a sensitivity analysis with respect to a cost. **Other sensitivities** In a similar fashion, it is possible to see how optimal decisions and profits change when payoffs or probabilities vary. This capability is especially important when you are using subjective probability estimates in your decision making, and it can be done in a quite straightforward fashion on a personal computer. The ability to perform such sensitivity analyses greatly enhances the value of decision trees in helping us to make important decisions.

Using Decision-Tree Analysis

Solving Christie Stem's problem was easy because the tree had only 11 nodes in it. But real-world decision analysis problems can be much more complex. There can be many more alternatives to consider

at each decision node and many more possible outcomes at each chance node. In addition, more realistic problems often involve longer sequences of decisions and chance events. (The trees get taller and bushier!) When solving a problem with a decision tree, remember to stop at a level of complexity that allows you to consider major consequences of future alternatives without becoming bogged down in too much detail.

Generally, decision-tree analysis requires the decision maker to proceed through the following six steps:

- 1. Define the problem in structured terms.** First, determine which factors are relevant to the solution. Then estimate probability distributions that are appropriate to describe future behavior of those factors. Collect financial data concerning conditional outcomes. *Decision-tree steps*
- 2. Model the decision process;** that is, construct a decision tree that illustrates all the alternatives involved in the problem. This step *structures* the problem in that it allows the entire decision process to be presented schematically and in an organized, step-by-step fashion. In this step, the decision maker chooses the number of periods into which the future is to be divided.
- 3. Apply the appropriate probability values and financial data** to each of the branches and subbranches of the decision tree. This will enable you to distinguish the probability value and conditional monetary value associated with each outcome.
- 4. “Solve” the decision tree.** Using the methodology we have illustrated, proceed to locate the particular branch of the tree that has the largest expected value or that maximizes the decision criterion, whatever it is.
- 5. Perform sensitivity analysis;** that is, determine how the solution reacts to changes in the inputs. Changing probability values and conditional financial values allows the decision maker to test both the magnitude and the direction of the reaction. This step allows experiments without real commitments or real mistakes and without disrupting operations.
- 6. List the underlying assumptions.** Explain the estimating techniques used to arrive at the probability distributions. What kinds of accounting and cost-finding assumptions underlie the conditional financial values used to arrive at a solution? Why has the future been divided into a certain number of periods? By making these assumptions explicit, you enable others to know what risks they are taking when they use the results of your decision-tree analysis. Use this step to specify limits under which the results obtained will be valid, and especially the conditions under which the decision will not be valid.

Decision-tree analysis is a technique managers use to structure and display alternatives and decision processes. It is popular because it

Advantages of the decision-tree approach

- Structures the decision process, guiding managers to approach decision making in an orderly, sequential fashion.
- Requires the decision maker to examine all possible outcomes, desirable and undesirable.
- Communicates the decision-making process to others, illustrating each assumption about the future.
- Allows a group to discuss alternatives by focusing on each financial figure, probability value, and underlying assumption one at a time; thus, a group can move in orderly steps toward a consensus decision, instead of debating a decision in its entirety.
- Can be used with a computer, so that many different sets of assumptions can be simulated and their effects on the final outcome observed.

HINTS & ASSUMPTIONS

Warning: Don't forget that the probabilities at each node of a decision tree must add up to 1.0. And don't forget that the important part of decision tree analysis is supplying the probabilities. These are far more difficult to ascertain than are the financial values. As we become more familiar with accounting and finance, we should feel more secure in estimating financial outcomes. But even when you become a financial whiz, you can still be uncomfortable and unable to "reach way down in your gut" and come up with reasonable probabilities of outcomes. The ability to attach reasonable subjective probabilities to outcomes in a consistent manner is why successful managers are paid more than successful bookkeepers even though both perform useful work for the organization. Finally, it shouldn't surprise you that companies actually use decision trees as a part of *expert systems* (systems written in advanced computer language that can handle symbols as well as numerical values), which actually *make* decisions by mimicking a decision maker's behavior as she solves a problem.

EXERCISES 17.6**Self-Check Exercise**

SC 17-4 Evelyn Parkhill is considering three possible ways to invest the \$200,000 she has just inherited.

- (1) Some of her friends are considering financing a combined laundromat, video-game arcade, and pizzeria, where the young singles in the area can meet and play while doing their laundry. This venture is highly risky and could result in either a major loss or a substantial gain within a year. Evelyn estimates that with probability 0.6, she will lose all of her money. However, with probability 0.4, she will make a \$200,000 profit.
- (2) She can invest in some new apartments that are being built in town. Within 1 year, this fairly conservative project will produce a profit of at least \$10,000, but it might yield \$15,000, \$20,000, \$25,000, or possibly even \$30,000. Evelyn estimates the probabilities of these five returns at 0.20, 0.30, 0.25, 0.20, and 0.05, respectively.
- (3) She can invest in some government securities that have a current yield of 8.25 percent.
 - (a) Construct a decision tree to help Evelyn decide how to invest her money.
 - (b) Which investment will maximize her expected 1-year profit?
 - (c) How high would the yield on the government bonds have to be before she would decide to invest in them?
 - (d) How much would she be willing to pay for perfect information about the success of the laundromat?
 - (e) How much would she be willing to pay for perfect information about the success of the apartments?

Applications

17-25 The Motor City Auto Company is planning to introduce a new automobile that features a radically new pollution-control system. It has two options. The first option is to build a new plant, anticipating full production in 3 years. The second option is to rebuild a small existing pilot plant for limited production for the coming model year. If the results of the limited production show promise at the end of the first year, full-scale production in a newly constructed plant would still be possible 3 years from now. If it decides to proceed with the pilot plant and later analysis shows that it is unattractive to go into full production, the pilot plant can still

be operated by itself at a small profit. The expected annual profits for various alternatives are as follows:

Production Facility	Consumer Acceptance	Annual Profit (\$ millions)
New plant	High	14
New plant	Low	-6
Pilot plant	High	2
Pilot plant	Low	1

Motor City's marketing-research division has estimated that there is a 50 percent probability that consumer acceptance will be high and 50 percent that it will be low. If the pilot plant is put into production, with a corresponding low-keyed advertising program, the researchers feel that the probabilities are 45 percent for high consumer acceptance and 55 percent for low acceptance. Further, they have estimated that if the pilot plant is built and consumer acceptance is found to be high, there is a 90 percent probability of high acceptance with full production. If consumer acceptance with the pilot models is found to be low, however, there is only a 10 percent probability of high eventual acceptance with full production. Which plant should be built?

17-26

Refer to Christie Stem's problem on p. 940 and in Figure 17-9.

- Suppose that the operating cost of the snowmaking equipment is actually 30 percent higher than Christie had estimated, that is, \$13,000 if the snowfall is heavy, \$65,000 if it is moderate, and \$117,000 if it is light. How will this affect Christie's optimal decision and expected profit?
- Answer the same questions if the actual operating cost is 20 percent higher than Christie's original estimate.
- At what percentage increase of the operating cost will Christie be indifferent between the optimal decisions in parts (a) and (b)? At this point, what will be her expected profit?

17-27

International Pictures is trying to decide how to distribute its new movie *Claws*. *Claws* is the story of an animal husbandry experiment at North Carolina State University that goes awry, with tragicomic results. An effort to breed meatier turkeys somehow produces an intelligent, 1,000-pound turkey that escapes from the lab and terrorizes the campus. In a surprise ending, the turkey is befriended by Coach Morey Robbins, who teaches it how to play basketball, and State goes on to win the NCAA championship. Because of the movie's controversial nature, it has the potential to be either a smash hit, a modest success, or a total bomb. International picture is trying to decide whether to release the picture for general distribution initially or to start out with a "limited first-run release" at a few selected theaters, followed by general distribution after 3 months. The company has estimated the following probabilities and conditional profits for *Claws*:

Level Success	Probability	Profits (\$ millions)	
		Limited Release	General Distribution
Smash	0.3	22	12
Modest	0.4	9	8
Bomb	0.3	-10	-2

- (a) Construct a decision tree to help International decide how to release *Claws*.
 (b) Which decision will maximize the expected profit?
 (c) How much would International pay for an absolutely reliable forecast of the movie's level of success?
 (d) International can run several sneak previews of *Claws* to get a better idea of the movie's ultimate level of success. Preview audiences rate movies as either good or excellent, but their opinions are not completely reliable. On the basis of past experience with previews, International has found that 90 percent of all smash successes were rated excellent (with 10 percent of them being rated good), 65 percent of all modest successes were rated excellent (with 35 percent of them being rated good), and 40 percent of all bombs were rated excellent (with 60 percent of them being rated good). If the cost of sneak previews would be about \$750,000, should *Claws* be previewed? How should International respond to the preview results? What is the maximum amount International should be willing to pay for the previews?

- 17-28** Sam Crawford, a junior business major, lives off campus and has just missed the bus that would have taken him to campus for his 9 A.M. test. It is now 8:45 A.M. and Sam has several options available to get him to campus: waiting for the next bus, walking, riding his bike, or driving his car. The bus is scheduled to arrive in 10 minutes, and it will take Sam exactly 20 minutes to get to his test from the time he gets on the bus. However, there is a 0.2 chance that the bus will be 5 minutes early, and a 0.3 chance that the bus will be 5 minutes late. If Sam walks, there is a 0.8 chance he will get to his test in 30 minutes, and a 0.2 chance he will get there in 35 minutes. If Sam rides his bike, he will get to the test in 25 minutes with probability 0.5, 30 minutes with probability 0.4, and there is a 0.1 chance of a flat tire, causing him to take 45 minutes. If Sam drives his car to campus, he will take 15 minutes to get to campus, but the time needed to park his car and get to his test is given by the following table:

Time to park & arrive (minutes)	10	15	20	25
Probability	0.30	0.45	0.15	0.10

- (a) Assuming that Sam wants to *minimize* his expected late time in getting to his test, draw the decision tree and determine his best option.
 (b) Suppose instead that Sam wants to *maximize* his expected utility as measured by the projected test score given below. Use the same decision tree to determine his optimal decision now.

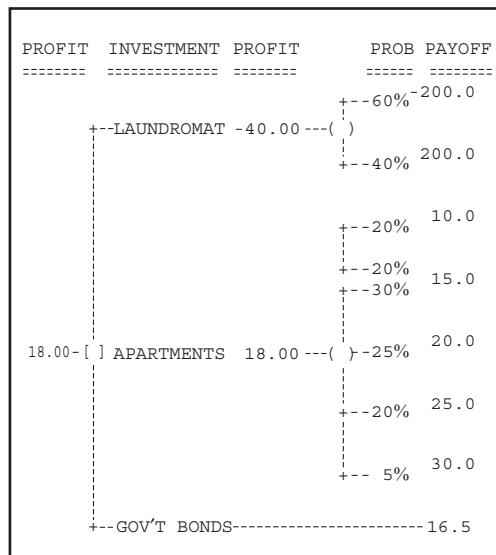
Arrival time	9:10	9:15	9:20	9:25	9:30
Projected test score	95	85	70	60	45

- 17-29** The North Carolina Airport Authority is trying to solve a difficult problem with the over-crowded Raleigh–Durham airport. There are three options to consider:
 (1) The airport could be totally redesigned and rebuilt at a cost of \$8.2 million. The present value of increased revenue from a new airport is in question. There is a 70 percent probability this present value would be \$11.0 million, a 20 percent probability the present value would be \$5 million, and a 10 percent probability the present value would be \$1.0 million, depending on whether the airport is a success, moderate success, or a failure.

- (2) The airport could be remodeled with a new runway for a cost of \$4.7 million. The present value of increased revenue would be \$6.0 million (with probability 0.8) or \$3.0 million (with probability 0.2).
- (3) They could do nothing with the airport and suffer a loss of revenue of either \$1 million (with probability 0.65) or \$4 million (with probability 0.35).
- Construct a decision tree to help the Airport Authority.
 - Which option will maximize the present value of profit?
 - How much would we be willing to pay for perfect information about success of a brand new airport?
 - How much would we be willing to pay for perfect information about success of a remodeled airport?

Worked-Out Answer to Self-Check Exercise

SC 17-4 (a)



- She should invest in the apartments.
- For the bonds to yield over \$18,000, they would have to pay more than a 9% rate of interest.
- With perfect information about the laundromat, she would invest in it if she knew it would be successful, but would invest in the apartments otherwise. Hence her expected return with perfect information is $0.6(18) + 0.4(200) = 90.8$, and so

$$\text{EVPI} = 90.8 - 18 = 72.8, \text{ i.e., } \$72,800$$

- With perfect information about the apartments, she would invest in them if their return is over \$16,500, but would buy government bonds otherwise. Hence, her expected return with perfect information is $0.5(16.5) + 0.25(20) + 0.20(25) + 0.05(30) = 19.75$, and so

$$\text{EVPI} = 19.75 - 18 = 1.75, \text{ i.e., } \$1,750$$

STATISTICS AT WORK

Loveland Computers

Case 17: Decision Theory A curious calm fell over Loveland Computers and Lee Azko began to think about scheduling a well-deserved day on the slopes. Both Walter Azko and Gratia Delaguardia had been away from the office for 2 days—the rumor mill had it that they were in New York, meeting with the investment bankers.

Lee found a message waiting on the answering machine at home: “Lee, this is your uncle. You can forget about skiing this weekend. And don’t go into the office. Gratia and I have a big decision to make. Come up to my house early tomorrow morning. I’ll fix you breakfast and you can help Gratia and me figure this one out.”

“Help yourself,” Walter began the next morning, indicating a large stack of pancakes. “You can probably guess where we’ve been this week.” Walter’s gesture indicated he was speaking of his partner and himself. “I know that it may seem strange for a company as big as this to still be a partnership. But in many ways it’s just a ‘mom and pop’ business”

“... with some pretty big numbers,” Gratia added.

“Well, there are all kinds of companies that got to be pretty big while they were still privately held.” Walter concluded.

“Most of the software companies—and some of your direct competitors,” Lee commented.

“So now we’re at a turning point,” the CEO of Loveland Computers continued. “These fellows in New York are prepared to make a substantial investment—and I mean substantial—in Loveland. But, as you’d expect, they want us to form a corporation and give them a 60 percent stake. I guess that’s pretty usual. Somewhere down the line, maybe 2 to 5 years from now, they’ll take the company public.”

“And you and Gratia—with 20 percent each—would be worth a fortune,” Lee said cheerfully, wondering if it was too early to ask for a bonus.

“But on the other hand,” Gratia cautioned, “we might be better off just staying as we are. Of course, it means that we’ll have to limit our growth to, well, maybe 25 percent per year.”

“As opposed to 50 to 100 percent annual growth in sales if we have enough capital behind us,” said Walter.

“Well, this one’s a ‘no-brainer,’ Nunc.” Lee was still dreaming of large bonuses. “Go for the gold. Take the money, expand all you want—new warehouses and more phone banks—and make a bigger profit.”

“It’s not as simple as that,” Gratia continued to have doubts. “The economy is flat at best. If there’s a rebound in the economy, expansion will pay off. But if the country continues for another year with very slow growth, then the only way we could expand our market share would be to seriously cut prices. So, we’d keep new facilities humming—but we’d be bringing much less money to the bottom line.”

“You mean you could sell more and earn less?” Lee was incredulous.

“Absolutely. It happens more often than you think.”

“In fact, I can’t be sure about the pricing structure of the whole industry,” Walter said, rejoining the conversation and stretching for the maple syrup. “Many industry experts are expecting some of the big names—IBM and Compaq—to give up their high-price strategy. If they accept a much lower margin on their machines they could greatly increase the number of computers they sell. And they both have manufacturing capability here in the U.S., so they may be able to increase production much faster than we can.”

“Give me that table napkin and a pen,” said Lee beginning to look more serious. “Let me see if I can sketch out your options.”

Study Questions: What is Lee drawing on the napkin? What is the action that the partners will—or will not—take after this discussion? What are the uncertainties they face? How good will these three people be at estimating the probabilities of various outcomes?

CHAPTER REVIEW

Terms Introduced in Chapter 17

Certainty The decision environment in which only one state of nature exists.

Conditional Profit The profit that would result from a given combination of decision alternative and state of nature.

Decision Point Branching point that requires a decision.

Decision Tree A graphic display of the decision environment, indicating decision alternatives, states of nature, probabilities attached to those states of nature, and conditional benefits and losses.

Expected Marginal Loss The marginal loss multiplied by the probability of not selling that unit.

Expected Marginal Profit The marginal profit multiplied by the probability of selling that unit.

