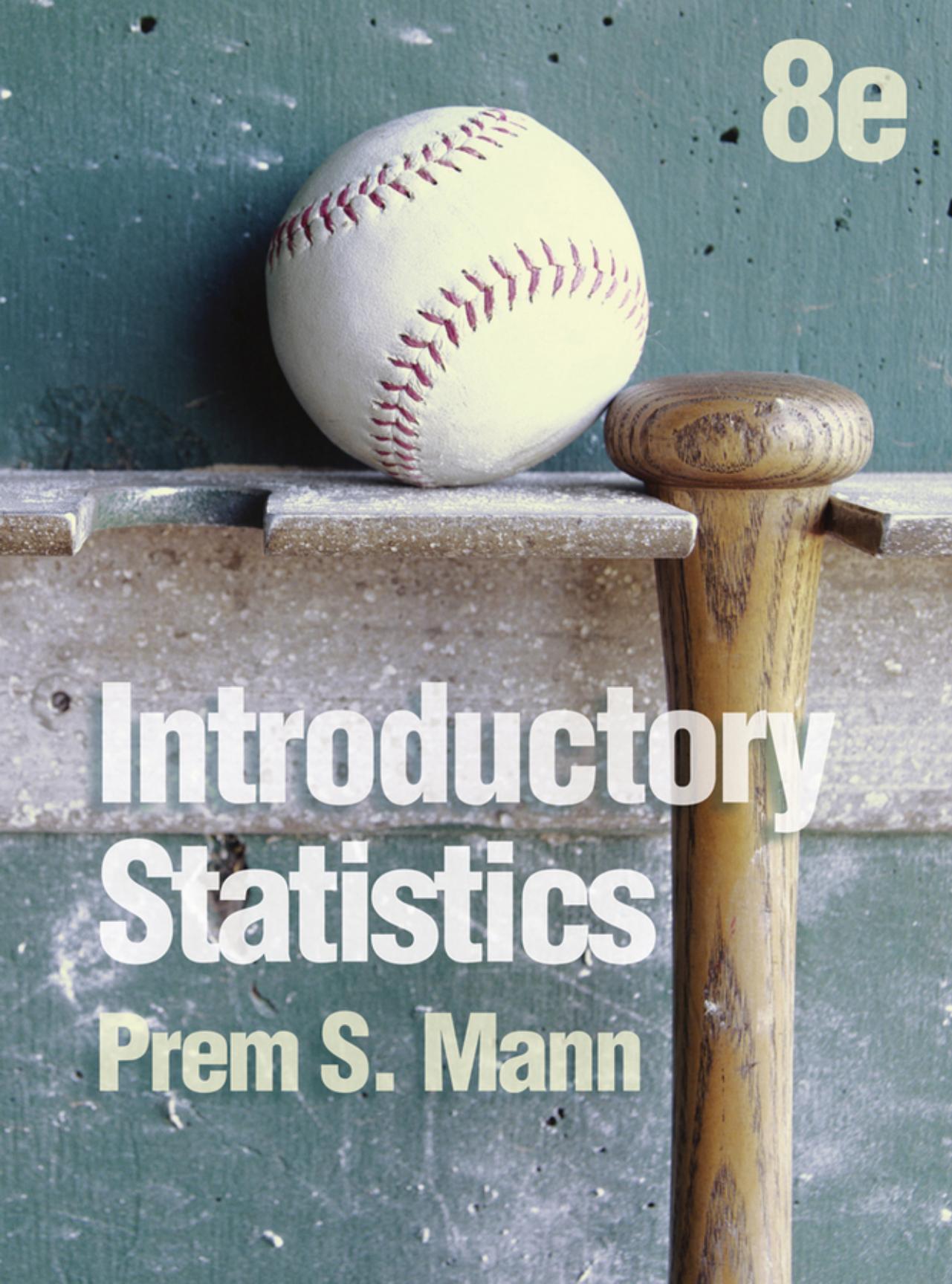


**8e**



**Introductory  
Statistics**  
**Prem S. Mann**



# WileyPLUS

**WileyPLUS is a research-based online environment for effective teaching and learning.**

WileyPLUS builds students' confidence because it takes the guesswork out of studying by providing students with a clear roadmap:

- what to do
- how to do it
- if they did it right

It offers interactive resources along with a complete digital textbook that help students learn more. With WileyPLUS, students take more initiative so you'll have greater impact on their achievement in the classroom and beyond.



Now available for



Blackboard

For more information, visit [www.wileyplus.com](http://www.wileyplus.com)

# WileyPLUS

**ALL THE HELP, RESOURCES, AND PERSONAL SUPPORT YOU AND YOUR STUDENTS NEED!**

[www.wileyplus.com/resources](http://www.wileyplus.com/resources)



2-Minute Tutorials and all of the resources you and your students need to get started



Student support from an experienced student user



Collaborate with your colleagues, find a mentor, attend virtual and live events, and view resources  
[www.WhereFacultyConnect.com](http://www.WhereFacultyConnect.com)



Pre-loaded, ready-to-use assignments and presentations created by subject matter experts



Technical Support 24/7  
FAQs, online chat, and phone support  
[www.wileyplus.com/support](http://www.wileyplus.com/support)



© Courtney Keating/Stockphoto

Your WileyPLUS Account Manager, providing personal training and support

Eighth Edition

# INTRODUCTORY STATISTICS

**PREM S. MANN**

EASTERN CONNECTICUT STATE UNIVERSITY

WITH CONTRIBUTIONS BY

**CHRISTOPHER JAY LACKE**

ROWAN UNIVERSITY

**WILEY**

**Vice President and Executive Publisher** Laurie Rosatone  
**Acquisitions Editor** Joanna Dingle  
**Project Editor** Ellen Keohane  
**Associate Content Editor** Beth Pearson  
**Editorial Program Assistant** Liz Baird  
**Senior Content Manager** Karoline Luciano  
**Senior Production Editor** Kerry Weinstein  
**Production Management Services** Aptara, Inc.  
**Photo Editor** Lisa Gee  
**Marketing Manager** Melanie Kurkjian  
**Senior Designer** Madelyn Lesure  
**Senior Product Designer** Thomas Kulesa  
**Editorial Operations Manager** Melissa Edwards  
**Media Production Specialist** Laura Abrams  
**Cover Designer** Madelyn Lesure  
**Cover photo:** © Lawrence Manning/Age Fotostock America, Inc.

This book was set in 10/12 Times Roman by Aptara®, Inc. and printed and bound by Courier-Kendallville. The cover was printed by Courier-Kendallville.

This book is printed on acid free paper. ∞

Copyright © 2013, 2010, 2007, 2004, 2001, John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc. 222 Rosewood Drive, Danvers, MA 01923, website [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, (201)748-6011, fax (201)748-6008, website <http://www.wiley.com/go/permissions>.