Expected Profit The sum of the conditional profits for a given decision alternative, each weighted by the probability that it will happen.

Expected Profit with Perfect Information The expected value of profit with perfect certainty about which of the states of nature will occur.

Expected-Value Criterion A criterion requiring the decision maker to calculate the expected value for each decision alternative (the sum of the weighted payoffs for that alternative in which the weights are the probability values assigned by the decision maker to the states of nature that can happen).

Expected Value of Perfect Information The difference between expected profit (under conditions of risk) and expected profit with perfect information.

Marginal Loss The loss incurred from stocking a unit that is not sold.

Marginal Profit The profit earned from selling one additional unit.

Minimum Probability The probability of selling at least an additional unit that must exist to justify stocking that unit.

Node The point at which a chance event or a decision takes place on a decision tree.

Obsolescence Loss The loss occasioned by stocking too many units and having to dispose of unsold units.

Opportunity Loss The profit that could have been earned if stock had been sufficient to supply a unit that was demanded.

Payoff The benefit that accrues from a given combination of a decision alternative and a state of nature.

Rollback Also called foldback; a method of using decision trees to find optimal alternatives. It involves working from right to left in the tree.

Salvage Value The value of an item after the initial selling period.

State of Nature A future event not under the control of the decision maker.

Utility The value of a certain outcome or payoff to someone; the pleasure or displeasure someone derives from an outcome.

Equations Introduced in Chapter 17

17-1

$$p(MP) = (1 - p)(ML)$$

p. 924

This equation describes the point at which the *expected marginal profit* from stocking and selling an additional unit, $p(MP)$, is equal to the *expected marginal loss* from stocking and

not selling the unit, $(1 - p)(ML)$. As long as $p(MP)$ is larger than $(1 - p)(ML)$, additional units should be stocked because the expected marginal profit from such a decision is greater than the expected marginal loss.

17-2

$$p^* = \frac{ML}{MP + ML}$$

p. 924

This is the *minimum probability equation*. The symbol p^* represents the minimum required probability of selling at least an additional unit to justify the stocking of that additional unit. As long as the probability of selling one additional unit is greater than p^* , the retailer should stock that unit. This equation is Equation 17-1 solved for p .

Review and Application Exercises

- 17-30** The Mountain Manufacturing Company is planning to produce dot-matrix printers for use with microcomputers. One problem it faces is a make-or-buy decision for the print heads. It can buy these units from a Japanese manufacturer for \$35 each or it can produce them at its own plant with variable costs of \$24 a unit. If it elects to produce the print heads itself, it will incur fixed costs of \$28,000 each year. Because of defective units, each printer requires 1.15 print heads. The company foresees annual demand for its printers to be normally distributed with mean $\mu = 3,000$ units and standard deviation $\sigma = 700$ units. What is the probability that the required usage of print heads will be sufficiently large to justify producing them rather than buying them? If it is company policy to make components only when there is better than a 60 percent chance that usage is 1.5 standard deviations above the make-or-buy break-even point, what should the decision be on this matter?
- 17-31** Sarah Peterson is going to open a health-food store, the Boysenberry Farms Organic Food Emporium. In planning for her initial stock, Sarah is trying to decide how many jars of Mrs. Miles' Currant Jelly to purchase. Mrs. Miles makes her currant jelly only once every 2 months, so it is necessary for Sarah to plan in advance how much she will need (there is no chance of reorder in the interim period). Sarah is torn between satisfying her customers and friends and losing money because of spoilage, since the jelly has only a 2 month shelf life. Sarah is sure that she will sell at least 10 jars during the period, and 18 different friends have promised that they will buy the jelly when it comes into stock. Sarah knows that the probability of selling more than 18 jars is practically nil and feels that sales will fall somewhere between 10 and 18 jars, despite what her friends have promised. Sarah has all the cost data and is planning a 50 percent markup on cost. As the problem stands now, can Sarah reach a solution to her problem by using decision theory?
- 17-32** For \$26.95, La Langouste offers an entrée consisting of two broiled spiny-lobster tails with drawn-butter garlic sauce. Because of federal health regulations, the lobsters, which are imported from the Yucatan Peninsula, cannot enter the United States if they are still alive. Accordingly, only refrigerated or frozen tails can be imported. The chef at La Langouste refuses to use frozen lobster tails and to maintain his establishment's reputation for serving only *haute cuisine*, he employs an agent to place freshly refrigerated lobster tails on a plane leaving the peninsula each day. Any tail not served the day it is shipped must be discarded. The chef wants to know how many tails the agent should ship each day. He wants to be able to satisfy his customers, but he realizes that always ordering enough to meet potential demand could involve substantial waste on days with low demand. He has calculated the cost of a single lobster tail at \$7.35, including transportation charges. Past records show the following distribution of daily demand for the lobster-tail entrée:

Number	18	19	20	21	22	23	24	25
Probability	0.07	0.09	0.11	0.16	0.20	0.15	0.14	0.08

- (a) If he wishes to maximize his daily expected profits on spiny-lobster tails, how many tails should the chef order?
- (b) If La Langouste adopted a policy that requires customers to order spiny lobster a day in advance, how much increase in profit could it expect to see?

17-33 Bay Lakes Lawn and Garden Care Company provides services for homeowners and small businesses. The firm is considering the purchase of a new fertilizer spreader at a cost of \$43.50. The spreader is estimated to save 8 minutes labor for every hour it is in use. Head lawn-care specialist Ralph Medlin estimates that the expected life of the spreader is only 48 hours due to corrosion and the odds are 7 to 5 that its life will be between 42 and 54 hours. If the company pays its gardening help \$12.50 an hour, what is the probability that the spreader will pay for itself before it is scrapped?

17-34 The luggage department of Madison Rhodes Department Store featured a special Day-After-Christmas Sale of Luggage on unsold Christmas merchandise. The luggage brand on sale was Imagemaker. The manager of the luggage department was planning his order. Because the store did not carry Imagemaker during the year, the manager wanted to avoid over-stocking; yet, because of a special price the manufacturer offered on the line, he also wanted to minimize stockouts. He was currently attempting to decide the number of women's tote bags to purchase. His estimate of the probable sales, based in part on past performance, is

Bags	32	33	35	35	36	37	38
Probability	0.10	0.14	0.15	0.20	0.17	0.13	0.11

The store is planning to sell the tote bag for \$42.75. The wholesale cost is \$26.00. How many bags should be ordered for the sale?

17-35 Archdale Stores, a chain of retailers specializing in men's fashions, is considering purchasing a batch of 5,700 neckties from Beau Charm Company. The batch of ties will cost Archdale \$16,500, and each tie will sell for \$3.50. Archdale's vice president of sales has stated that he thinks the chain could sell 5,000 ties, and the odds are 2 to 3 that the actual sales will be within 200 of his estimate. Leftover ties are worthless.

- (a) What is the probability that Archdale will at least break even on the necktie sales?
- (b) What is the probability that Archdale can earn 10 percent or more on its inventory investment?

17-36 Barry Roberts, chief corporate counsel for Triangle Electronics, has just learned that a competitor has filed two related patent infringement suits against Triangle. The first of these will be heard in Superior Court in 3 months, and the second is scheduled for 6 months thereafter. Barry estimates that the first trial will take no longer than 4 months to complete. The options available to Triangle in each case are to settle out of court or to let the trial take place. Preparing to try either suit alone will cost \$7,500, but some of the legal preparation on the first suit will help on the second, so the cost of preparing to try both suits will be only \$12,000. Barry estimates that it will cost Triangle \$75,000 to settle the first suit out of court and \$45,000 to settle the second. Of course, settling out of court enables Triangle to avoid the trial preparation costs. If the suits go to trial and Triangle wins, they will incur no further costs. However, Barry estimates that losing the first will result in additional costs of \$150,000, and losing the

second will cost approximately \$90,000. He feels that Triangle has a 60 percent chance of winning the first suit. The chance of winning the second suit depends on the resolution of the first: 40 percent if it is settled out of court, 80 percent if it is tried and won, and 10 percent if it is tried and lost.

- Construct Barry's decision tree for deciding how to proceed.
- What should Barry do to minimize Triangle's expected cost?
- Barry could run a mock trial to get a better idea of the probability of winning the first suit. How much should Triangle be willing to pay if Barry can arrange for an absolutely reliable mock trial?
- How would Barry's decision in part (b) change if the cost of settling the second suit were only \$20,000? What if that cost were \$90,000?

17-37 Optometrics Village owns a regional chain of eyecare shops and its managers were considering adding prescription underwater goggles for customers who like to scuba or snorkel. A marketing consulting firm has estimated annual demand at 4,000 pairs with a standard deviation of 450 if the price is set at \$130 per pair; at a price of \$140 per pair, the estimated annual demand is 3,200 with a standard deviation of 300 pairs. The investment required for lens grinding equipment is \$500,000 and there are fixed costs of \$125,000 per year. The variable cost for each pair of goggles is \$80. The Board of Directors for Optometrics has set a "hurdle" annual rate of return for new ventures at 13 percent, and the managing director wants at least a 60 percent chance of meeting that target. Should they proceed with this venture? If so, which price is most likely to meet the hurdle rate of return on their investment?

17-38 At the Campus Set, a clothing store for stylish young moderns, manager Judy Sommers is ordering the season's bathing suits from Jamaican Swimwear. As in past years, she is ordering mostly two-piece suits, but she does plan to carry some one-piece suits. From past experience, she estimates demand for the latter:

Units demanded	19	20	21	22	23	24	25
Probability	0.05	0.18	0.21	0.22	0.16	0.10	0.08

The one-piece suits will retail for \$43.95; Judy's cost is \$21.50. Any suits left at the end of the season go on sale for \$19.95 and are certain to sell at that price. Use marginal analysis to determine the number of one-piece suits Judy should order.

17-39 Flint City Appliance Sales is planning for its big Founder's Day Weekend Sale. As a special offer, the store is selling a Royalty washer-dryer combination for only \$600. Royalty has recently informed its distributors that a product innovation will make existing washer-dryer combinations virtually obsolete, and therefore it is offering stores its current first-line washer-dryer combination for only \$325. Although the manager of Flint City does not believe all of Royalty's talk of obsolescence, he does know that any new gadget that Royalty puts on its newer machines will make his older machines very difficult to sell. Therefore, he wants to be very careful about the number of machines he orders for the Founder's Day Sale. His estimate of the demand for the washer-dryer combinations during the sale is

Units demanded	6	7	8	9	10	11
Probability	0.04	0.12	0.30	0.24	0.18	0.12

Use marginal analysis to determine how many more washer-dryer combinations should be ordered for the sale if Flint City already has two in stock.

- 17-40** Steel-Fab Manufacturing is a competitor of the Enduro Company (Exercise 17-17) in the structural-steel components market. Unlike Enduro, Steel-Fab is publicly held and is also financed in part by a bond issue. Accordingly, the company has adopted a 9 percent cutoff rate of return. Below the 9 percent level, the firm's utility curve steepens as the return moves farther away. Above the 9 percent level, the firm's utility grows at a slower rate because of the accompanying risk involved with higher rates of return. The utility for 15 percent is only slightly higher than for 14 percent. Steel-Fab is considering a \$300,000 project. Plot the firm's utility curve.

- 17-41** A textile mill must decide whether to extend \$150,000 credit to a new customer that manufactures dresses. The mill's prior experience with a number of dress manufacturers has led it to classify such customers as follows: 25 percent are poor risks, 45 percent are average risks, and 30 percent are good risks. Expected profits on this order (if credit is extended to the dress manufacturer) are – \$20,000 if it turns out to be a poor risk, \$18,000 if it turns out to be an average risk, and \$25,000 if it turns out to be a good risk. Draw a decision tree to determine whether the mill should extend credit to this manufacturer.

- 17-42** For \$750, the textile mill in Exercise 17-41 can purchase a comprehensive credit analysis and rating of the manufacturer. The rating, in increasing order of creditworthiness, will be C, B, or A. The credit agency's reliability is summed up in the following table, whose entries are the probabilities (from past experience) of the agency's rating of the dress manufacturer, given the true credit category in which the manufacturer belongs.

Agency Rating	True Category		
	Poor	Average	Good
A	0.1	0.1	0.6
B	0.2	0.8	0.3
C	0.7	0.1	0.1

- (a) Use Bayes' Theorem and a decision tree to determine whether the mill should purchase the credit rating.
- (b) If it does purchase the rating, how will this affect the decision to grant credit to the dress manufacturer?
- (c) What is the maximum amount the mill will be willing to pay for the credit report?
- (d) What would the mill be willing to pay for an absolutely reliable credit rating of the manufacturer?

- 17-43** John Silver can use his boat, the *Jolly Roger*, for either commercial tuna fishing or sport fishing. For the latter, he rents it out at a daily charge of \$500. In a fishing season with good weather, he averages 150 rental days. However, if the weather is bad, he averages only 105 rental days. For each day the boat is rented, John estimates he incurs variable costs of about \$135. When the weather is good, the revenues from fishing for tuna exceed the variable costs of that operation by \$50,000, whereas in seasons with bad weather, the profit contribution from tuna fishing is only \$43,000. At the beginning of the 1997 season, John feels that the odds are about 7 to 3 in favor of good weather for the season.

- (a) Use a decision tree to help John decide how to use the *Jolly Roger* during the 1997 fishing season.
- (b) How much would John pay for a perfectly reliable long-range weather forecast for the season?

John's good friend, Jim Hawkins, runs a private weather forecasting service that has been 90 percent accurate in the past. In 90 percent of all seasons that had good weather, Jim had forecast good weather, and likewise in 90 percent of all seasons when the weather proved to be bad, Jim's forecast had been for bad weather. Jim usually sells his forecast for \$1,000, but because John is a good friend, Jim is willing to sell it to him for only \$400.

- (c) Expand your decision tree to help John decide whether he should buy Jim's forecast. How will the forecast affect his use of the boat during the 1997 season?
- (d) Would John buy Jim's forecast if they weren't friends? Explain. What is the maximum amount John would be willing to pay for the forecast?

17-44 Robert Ingwersen of Tungsten Products has approached both the Enduro Manufacturing Company and Steel-Fab Manufacturing about the possibility of a joint venture with one of them. In this venture, a tungsten alloy is used in place of certain steel alloys. Tungsten Products has the technological expertise but not the production capabilities. The joint venture will be a 50–50 split and will cost each company \$500,000 in capital investment.

- (a) If the expected first-year profit on the project is \$80,000, would either or both firms accept the offer?
- (b) Superimpose the graphs from Exercises 17-17 and 17-40, adjusting the coordinates, and show the area where Euduro would accept a project and Steel-Fab would not.
- (c) If the expected first-year profit on the project was \$110,000, would either of the firms accept it? How much would Steel-Fab bid for a 50 percent share of the \$110,000?

17-45 Marty Tait is a housing developer who is considering building a *spec* house, so called because there is no particular buyer lined up so the venture is speculative. The lot overlooks the Golden Gate Bridge, so it is expensive. The hill-side location means that it will require substantial foundation work. But the view is spectacular and the selling price of the house should be high. If the house is sold quickly upon completion, Marty makes a good profit above the contractor's fee he charges for each deal. But if it takes too long to market the house after completion, his profit is eaten up by interest on the construction loan and price reductions to sell the property. Marty works closely with a real estate agent, who has estimated the chances that the house will sell in 30, 60, and 90 days after completion. The payoffs and probabilities are given in the following table. Should Marty go ahead and build the house?

Days to sell	Probability	Playoffs (loss)	
		Build	Don't Build
30	0.20	\$71,000	\$0
60	0.30	\$26,000	\$0
90	0.50	(\$42,000)	\$0

17-46 Stanley Glass, the owner of a chain of family amusement centers in Ohio, plans to open another center in Cincinnati. He must decide whether it should have 20, 25, or 35 video games. He expects that the demand may be either high, medium, or low, and he has determined probabilities associated with each level. The probabilities and payoffs are as follows:

Event	Probability	20 Games	25 Games	35 Games
High demand	0.55	\$12,600	\$18,000	\$23,000
Medium demand	0.30	11,000	16,200	15,000
Low demand	0.15	10,600	8,500	7,100

- (a) Without further information about demand, what should Mr. Glass choose to do?
 (b) What is the maximum amount he would be willing to pay for perfectly reliable information?

17-47

The new engineering school at a small southern university is currently deciding which textbooks to use in its undergraduate courses. The department chairpersons want to know whether to use textbooks written by professors within the university (university textbooks) or those written by professors from other institutions (outside textbooks). It has been rumored that the school's administrators are pushing for more support for the university and may require that departments use university textbooks whenever possible. If this requirement is passed, and if the department has decided to purchase outside textbooks, the switch to university textbooks will prove to be quite costly. The university's preliminary payoff table follows (payoffs are in thousands of dollars).

Event	Probability	Use University Texts	Use Outside Texts
Requirement passed	0.70	\$ 8	\$13
Requirement not passed	0.30	16	13

- (a) Compute the expected payoff for each of the two decisions.
 (b) Which decision should the engineering school choose?

17-48

Allyson Smith, assistant manager of Records and Tapes Unlimited, plans to sell a weekly music magazine. She is aware that if the magazine does not sell within the week of publication, it is considered to be worthless to the store. Allyson speculates, based on past sales data, about how well the magazine would sell; her weekly sales and probability estimates are as follows:

No. of magazines	500	600	700	800	900
Probability	0.10	0.12	0.15	0.33	0.30

The magazine has a production cost of \$0.70 each, but Records and Tapes Unlimited plans to sell it for \$1.50 each. Determine the optimal number of magazines that the store should order, using the expected-value decision criterion.

17-49

The women of Alpha Zeta sorority at a small midwestern college are getting ready to participate in the school's annual 3-day spring celebration. As in previous years, the sorority will run a soda booth, selling drinks for \$0.75 a cup. When initial setup and material costs are deducted, the sorority incurs a cost of \$0.35 for each (8 oz) cup of soda. Data collected from last year's celebration indicate that total soda sales are normally distributed with mean 960 and standard deviation of 140. Determine the amount of soda (in ounces) that the women should purchase.

17-50

The chief administrator of a chain of convalescent homes wants to open a new facility in southern California. His decision to build a 50-, 75-, or 150-bed facility will be based on

whether expected demand is low, medium, or high. Based on past experience, he constructs the following table of short-range profits:

Event	Probability	50-Bed	75-Bed	150-Bed
Low demand	0.2	\$41,000	-\$12,000	-\$53,000
Medium demand	0.3	52,000	68,000	-24,000
High demand	0.5	65,000	80,000	117,000

- (a) What size facility should the administrator decide to build?
- (b) Calculate the expected profit with perfect information.
- (c) Use your answer to part (b) to calculate the administrator's expected value of perfect information.

17-51 University Gear Sweatshop is a clothing store that caters to the students of a college known for its fantastic football record. Janet Sawyer, the store's manager, is deciding whether to order more sweatshirts printed with the team's name and mascot. If the team loses the championship this year, the extra sweatshirts won't sell very well, but if the team wins, she expects to be able to make a high profit on the shirts. The local paper is predicting a 65 percent chance that the team will win the championship. Sawyer has constructed the following payoff table (for the additional sweatshirts):

Event	Stock Additional Shirts	Don't Stock Shirts
Team wins	\$6,110	\$0
Team loses	\$1,500	\$0

What course of action should Ms. Sawyer take?

17-52 A local telephone distributor, Phones and More, plans to offer a special deal this week on its remote-activated answering machine. The store needs to decide how many "standard" and how many "remote" answering machines to order from the manufacturer. Based on prior experience, the management estimates the sales of the remote machine as given in the following table.

Sales	15	16	17	18	19	20	21
Probability	0.12	0.17	0.26	0.23	0.15	0.05	0.02

The retail price of the remote machine is \$89.95, but Phones and More's cost will be \$75.50. Use marginal analysis to determine the number of remote machines that the distributor should order.

17-53 Trade talks have broken down and there is a strong possibility of punitive tariffs being assessed on imported luxury cars. The owner of Motors is considering doubling the usual monthly import order; if tariffs are imposed, the firm will make a windfall profit on cars already in the country. But if tariffs are not imposed, the holding costs (chiefly interest on the company's line of credit) will reduce profit. The following table gives the owner's best estimate of the probabilities and payoffs:

Ordering Decision			
Event	Probability	Double	Don't Double
Tariffs imposed	0.15	\$240,000	\$100,000
No tariff	0.85	\$20,000	\$80,000

What should the owner do?

- 17-54** Technology stocks often show great price volatility, depending on whether Wall Street analysts perceive that the company's next product will be successful. At the end of the first quarter of 1994, an investment group considered its position in the stock of Digital Equipment Corporation (DEC), which was trading at \$31.50, down almost 50 percent from the group's cost basis.

The group had an investment horizon of January 1995, and debated whether to sell the stock. A consensus of expert opinion was that the most likely (expected) January 1995 price of DEC stock was \$35 per share, but it might drift lower (say to \$25). There was some hope that the stock could be trading as high as \$50, on the strength of the new Alpha chip, a fast proprietary design around which DEC was launching a new line of computers.

The investment group had substantial cash reserves on which they expected to earn 8 percent in the 9 months leading up to January 1995. Proceeds from selling the DEC stock could be added to these cash reserves.

In addition to holding the stock until January 1995 or selling it now and placing the proceeds into their cash reserves, the investors could reinvest the proceeds in LEAPs (long-term options) on DEC stock. A LEAP is the right to buy a stock in the future at a set price. In March 1994, the cost of a LEAP giving the right to buy one share of DEC stock for \$30 was \$6. This LEAP would expire in January of 1995. If the price of DEC stock at that time was above \$30, the investors would exercise the LEAP and then sell the DEC stock. If the price of DEC stock was below \$30, then the LEAP would expire, with no further value.

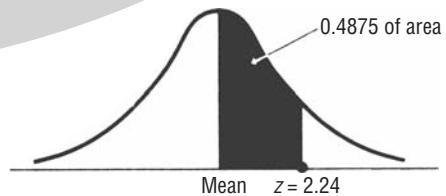
In the following, ignore tax consequences and assume that transaction fees are negligible because of the large number of shares involved. The investors have 100 shares of DEC stock, so if they sell them now at \$31.50 per share, they can use the proceeds of \$3,150 to buy LEAPs on 525 (= 3150/6) DEC shares.

- (a) How much will the investors have in January 1995, if they sell their stock now and place the proceeds into their cash reserves?
- (b) Suppose they estimate probabilities of 0.25, 0.50, and 0.25 that DEC stock will be selling for \$25, \$35, and \$50 in January 1995. How much will they expect to receive if they
 - i. hold their stock until January 1995 before selling?
 - ii. sell their stock now, buy LEAPs, and liquidate them (exercise them or let them expire) in January 1995?
- (c) What strategy do you recommend? Why?

This page is intentionally left blank.

Appendix Tables

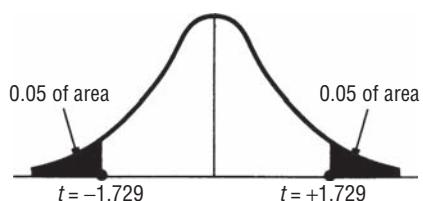
EXAMPLE: TO FIND THE AREA UNDER THE CURVE BETWEEN THE MEAN AND A POINT 2.24 STANDARD DEVIATIONS TO THE RIGHT OF THE MEAN, LOOK UP THE VALUE OPPOSITE 2.2 AND UNDER 0.04 IN THE TABLE; 0.4875 OF THE AREA UNDER THE CURVE LIES BETWEEN THE MEAN AND A z VALUE OF 2.24.



APPENDIX TABLE 1 AREAS UNDER THE STANDARD NORMAL PROBABILITY DISTRIBUTION BETWEEN THE MEAN AND POSITIVE VALUES OF z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

EXAMPLE: TO FIND THE VALUE OF t THAT CORRESPONDS TO AN AREA OF 0.10 IN BOTH TAILS OF THE DISTRIBUTION COMBINED, WHEN THERE ARE 19 DEGREES OF FREEDOM, LOOK UNDER THE 0.10 COLUMN, AND PROCEED DOWN TO THE 19 DEGREES OF FREEDOM ROW; THE APPROPRIATE t VALUE THERE IS 1.729.



APPENDIX TABLE 2 AREAS IN BOTH TAILS COMBINED FOR STUDENT'S t DISTRIBUTION

Degrees of Freedom	0.10	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	0.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
Normal Distribution	1.645	1.960	2.326	2.576

APPENDIX TABLE 3 BINOMIAL PROBABILITIES

FOR A GIVEN COMBINATION OF n AND p , ENTRY INDICATES THE PROBABILITY OF OBTAINING A SPECIFIED VALUE OF r . TO LOCATE ENTRY: WHEN $p \leq 0.50$, READ p ACROSS THE TOP AND BOTH n AND r DOWN THE LEFT MARGIN; WHEN $p \geq 0.50$, READ p ACROSS THE BOTTOM AND BOTH n AND r UP THE RIGHT MARGIN.

		P																			
n	r	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	r	n
2	0	0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281	0.8100	0.7921	0.7744	0.7569	0.7396	0.7225	0.7056	0.6889	0.6724	2	
1	0.0198	0.0392	0.0582	0.0768	0.0950	0.1128	0.1302	0.1472	0.1638	0.1800	0.1958	0.2112	0.2262	0.2408	0.2550	0.2688	0.2822	0.2952	1		
2	0.0001	0.0004	0.0009	0.0016	0.0025	0.0036	0.0049	0.0064	0.0081	0.0100	0.0121	0.0144	0.0169	0.0196	0.0225	0.0256	0.0289	0.0324	0		
3	0	0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536	0.7290	0.7050	0.6815	0.6585	0.6361	0.6141	0.5927	0.5718	0.5514	3	
1	0.0294	0.0576	0.0847	0.1106	0.1354	0.1590	0.1816	0.2031	0.2236	0.2430	0.2614	0.2788	0.2952	0.3106	0.3251	0.3387	0.3513	0.3631	2		
2	0.0003	0.0012	0.0026	0.0046	0.0071	0.0102	0.0137	0.0177	0.0221	0.0270	0.0323	0.0380	0.0441	0.0506	0.0574	0.0645	0.0720	0.0797	1		
3	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0007	0.0010	0.0013	0.0017	0.0022	0.0027	0.0034	0.0041	0.0049	0.0058	0		
4	0	0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857	0.6561	0.6274	0.5997	0.5729	0.5470	0.5220	0.4979	0.4746	0.4521	4	
1	0.0388	0.0753	0.1095	0.1416	0.1715	0.1993	0.2252	0.2492	0.2713	0.2916	0.3102	0.3271	0.3424	0.3562	0.3685	0.3793	0.3888	0.3970	3		
2	0.0006	0.0023	0.0051	0.0088	0.0135	0.0191	0.0254	0.0325	0.0402	0.0486	0.0575	0.0669	0.0767	0.0870	0.0975	0.1084	0.1195	0.1307	2		
3	0.0000	0.0000	0.0001	0.0002	0.0005	0.0008	0.0013	0.0019	0.0027	0.0036	0.0047	0.0061	0.0076	0.0094	0.0115	0.0138	0.0163	0.0191	1		
4	-	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0005	0.0007	0.0008	0.0010	0		
5	0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240	0.5905	0.5584	0.5277	0.4984	0.4704	0.4437	0.4182	0.3939	0.3707	5	
1	0.0480	0.0922	0.1328	0.1699	0.2036	0.2342	0.2618	0.2866	0.3086	0.3280	0.3451	0.3598	0.3724	0.3829	0.3915	0.3983	0.4034	0.4069	4		
2	0.0010	0.0038	0.0082	0.0142	0.0214	0.0299	0.0394	0.0498	0.0610	0.0729	0.0853	0.0981	0.1113	0.1247	0.1382	0.1517	0.1652	0.1786	3		
3	0.0000	0.0001	0.0003	0.0006	0.0011	0.0019	0.0030	0.0043	0.0060	0.0081	0.0105	0.0134	0.0166	0.0203	0.0244	0.0289	0.0338	0.0392	2		
4	-	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0007	0.0009	0.0012	0.0017	0.0022	0.0028	0.0035	0.0043	1		
5	-	-	-	-	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0			
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679	0.5314	0.4970	0.4644	0.4336	0.4046	0.3771	0.3513	0.3269	0.3040	6	
1	0.0571	0.1085	0.1546	0.1957	0.2321	0.2642	0.2922	0.3164	0.3370	0.3543	0.3685	0.3800	0.3888	0.3952	0.3993	0.4015	0.4018	0.4040	5		
2	0.0014	0.0055	0.0120	0.0204	0.0305	0.0422	0.0550	0.0688	0.0833	0.0984	0.1139	0.1295	0.1452	0.1608	0.1762	0.1912	0.2057	0.2197	4		
3	0.0000	0.0002	0.0005	0.0011	0.0021	0.0036	0.0055	0.0080	0.0110	0.0146	0.0188	0.0236	0.0289	0.0349	0.0415	0.0486	0.0562	0.0643	3		
4	-	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008	0.0012	0.0017	0.0024	0.0032	0.0043	0.0055	0.0069	0.0086	0.0106	2		
5	-	-	-	-	-	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0005	0.0007	0.0009	0			
6	-	-	-	-	-	-	-	-	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0			
7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168	0.4783	0.4423	0.4087	0.3773	0.3479	0.3206	0.2951	0.2714	0.2493	7	
1	0.0659	0.1240	0.1749	0.2192	0.2573	0.2897	0.3170	0.3396	0.3578	0.3720	0.3827	0.3901	0.3946	0.3965	0.3960	0.3935	0.3891	0.3830	6		
2	0.0020	0.0076	0.0162	0.0274	0.0406	0.0555	0.0716	0.0886	0.1061	0.1240	0.1419	0.1596	0.1769	0.1936	0.2097	0.2248	0.2391	0.2523	5		
3	0.0000	0.0003	0.0008	0.0019	0.0036	0.0059	0.0090	0.0128	0.0175	0.0230	0.0292	0.0363	0.0441	0.0525	0.0617	0.0714	0.0816	0.0923	4		
4	-	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0011	0.0017	0.0026	0.0036	0.0049	0.0066	0.0086	0.0109	0.0136	0.0167	0.0203	3		
5	-	-	-	-	-	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0006	0.0008	0.0012	0.0016	0.0021	0.0027	2		
6	-	-	-	-	-	-	-	-	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0			
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0000	0.0000	0			