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at [www.wiley.com/go/returnlabel](http://www.wiley.com/go/returnlabel). Outside of the United States, please contact your local representative.

*Library of Congress Cataloging in Publication Data*

ISBN 978-0-470-90410-7 (cloth)  
ISBN 978-1-118-17224-7 (Binder Ready Version )

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

**To the memory of my parents**

# PREFACE

---

**Introductory Statistics** is written for a one- or two-semester first course in applied statistics. This book is intended for students who do not have a strong background in mathematics. The only prerequisite for this text is knowledge of elementary algebra.

Today, college students from almost all fields of study are required to take at least one course in statistics. Consequently, the study of statistical methods has taken on a prominent role in the education of students from a variety of backgrounds and academic pursuits. From the first edition, the goal of **Introductory Statistics** has been to make the subject of statistics interesting and accessible to a wide and varied audience. Three major elements of this text support this goal:

1. Realistic content of its examples and exercises, drawing from a comprehensive range of applications from all facets of life
2. Clarity and brevity of presentation
3. Soundness of pedagogical approach

These elements are developed through the interplay of a variety of significant text features.

The feedback received from the users of the seventh edition (and earlier editions) of **Introductory Statistics** has been very supportive and encouraging. Positive experiences reported by instructors and students have served as evidence that this text offers an interesting and accessible approach to statistics—the author’s goal from the very first edition. The author has pursued the same goal through the refinements and updates in this eighth edition, so that **Introductory Statistics** can continue to provide a successful experience in statistics to a growing number of students and instructors.

## New to the Eighth Edition

---

The following are some of the changes made in the eighth edition:

- A large number of the examples and exercises are new or revised, providing contemporary and varied ways for students to practice statistical concepts.
- Factorials, combinations, and permutations have been moved from Chapter 5 to Section 4.6 of Chapter 4.
- All case studies are new or revised, drawing on current uses of statistics in areas of student interest.
- New chapter opening images and questions incorporate real data in familiar situations.
- New data are integrated throughout, reinforcing the vibrancy of statistics and the relevance of statistics to student lives right now.
- The *Technology Instruction* sections have been updated to support the use of the latest versions of TI-84/84+, Minitab, and Excel.
- Many of the *Technology Assignments* at the end of each chapter are either new or have been updated.

- Many *Uses and Misuses* sections are either new or have been updated.
- Many *Decide for Yourself* sections are either new or have been updated.
- Several new Mini-projects have been added to this edition.
- The data sets posted on the book companion Web site and WileyPLUS have been updated.
- In Chapter 6, a new appendix on normal quantile plots has been added.

## Hallmark Features of this Text

**Clear and Concise Exposition** The explanation of statistical methods and concepts is clear and concise. Moreover, the style is user-friendly and easy to understand. In chapter introductions and in transitions from section to section, new ideas are related to those discussed earlier.

**Examples** The text contains a wealth of examples, 219 in 15 chapters and Appendix A. The examples are usually presented in a format showing a problem and its solution. They are well sequenced and thorough, displaying all facets of concepts. Furthermore, the examples capture students' interest because they cover a wide variety of relevant topics. They are based on situations that practicing statisticians encounter every day. Finally, a large number of examples are based on real data taken from sources such as books, government and private data sources and reports, magazines, newspapers, and professional journals.

**Solutions** A clear, concise solution follows each problem presented in an example. When the solution to an example involves many steps, it is presented in a step-by-step format. For instance, examples related to tests of hypothesis contain five steps that are consistently used to solve such examples in all chapters. Thus, procedures are presented in the concrete settings of applications rather than as isolated abstractions. Frequently, solutions contain highlighted remarks that recall and reinforce ideas critical to the solution of the problem. Such remarks add to the clarity of presentation.

**Margin Notes for Examples** A margin note appears beside each example that briefly describes what is being done in that example. Students can use these margin notes to assist them as they read through sections and to quickly locate appropriate model problems as they work through exercises.

**Frequent Use of Diagrams** Concepts can often be made more understandable by describing them visually with the help of diagrams. This text uses diagrams frequently to help students understand concepts and solve problems. For example, tree diagrams are used extensively in Chapters 4 and 5 to assist in explaining probability concepts and in computing probabilities. Similarly, solutions to all examples about tests of hypothesis contain diagrams showing rejection regions, nonrejection regions, and critical values.

**Highlighting** Definitions of important terms, formulas, and key concepts are enclosed in colored boxes so that students can easily locate them.

**Cautions** Certain items need special attention. These may deal with potential trouble spots that commonly cause errors, or they may deal with ideas that students often overlook. Special emphasis is placed on such items through the headings *Remember*, *An Observation*, or *Warning*. An icon is used to identify such items.

**Case Studies** Case studies, which appear in almost all chapters, provide additional illustrations of the applications of statistics in research and statistical analysis. Most of these case studies are based on articles or data published in journals, magazines, newspapers, or Web sites. All case studies are based on real data.

### Style and Pedagogy

### Thorough Examples

### Step-by-Step Solutions

### Enlightening Pedagogy

### Realistic Applications

**Abundant Exercises**

**Exercises and Supplementary Exercises** The text contains an abundance of exercises (excluding Technology Assignments)—1542 in 15 chapters and Appendix A. Moreover, a large number of these exercises contain several parts. Exercise sets appearing at the end of each section (or sometimes at the end of two or three sections) include problems on the topics of that section. These exercises are divided into two parts: **Concepts and Procedures** that emphasize key ideas and techniques, and **Applications** that use these ideas and techniques in concrete settings. Supplementary exercises appear at the end of each chapter and contain exercises on all sections and topics discussed in that chapter. A large number of these exercises are based on real data taken from varied data sources such as books, government and private data sources and reports, magazines, newspapers, and professional journals. Not merely do the exercises given in the text provide practice for students, but the real data contained in the exercises provide interesting information and insight into economic, political, social, psychological, and other aspects of life. The exercise sets also contain many problems that demand critical thinking skills. The answers to selected odd-numbered exercises appear in the *Answers* section at the back of the book. **Optional exercises** are indicated by an asterisk (\*).

**Challenging Problems**

**Advanced Exercises** All chapters (except Chapters 1 and 14) have a set of exercises that are of greater difficulty. Such exercises appear under the heading *Advanced Exercises* after the *Supplementary Exercises*.

**Misconceptions and Pitfalls**

**Uses and Misuses** This feature towards the end of each chapter (before the Glossary) points out common misconceptions and pitfalls students will encounter in their study of statistics and in everyday life. Subjects highlighted include such diverse topics as the use of the word *average* and do not feed the animals.

**Open-ended Problems**

**Decide for Yourself** This feature appears towards the end of each chapter (except Chapter 1) just before the Technology Instruction section. In this section, a real-world problem is discussed, and questions are raised about this problem that readers are required to answer.

**Summary and Review**

**Glossary** Each chapter has a glossary that lists the key terms introduced in that chapter, along with a brief explanation of each term. Almost all the terms that appear in boldface type in the text are in the glossary.

**Testing Yourself**

**Self-Review Tests** Each chapter contains a *Self-Review Test*, which appears immediately after the *Supplementary* and *Advanced Exercises*. These problems can help students test their grasp of the concepts and skills presented in respective chapters and monitor their understanding of statistical methods. The problems marked by an asterisk (\*) in the *Self-Review Tests* are **optional**. The answers to almost all problems of the *Self-Review Tests* appear in the *Answer* section.

**Key Formulas**

**Formula Card** A formula card that contains key formulas from all chapters and the normal distribution and *t* distribution tables is included at the beginning of the book.

**Technology Usage**

**Technology Usage** At the end of each chapter is a section covering uses of three major technologies of statistics and probability: the TI-84, Minitab, and Excel. For each technology, students are guided through performing statistical analyses in a step-by-step fashion, showing them how to enter, revise, format, and save data in a spreadsheet, workbook, or named and unnamed lists, depending on the technology used. Illustrations and screen shots demonstrate the use of these technologies. Additional detailed technology instruction is provided in the technology manuals that are online at [www.wiley.com/college/mann](http://www.wiley.com/college/mann).

**Technology Assignments**

**Technology Assignments** Each chapter contains a few technology assignments that appear at the end of the chapter. These assignments can be completed using any of the statistical software.

**Mini-projects**

**Mini-projects** Each chapter contains a few Mini-projects that appear just before the Decide it Yourself sections. These Mini-projects are like very comprehensive exercises or ask students to perform their own surveys and experiments. They provide practical applications of statistical concepts to real life.

**Data Sets** A large number of data sets appear on the book companion Web site at [www.wiley.com/college/mann](http://www.wiley.com/college/mann). These data sets include thirteen large data sets. These thirteen large data sets are collected from various sources and they contain information on several variables. Many exercises and assignments in the text are based on these data sets. These large data sets can also be used for instructor-driven analyses using a wide variety of statistical software packages as well as the TI-84. **These data sets are available on the Web site of the text in eight formats including Minitab<sup>1</sup>, Excel, and SPSS.**

**Data Sets**

**Statistical Animations** In relevant places throughout the text, an icon alerts students to the availability of a statistical animation. These animations illustrate statistical concepts in the text, and can be found on the companion Web site.

**Statistical Animations**

## GAISE Report Recommendations Adopted

In 2003, the American Statistical Association (ASA) funded the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project to develop ASA-endorsed guidelines for assessment and instruction in statistics for the introductory college statistics course. The report, which can be found at [www.amstat.org/education/gaise](http://www.amstat.org/education/gaise), resulted in the following series of recommendations for the first course in statistics and data analysis.

1. Emphasize statistical literacy and develop statistical thinking.
2. Use real data.
3. Stress conceptual understanding rather than mere knowledge of procedures.
4. Foster active learning in the classroom.
5. Use technology for developing concepts and analyzing data.
6. Use assessments to improve and evaluate student learning.

Here are a few examples of how this Introductory Statistics text can assist in helping you, the instructor, in meeting the GAISE recommendations.

1. Many of the exercises require interpretation, not just answers in terms of numbers. Graphical and numeric summaries are combined in some exercises in order to emphasize looking at the whole picture, as opposed to using just one graph or one summary statistic.
2. The *Decide for Yourself* and *Uses and Misuses* features help to develop statistical thinking and conceptual understanding.
3. All of the data sets in the exercises and in Appendix B are available on the book's Web site. They have been formatted for a variety of statistical software packages. This eliminates the need to enter data into the software. A variety of software instruction manuals also allows the instructor to spend more time on concepts, and less time teaching how to use technology.
4. The Mini-projects help students to generate their own data by performing an experiment and/or taking random samples from the large data sets mentioned in Appendix B.

We highly recommend that all statistics instructors take the time to read the GAISE report. There is a wealth of information in this report that can be used by everyone.

## Web Site

<http://www.wiley.com/college/mann>

After you go to the page exhibited by the above URL, click on *Visit the Companion Sites*. Then click on the site that applies to you out of the two choices. This Web site provides additional resources for instructors and students. The following items are available for instructors on this Web site:

- Formula Card
- Statistical Animations
- Printed Test Bank

<sup>1</sup>Minitab is a registered trademark of Minitab, Inc., Quality Plaza, 1829 Pine Hall Road, State College, PA 16801–3008. Phone: 814-238-3280.

- Computerized Test Bank
- Instructor's Solutions Manual
- PowerPoint Slides
- Data Sets (see Appendix B for a complete list of these data sets)
- Chapter 14: Multiple Regression
- Chapter 15: Nonparametric Methods
- Technology Resource Manuals.
  - TI Graphing Calculator Manual
  - Minitab Manual
  - Excel Manual

These manuals provide step-by-step instructions, screen captures, and examples for using technology in the introductory statistics course. Also provided are exercise lists and indications of which exercises from the text best lend themselves to the use of the package presented.

## Using WileyPLUS

**SUCCESS:** *WileyPLUS* helps to ensure that each study session has a positive outcome by putting students in control. Through instant feedback and study objective reports, students know **if they did it right** and where to focus next, so they achieve the strongest results.

With *WileyPLUS*, our efficacy research shows that students improve their outcomes by as much as one letter grade. *WileyPLUS* helps students take more initiative, so you'll have greater impact on their achievement in the classroom and beyond.

### What do students receive with *WileyPLUS*?

- The complete digital textbook, saving students up to 60% off the cost of a printed text.
- Question assistance, including links to relevant sections in the online digital textbook.
- Immediate feedback and proof of progress, 24/7.
- Integrated, multimedia resources—including animations and videos—that provide multiple study paths and encourage more active learning.

### What do instructors receive with *WileyPLUS*?

- Reliable resources that reinforce course goals inside and outside of the classroom.
- The ability to easily identify those students who are falling behind.
- Media-rich course materials and assessment content, including Instructor's Solutions Manual, PowerPoint Slides, Learning Objectives, Computerized and Printed Test Banks, and much more.

[www.wileyplus.com](http://www.wileyplus.com). Learn More.

## Supplements

The following supplements are available to accompany this text:

- **Instructor's Solutions Manual (ISBN 978-1-118-50414-7)** This manual contains compete solutions to all of the exercises in the text.
- **Printed Test Bank** The printed copy of the test bank contains a large number of multiple-choice questions, essay questions, and quantitative problems for each chapter. It can be downloaded and printed from *WileyPLUS* or from [www.wiley.com/college/mann](http://www.wiley.com/college/mann).