<i>n</i>	<i>r</i>	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	<i>p</i>	<i>n</i>
12	0	0.8864	0.7847	0.6938	0.6127	0.5404	0.4759	0.4186	0.3677	0.3225	0.2824	0.2470	0.2157	0.1880	0.1637	0.1422	0.1234	0.1069	0.0924	12	
1	1	0.1074	0.1922	0.2575	0.3064	0.3413	0.3645	0.3781	0.3837	0.3827	0.3766	0.3663	0.3529	0.3372	0.3197	0.3012	0.2821	0.2627	0.2434	11	
2	0	0.0060	0.0216	0.0438	0.0702	0.0988	0.1280	0.1565	0.1835	0.2082	0.2301	0.2490	0.2647	0.2771	0.2863	0.2955	0.2960	0.2939	0.2939	10	
3	0	0.0002	0.0015	0.0045	0.0098	0.0173	0.0279	0.0393	0.0532	0.0686	0.0852	0.1026	0.1203	0.1380	0.1553	0.1720	0.1876	0.2021	0.2151	9	
4	0	0.0000	0.0003	0.0009	0.0021	0.0039	0.0067	0.0104	0.0153	0.0213	0.0285	0.0369	0.0464	0.0569	0.0683	0.0804	0.0931	0.1062	0.1062	8	
5	-	0.0000	0.0000	0.0001	0.0002	0.0004	0.0008	0.0014	0.0024	0.0038	0.0056	0.0081	0.0111	0.0148	0.0193	0.0245	0.0305	0.0373	0.0373	7	
6	-	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0008	0.0013	0.0019	0.0028	0.0040	0.0054	0.0073	0.0096	0.0096	0.0096	6	
7	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0004	0.0006	0.0009	0.0013	0.0018	0.0018	0.0018	0.0018	5	
8	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0002	0.0002	4	
9	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3	
10	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2	
11	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1	
12	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	12	
13	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	15	
14	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	15	
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430	0.2059	0.1741	0.1470	0.1238	0.1041	0.0874	0.0731	0.0611	0.0510	15	
1	1	0.1303	0.2261	0.2938	0.3636	0.3988	0.3658	0.3348	0.3148	0.2903	0.2734	0.2496	0.2273	0.2093	0.2297	0.2442	0.2312	0.2090	0.1878	14	
2	0	0.0092	0.0323	0.0636	0.0988	0.1348	0.1691	0.2003	0.2273	0.2496	0.2669	0.2793	0.2870	0.2903	0.2897	0.2787	0.2692	0.2578	0.2578	13	
3	0	0.0004	0.0029	0.0085	0.0178	0.0307	0.0468	0.0653	0.0857	0.1070	0.1285	0.1496	0.1696	0.1880	0.2044	0.2184	0.2300	0.2389	0.2452	12	
4	0	0.0000	0.0002	0.0008	0.0022	0.0049	0.0090	0.0148	0.0223	0.0317	0.0428	0.0555	0.0694	0.0843	0.0998	0.1156	0.1314	0.1468	0.1615	11	
5	-	0.0000	0.0001	0.0002	0.0006	0.0013	0.0024	0.0043	0.0069	0.0105	0.0151	0.0208	0.0277	0.0357	0.0449	0.0551	0.0662	0.0780	0.0810	10	
6	-	0.0000	0.0000	0.0000	0.0001	0.0003	0.0006	0.0011	0.0019	0.0031	0.0047	0.0069	0.0097	0.0132	0.0175	0.0226	0.0285	0.0285	0.0285	9	
7	-	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0008	0.0013	0.0020	0.0030	0.0043	0.0059	0.0081	0.0112	0.0118	0.0118	0.0118	8	
8	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7	
9	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6	
10	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	5	
11	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	4	
12	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3	
13	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2	
14	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1	
15	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	15	
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516	0.1216	0.0972	0.0776	0.0617	0.0490	0.0388	0.0306	0.0241	0.0189	20	
1	1	0.1652	0.2725	0.3364	0.3703	0.3526	0.3282	0.3000	0.2702	0.2403	0.2115	0.1844	0.1595	0.1368	0.1165	0.0986	0.0829	0.0730	0.0730	18	
2	0	0.0159	0.0528	0.0988	0.1458	0.1887	0.2246	0.2521	0.2711	0.2818	0.2852	0.2822	0.2740	0.2618	0.2466	0.2293	0.2109	0.1919	0.1730	18	
3	0	0.0010	0.0065	0.0183	0.0364	0.0596	0.0860	0.1139	0.1414	0.1672	0.1901	0.2093	0.2242	0.2347	0.2409	0.2428	0.2410	0.2358	0.2278	17	
4	0	0.0000	0.0006	0.0024	0.0065	0.0133	0.0233	0.0364	0.0523	0.0703	0.0898	0.1099	0.1299	0.1491	0.1666	0.1821	0.1951	0.2053	0.2125	16	
5	-	0.0000	0.0002	0.0009	0.0049	0.0099	0.0148	0.0222	0.0319	0.0435	0.0567	0.0713	0.0868	0.1028	0.1189	0.1345	0.1493	0.15	15		
6	-	0.0000	0.0001	0.0003	0.0008	0.0017	0.0032	0.0055	0.0089	0.0134	0.0193	0.0266	0.0353	0.0454	0.0566	0.0689	0.0819	0.0914	0.0914	14	
7	-	0.0000	0.0000	0.0001	0.0002	0.0005	0.0011	0.0020	0.0033	0.0053	0.0080	0.0115	0.0160	0.0216	0.0282	0.0360	0.0430	0.0500	0.0500	13	
8	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	12	
9	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	11	
10	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	10	
11	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	9	
12	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
13	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7	
14	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6	
15	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	5	
16	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	4	
17	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3	
18	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2	
19	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1	
20	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	

<i>n</i>	<i>r</i>	<i>p</i>	0.19	0.20	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.30	0.31	0.32	0.33	0.34	0.35	0.36	<i>r</i>	<i>n</i>	
2	0	0.6561	0.6400	0.6241	0.6084	0.5929	0.5776	0.5625	0.5476	0.5329	0.5184	0.5041	0.4900	0.4761	0.4624	0.4489	0.4356	0.4225	0.4096	0.2			
	1	0.3078	0.3200	0.3318	0.3432	0.3542	0.3648	0.3750	0.3848	0.3942	0.4032	0.4118	0.4200	0.4278	0.4352	0.4422	0.4488	0.4550	0.4608	1			
3	0	0.5314	0.5120	0.4930	0.4746	0.4565	0.4390	0.4219	0.4052	0.3890	0.3732	0.3579	0.3430	0.3285	0.3144	0.3008	0.2875	0.2746	0.2621	0.2			
	1	0.3740	0.3840	0.3932	0.4015	0.4091	0.4159	0.4219	0.4271	0.4316	0.4355	0.4386	0.4410	0.4428	0.4439	0.4444	0.4443	0.4443	0.4424	2			
4	0	0.0877	0.0960	0.1045	0.1133	0.1222	0.1313	0.1406	0.1501	0.1597	0.1693	0.1791	0.1890	0.1989	0.2089	0.2189	0.2289	0.2389	0.2488	1			
	3	0.0069	0.0080	0.0093	0.0106	0.0122	0.0138	0.0156	0.0176	0.0197	0.0220	0.0244	0.0270	0.0298	0.0328	0.0359	0.0393	0.0429	0.0467	0	3		
5	0	0.4305	0.4096	0.3895	0.3702	0.3515	0.3336	0.3164	0.2999	0.2840	0.2687	0.2541	0.2401	0.2267	0.2138	0.2015	0.1897	0.1785	0.1678	4			
	1	0.4039	0.4096	0.4142	0.4176	0.4200	0.4214	0.4219	0.4214	0.4201	0.4180	0.4152	0.4116	0.4074	0.4025	0.3970	0.3910	0.3845	0.3775	3			
6	0	0.1421	0.1536	0.1651	0.1767	0.1882	0.1996	0.2109	0.2221	0.2331	0.2439	0.2544	0.2646	0.2745	0.2841	0.2933	0.3021	0.3105	0.3185	2			
	3	0.0222	0.0256	0.0293	0.0332	0.0375	0.0420	0.0469	0.0520	0.0575	0.0632	0.0693	0.0756	0.0822	0.0891	0.0963	0.1038	0.1115	0.1194	1			
7	0	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039	0.0046	0.0053	0.0061	0.0071	0.0081	0.0092	0.0105	0.0119	0.0134	0.0150	0.0168	0	4		
	1	0.3487	0.3277	0.3077	0.2887	0.2707	0.2536	0.2373	0.2219	0.2073	0.1935	0.1804	0.1681	0.1564	0.1454	0.1350	0.1252	0.1160	0.1074	5			
8	0	1.0489	0.4096	0.4090	0.4090	0.4072	0.4043	0.4003	0.3955	0.3898	0.3834	0.3762	0.3685	0.3601	0.3513	0.3421	0.3325	0.3226	0.3124	0.3020	4		
	2	0.1919	0.2048	0.2174	0.2297	0.2415	0.2529	0.2637	0.2739	0.2836	0.2926	0.3010	0.3087	0.3157	0.3220	0.3275	0.3323	0.3364	0.3397	3			
9	0	0.0450	0.0512	0.0578	0.0648	0.0721	0.0798	0.0879	0.0962	0.1049	0.1138	0.1229	0.1323	0.1418	0.1515	0.1613	0.1712	0.1811	0.1911	2			
	1	0.0053	0.0064	0.0077	0.0091	0.0108	0.0126	0.0146	0.0169	0.0194	0.0221	0.0251	0.0283	0.0319	0.0357	0.0397	0.0441	0.0488	0.0537	1			
10	0	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008	0.0010	0.0012	0.0014	0.0017	0.0021	0.0024	0.0029	0.0034	0.0039	0.0045	0.0053	0.0060	0	5		
	1	0.2824	0.2621	0.2431	0.2252	0.2084	0.1927	0.1780	0.1642	0.1513	0.1393	0.1281	0.1176	0.1079	0.0989	0.0905	0.0827	0.0754	0.0687	6			
11	0	1.3975	0.3932	0.3877	0.3811	0.3735	0.3651	0.3560	0.3462	0.3358	0.3251	0.3139	0.3025	0.2909	0.2792	0.2673	0.2555	0.2437	0.2319	5			
	2	0.2331	0.2458	0.2577	0.2687	0.2789	0.2882	0.2966	0.3041	0.3105	0.3160	0.3206	0.3241	0.3267	0.3284	0.3292	0.3280	0.3261	0.3261	4			
12	0	3.0729	0.0819	0.0913	0.1011	0.1111	0.1214	0.1318	0.1424	0.1531	0.1639	0.1746	0.1852	0.1957	0.2061	0.2162	0.2260	0.2355	0.2446	3			
	1	0.0128	0.0154	0.0182	0.0214	0.0249	0.0287	0.0330	0.0375	0.0425	0.0478	0.0535	0.0595	0.0660	0.0727	0.0799	0.0873	0.0951	0.1032	2			
13	0	5.0012	0.0015	0.0019	0.0024	0.0030	0.0036	0.0044	0.0053	0.0063	0.0074	0.0087	0.0102	0.0119	0.0137	0.0157	0.0180	0.0205	0.0232	1			
	1	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0006	0.0007	0.0009	0.0011	0.0013	0.0015	0.0018	0.0022	0	6		
14	0	7.02288	0.2097	0.1920	0.1757	0.1605	0.1465	0.1335	0.1215	0.1105	0.1003	0.0910	0.0824	0.0745	0.0672	0.0606	0.0546	0.0490	0.0440	7			
	1	0.3756	0.3670	0.3573	0.3468	0.3356	0.3237	0.3115	0.2989	0.2860	0.2731	0.2600	0.2471	0.2342	0.2215	0.2090	0.1967	0.1848	0.1732	6			
15	2	0.2643	0.2753	0.2850	0.2935	0.3007	0.3067	0.3115	0.3150	0.3174	0.3186	0.3186	0.3177	0.3156	0.3127	0.3088	0.3040	0.2985	0.2922	5			
	3	0.1033	0.1147	0.1263	0.1379	0.1497	0.1614	0.1730	0.1845	0.1956	0.2065	0.2169	0.2269	0.2363	0.2452	0.2535	0.2610	0.2679	0.2740	4			
16	4	0.0242	0.0287	0.0336	0.0389	0.0447	0.0510	0.0577	0.0648	0.0724	0.0803	0.0886	0.0972	0.1062	0.1154	0.1248	0.1345	0.1442	0.1541	3			
	5	0.0034	0.0043	0.0054	0.0066	0.0080	0.0097	0.0115	0.0137	0.0161	0.0187	0.0217	0.0250	0.0286	0.0326	0.0369	0.0416	0.0466	0.0520	2			
17	6	0.0003	0.0004	0.0005	0.0006	0.0008	0.0010	0.0013	0.0016	0.0020	0.0024	0.0030	0.0036	0.0043	0.0051	0.0061	0.0071	0.0084	0.0098	1			
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0008	0	7			
<hr/>																							
<hr/>																							
<hr/>																							
<hr/>																							

<i>n</i>	<i>r</i>	<i>p</i>	0.19	0.20	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.30	0.31	0.32	0.33	0.34	0.35	0.36	<i>r</i>	<i>n</i>
8	0	0.1853	0.1678	0.1517	0.1370	0.1236	0.1113	0.1001	0.0899	0.0806	0.0722	0.0646	0.0576	0.0514	0.0457	0.0406	0.0360	0.0319	0.0281	0.0281	8	
1	0.3477	0.3355	0.3226	0.3092	0.2953	0.2812	0.2670	0.2527	0.2447	0.2347	0.2110	0.1977	0.1847	0.1721	0.1597	0.1484	0.1373	0.1267	0.1267	0.1267	7	
2	0.2855	0.2936	0.3002	0.3052	0.3108	0.3115	0.3108	0.3089	0.3058	0.3017	0.2965	0.2904	0.2835	0.2758	0.2675	0.2587	0.2494	0.2494	0.2494	0.2494	6	
3	0.1339	0.1468	0.1596	0.1722	0.1844	0.1963	0.2076	0.2184	0.2285	0.2379	0.2464	0.2541	0.2609	0.2668	0.2717	0.2756	0.2786	0.2805	0.2805	0.2805	5	
4	0.0393	0.0459	0.0530	0.0607	0.0689	0.0775	0.0865	0.0959	0.1056	0.1156	0.1258	0.1361	0.1465	0.1569	0.1673	0.1775	0.1875	0.1973	0.1973	0.1973	4	
5	0.0074	0.0092	0.0113	0.0137	0.0165	0.0196	0.0231	0.0270	0.0313	0.0360	0.0411	0.0467	0.0527	0.0591	0.0659	0.0732	0.0808	0.0888	0.0888	0.0888	3	
6	0.0009	0.0011	0.0015	0.0019	0.0025	0.0031	0.0038	0.0047	0.0058	0.0070	0.0084	0.0100	0.0118	0.0139	0.0162	0.0188	0.0217	0.0250	0.0250	0.0250	2	
7	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008	0.0010	0.0012	0.0015	0.0019	0.0023	0.0028	0.0033	0.0040	0.0040	0.0040	1	
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	
9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	
10	0	0.1501	0.1342	0.1199	0.1069	0.0952	0.0846	0.0751	0.0665	0.0589	0.0520	0.0458	0.0404	0.0355	0.0311	0.0272	0.0238	0.0207	0.0180	0.0180	9	
1	0.3169	0.3020	0.2867	0.2713	0.2558	0.2404	0.2253	0.2104	0.1960	0.1820	0.1685	0.1556	0.1433	0.1317	0.1206	0.1102	0.1004	0.0912	0.0912	8		
2	0.2973	0.3020	0.3049	0.3061	0.3056	0.3037	0.3003	0.2957	0.2899	0.2831	0.2754	0.2668	0.2576	0.2478	0.2376	0.2270	0.2162	0.2052	0.2052	7		
3	0.1627	0.1762	0.1891	0.2014	0.2130	0.2238	0.2336	0.2424	0.2502	0.2569	0.2624	0.2668	0.2701	0.2721	0.2731	0.2729	0.2716	0.2693	0.2693	6		
4	0.0573	0.0661	0.0754	0.0852	0.0954	0.1060	0.1168	0.1278	0.1388	0.1499	0.1608	0.1715	0.1820	0.1921	0.2017	0.2109	0.2194	0.2272	0.2272	5		
5	0.0134	0.0165	0.0200	0.0240	0.0285	0.0335	0.0389	0.0449	0.0513	0.0583	0.0657	0.0735	0.0818	0.0904	0.0994	0.1086	0.1181	0.1278	0.1278	4		
6	0.0021	0.0028	0.0036	0.0045	0.0057	0.0070	0.0087	0.0105	0.0127	0.0151	0.0179	0.0210	0.0245	0.0284	0.0326	0.0373	0.0424	0.0479	0.0479	3		
7	0.0002	0.0003	0.0004	0.0005	0.0007	0.0010	0.0012	0.0016	0.0020	0.0025	0.0031	0.0039	0.0047	0.0057	0.0069	0.0082	0.0098	0.0116	0.0116	2		
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003	0.0004	0.0005	0.0007	0.0008	0.0011	0.0013	0.0016	0.0016	1		
9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9		
10	0	0.1216	0.1074	0.0947	0.0834	0.0733	0.0643	0.0563	0.0492	0.0430	0.0374	0.0326	0.0282	0.0245	0.0211	0.0182	0.0157	0.0135	0.0115	0.0115	10	
1	0.2852	0.2684	0.2517	0.2351	0.2188	0.2030	0.1877	0.1730	0.1590	0.1456	0.1330	0.1211	0.1099	0.0995	0.0898	0.0808	0.0725	0.0649	0.0649	9		
2	0.3010	0.3020	0.3011	0.2984	0.2942	0.2885	0.2816	0.2735	0.2646	0.2548	0.2444	0.2335	0.2235	0.2107	0.1990	0.1873	0.1757	0.1642	0.1642	8		
3	0.1883	0.2013	0.2134	0.2244	0.2343	0.2429	0.2563	0.2609	0.2642	0.2662	0.2668	0.2662	0.2664	0.2664	0.2664	0.2573	0.2522	0.2462	0.2462	7		
4	0.0773	0.0881	0.0993	0.1108	0.1225	0.1343	0.1460	0.1576	0.1689	0.1798	0.1903	0.2001	0.2093	0.2177	0.2253	0.2320	0.2377	0.2424	0.2424	6		
5	0.0218	0.0264	0.0317	0.0375	0.0439	0.0509	0.0584	0.0664	0.0750	0.0839	0.0933	0.1029	0.1128	0.1229	0.1332	0.1434	0.1536	0.1636	0.1636	5		
6	0.0043	0.0055	0.0070	0.0088	0.0109	0.0134	0.0162	0.0195	0.0231	0.0272	0.0317	0.0368	0.0422	0.0482	0.0547	0.0616	0.0689	0.0767	0.0767	4		
7	0.0006	0.0008	0.0011	0.0014	0.0019	0.0024	0.0031	0.0039	0.0049	0.0060	0.0074	0.0090	0.0108	0.0130	0.0154	0.0181	0.0212	0.0247	0.0247	3		
8	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0007	0.0009	0.0011	0.0014	0.0018	0.0023	0.0028	0.0035	0.0043	0.0052	0.0052	0.0052	2		
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0006	1		
10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10		
12	0	0.0798	0.0687	0.0591	0.0507	0.0434	0.0371	0.0317	0.0270	0.0229	0.0194	0.0164	0.0138	0.0116	0.0098	0.0082	0.0068	0.0057	0.0047	0.0047	12	
1	0.2245	0.2062	0.1885	0.1717	0.1557	0.1407	0.1267	0.1137	0.1016	0.0906	0.0804	0.0712	0.0628	0.0552	0.0484	0.0422	0.0368	0.0319	0.0319	11		
2	0.2897	0.2835	0.2756	0.2663	0.2558	0.2444	0.2393	0.2197	0.2068	0.1937	0.1807	0.1678	0.1552	0.1429	0.1310	0.1197	0.1088	0.0986	0.0986	10		
3	0.2265	0.2362	0.2442	0.2503	0.2547	0.2573	0.2581	0.2573	0.2549	0.2511	0.2460	0.2397	0.2324	0.2241	0.2151	0.2055	0.1954	0.1849	0.1849	9		
4	0.1195	0.1329	0.1460	0.1589	0.1712	0.1828	0.1936	0.2034	0.2122	0.2197	0.2261	0.231	0.2349	0.2373	0.2384	0.2367	0.2340	0.2340	0.2340	8		
5	0.0449	0.0532	0.0621	0.0717	0.0818	0.0924	0.1032	0.1143	0.1255	0.1367	0.1477	0.1585	0.1688	0.1787	0.1879	0.1963	0.2039	0.2106	0.2106	7		
6	0.0123	0.0155	0.0193	0.0236	0.0285	0.0340	0.0401	0.0469	0.0542	0.0620	0.0704	0.0792	0.0885	0.0981	0.1079	0.1180	0.1281	0.1382	0.1382	6		
7	0.0025	0.0033	0.0044	0.0057	0.0073	0.0092	0.0115	0.0141	0.0172	0.0207	0.0246	0.0291	0.0341	0.0396	0.0456	0.0521	0.0591	0.0666	0.0666	5		
8	0.0004	0.0005	0.0007	0.0010	0.0014	0.0018	0.0024	0.0031	0.0040	0.0050	0.0063	0.0078	0.0096	0.0116	0.0140	0.0168	0.0199	0.0234	0.0234	4		
9	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0005	0.0007	0.0009	0.0011	0.0015	0.0019	0.0024	0.0031	0.0038	0.0048	0.0059	0.0059	0.0059	3		
10	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0006	0.0006	0.0010	0.0010	2		
11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1		
12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12		