- **Computerized Test Bank** All of the questions in the Printed Test Bank are available electronically and can be obtained from the publisher.
- **Student Solutions Manual (ISBN 978-1-118-50410-9).** This manual contains complete solutions to all of the odd-numbered exercises in the text.

## Acknowledgments

I thank the following reviewers of this and/or previous editions of this book, whose comments and suggestions were invaluable in improving the text.

James Adcock <i>University of Western Ontario</i>	C. K. Chauhan <i>Indiana-Purdue University at Fort Wayne</i>
Alfred A. Akinsete <i>Marshall University</i>	Jerry Chen <i>Suffolk County Community College</i>
Scott S. Albert <i>College of DuPage</i>	Dianna Cichocki <i>Erie Community College</i>
Michael R. Allen <i>Tennessee Technological University</i>	James Curl <i>Modesto Community College</i>
Raid Amin <i>University of West Florida</i>	Gregory Daubenmire <i>Las Positas Community College</i>
Gurdial Arora <i>Xavier University of Louisiana</i>	Robert M. Davis <i>Alamance Community College</i>
Peter Arvanites <i>Rockland Community College</i>	Joe DeMaio <i>Kennesaw State University</i>
K. S. Asal <i>Broward Community College</i>	Kevin Dennis <i>Saint Mary's University of Minnesota</i>
Louise Audette <i>Manchester Community College</i>	Mihaela Dobrescu <i>Christopher Newport University</i>
Joleen Beltrami <i>University of the Incarnate Word</i>	Fred H. Dorner <i>Trinity University, San Antonio</i>
Nicole Betsinger <i>Arapahoe Community College</i>	William D. Ergle <i>Roanoke College, Salem, Virginia</i>
Cornelia Bica <i>Northern Alberta Institute of Technology</i>	Ruby Evans <i>Santa Fe Community College</i>
Patricia J. Blus <i>National-Louis University</i>	Ronald Ferguson <i>San Antonio College</i>
Joan Bookbinder <i>Johnson &amp; Wales University</i>	James C. Ford <i>Anda Gadidov</i>
Christine H. Brady <i>Suffolk County Community College</i>	<i>Kennesaw State University</i>
Dean Burbank <i>Gulf Coast Community College</i>	Jason Gershman <i>Nova Southeastern University</i>
Helen Burn <i>Highline Community College</i>	Frank Goulard <i>Portland Community College</i>
Gerald Busald <i>San Antonio College</i>	Robert Graham <i>Jacksonville State University, Jacksonville, Alabama</i>
Ferry Butar Butar <i>Sam Houston State University</i>	Larry Griffey <i>Florida Community College, Jacksonville</i>
Peter A. Carlson <i>Delta College</i>	Arjun K. Gupta <i>Bowling Green State University</i>
Jayanta Chandra <i>University of Notre Dame</i>	David Gurney <i>Southeastern Louisiana University</i>

- Daesung Ha  
*Marshall University*  
John Haussermann  
*Monterey Peninsular College*  
A. Eugene Hileman  
*Northeastern State University, Tahlequah,  
Oklahoma*  
John G. Horner  
*Cabrillo College*  
Virginia Horner  
*Diablo Valley College*  
Ina Parks S. Howell  
*Florida International University*  
Tanya Huffman  
*Florida Gulf Coast University*  
Shana Irwin  
*University of North Texas*  
Gary S. Itzkowitz  
*Rowan State College*  
Joanna Jeneralczuk  
*University of Massachusetts, Amherst*  
Jean Johnson  
*Governors State University*  
Eryn M. Kalbfleisch  
*University of Akron*  
Michael Karelius  
*American River College, Sacramento*  
Dix J. Kelly  
*Central Connecticut State University*  
Parviz Khalili  
*Christopher Newport University*  
Jong Sung Kim  
*Portland State University*  
Hoon Kim  
*California State Polytechnic University,  
Pomona*  
Jong Sung Kim  
*Portland State University*  
Linda Kohl  
*University of Michigan, Ann Arbor*  
Martin Kotler  
*Pace University, Pleasantville, New York*  
Marlene Kovaly  
*Florida Community College, Jacksonville*  
Hillel Kumin  
*University of Oklahoma*  
Carlos de la Lama  
*San Diego City College*  
Yingfu (Frank) Li  
*University of Houston, Clear Lake*  
Rita Lindsay  
*Indian River State College*  
Reginald Luke  
*Middlesex County College*  
Gaurab Mahapatra  
*University of Akron*  
Vinod P. Manglik  
*Elizabeth City State University*  
Christopher Mansfield  
*Durham Technical Community College*  
Richard McGowan  
*University of Scranton*  
Paul F. Messina  
*University of the Incarnate Word*  
Daniel S. Miller  
*Central Connecticut State University*  
Dorothy Miners  
*Brock University*  
Nutan Mishra  
*University of South Alabama*  
Satya N. Mishra  
*University of South Alabama*  
Jeffrey Mock  
*Diablo Valley College*  
Hojin Moon  
*California State University,  
Long Beach*  
Luis Moreno  
*Broome Community College,*  
Robert A. Nagy  
*University of Wisconsin, Green Bay*  
Sharon Navard  
*The College of New Jersey*  
Nhu T. Nguyen  
*New Mexico State University*  
Paul T. Nkansah  
*Florida Agricultural and Mechanical  
University*  
Alan Olinsky  
*Bryant University*  
Joyce Oster  
*Johnson and Wales University*  
Lindsay Packer  
*College of Charleston*  
Mary Parker  
*Austin Community College*  
Roger Peck  
*University of Rhode Island, Kingston*  
Julie Peschke  
*Athabasca University*  
Chester Piascik  
*Bryant College, Smithfield*  
Joseph Pigeon  
*Villanova University*  
Cristina Popescue  
*Grant MacEwan College*  
Ramaswamy Radhakrishnan  
*Illinois State University*  
Aaron Robertson  
*Colgate University*  
Gerald Rogers  
*New Mexico State University, Las Cruces*