<i>n</i>	<i>r</i>	0.37	0.38	0.39	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50	<i>r</i>	<i>n</i>	
<i>P</i>																		
2	0	0.3969	0.3844	0.3721	0.3600	0.3481	0.3364	0.3249	0.3136	0.3025	0.2916	0.2809	0.2704	0.2601	0.2500	2		
1	1	0.4662	0.4712	0.4758	0.4800	0.4838	0.4872	0.4902	0.4928	0.4950	0.4968	0.4982	0.4992	0.4998	0.5000	1		
2	2	0.1369	0.1444	0.1521	0.1600	0.1681	0.1764	0.1849	0.1936	0.2025	0.2116	0.2209	0.2304	0.2401	0.2500	0	2	
3	0	0.2500	0.2383	0.2270	0.2160	0.2054	0.1951	0.1852	0.1756	0.1664	0.1575	0.1489	0.1406	0.1327	0.1250	3		
1	1	0.4406	0.4382	0.4354	0.4320	0.4282	0.4239	0.4191	0.4140	0.4084	0.4024	0.3961	0.3894	0.3823	0.3750	2		
2	2	0.2587	0.2686	0.2783	0.2880	0.2975	0.3069	0.3162	0.3252	0.3341	0.3428	0.3512	0.3594	0.3674	0.3750	1		
3	3	0.0507	0.0549	0.0593	0.0640	0.0689	0.0741	0.0795	0.0852	0.0911	0.0973	0.1038	0.1106	0.1176	0.1250	0	3	
4	0	0.1575	0.1478	0.1385	0.1296	0.1212	0.1132	0.1056	0.0983	0.0915	0.0850	0.0789	0.0731	0.0677	0.0625	4		
1	1	0.3701	0.3623	0.3541	0.3456	0.3368	0.3278	0.3185	0.3091	0.2995	0.2897	0.2799	0.2700	0.2600	0.2500	3		
2	2	0.3260	0.3330	0.3396	0.3456	0.3511	0.3560	0.3604	0.3643	0.3675	0.3702	0.3723	0.3738	0.3747	0.3750	2		
3	3	0.1276	0.1361	0.1447	0.1536	0.1627	0.1719	0.1813	0.1908	0.2005	0.2102	0.2201	0.2300	0.2400	0.2500	1		
4	4	0.0187	0.0209	0.0231	0.0256	0.0283	0.0311	0.0342	0.0375	0.0410	0.0448	0.0488	0.0531	0.0576	0.0625	0	4	
5	0	0.0992	0.0916	0.0845	0.0778	0.0715	0.0656	0.0602	0.0551	0.0503	0.0459	0.0418	0.0380	0.0345	0.0312	5		
1	1	0.2914	0.2808	0.2700	0.2592	0.2484	0.2376	0.2270	0.2164	0.2059	0.1956	0.1854	0.1755	0.1657	0.1562	4		
2	2	0.3423	0.3441	0.3452	0.3456	0.3452	0.3442	0.3424	0.3400	0.3369	0.3332	0.3289	0.3240	0.3185	0.3125	3		
3	3	0.2010	0.2109	0.2207	0.2304	0.2399	0.2492	0.2583	0.2671	0.2757	0.2838	0.2916	0.2990	0.3060	0.3125	2		
4	4	0.0590	0.0646	0.0706	0.0768	0.0834	0.0902	0.0974	0.1049	0.1128	0.1209	0.1293	0.1380	0.1470	0.1562	1		
5	5	0.0069	0.0079	0.0090	0.0102	0.0116	0.0131	0.0147	0.0165	0.0185	0.0206	0.0229	0.0255	0.0282	0.0312	0	5	
6	0	0.0625	0.0568	0.0515	0.0467	0.0422	0.0381	0.0343	0.0308	0.0277	0.0248	0.0222	0.0198	0.0176	0.0156	6		
1	1	0.2203	0.2089	0.1976	0.1866	0.1759	0.1654	0.1552	0.1454	0.1359	0.1267	0.1179	0.1095	0.1014	0.0937	5		
2	2	0.3235	0.3201	0.3159	0.3110	0.3055	0.2994	0.2928	0.2856	0.2780	0.2699	0.2615	0.2527	0.2436	0.2344	4		
3	3	0.2533	0.2616	0.2693	0.2765	0.2831	0.2891	0.2945	0.2992	0.3032	0.3065	0.3091	0.3110	0.3121	0.3125	3		
4	4	0.1116	0.1202	0.1291	0.1382	0.1475	0.1570	0.1666	0.1763	0.1861	0.1958	0.2056	0.2153	0.2249	0.2344	2		
5	5	0.0262	0.0295	0.0330	0.0369	0.0410	0.0455	0.0503	0.0554	0.0609	0.0667	0.0729	0.0795	0.0864	0.0937	1		
6	6	0.0026	0.0030	0.0035	0.0041	0.0048	0.0055	0.0063	0.0073	0.0083	0.0095	0.0108	0.0122	0.0138	0.0156	0	6	
7	7	0	0.0394	0.0352	0.0314	0.0280	0.0249	0.0221	0.0195	0.0173	0.0152	0.0134	0.0117	0.0103	0.0090	0.0078	7	
1	1	0.1619	0.1511	0.1407	0.1306	0.1211	0.1119	0.1032	0.0950	0.0872	0.0798	0.0729	0.0664	0.0604	0.0547	6		
2	2	0.2833	0.2778	0.2698	0.2613	0.2524	0.2431	0.2336	0.2239	0.2140	0.2040	0.1940	0.1840	0.1740	0.1641	5		
3	3	0.2793	0.2838	0.2875	0.2903	0.2923	0.2934	0.2937	0.2932	0.2918	0.2897	0.2867	0.2830	0.2786	0.2734	4		
4	4	0.1640	0.1739	0.1838	0.1935	0.2031	0.2125	0.2216	0.2304	0.2388	0.2468	0.2543	0.2612	0.2676	0.2734	3		
5	5	0.0578	0.0640	0.0705	0.0774	0.0847	0.0923	0.1003	0.1086	0.1172	0.1261	0.1353	0.1447	0.1543	0.1641	2		
6	6	0.0113	0.0131	0.0150	0.0172	0.0196	0.0223	0.0252	0.0284	0.0320	0.0358	0.0400	0.0445	0.0494	0.0547	1		
7	7	0.0009	0.0011	0.0014	0.0016	0.0019	0.0023	0.0027	0.0032	0.0037	0.0044	0.0051	0.0059	0.0068	0.0078	0	7	

<i>n</i>	<i>r</i>	<i>p</i>															
		0.37	0.38	0.39	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50	<i>r</i>	<i>n</i>
8	0	0.0248	0.0218	0.0192	0.0168	0.0147	0.0128	0.0111	0.0097	0.0084	0.0072	0.0062	0.0053	0.0046	0.0039	8	
1	0.1166	0.1071	0.0981	0.0896	0.0816	0.0742	0.0672	0.0608	0.0548	0.0493	0.0442	0.0395	0.0325	0.0312	7		
2	0.2397	0.2297	0.2194	0.2090	0.1985	0.1880	0.1776	0.1672	0.1569	0.1469	0.1371	0.1275	0.1183	0.1094	6		
3	0.2815	0.2815	0.2806	0.2787	0.2759	0.2723	0.2679	0.2627	0.2568	0.2503	0.2431	0.2355	0.2273	0.2187	5		
4	0.2067	0.2157	0.2242	0.2322	0.2397	0.2465	0.2526	0.2580	0.2627	0.2665	0.2695	0.2717	0.2730	0.2754	4		
5	0.0971	0.1058	0.1147	0.1239	0.1332	0.1428	0.1525	0.1622	0.1719	0.1816	0.1912	0.2006	0.2098	0.2187	3		
6	0.0285	0.0324	0.0367	0.0413	0.0463	0.0517	0.0575	0.0637	0.0703	0.0774	0.0848	0.0926	0.1008	0.1094	2		
7	0.0048	0.0057	0.0067	0.0079	0.0092	0.0107	0.0124	0.0143	0.0164	0.0188	0.0215	0.0244	0.0277	0.0312	1		
8	0.0004	0.0004	0.0005	0.0007	0.0008	0.0010	0.0012	0.0014	0.0017	0.0020	0.0024	0.0028	0.0033	0.0039	0		
9	0	0.0156	0.0135	0.0117	0.0101	0.0087	0.0074	0.0064	0.0054	0.0046	0.0039	0.0033	0.0028	0.0023	0.0020	9	
1	0.0826	0.0747	0.0673	0.0605	0.0542	0.0484	0.0431	0.0383	0.0339	0.0299	0.0263	0.0231	0.0202	0.0176	8		
2	0.1941	0.1831	0.1721	0.1612	0.1506	0.1402	0.1301	0.1204	0.1110	0.1020	0.0934	0.0853	0.0776	0.0703	7		
3	0.2660	0.2618	0.2567	0.2508	0.2442	0.2369	0.2291	0.2207	0.2119	0.2027	0.1933	0.1837	0.1739	0.1641	6		
4	0.2344	0.2407	0.2462	0.2508	0.2545	0.2573	0.2592	0.2601	0.2600	0.2590	0.2571	0.2543	0.2506	0.2461	5		
5	0.1376	0.1475	0.1574	0.1672	0.1769	0.1863	0.1955	0.2044	0.2128	0.2207	0.2280	0.2347	0.2408	0.2641	4		
6	0.0539	0.0603	0.0671	0.0743	0.0819	0.0900	0.0983	0.1070	0.1160	0.1253	0.1348	0.1445	0.1542	0.1641	3		
7	0.0136	0.0158	0.0184	0.0212	0.0244	0.0279	0.0318	0.0360	0.0407	0.0458	0.0512	0.0571	0.0635	0.0703	2		
8	0.0020	0.0024	0.0029	0.0035	0.0042	0.0051	0.0060	0.0071	0.0083	0.0097	0.0114	0.0132	0.0153	0.0176	1		
9	0.0001	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0008	0.0009	0.0011	0.0014	0.0016	0.0020	0	9		
10	0	0.0098	0.0084	0.0071	0.0060	0.0051	0.0043	0.0036	0.0030	0.0025	0.0021	0.0017	0.0014	0.0010	0.0010	10	
1	0.0578	0.0514	0.0456	0.0403	0.0355	0.0312	0.0273	0.0238	0.0207	0.0180	0.0155	0.0133	0.0114	0.0098	9		
2	0.1529	0.1419	0.1312	0.1209	0.1111	0.1017	0.0927	0.0843	0.0763	0.0688	0.0619	0.0554	0.0494	0.0439	8		
3	0.2394	0.2319	0.2237	0.2150	0.2058	0.1963	0.1865	0.1765	0.1665	0.1564	0.1464	0.1364	0.1267	0.1172	7		
4	0.2461	0.2487	0.2503	0.2508	0.2503	0.2488	0.2462	0.2427	0.2384	0.2331	0.2271	0.2204	0.2130	0.2051	6		
5	0.1734	0.1829	0.1920	0.2007	0.2087	0.2162	0.2229	0.2289	0.2340	0.2383	0.2417	0.2441	0.2456	0.2461	5		
6	0.0849	0.0934	0.1023	0.1115	0.1209	0.1304	0.1401	0.1499	0.1596	0.1692	0.1786	0.1878	0.1966	0.2051	4		
7	0.0285	0.0327	0.0374	0.0425	0.0480	0.0540	0.0604	0.0673	0.0746	0.0824	0.0905	0.0991	0.1080	0.1172	3		
8	0.0063	0.0075	0.0090	0.0106	0.0125	0.0147	0.0171	0.0198	0.0229	0.0263	0.0301	0.0343	0.0389	0.0439	2		
9	0.0008	0.0010	0.0013	0.0016	0.0019	0.0024	0.0035	0.0042	0.0050	0.0059	0.0070	0.0083	0.0098	0.0116	1		
10	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0008	0.0010	0		

<i>n</i>	<i>r</i>	<i>p</i>	<i>n</i>
12	0	0.0039	0.37
	1	0.0276	0.38
2	0.0890	0.0032	0.39
3	0.1742	0.0237	0.40
4	0.2302	0.0800	0.41
5	0.2163	0.0204	0.42
6	0.1482	0.0639	0.43
7	0.0746	0.1526	0.44
8	0.0274	0.2254	0.45
9	0.0071	0.2210	0.46
10	0.0013	0.0086	0.47
11	0.0001	0.0002	0.48
12	0.0000	0.0000	0.49
15	0	0.0010	0.50
	1	0.0086	0.51
2	0.0354	0.0071	0.52
3	0.0904	0.0259	0.53
4	0.1587	0.0716	0.54
5	0.2051	0.1374	0.55
6	0.2008	0.1933	0.56
7	0.1516	0.2040	0.57
8	0.0890	0.1608	0.58
9	0.0407	0.1082	0.59
10	0.0143	0.0538	0.60
11	0.0038	0.0206	0.61
12	0.0007	0.0010	0.62
13	0.0001	0.0002	0.63
14	0.0000	0.0000	0.64
15	—	—	0.65
20	0	0.0001	0.66
	1	0.0011	0.67
2	0.0064	0.0050	0.68
3	0.0224	0.0185	0.69
4	0.0559	0.0482	0.70
5	0.1051	0.0945	0.71
6	0.1543	0.1447	0.72
7	0.1812	0.1774	0.73
8	0.1730	0.1767	0.74
9	0.1354	0.1444	0.75
10	0.0875	0.0974	0.76
11	0.0467	0.0542	0.77
12	0.0206	0.0249	0.78
13	0.0074	0.0094	0.79
14	0.0022	0.0029	0.80
15	0.0005	0.0007	0.81
16	0.0001	0.0002	0.82
17	0.0000	0.0000	0.83
18	—	—	0.84
19	—	—	0.85
20	—	—	0.86

APPENDIX TABLE 4(a) VALUES OF $e^{-\lambda}$ FOR COMPUTING POISSON PROBABILITIES

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0.1	0.90484	2.6	0.07427	5.1	0.00610	7.6	0.00050
0.2	0.81873	2.7	0.06721	5.2	0.00552	7.7	0.00045
0.3	0.74082	2.8	0.06081	5.3	0.00499	7.8	0.00041
0.4	0.67032	2.9	0.05502	5.4	0.00452	7.9	0.00037
0.5	0.60653	3.0	0.04979	5.5	0.00409	8.0	0.00034
0.6	0.54881	3.1	0.04505	5.6	0.00370	8.1	0.00030
0.7	0.49659	3.2	0.04076	5.7	0.00335	8.2	0.00027
0.8	0.44933	3.3	0.03688	5.8	0.00303	8.3	0.00025
0.9	0.40657	3.4	0.03337	5.9	0.00274	8.4	0.00022
1.0	0.36788	3.5	0.03020	6.0	0.00248	8.5	0.00020
1.1	0.33287	3.6	0.02732	6.1	0.00224	8.6	0.00018
1.2	0.30119	3.7	0.02472	6.2	0.00203	8.7	0.00017
1.3	0.27253	3.8	0.02237	6.3	0.00184	8.8	0.00015
1.4	0.24660	3.9	0.02024	6.4	0.00166	8.9	0.00014
1.5	0.22313	4.0	0.01832	6.5	0.00150	9.0	0.00012
1.6	0.20190	4.1	0.01657	6.6	0.00136	9.1	0.00011
1.7	0.18268	4.2	0.01500	6.7	0.00123	9.2	0.00010
1.8	0.16530	4.3	0.01357	6.8	0.00111	9.3	0.00009
1.9	0.14957	4.4	0.01228	6.9	0.00101	9.4	0.00008
2.0	0.13534	4.5	0.01111	7.0	0.00091	9.5	0.00007
2.1	0.12246	4.6	0.01005	7.1	0.00083	9.6	0.00007
2.2	0.11080	4.7	0.00910	7.2	0.00075	9.7	0.00006
2.3	0.10026	4.8	0.00823	7.3	0.00068	9.8	0.00006
2.4	0.09072	4.9	0.00745	7.4	0.00061	9.9	0.00005
2.5	0.08208	5.0	0.00674	7.5	0.00055	10.0	0.00005

APPENDIX TABLE 4(b) DIRECT VALUES FOR DETERMINING POISSON PROBABILITIES

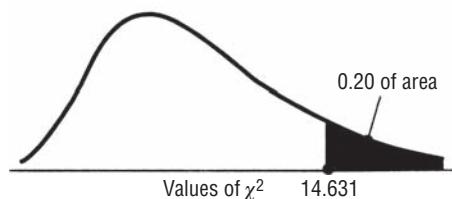
FOR A GIVEN VALUE OF λ , ENTRY INDICATES THE PROBABILITY OF OBTAINING A SPECIFIED VALUE OF X.

X	λ									
X	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.0984	0.0915	0.0850	0.0789	0.0733
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615	0.1539	0.1465
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046	0.2001	0.1954
4	0.1734	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944	0.1951	0.1954
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477	0.1522	0.1563
6	0.0555	0.0608	0.0662	0.0716	0.0771	0.0826	0.0881	0.0936	0.0989	0.1042
7	0.0246	0.0278	0.0312	0.0348	0.0385	0.0425	0.0466	0.0508	0.0551	0.0595
8	0.0095	0.0111	0.0129	0.0148	0.0169	0.0191	0.0215	0.0241	0.0269	0.0298
9	0.0033	0.0040	0.0047	0.0056	0.0066	0.0076	0.0089	0.0102	0.0116	0.0132
10	0.0010	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039	0.0045	0.0053
11	0.0003	0.0004	0.0005	0.0006	0.0007	0.0009	0.0011	0.0013	0.0016	0.0019
12	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006
13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
X	λ									
X	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0679	0.0630	0.0583	0.0540	0.0500	0.0462	0.0427	0.0395	0.0365	0.0337
2	0.1393	0.1323	0.1254	0.1188	0.1125	0.1063	0.1005	0.0948	0.0894	0.0842
3	0.1904	0.1852	0.1798	0.1743	0.1687	0.1631	0.1574	0.1517	0.1460	0.1404
4	0.1951	0.1944	0.1933	0.1917	0.1898	0.1875	0.1849	0.1820	0.1789	0.1755
5	0.1600	0.1633	0.1662	0.1687	0.1708	0.1725	0.1738	0.1747	0.1753	0.1755
6	0.1093	0.1143	0.1191	0.1237	0.1281	0.1323	0.1362	0.1398	0.1432	0.1462
7	0.0640	0.0686	0.0732	0.0778	0.0824	0.0869	0.0914	0.0959	0.1022	0.1044
8	0.0328	0.0360	0.0393	0.0428	0.0463	0.0500	0.0537	0.0575	0.0614	0.0653
9	0.0150	0.0168	0.0188	0.0209	0.0232	0.0255	0.0280	0.0307	0.0334	0.0363
10	0.0061	0.0071	0.0081	0.0092	0.0104	0.0118	0.0132	0.0147	0.0164	0.0181
11	0.0023	0.0027	0.0032	0.0037	0.0043	0.0049	0.0056	0.0064	0.0073	0.0082
12	0.0008	0.0009	0.0011	0.0014	0.0016	0.0019	0.0022	0.0026	0.0030	0.0034
13	0.0002	0.0003	0.0004	0.0004	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013
14	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005
15	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
X	λ									
X	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0311	0.0287	0.0265	0.0244	0.0225	0.0207	0.0191	0.0176	0.0162	0.0149
2	0.0793	0.0746	0.0701	0.0659	0.0618	0.0580	0.0544	0.0509	0.0477	0.0446
3	0.1348	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.0985	0.0938	0.0892
4	0.1719	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339
5	0.1753	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606
6	0.1490	0.1515	0.1537	0.1555	0.1571	0.1584	0.1594	0.1601	0.1605	0.1606
7	0.1086	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377
8	0.0692	0.0731	0.0771	0.0810	0.0849	0.0887	0.0925	0.0962	0.0998	0.1033
9	0.0392	0.0423	0.0454	0.0486	0.0519	0.0552	0.0586	0.0620	0.0654	0.0688
10	0.0200	0.0220	0.0241	0.0262	0.0285	0.0309	0.0334	0.0359	0.0386	0.0413
11	0.0093	0.0104	0.0116	0.0129	0.0143	0.0157	0.0173	0.0190	0.0207	0.0225
12	0.0039	0.0045	0.0051	0.0058	0.0065	0.0073	0.0082	0.0092	0.0102	0.0113
13	0.0015	0.0018	0.0021	0.0024	0.0028	0.0032	0.0036	0.0041	0.0046	0.0052
14	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013	0.0015	0.0017	0.0019	0.0022
15	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009
16	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0003
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001

X	λ									
	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0137	0.0126	0.0116	0.0106	0.0098	0.0090	0.0082	0.0076	0.0070	0.0064
2	0.0417	0.0390	0.0364	0.0340	0.0318	0.0296	0.0276	0.0258	0.0240	0.0223
3	0.0848	0.0806	0.0765	0.0726	0.0688	0.0652	0.0617	0.0584	0.0552	0.0521
4	0.1294	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.0992	0.0952	0.0912
5	0.1579	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277
6	0.1605	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490
7	0.1399	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490
8	0.1066	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304
9	0.0723	0.0757	0.0791	0.0825	0.0858	0.0891	0.0923	0.0954	0.0985	0.1014
10	0.0441	0.0469	0.0498	0.0528	0.0558	0.0588	0.0618	0.0649	0.0679	0.0710
11	0.0245	0.0265	0.0285	0.0307	0.0330	0.0353	0.0377	0.0401	0.0426	0.0452
12	0.0124	0.0137	0.0150	0.0164	0.0179	0.0194	0.0210	0.0227	0.0245	0.0264
13	0.0058	0.0065	0.0073	0.0081	0.0089	0.0098	0.0108	0.0119	0.0130	0.0142
14	0.0025	0.0029	0.0033	0.0037	0.0041	0.0046	0.0052	0.0058	0.0064	0.0071
15	0.0010	0.0012	0.0014	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0033
16	0.0004	0.0005	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0013	0.0014
17	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006
18	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001
X	λ									
	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0059	0.0054	0.0049	0.0045	0.0041	0.0038	0.0035	0.0032	0.0029	0.0027
2	0.0208	0.0194	0.0180	0.0167	0.0156	0.0145	0.0134	0.0125	0.0116	0.0107
3	0.0492	0.0464	0.0438	0.0413	0.0389	0.0366	0.0345	0.0324	0.0305	0.0286
4	0.0874	0.0836	0.0799	0.0764	0.0729	0.0696	0.0663	0.0632	0.0602	0.0573
5	0.1241	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.0986	0.0951	0.0916
6	0.1468	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221
7	0.1489	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396
8	0.1321	0.1337	0.1351	0.1363	0.1373	0.1382	0.1388	0.1392	0.1395	0.1396
9	0.1042	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241
10	0.0740	0.0770	0.0800	0.0829	0.0858	0.0887	0.0914	0.0941	0.0967	0.0993
11	0.0478	0.0504	0.0531	0.0558	0.0585	0.0613	0.0640	0.0667	0.0695	0.0722
12	0.0283	0.0303	0.0323	0.0344	0.0366	0.0388	0.0411	0.0434	0.0457	0.0481
13	0.0154	0.0168	0.0181	0.0196	0.0211	0.0227	0.0243	0.0260	0.0278	0.0296
14	0.0078	0.0086	0.0095	0.0104	0.0113	0.0123	0.0134	0.0145	0.0157	0.0169
15	0.0037	0.0041	0.0046	0.0051	0.0057	0.0062	0.0069	0.0075	0.0083	0.0090
16	0.0016	0.0019	0.0021	0.0024	0.0026	0.0030	0.0033	0.0037	0.0041	0.0045
17	0.0007	0.0008	0.0009	0.0010	0.0012	0.0013	0.0015	0.0017	0.0019	0.0021
18	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
19	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0004
20	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

X	λ									
	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0025	0.0023	0.0021	0.0019	0.0017	0.0016	0.0014	0.0013	0.0012	0.0011
2	0.0100	0.0092	0.0086	0.0079	0.0074	0.0068	0.0063	0.0058	0.0054	0.0050
3	0.0269	0.0252	0.0237	0.0222	0.0208	0.0195	0.0183	0.0171	0.0160	0.0150
4	0.0544	0.0517	0.0491	0.0466	0.0443	0.0420	0.0398	0.0377	0.0357	0.0337
5	0.0882	0.0849	0.0816	0.0784	0.0752	0.0722	0.0692	0.0663	0.0635	0.0607
6	0.1191	0.1160	0.1128	0.1097	0.1066	0.1034	0.1003	0.0972	0.0941	0.0911
7	0.1378	0.1358	0.1338	0.1317	0.1294	0.1271	0.1247	0.1222	0.1197	0.1171
8	0.1395	0.1392	0.1388	0.1382	0.1375	0.1366	0.1356	0.1344	0.1332	0.1318
9	0.1256	0.1269	0.1280	0.1290	0.1299	0.1306	0.1311	0.1315	0.1317	0.1318
10	0.1017	0.1040	0.1063	0.1084	0.1104	0.1123	0.1140	0.1157	0.1172	0.1186
11	0.0749	0.0776	0.0802	0.0828	0.0853	0.0878	0.0902	0.0925	0.0948	0.0970
12	0.0505	0.0530	0.0555	0.0579	0.0604	0.0629	0.0654	0.0679	0.0703	0.0728
13	0.0315	0.0334	0.0354	0.0374	0.0395	0.0416	0.0438	0.0459	0.0481	0.0504
14	0.0182	0.0196	0.0210	0.0225	0.0240	0.0256	0.0272	0.0289	0.0306	0.0324
15	0.0098	0.0107	0.0116	0.0126	0.0136	0.0147	0.0158	0.0169	0.0182	0.0194
16	0.0050	0.0055	0.0060	0.0066	0.0072	0.0079	0.0086	0.0093	0.0101	0.0109
17	0.0024	0.0026	0.0029	0.0033	0.0036	0.0040	0.0044	0.0048	0.0053	0.0058
18	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019	0.0021	0.0024	0.0026	0.0029
19	0.0005	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014
20	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0005	0.0006
21	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
X	λ									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.0378
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

EXAMPLE: IN A CHI-SQUARE DISTRIBUTION WITH 11 DEGREES OF FREEDOM, TO FIND THE CHI-SQUARE VALUE FOR 0.20 OF THE AREA UNDER THE CURVE (THE COLORED AREA IN THE RIGHT TAIL) LOOK UNDER THE 0.20 COLUMN IN THE TABLE AND THE 11 DEGREES OF FREEDOM ROW; THE APPROPRIATE CHI-SQUARE VALUE IS 14.631.



APPENDIX TABLE 5 AREA IN THE
RIGHT TAIL OF A CHI-SQUARE (χ^2)
DISTRIBUTION

Degrees of Freedom	Area in Right Tail				
	0.99	0.975	0.95	0.90	0.800
1	0.00016	0.00098	0.00398	0.0158	0.0642
2	0.0201	0.0506	0.103	0.211	0.446
3	0.115	0.216	0.352	0.584	1.005
4	0.297	0.484	0.711	1.064	1.649
5	0.554	0.831	1.145	1.610	2.343
6	0.872	1.237	1.685	2.204	3.070
7	1.239	1.690	2.167	2.833	3.822
8	1.646	2.180	2.733	3.490	4.594
9	2.088	2.700	3.325	4.168	5.380
10	2.558	3.247	3.940	4.865	6.179
11	3.053	3.816	4.575	5.578	6.989
12	3.571	4.404	5.226	6.304	7.807
13	4.107	5.009	5.892	7.042	8.634
14	4.660	5.629	6.571	7.790	9.467
15	5.229	6.262	7.261	8.547	10.307
16	5.812	6.908	7.962	9.312	11.152
17	6.408	7.564	8.672	10.085	12.002
18	7.015	8.231	9.390	10.865	12.857
19	7.633	8.907	10.117	11.651	13.716
20	8.260	9.591	10.851	12.443	14.578
21	8.897	10.283	11.591	13.240	15.445
22	9.542	10.982	12.338	14.041	16.314
23	10.196	11.689	13.091	14.848	17.187
24	10.856	12.401	13.848	15.658	18.062
25	11.524	13.120	14.611	16.473	18.940
26	12.198	13.844	15.379	17.292	19.820
27	12.879	14.573	16.151	18.114	20.703
28	13.565	15.308	16.928	18.939	21.588
29	14.256	16.047	17.708	19.768	22.475
30	14.953	16.791	18.493	20.599	23.364

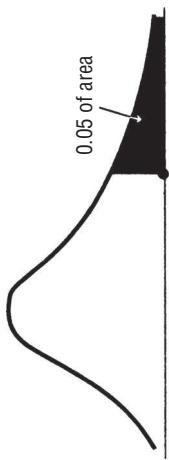
Note: If v , the number of degrees of freedom, is greater than 30, we can approximate χ^2_α , the chi-square value leaving α of the area the right tail, by

$$\chi^2_\alpha = v \left(1 - \frac{2}{9v} + z_\alpha \sqrt{\frac{2}{9v}} \right)^3$$

where z_α is the standard normal value (from Appendix Table 1) that leaves α of the area in the right tail.