Lisa Rombes	Kagba Suaray
<i>Washtenaw Community College</i>	<i>California State University, Long Beach</i>
Emily Ross	Arnavaz P. Taraporevala
<i>University of Missouri, St. Louis</i>	<i>New York City College of Technology</i>
Said E. Said	Bruce Trumbo
<i>East Carolina University</i>	<i>California State University, Hayward</i>
Juana Sanchez	Deanna Voehl
<i>UCLA</i>	<i>Indian River State College</i>
Brunilda Santiago	Vasant Waikar
<i>Indian River State College</i>	<i>Miami University</i>
Iris Schneider	Bin Wang
<i>Pace University</i>	<i>University of South Alabama</i>
Phillis Schumacher	Jean Weber
<i>Bryant College, Smithfield</i>	<i>University of Arizona, Tucson</i>
Kathryn Schwartz	Terry Wilson
<i>Scottsdale Community College</i>	<i>San Jacinto College, Pasadena</i>
Ronald Schwartz	James Wright
<i>Wilkes University, Wilkes-Barre</i>	<i>Bucknell University</i>
Sean Simpson	Xin Yan
<i>Westchester Community College</i>	<i>University of Missouri, Kansas City</i>
Satyanand Singh	K. Paul Yoon
<i>New York City College of Technology</i>	<i>Fairleigh Dickinson University, Madison</i>
David Stark	Zhiyi Zhang
<i>University of Akron</i>	<i>University of North Carolina</i>
Larry Stephens	
<i>University of Nebraska, Omaha</i>	

I express my thanks to the following for their contributions to earlier editions of this book that made it better in many ways: Gerald Geissert (formerly of Eastern Connecticut State University), Daniel S. Miller (Central Connecticut State University), and David Santana-Ortiz (Rand Organization).

I extend my special thanks to Christopher Lacke of Rowan University, who contributed to this edition in many significant ways. Without his help, this book would not be in this form. I take this opportunity to thank Beverly Fusfield for working on the solutions manuals and preparing the answer section, Sandra Zirkes for checking the text and answer section for accuracy, and Ann Ostberg and Dan Miller for checking the solutions for accuracy. I also thank Andrea Boito, Sean Simpson, and Doug Tyson for their work on the technology manuals and Hoon Kim for his work on the PowerPoint lecture slides. I would also like to thank Todd Hoff for his work on the test bank. In addition, I thank Eastern Connecticut State University for all the support I received.

It is of utmost importance that a textbook be accompanied by complete and accurate supplements. I take pride in mentioning that the supplements prepared for this text possess these qualities and much more. I thank the authors of all these supplements.

It is my pleasure to thank all the professionals at John Wiley & Sons with whom I enjoyed working during this revision. Among them are Laurie Rosatone (Vice President and Executive Publisher), Joanna Dingle (Acquisitions Editor), Jackie Henry (Full Service Manager), Madelyn Lesure (Senior Designer), Lisa Gee (Senior Photo Editor), Karoline Luciano (Senior Content Manager), Kerry Weinstein (Senior Production Editor), Ellen Keohane (Project Editor), Beth Pearson (Associate Content Editor), Elizabeth Baird (Editorial Program Assistant), Laura Abrams (Media Assistant), Thomas Kulesa (Senior Product Designer), and Melanie Kurkjian (Marketing Manager). I extend my most heartfelt thanks to Ellen Keohane, whose support and guidance was of immense help during this revision. I also thank Lisa Torri (Art Development Editor) for her work on the case study art.

Any suggestions from readers for future revisions would be greatly appreciated. Such suggestions can be sent to the author at [mann@easternct.edu](mailto:mann@easternct.edu) or [premann@yahoo.com](mailto:premann@yahoo.com).

Prem S. Mann  
Willimantic, CT  
September 2012

# **CONTENTS**

---

## **CHAPTER 1 Introduction 1**

### **1.1 Statistics and Types of Statistics 2**

*Case Study 1–1* How Much Did Companies Spend on Ads in 2011? **3**

*Case Study 1–2* How Women Rate Their Lives **4**

### **1.2 Population Versus Sample 5**

*Case Study 1–3* Are We Becoming Less “Green?” **7**

### **1.3 Basic Terms 8**

### **1.4 Types of Variables 10**

### **1.5 Cross-Section Versus Time-Series Data 12**

### **1.6 Sources of Data 14**

### **1.7 Summation Notation 15**

Uses and Misuses / Glossary / Supplementary Exercises / Self-Review Test / Mini-Project/  
Decide for Yourself / Technology Instruction / Technology Assignments

## **CHAPTER 2 Organizing and Graphing Data 28**

### **2.1 Organizing and Graphing Qualitative Data 29**

*Case Study 2–1* Will Today’s Children Be Better Off Than Their Parents? **32**

*Case Study 2–2* Employees’ Overall Financial Stress Levels **33**

### **2.2 Organizing and Graphing Quantitative Data 36**

*Case Study 2–3* How Long Does Your Typical One-Way Commute Take? **42**

*Case Study 2–4* How Much Does It Cost to Insure a Car? **43**

*Case Study 2–5* How Many Cups of Coffee Do You Drink a Day? **46**

### **2.3 Cumulative Frequency Distributions 54**

### **2.4 Stem-and-Leaf Displays 57**

### **2.5 Dotplots 62**

Uses and Misuses / Glossary /Supplementary Exercises / Advanced Exercises / Self-Review Test /  
Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

## **CHAPTER 3 Numerical Descriptive Measures 85**

### **3.1 Measures of Central Tendency for Ungrouped Data 86**

*Case Study 3–1* Average NFL Ticket Prices in the Secondary Market **89**

Case Study 3–2 Average Is Over **90**

Case Study 3–3 Education Pays **92**

### 3.2 Measures of Dispersion for Ungrouped Data **99**

### 3.3 Mean, Variance, and Standard Deviation for Grouped Data **106**

### 3.4 Use of Standard Deviation **113**

Case Study 3–4 Does Spread Mean the Same as Variability and Dispersion? **116**

### 3.5 Measures of Position **118**

### 3.6 Box-and-Whisker Plot **123**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Appendix 3.1 / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

## CHAPTER 4 Probability **146**

### 4.1 Experiment, Outcome, and Sample Space **147**

### 4.2 Calculating Probability **152**

### 4.3 Marginal Probability, Conditional Probability, and Related Probability Concepts **158**

Case Study 4–1 Do You Worry About Your Weight? **162**

### 4.4 Intersection of Events and the Multiplication Rule **170**

### 4.5 Union of Events and the Addition Rule **179**

### 4.6 Counting Rule, Factorials, Combinations, and Permutations **187**

Case Study 4–2 Probability of Winning a Mega Millions Lottery Jackpot **192**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

## CHAPTER 5 Discrete Random Variables and Their Probability Distributions **209**

### 5.1 Random Variables **210**

### 5.2 Probability Distribution of a Discrete Random Variable **212**

### 5.3 Mean and Standard Deviation of a Discrete Random Variable **219**

Case Study 5–1 \$1,000 Downpour **221**

### 5.4 The Binomial Probability Distribution **226**

### 5.5 The Hypergeometric Probability Distribution **239**

### 5.6 The Poisson Probability Distribution **242**

Case Study 5–2 Global Birth and Death Rates **246**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

## CHAPTER 6 Continuous Random Variables and the Normal Distribution **264**

### 6.1 Continuous Probability Distribution and the Normal Probability Distribution **265**

Case Study 6–1 Distribution of Time Taken to Run a Road Race **269**

### 6.2 Standardizing a Normal Distribution **281**

### 6.3 Applications of the Normal Distribution **287**

### 6.4 Determining the $z$ and $x$ Values When an Area Under the Normal Distribution Curve Is Known **292**

**6.5 The Normal Approximation to the Binomial Distribution 297**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Appendix 6.1 / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 7 Sampling Distributions 320**

**7.1 Sampling Distribution, Sampling Error, and Nonsampling Errors 321**

**7.2 Mean and Standard Deviation of  $\bar{x}$  326**

**7.3 Shape of the Sampling Distribution of  $\bar{x}$  330**

**7.4 Applications of the Sampling Distribution of  $\bar{x}$  336**

**7.5 Population and Sample Proportions; and Mean, Standard Deviation, and Shape of the Sampling Distribution of  $\hat{p}$  341**

**7.6 Applications of the Sampling Distribution of  $\hat{p}$  348**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 8 Estimation of the Mean and Proportion 360**

**8.1 Estimation, Point Estimate, and Interval Estimate 361**

**8.2 Estimation of a Population Mean:  $\sigma$  Known 364**

*Case Study 8-1* How Much Did Registered Nurses Earn in 2011? 370

**8.3 Estimation of a Population Mean:  $\sigma$  Not Known 374**

**8.4 Estimation of a Population Proportion: Large Samples 383**

*Case Study 8-2* Do You Bring Your Lunch From Home? 386

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 9 Hypothesis Tests About the Mean and Proportion 404**

**9.1 Hypothesis Tests: An Introduction 405**

**9.2 Hypothesis Tests About  $\mu$ :  $\sigma$  Known 413**

*Case Study 9-1* Average Student Debt for the Class of 2010 422

**9.3 Hypothesis Tests About  $\mu$ :  $\sigma$  Not Known 427**

**9.4 Hypothesis Tests About a Population Proportion: Large Samples 437**

*Case Study 9-2* Is Raising Taxes on the Rich Fair? 443

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 10 Estimation and Hypothesis Testing: Two Populations 462**

**10.1 Inferences About the Difference Between Two Population Means for Independent Samples:  $\sigma_1$  and  $\sigma_2$  Known 463**

**10.2 Inferences About the Difference Between Two Population Means for Independent Samples:  $\sigma_1$  and  $\sigma_2$  Unknown but Equal 470**

*Case Study 10-1* One-Way Commute Times For Six Cities 476

**10.3 Inferences About the Difference Between Two Population Means for Independent Samples:  $\sigma_1$  and  $\sigma_2$  Unknown and Unequal 480**

**10.4 Inferences About the Difference Between Two Population Means for Paired Samples 487**

**10.5 Inferences About the Difference Between Two Population Proportions for Large and Independent Samples 496**

**Case Study 10–2 Do You Worry About Your Weight? 501**

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 11 Chi-Square Tests 521**

11.1 The Chi-Square Distribution 522

11.2 A Goodness-of-Fit Test 525

**Case Study 11–1 Are People on Wall Street Honest and Moral? 530**

11.3 A Test of Independence or Homogeneity 534

11.4 Inferences About the Population Variance 546

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 12 Analysis of Variance 566**

12.1 The F Distribution 567

12.2 One-Way Analysis of Variance 569

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 13 Simple Linear Regression 591**

13.1 Simple Linear Regression 592

**Case Study 13–1 Regression of Weights on Heights for NFL Players 601**

13.2 Standard Deviation of Errors and Coefficient of Determination 608

13.3 Inferences About  $B$  614

13.4 Linear Correlation 620

13.5 Regression Analysis: A Complete Example 626

13.6 Using the Regression Model 633

Uses and Misuses / Glossary / Supplementary Exercises / Advanced Exercises / Self-Review Test / Mini-Projects / Decide for Yourself / Technology Instruction / Technology Assignments

**CHAPTER 14 Multiple Regression 651**

This chapter is not included in this text but is available for download from WileyPLUS or from [www.wiley.com/college/mann](http://www.wiley.com/college/mann).

**CHAPTER 15 Nonparametric Methods 652**

This chapter is not included in this text but is available for download from WileyPLUS or from [www.wiley.com/college/mann](http://www.wiley.com/college/mann).

**APPENDIX A Sample Surveys, Sampling Techniques, And Design Of Experiments A1**

A.1 Sources of Data A1

A.2 Sample Surveys and Sampling Techniques A3

A.3 Design of Experiments A9

Advanced Exercises/Glossary

**APPENDIX B Explanation Of Data Sets B1**

Data Set I: City Data B1

- Data Set II: Data on States **B3**  
Data Set III: NFL Data **B3**  
Data Set IV: Beach to Beacon 10k Road Race Data **B3**  
Data Set V: Sample of 500 Observations Selected  
From Beach to Beacon 10k Road Race Data **B4**  
Data Set VI: Data on Movies **B4**  
Data Set VII: Standard & Poor's 100 Index Data **B4**  
Data Set VIII: McDonald's Data **B4**  
Data Set IX: Candidate Data **B5**  
Data Set X: Kickers2010 Data **B6**  
Data Set XI: Billboard Data **B6**  
Data Set XII: Motorcycle Data **B6**  
Data Set XIII: Simulated Data **B7**

**APPENDIX C Statistical Tables** **C1**

- Table I Table of Binomial Probabilities **C2**  
Table II Values of  $e^{-\lambda}$  **C11**  
Table III Table of Poisson Probabilities **C13**  
Table IV Standard Normal Distribution Table **C19**  
Table V The *t* Distribution Table **C21**  
Table VI Chi-Square Distribution Table **C23**  
Table VII The *F* Distribution Table **C24**

*Tables VIII through XII (listed below) are available from WileyPLUS or from [www.wiley.com/college/mann](http://www.wiley.com/college/mann).*

- Table VIII Critical Values of *X* for the Sign Test  
Table IX Critical Values of *T* for the Wilcoxon Signed-Rank Test  
Table X Critical Values of *T* for the Wilcoxon Rank Sum Test  
Table XI Critical Values for the Spearman Rho Rank Correlation Coefficient Test  
Table XII Critical Values for a Two-Tailed Runs Test with *A* = .05

**ANSWERS TO SELECTED ODD-NUMBERED EXERCISES AND SELF REVIEW TESTS** **AN1**

**INDEX** **I1**



© Carol Thacker/Stockphoto

## Introduction<sup>1</sup>

If you are a woman, are you thriving? Or are you struggling? Or, even worse, are you suffering? A global poll of women conducted by Gallup found that while 24% of women in the world are thriving, 63% are struggling and 13% are suffering. (See Case Study 1–2.)