	Area in Right Tail					Degrees of Freedom
0.20	0.10	0.05	0.025	0.01		
1.642	2.706	3.841	5.024	6.635		1
3.219	4.605	5.991	7.378	9.210		2
4.642	6.251	7.815	9.348	11.345		3
5.989	7.779	9.488	11.143	13.277		4
7.289	9.236	11.070	12.833	15.086		5
8.558	10.645	12.592	14.449	16.812		6
9.803	12.017	14.067	16.013	18.475		7
11.030	13.362	15.507	17.535	20.090		8
12.242	14.684	16.919	19.023	21.666		9
13.442	15.987	18.307	20.483	23.209		10
14.631	17.275	19.675	21.920	24.725		11
15.812	18.549	21.026	23.337	26.217		12
16.985	19.812	22.362	24.736	27.688		13
18.151	21.064	23.685	26.119	29.141		14
19.311	22.307	24.996	27.488	30.578		15
20.465	23.542	26.296	28.845	32.000		16
21.615	24.769	27.587	30.191	33.409		17
22.760	25.989	28.869	31.526	34.805		18
23.900	27.204	30.144	32.852	36.191		19
25.038	28.412	31.410	34.170	37.566		20
26.171	29.615	32.671	35.479	38.932		21
27.301	30.813	33.924	36.781	40.289		22
28.429	32.007	35.172	38.076	41.638		23
29.553	33.196	36.415	39.364	42.980		24
30.675	34.382	37.652	40.647	44.314		25
31.795	35.563	38.885	41.923	45.642		26
32.912	36.741	40.113	43.194	46.963		27
34.027	37.916	41.337	44.461	48.278		28
35.139	39.087	42.557	45.722	49.588		29
36.250	40.256	43.773	46.979	50.892		30

EXAMPLE: IN AN F DISTRIBUTION WITH 15 DEGREES OF FREEDOM FOR THE NUMERATOR AND 6 DEGREES OF FREEDOM FOR THE DENOMINATOR, TO FIND THE F VALUE FOR 0.05 OF THE AREA UNDER THE CURVE LOOK UNDER THE 15 DEGREES OF FREEDOM COLUMN AND ACROSS THE 6 DEGREES OF FREEDOM ROW; THE APPROPRIATE F VALUE IS 3.94.



3.94

APPENDIX TABLE 6(a)
VALUES OF F
FOR F DISTRIBUTIONS WITH 0.05 OF
THE AREA IN THE RIGHT TAIL

	Degrees of Freedom for Numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.72	5.69	5.66	5.63	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.70	3.67	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Degrees of Freedom for Denominator

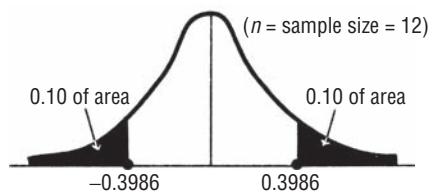


EXAMPLE: IN AN F DISTRIBUTION WITH 7 DEGREES OF FREEDOM FOR THE NUMERATOR AND 5 DEGREES OF FREEDOM FOR THE DENOMINATOR, TO FIND THE F VALUE FOR 0.01 OF THE AREA UNDER THE CURVE LOOK UNDER THE 7 DEGREES OF FREEDOM COLUMN AND ACROSS THE 5 DEGREES OF FREEDOM ROW; THE APPROPRIATE F VALUE IS 10.5.

APPENDIX TABLE 6(b) VALUES OF F FOR F DISTRIBUTIONS WITH 0.01 OF THE AREA IN THE RIGHT TAIL

		Degrees of Freedom for Numerator																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Degrees of Freedom for Denominator	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366	
2	98.5	99.0	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1	
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.2	14.0	13.9	13.8	13.7	13.6	13.5	13.5	13.5	
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.37	3.27	3.18	3.09	3.00	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.01		
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38	
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00	

Degrees of Freedom for Denominator



EXAMPLE: FOR A TWO-TAILED TEST OF SIGNIFICANCE AT THE 0.20 LEVEL, WITH $n = 12$, THE APPROPRIATE VALUE FOR r_s CAN BE FOUND BY LOOKING UNDER THE 0.20 COLUMN AND ACROSS THE 12 ROW; THE APPROPRIATE r_s VALUE IS 0.3986.

APPENDIX TABLE 7 VALUES FOR SPEARMAN'S RANK CORRELATION (r_s) FOR COMBINED AREAS IN BOTH TAILS

<i>n</i>	0.20	0.10	0.05	0.02	0.01	0.002
4	0.8000	0.8000				
5	0.7000	0.8000	0.9000	0.9000		
6	0.6000	0.7714	0.8286	0.8857	0.9429	
7	0.5357	0.6786	0.7450	0.8571	0.8929	0.9643
8	0.5000	0.6190	0.7143	0.8095	0.8571	0.9286
9	0.4667	0.5833	0.6833	0.7667	0.8167	0.9000
10	0.4424	0.5515	0.6364	0.7333	0.7818	0.8667
11	0.4182	0.5273	0.6091	0.7000	0.7455	0.8364
12	0.3986	0.4965	0.5804	0.6713	0.7273	0.8182
13	0.3791	0.4780	0.5549	0.6429	0.6978	0.7912
14	0.3626	0.4593	0.5341	0.6220	0.6747	0.7670
15	0.3500	0.4429	0.5179	0.6000	0.6536	0.7464
16	0.3382	0.4265	0.5000	0.5824	0.6324	0.7265
17	0.3260	0.4118	0.4853	0.5637	0.6152	0.7083
18	0.3148	0.3994	0.4716	0.5480	0.5975	0.6904
19	0.3070	0.3895	0.4579	0.5333	0.5825	0.6737
20	0.2977	0.3789	0.4451	0.5203	0.5684	0.6586
21	0.2909	0.3688	0.4351	0.5078	0.5545	0.6455
22	0.2829	0.3597	0.4241	0.4963	0.5426	0.6318
23	0.2767	0.3518	0.4150	0.4852	0.5306	0.6186
24	0.2704	0.3435	0.4061	0.4748	0.5200	0.6070
25	0.2646	0.3362	0.3977	0.4654	0.5100	0.5962
26	0.2588	0.3299	0.3894	0.4564	0.5002	0.5856
27	0.2540	0.3236	0.3822	0.4481	0.4915	0.5757
28	0.2490	0.3175	0.3749	0.4401	0.4828	0.5660
29	0.2443	0.3113	0.3685	0.4320	0.4744	0.5567
30	0.2400	0.3059	0.3620	0.4251	0.4665	0.5479

APPENDIX TABLE 8 CRITICAL VALUES OF D IN THE KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

Sample Size (n)	Level of Significance for $D = \text{Maximum } F_e - F_o $				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Over 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Note: The values of D given in the table are critical values associated with selected values of n . Any value of D that is greater than or equal to the tabulated value is significant at the indicated level of significance.

APPENDIX TABLE 9 CONTROL CHART FACTORS

Sample size, n	Factors for \bar{x} Charts		Factors for R Charts		
	$d_2 = \frac{R}{\sigma}$	$A_2 = \frac{3}{d_2\sqrt{n}}$	$d_3 = \frac{\sigma_R}{\sigma}$	$D_3 = 1 - \frac{3d_3}{d_2}$	$D_4 = 1 + \frac{3d_3}{d_2}$
2	1.128	1.881	0.853	0	3.269
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.779	0.283	1.717
13	3.336	0.249	0.770	0.308	1.692
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.609
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.414	1.586
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.452	1.548
25	3.931	0.153	0.708	0.460	1.540

Note: If $1 - 3d_3/d_2 < 0$, then $D_3 = 0$.

Bibliography

Data Analysis and Presentation

CLEVELAND, W. S., *The Elements of Graphing Data*, rev. ed., Murray Hill, NJ, AT&T Bell Laboratories, 1994.

EVERITT, B. S., AND G. DUNN. *Advanced Methods of Data Exploration and Modelling*, London, Heinemann Education Books, Ltd., 1983.

TUFTE, E. R., *The Visual Display of Quantitative Information*, Cheshire, CT, Graphics Press, 1983.

TUKEY, J. W., *Understanding Robust and Exploratory Data Analysis*, New York, John Wiley & Sons, 1983.

History of Statistics

STIGLER, S. M., *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, MA, Belknap Press, 1986.

Introductory Statistics

BERENSON, M. L., AND D. M. LEVINE, *Basic Business Statistics: Concepts and Applications*, 6th ed., Englewood Cliffs, NJ, Prentice Hall, 1996.

FREUND, J. E., F. J. WILLIAMS, AND B. M. PERLES, *Elementary Business Statistics*, 6th ed., Englewood Cliffs, NJ, Prentice Hall, 1993.

MCCLAVE, J. T., AND P. G. BENSON, *Statistics for Business and Economics*, 6th ed., Englewood Cliffs, NJ, Prentice Hall, 1994.

Nonparametric Statistics

CONOVER, W. J., *Practical Nonparametric Statistics*, 2d ed., New York, John Wiley & Sons, 1980.

GIBBONS, J. D., AND S. CHAKRABORTI, *Nonparametric Statistical Inference*, 3d ed., New York, Marcel Dekker, 1992.

Probability

HOGG, R. V., AND E. A. TANIS, *Probability and Statistical Inference*, 5th ed., Englewood Cliffs, NJ, Prentice Hall, 1997.

ROWNTREE, D., *Probability*, New York, Charles Scribner's Sons, 1984.

Quality and Quality Control

DEMING, W. E., *Out of the Crisis*, Cambridge, MA, MIT Center for Advanced Engineering Study, 1986.

GITLOW, H., S. GITLOW, A. OPPENHEIM, AND R. OPPENHEIM, 2d ed., *Quality Management: Tools and Methods for Improvement*, Homewood, IL, Richard D. Irwin, Inc., 1995.

GRANT, E. L., AND R. S. LEAVENWORTH, *Statistical Quality Control*, 7th ed., New York, McGraw-Hill Book Co., 1996.

ISHIKAWA, K., *Guide to Quality Control*, 2d ed., White Plains, NY, Kraus International Publications, 1986.

Regression and Analysis of Variance

BERRY, W. D., *Multiple Regression in Practice* Beverly Hills, Sage Publications, 1985.

KLEINBAUM, D. G., L. L. KUPPER, AND K. E. MULLER, *Applied Regression Analysis and Other Multivariable Methods*, 2d ed. Boston, PWS-Kent Publishing Co., 1988.

MENDENHALL, W. AND T. SINCICH, *A Second Course in Statistics: Regression Analysis*, 5th ed., Englewood Cliffs, NJ, Prentice Hall, 1996.

NETER, J., W. WASSERMAN, AND M. H. KUTNER, *Applied Linear Statistical Models*, 2d ed., Homewood, IL, Richard D. Irwin, Inc., 1985.

Sampling

GUY, D. M., D. R. CARMICHAEL, AND O. R. WHITTINGTON, *Audit Sampling: An Introduction*, 3rd ed., New York, John Wiley & Sons, 1994.

SCHAEFER, R. L., W. MENDENHALL, AND L. OTT, *Elementary Survey Sampling*, Boston, 5th ed., Duxbury Press, 1996.

Special Topics in Statistics

HUFF, D., *How to Lie with Statistics*, New York, W. W. Norton & Co., 1954.

JAFFE, A. J., *Misused Statistics: Straight Talk for Twisted Numbers*, New York, Marcel Dekker, 1987.

MADANSKY, A., *Prescriptions for Working Statisticians*, New York, Springer-Verlag, 1988.

Statistical Decision Theory

COOK, T. M., AND R. A. RUSSELL, *Introduction to Management Science*, 5th ed., Englewood Cliffs, NJ, Prentice Hall, 1993.

HILLIER, F. S., AND G. J. LIEBERMAN, *Introduction to Operations Research*, 6th ed., New York, McGraw-Hill Book Co., 1995.

LEVIN, R. I., D. S. RUBIN, J. P. STINSON, AND E. S. GARDNER, JR., *Quantitative Approaches to Management*, 8th ed. New York, McGraw-Hill Book Co., 1992.

Statistical Software

MINITAB, INC., *MINITAB User's Guide: Release 10 Xtra*, State College, PA, 1995.

SAS INSTITUTE, INC., *SAS Introductory Guide for Personal Computers*, Release 6.03 ed., Cary, NC, 1988.

Time Series

BOWERMAN, B. L. AND R. T. O'CONNELL, *Forecasting and Time Series: An Applied Approach*, 3d ed., Boston, Duxbury Press, 1993.

FARNUM, N. R., AND L. W. STANTON, *Quantitative Forecasting Methods*, Boston, PWS-Kent Publishing Co., 1989.

MILLS, T. C., *Time Series Techniques for Economists*, Cambridge, Cambridge University Press, 1990.

This page is intentionally left blank.

Index

A

α (the Greek letter, alpha) 356

- a priori probability 159, 199
- acceptable quality level (AQL) 515, 523
- acceptance sampling 514–518, 523
- Aiding Insight, 11, 945
- Ali, Muhammad 154
- alpha (α) 356, 387, 418
- alternative hypothesis 385
- analysis

- marginal 922
- sensitivity 946
- trend 820–821

analysis of variance (ANOVA) 555–563, 566–571, 598

- for the regression 682
- using the computer for 564

AOQ curve 518, 523

AQL (acceptable quality level) 515, 523

arithmetic mean 77–79

- advantages and disadvantages of 82

assignable variation 483, 523

attributes 501, 523

average deviation measures 119

average of the sample ranges 489

average outgoing quality (AOQ) curve 518, 523

B

β (the Greek letter beta) 395, 418

Bayes' theorem 189–190, 942–946

Bayesian decision theory 912

Bayes, Reverend Thomas 190, 912

Bernoulli, Jacob 154

beta (β) 387, 418

between-column variance 557, 561, 598

Bills of Mortality 3

bimodal distribution 106, 141

binomial distribution 106, 225–235

- formula 226

measures of central tendency and dispersion for 233–234

binomial probability tables 227, 234

bivariate frequency distributions 30

boxplot 141

Buede, Dennis 945

C

c (acceptance number) 514, 523

categorical variable 501

cause-and-effect diagram 509, 523

census 278, 320

central limit theorem 307–309

central tendency 74–140

chance events 940

chance node 941

characteristic probability 226

Charlemagne 3

chart

- control 496, 523

- p 501–505, 523

- Pareto 510, 523

- R 496, 523

- x 484–491, 524

Chebyshev's theorem 122

Chebyshev, P.L. 122

chi-square 532–550

chi-square distribution 537

chi-square statistic 536

class

- continuous 22

- discrete 22

- equal 27

- median 98–99

- open-ended 22, 60

classical probability 158–159, 199

cluster sampling 284–285, 321

clusters 285, 321

coding 81, 141, 822–823
 coefficient of correlation 651
 coefficient of determination 643–644
 coefficient of multiple determination, R^2 692
 coefficient of variation 132, 133
 collectively exhaustive events 156, 199
 combined proportion of successes 457, 470
 common variation 523
 computed t 702, 735
 computed f , 705, 735
 conditional probability 176, 181, 199
 conditional profit 915
 table 923
 confidence interval 341–342
 confidence interval for σ^2 584
 confidence levels 341
 relationship with confidence interval 341–342
 confidence limit 364–365
 lower limit 365
 upper limit 365
 consistent estimator 330–331
 Consumer Price Index (CPI) 870, 872, 900
 consumer's risk 515–516
 contingency table 534, 598
 continuity correction factor, 258, 265
 continuous class 22
 continuous data 22, 27
 continuous distributions 24, 922
 continuous probability distribution 212, 246, 265
 continuous quality improvement (CQI) 511, 523
 continuous random variable 246, 265
 control chart 496, 523
 control limits 495–496
 for an R chart 496
 correlation 643–657
 perfect 644–645
 correlation analysis 643–652
 CPI (Consumer Price Index) 870, 872, 900
 CQI (continuous quality improvement) 511, 523
 critical value 398–399
 cumulative frequency distribution 41, 59
 curve
 AOQ 518, 523
 average outgoing quality 518, 523
 OC 517, 523
 operating characteristic 517, 523
 second-degree 826
 curvilinear relationship 613–614, 667
 cyclical fluctuation 819, 860
 cyclical variation 832–836

D

de Fermat, Pierre 154
 de Moivre, Abraham 154
 deciles 115, 141
 Decision Analysis Using Spreadsheets 945
 decision node 941
 decision point 953
 decision theory 4, 912–947
 decision tree 939–947, 953
 on the personal computer 945
 steps 947
 decomposing the total variation in Y 704
 degrees of freedom 355, 371
 denominator 562
 in a chi-square test of independence 533
 in a goodness-of-fit test 550
 numerator 552
 Deming, W. Edwards 482
 denominator degrees-of-freedom 562
 dependence 180
 dependent samples 445–449, 470
 dependent variable 612, 667
 descriptive statistics 4
 deseasonalization 860
 diagram
 cause-and-effect 509, 523
 fishbone 509, 523
 Ishikawa 509, 523
 direct relationship 611, 667
 discrete classes 22
 discrete random variable 214, 215, 265
 dispersion 74
 average deviation measures 119–127
 importance of 111–112
 relative 132–133
 useful measures of 113–114
 Disraeli, Benjamin 3
 distance measure 141
 distribution
 appropriate 548
 bimodal 106, 141
 binomial 206, 225–235
 chi-square 537
 continuous 24, 922–923
 continuous probability 212, 246, 265
 discrete probability 212, 265
 F 561–564, 598
 frequency 18–22
 hypergeometric 516, 523
 multimodal 106

normal 246–259, 265
 Poisson 238–243, 265
 probability 210–212, 238–243, 265
 sampling 296–297, 300–309
 standard normal probability 250, 265, 926
 Student's t 354
 Dodge, Harold F. 514
Domesday Book 3
 double-sampling 514
 dummy variable 720, 722

E

EDA (exploratory data analysis) 141
 equal classes 27
 equations
 normal 681
 error
 type I 387, 419
 type II 387, 419
 estimate of between-column variance 559
 estimate of within-column variance 560
 estimated standard error of the difference between two means 428
 estimating equation 610, 629
 estimating equation describing the relationship among three variables 680
 estimating line 620–621
 estimation 329
 estimator
 consistent 330, 371
 efficient 330, 371
 sufficient 330, 371
 unbiased 330, 371
 event 155
 chance 940
 collectively exhaustive 156, 199
 mutually exclusive 156, 166–167, 199
 EVPI (expected value of perfect information) 919
 EVSI (expected value of sample information) 945
 expected frequencies 534
 expected gain 222
 expected loss 222
 expected marginal loss 924, 953
 expected marginal profit 924, 953
 expected profit 913–919, 953
 with perfect information 918
 expected value 216, 217
 in decision making 220–223
 of a random variable 215–216
 expected value of perfect information (EVPI) 919

expected value of sample information (EVSI) 945
 expected value criterion 945
 experiment 155, 156
 experimental design 292–293
 exploratory data analysis (EDA) 141

F

F distribution 561–562, 598
 F ratio 561–563, 590, 598
 F statistic hypothesis test 561
 F table 563
 F test 564–565
 factorial experiment 294, 321
 finite population 314, 321
 finite population multiplier 314–315
 first-in, first-out (FIFO) 858
 fishbone diagram 509, 523
 fitting a curve to the data 725
 Food and Drug Administration 2
 forecasting 858
 fractile 114, 141
 deciles 115, 141
 percentile 115
 quartile 115
 frequencies
 expected 534, 535
 observed 534, 535, 598
 frequency curve 41, 60
 frequency distribution 19–20, 41
 bivariate 30
 cumulative 41
 relative 20
 frequency polygon 39, 40
 frequency table 19

G

Galton, Sir Francis 610
 garbage in, garbage out (GIGO) 15
 Gardner, Everette S., Jr. 517
 Gauss, Karl 246
 general form for a fitted second-degree curve 826
 geometric mean 92–94, 141
 Gombauld, Antoine 154
 goodness-of-fit test 548, 598
 Good, Richard 161
 Gossett, W.S. 354
 grand mean 487, 598
 from several samples of the same size 487
 Graunt, Captain John 3

H **H_1 (H sub-one)** 385

Henry VII 3

histogram 38, 60

 H_0 (H sub-zero) 385*How to Lie with Statistics* 3

Huff, Darrell 3

hypergeometric distribution 516, 523

hypothesis 385, 419

alternative 385

null 385, 419

test 402, 460, 466

 F 561

hypothesis testing 393–399, 402–404, 405–409

measuring the power of 402–405

of means 393–402, 411

of proportions 405–411

using the standardized scale 395

I**independence** 176

test of 565

independent variables 611–612, 668

inferential statistics 4

infinite population 282

inherent variation 483, 523

intercept

of the trend line for coded time values 824

 Y 617, 624, 668

interfractile range 114, 141

interquartile range 115, 141

interval estimate 329, 338, 371

interval prediction 633

inverse relationship 611

irregular variation 847

Ishikawa diagram 509, 523

Ishikawa, Kaoru 509

J**joint probability** 172, 176, 183, 199

Jones, J. Morgan 945

judgment sampling 279, 321

Juran, Joseph M. 481

K**K statistic** 765

Kolmogorov-Smirnov test 793, 797, 801

Kolmogorov, A. N. 793

Koopman, Bernard 161

Kruskal-Wallis test 758–767, 801

K-S statistic 793, 794

Kurtosis 76, 141

L**Largrange, Joseph** 154

last-in, first-out (LIFO) 858

Latin square 296, 321

law of diminishing returns 316

LCL (lower confidence limit) 341

LCL (lower control limit) 489

least-squares method 623, 624, 670

left-tailed test 390

less-than ogive 41–42

LIFO 858

linear relationship 613–615

loss

expected 222

expected marginal 924, 953

marginal 922–923, 953

obsolescence 221, 953

opportunity 222, 953

lot tolerance percent defective (LTPD) 515, 523

Lotus 1-2-3 517

lower confidence limit (LCL) 341

lower control limit (LCL) 489

lower-tail value of F for two-tailed tests 593

lower-tailed test 390, 418

LTPD (lot tolerance percent defective) 515, 523

M **μ (population mean)** 279 μ (the Greek letter mu) 77

Mann-Whitney U test 758–767, 801

margin of error 2

marginal analysis 922

marginal loss (ML) 922–923, 953

marginal probability 165, 176, 183–184

marginal profit (MP) 922, 924

mean 77

arithmetic 77–83

compared to median and mode 107

geometric 92–94, 141

grand 487–488, 598

modified 842

sample arithmetic 79

weighted 87–90

of the sampling distribution of the proportion 350

measures of central tendency 77–83

measures of dispersion 111

measures of location 74

measures of central tendency 77–141
 median 96–101, 141
 advantages and disadvantages of 100–101
 compared to mean and mode 107
 class 98, 141
 Mendel, Gregor 543
 methods of constructing an index
 Laspeyres 880–881,
 Paasche 882–883
 Michelangelo 481
 minimum probability 924, 953
 minimum probability required to stock another unit 924
 Minitab 30
 ML (marginal loss) 922–923, 953
 mode 104–105, 141
 advantages and disadvantages of 106
 compared to mean and median 107
 modeling 717–728
 modeling techniques 717–728
 modified mean 842, 843
Monthly Labor Review 324
Moody's 873
 MP (marginal profit) 922–924
 mu (μ) 77
 multicollinearity 706–709
 multimodal distributions 106
 multiple regression 610, 668, 678–692, 706
 equation 689
 mutually exclusive events 156, 166–167, 199

N

N (population size) 314
n (sample size) 293, 307
Natural and Political Observations ...Made Upon the Bills of Morality 3
 node 941
 chance 941
 decision 941
 nonnormal population 310
 nonparametric measures 748–749
 nonparametric statistics 748
 nonparametric tests 748
 nonrandom sampling 289–290
 normal distribution 248–259, 265
 as an approximation of the binomial distribution 257–258
 shortcoming of 257
 normal equations 681
 normal population 302

null hypothesis 385, 419
 numerator degrees-of-freedom 562

O

observed frequencies 534–535, 598
 observed value 396
 obsolescence loss 221, 953
 OC (operating characteristic) curve 517, 523
 ogive 42–43, 60
 Old Testament 3
 one-sample runs test 772–773, 801
 one-sample tests 381–415
 one-tailed test 390
 of means 397
 of means using the distribution 414
 of proportions 407–408
 open-ended class 60
 operating characteristic (OC) curve 517, 523
 opportunity loss 222, 953
 optimal stock action 917
OR/MS Today 945
 outliers 485, 523
 out-of-control 485, 523

P

p chart 501–505, 523
 center line for 503
 control line for 503
 estimate of 503
 Paasche method 882–883
 Paasche price index 882
 paired difference test 448, 470
 paired samples 445, 470
 parameters 77, 141, 278–279, 321
 symbols for 279
 parametric statistics 748
 Pareto chart 510, 523
 Pareto, Vilfredo 510
 Pascal, Blaise 154
 payoff 953
 percent of trend 833
 percentage relative 870, 902
 percentiles 115, 142
 perfect correlation 644
 point estimate 329, 331–332
 Poisson distribution 238–244, 265
 as an approximation of the binomial 243
 formula 239
 Poisson, Simeon Denis 238
 pooled estimate of σ^2 435, 470

population
 finite 314, 321
 infinite 282, 321
 nonnormal 310
 normal 302
 population arithmetic mean 78
 population mean 337, 338
 population mean, μ 328
 population parameter 329
 population proportion 334–350
 population regression
 equation 698
 line 657, 658
 plane plus random disturbance 699
 population size, N 314
 population standard deviation 124
 estimate of 345
 population variance 119, 557–558,
 582–583
 posterior probability 193, 199
 power curve 402, 419
 power of the hypothesis test 402, 419
 precision 313, 314, 321
 prediction interval 634
 price index 870, 902
 Paasche 882
 prob value 465–466, 467, 470
 probability
 a priori 159, 199
 characteristic 226
 classical 158–159, 199
 conditional 176, 181, 199
 joint 172, 176, 199
 marginal 165, 176, 183–184
 minimum 924, 953
 posterior 193, 199
 rules 165
 subjective 161, 199
 unconditional 165, 171
 probability distribution 210–217
 choosing 263
 continuous 246–247
 creating 214
 discrete 212
 standard normal 250, 252
 types 212
 probability sampling 279, 321
 probability tree 172, 199
 producer's risk 515, 523

profit
 conditional 915
 expected 913–914, 918
 expected marginal 924
 expected with perfect information 918
 marginal 922
 proportion
 population 328, 334
 p -value 465–470

Q
 quadratic regression model 727
 qualitative data 727
 qualitative variable 501, 523
 quantitative data 106
 quantity index 870, 874, 895
 quartiles 115, 142

R
r(number of runs) 774
R chart 496, 523
 control limits for 496
R (average of the sample ranges) 489
 Ramsey, Frank 161
 random sampling 281–285, 321
 random variable 215–216
 standardizing 251
 range
 interfractile 114, 141
 interquartile 115, 141
 rank correlation 781–782, 784, 801
 rank correlation coefficient 782, 801
 rank sum test 758
 ratio
 F 561–563, 590, 598
 ratio-to-moving-average method 839, 860
 raw data 17, 60
 raw scale 395, 419
 regression analysis 610
 regression coefficient
 standard error for 659, 700
 regression line 617–618
 regression model
 quadratic 727
 second-degree 727
 relative cyclical residual 833, 834, 860
 relative frequency distribution 20, 60
 relative frequency histogram 38
 relative frequency of occurrence 159, 160, 161, 199
 relative frequency polygon 40

- relatives method
 unweighted 888, 902
 weighted 888–889, 902
- replacement 281
- representative sample 16, 60
- residual method 832, 836, 860
- response variable 293
- right-tailed test 391
- risk
 consumer's 515, 523
 producer's 515, 523
- rollback 941, 953
- runs
 one sample 772–774, 801
 theory of 773, 802
- S**
- σ (population standard deviation) 124**
- salvage value 953
- sample
 representative 16, 60
 standard deviation 124–125
- sample arithmetic mean of grouped data 79
- sample mean \pm 302
- sample median of grouped data 100
- sample size (n) 293, 313, 354–355
- sample space 155, 200
- sample standard deviation, s 124–125
- samples
 dependent 445, 470
 paired 445, 470
- sampling
 acceptance 514–518, 523
 cluster 284–285, 321
 distributions 296–298, 300, 309
 double 514
 judgment 289, 321
 nonrandom 289–290
 probability 279, 321
 random 281–285
 simple random 281–282, 321
 single 514
 stratified 284, 322
 systematic 283–284, 322
 with replacement 321
 without replacement 321
- sampling distribution
 of a statistic 321
 of the mean 297, 301, 321
- of the proportion 296, 350
- sampling error 297, 321
- sampling fraction 315, 321
- Savage, Leonard 161
- Scale
 raw 402, 419
 standardized 397, 419
- scatter diagram 612–615
- seasonal variation 838–844
- second-degree curve 826
- second-degree equation 826, 827, 828, 860
- second-degree regression model 727
- secular trend 819, 821, 860
- sensitivity analysis 945
- Shewhart, Walter 482, 514
- sigma (σ) 78
- sigma hat 361
- sign test 748, 750–757, 802
- significance level 385, 387, 412, 428
- Simon, Pierre 154
- simple arithmetic mean 93
- simple random sampling 281–282, 321
- simple regression 610–615, 618–661
- Sinclair, Sir John 3
- single-sampling 514
- skewness 75, 142
- slope 611, 668
 of a straight line 618
 of the best-fitting regression line 624
 of the population regression line 658
 of the trend line for coded time values 824
- Smirnov, N. V. 793
- software packages for statistical analysis 29
- SPC (statistical process control) 482–483, 523
- special cause variation 483, 523
- Spinks, Leon 154
- SPSS 136
- standard deviation 119, 120, 124, 142
- standard error 303, 346
 of b 659
 standard error of the difference between two means 428
 standard error of the mean 303, 346
 of a finite population 346
 of an infinite population 357
- standard error of the regression coefficient 659
- standard error of the statistic 297
- standard normal probability distribution 250, 265
- standard score 122, 127

standardized regression coefficient 722
 standardized scale 419
 standardizing the sample mean 304
Statistical Account of Scotland 1791–1799 3
 statistical dependence 180, 199
 statistical independence 171, 199
 statistical inference 4, 307, 321, 328
 statistical process control (SPC) 482–483, 523
 statistics
 descriptive 4
 inferential 4
 K 765
 K-S 793
 nonparametric 748
 origin of the word 3
 parametric 748
 r 774
 subdivisions within 4
 summary 74
 U 760–761
 stem and leaf display 142
 straight line equation 850
 strata 284, 321
 stratified sampling 284, 322
 student's t distribution 354, 371
 subjective probability 161, 199
 summary statistics 74, 136
 sums of squares 704
 systematic sampling 283–284, 322

T

t distribution 435
 table
 contingency 534, 598
 random digits 282
 z 356
 tests
 goodness-of-fit 548, 598
 Kolmogorov–Smirnov 773–779, 801
 Kruskal-Wallis 758–767, 801
 left-tailed 390
 lower-tailed 342, 418
 Mann-Whitney 758–767, 801
 nonparametric 748
 of independence 533, 598
 one-sample runs 772–773, 801
 paired difference 448, 470
 power of the hypothesis 402, 419
 rank correlation 781–782

rank sum 758
 right-tailed 391
 sign 748, 750–757, 802
 two-tailed 389, 393, 412, 456
 upper-tailed 391, 419
The European Journal of Operations Research 945
 theory of runs 773, 802
 three different sums of squares 704
 time series 818–867
 time series analysis 858
 total quality management (TQM) 480, 509
 transformations 735
 trend analysis 820, 822
 trend line 820
 two population variances 589
 two-sample tests 426–468
 two-tailed test 389–393, 412, 456
 of means 393–395
 of means using the *t* distribution 412
 type I error 405, 387–388, 419
 producer's risk 515
 type II error 387–388, 419

U

U statistic 760, 761
 alternate formula 762
 mean 760
 standard error 760
 UCL (upper confidence limit) 341
 UCL (upper control limit) 489
 unbiased estimator 320, 371
 unconditional probability 165
 ungrouped data 78
 unweighted aggregates index 874–876, 902
 unweighted average of relatives method 888, 902
 unweighted average of relatives price index 888
 upper-tailed test 391, 419
 utility 931–933, 953

V

value
 critical 406, 418
 expected 931–932, 953
 expected, of perfect information 919, 953
 observed 396
 salvage 953
 Z 252
 value index 896

variable
categorical 501
continuous random 246, 265
dependent 612, 667
discrete random 246, 265
dummy 720, 722
estimating equation 680
expected value of a random 220, 265
independent 611–612, 668
qualitative 501, 523
quantitative 501, 523
random 214–216
response 293
variance
analysis of 555–563, 566–571, 598
ANOVA 555–563, 566–571, 598
between-column 557, 561, 598
one tailed test 586–587
within-column 559–560
variation
assignable 483, 523
common 523
cyclical 832–836
inherent 483, 523
irregular 847
seasonal 819, 838–844

special cause 483, 523
Venn diagram 165, 199
Venn, John 165

W

weighted average
of relative method 888–891
of relatives price index 890
of relative quantity index 895
weighted mean 87–89, 142
width of class intervals 28
William the Conqueror 3
within-column variance 559, 560

X
x chart 484–491, 524

Y
Y values 643–644
Y-intercept 617
of the best-fitting regression line 624

Z
z table 356
z values 252
zero defects 482
Zimmerman, E. A. W. 3