The study of statistics has become more popular than ever over the past four decades or so. The increasing availability of computers and statistical software packages has enlarged the role of statistics as a tool for empirical research. As a result, statistics is used for research in almost all professions, from medicine to sports. Today, college students in almost all disciplines are required to take at least one statistics course. Almost all newspapers and magazines these days contain graphs and stories on statistical studies. After you finish reading this book, it should be much easier to understand these graphs and stories.

Every field of study has its own terminology. Statistics is no exception. This introductory chapter explains the basic terms of statistics. These terms will bridge our understanding of the concepts and techniques presented in subsequent chapters.

### 1.1 Statistics and Types of Statistics

[Case Study 1–1 How Much Did Companies Spend on Ads in 2011?](#)

[Case Study 1–2 How Women Rate Their Lives](#)

### 1.2 Population Versus Sample

[Case Study 1–3 Are We Becoming Less “Green?”](#)

### 1.3 Basic Terms

### 1.4 Types of Variables

### 1.5 Cross-Section Versus Time-Series Data

### 1.6 Sources of Data

### 1.7 Summation Notation

<sup>1</sup>Appendix A discusses many concepts in more detail and introduces many new topics. Interested instructors can combine the coverage of Appendix A with Chapter 1.

## 1.1 Statistics and Types of Statistics

In this section we will learn about statistics and types of statistics.

### 1.1.1 What Is Statistics?

The word **statistics** has two meanings. In the more common usage, *statistics* refers to numerical facts. The numbers that represent the income of a family, the age of a student, the percentage of passes completed by the quarterback of a football team, and the starting salary of a typical college graduate are examples of statistics in this sense of the word. A 1988 article in *U.S. News & World Report* mentioned that “Statistics are an American obsession.”<sup>2</sup> During the 1988 baseball World Series between the Los Angeles Dodgers and the Oakland A’s, the then NBC commentator Joe Garagiola reported to the viewers numerical facts about the players’ performances. In response, fellow commentator Vin Scully said, “I love it when you talk statistics.” In these examples, the word *statistics* refers to numbers.

The following examples present some statistics:

1. In a Pew Research Center survey, about 86% of college graduates said that college was a good investment for them.
2. According to a study by the National Center for Atmospheric Research, expenses related to weather amount to approximately \$485 billion each year in the United States.
3. According to the American Time Use survey, Americans watch TV each weekday for an average of 2.31 hours.
4. About 50% of U.S. adults shopped online on Cyber Monday 2011 (the Monday after Thanksgiving Day 2011). They spent a total of \$1.2 billion.
5. About 35.8 million people visited [walmart.com](http://walmart.com) in June 2011, which is one-third of Amazon’s visitors during the same month.
6. Total compensation (base salary, cash bonuses, perks, stock awards, and option awards) of Viacom’s CEO Philippe P. Dauman was \$84.5 million in 2010.

The second meaning of *statistics* refers to the field or discipline of study. In this sense of the word, *statistics* is defined as follows.

#### Definition

**Statistics** *Statistics* is the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses.

Every day we make decisions that may be personal, business related, or of some other kind. Usually these decisions are made under conditions of uncertainty. Many times, the situations or problems we face in the real world have no precise or definite solution. Statistical methods help us make scientific and intelligent decisions in such situations. Decisions made by using statistical methods are called *educated guesses*. Decisions made without using statistical (or scientific) methods are *pure guesses* and, hence, may prove to be unreliable. For example, opening a large store in an area with or without assessing the need for it may affect its success.

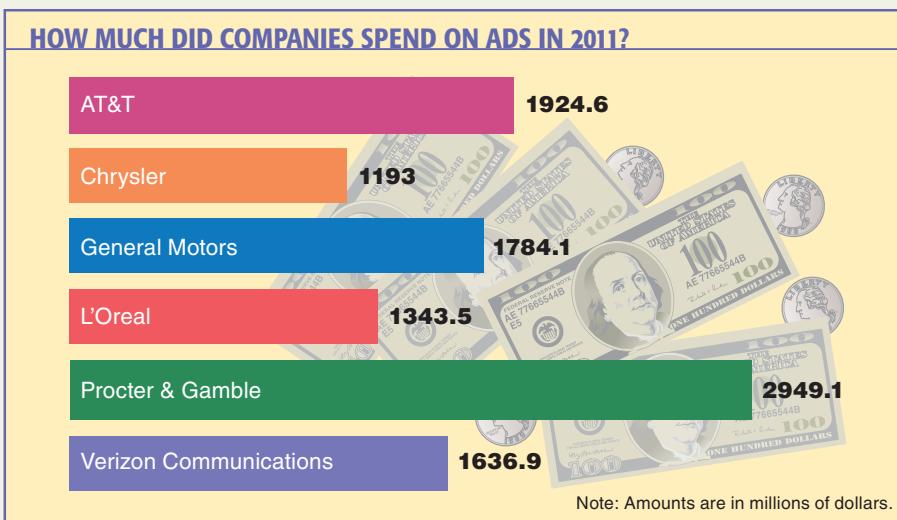
Like almost all fields of study, statistics has two aspects: theoretical and applied. *Theoretical* or *mathematical statistics* deals with the development, derivation, and proof of statistical theorems, formulas, rules, and laws. *Applied statistics* involves the applications of those theorems, formulas, rules, and laws to solve real-world problems. This text is concerned with applied statistics and not with theoretical statistics. By the time you finish studying this book, you will have learned how to think statistically and how to make educated guesses.

### 1.1.2 Types of Statistics

Broadly speaking, applied statistics can be divided into two areas: *descriptive statistics* and *inferential statistics*.

<sup>2</sup>“The Numbers Racket: How Polls and Statistics Lie,” *U.S. News & World Report*, July 11, 1988, pp. 44–47.

## HOW MUCH DID COMPANIES SPEND ON ADS IN 2011?



Data source: WPP Kantar Media.

The accompanying chart shows the expenditures incurred by six companies on advertising in 2011. As the chart shows, AT&T spent \$1924.6 million on advertising in 2011. Of these six companies, Procter & Gamble spent the most on advertising in 2011, \$2949.1 million. This chart describes the data on the 2011 advertising expenditures by these six companies as collected and, hence, is an example of descriptive statistics.

*Data Source:* <http://www.wpp.com/wpp/press/press/default.htm?guid=%7Bf3a07742-eac4-4c92-bb44-f9a8973b014b%7D>.

## Descriptive Statistics

Suppose we have information on the test scores of students enrolled in a statistics class. In statistical terminology, the whole set of numbers that represents the scores of students is called a **data set**, the name of each student is called an **element**, and the score of each student is called an **observation**. (These terms are defined in more detail in Section 1.3.)

Many data sets in their original forms are usually very large, especially those collected by federal and state agencies. Consequently, such data sets are not very helpful in drawing conclusions or making decisions. It is easier to draw conclusions from summary tables and diagrams than from the original version of a data set. So, we reduce data to a manageable size by constructing tables, drawing graphs, or calculating summary measures such as averages. The portion of statistics that helps us do this type of statistical analysis is called **descriptive statistics**.

### Definition

**Descriptive Statistics** *Descriptive statistics* consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.

Chapters 2 and 3 discuss descriptive statistical methods. In Chapter 2, we learn how to construct tables and how to graph data. In Chapter 3, we learn how to calculate numerical summary measures, such as averages.

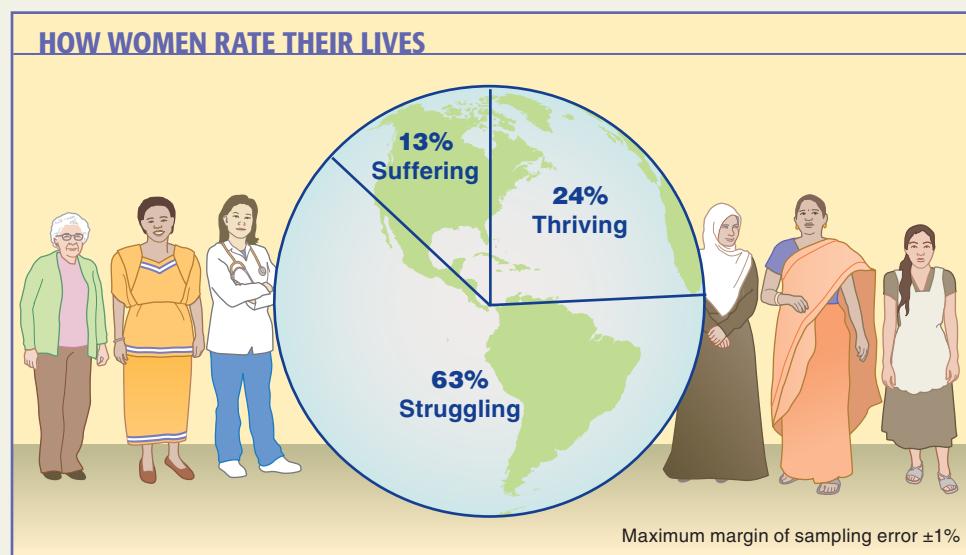
Case Study 1–1 presents an example of descriptive statistics.

## Inferential Statistics

In statistics, the collection of all elements of interest is called a **population**. The selection of a number of elements from this population is called a **sample**. (Population and sample are discussed in more detail in Section 1.2.)

A major portion of statistics deals with making decisions, inferences, predictions, and forecasts about populations based on results obtained from samples. For example, we may make some decisions about the political views of all college and university students based on the political

## HOW WOMEN RATE THEIR LIVES



Data source: Gallup poll of adult women aged 15 and older conducted during 2011 in 147 countries and areas.

During 2011, the Gallup polling agency conducted a poll of 191,317 adults (men and women) of age 15 years and older in 147 countries and areas to find out, among other things, how women rated their lives. As the accompanying chart shows, 24% of these women rated their lives as thriving, 63% said they were struggling, and 13% said they were suffering. As mentioned in the chart, the maximum margin of sampling error was less than  $\pm 1\%$ . In Chapter 8, we will discuss the concept of margin of error, which can be combined with these percentages while making inferences.

**Data Source:** <http://www.gallup.com/poll/155462/Women-Men-Worldwide-Equally-Likely-Thriving.aspx>.

views of 1000 students selected from a few colleges and universities. As another example, we may want to find the starting salary of a typical college graduate. To do so, we may select 2000 recent college graduates, find their starting salaries, and make a decision based on this information. The area of statistics that deals with such decision-making procedures is referred to as **inferential statistics**. This branch of statistics is also called *inductive reasoning* or *inductive statistics*.

### Definition

**Inferential Statistics** *Inferential statistics* consists of methods that use sample results to help make decisions or predictions about a population.

Case Study 1–2 presents an example of inferential statistics. It shows the results of a survey in which people were asked about their feeling toward their jobs.

Chapters 8 through 15 and parts of Chapter 7 deal with inferential statistics.

**Probability**, which gives a measurement of the likelihood that a certain outcome will occur, acts as a link between descriptive and inferential statistics. Probability is used to make statements about the occurrence or nonoccurrence of an event under uncertain conditions. Probability and probability distributions are discussed in Chapters 4 through 6 and parts of Chapter 7.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 1.1 Briefly describe the two meanings of the word *statistics*.
- 1.2 Briefly explain the types of statistics.

## 1.2 Population Versus Sample

We will encounter the terms *population* and *sample* on almost every page of this text.<sup>3</sup> Consequently, understanding the meaning of each of these two terms and the difference between them is crucial.

Suppose a statistician is interested in knowing the following:

1. The percentage of all voters in a city who will vote for a particular candidate in an election
2. Last year's gross sales of all companies in New York City
3. The prices of all houses in California

In these examples, the statistician is interested in *all* voters in a city, *all* companies in New York City, and *all* houses in California. Each of these groups is called the population for the respective example. In statistics, a population does not necessarily mean a collection of people. It can, in fact, be a collection of people or of any kind of item such as houses, books, television sets, or cars. The population of interest is usually called the **target population**.

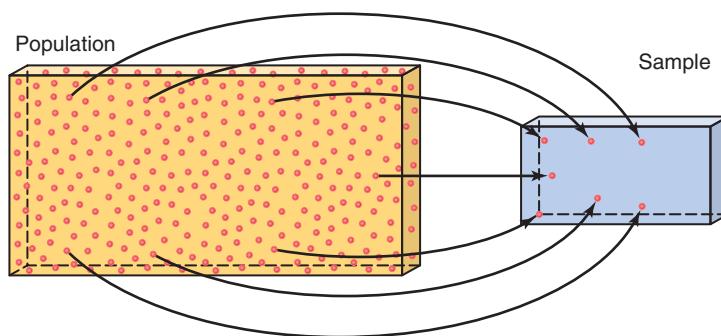
### Definition

**Population or Target Population** A *population* consists of all elements—individuals, items, or objects—whose characteristics are being studied. The population that is being studied is also called the *target population*.

Most of the time, decisions are made based on portions of populations. For example, the election polls conducted in the United States to estimate the percentages of voters who favor various candidates in any presidential election are based on only a few hundred or a few thousand voters selected from across the country. In this case, the population consists of all registered voters in the United States. The sample is made up of a few hundred or few thousand voters who are included in an opinion poll. Thus, the collection of a number of elements selected from a population is called a **sample**. Figure 1.1 illustrates the selection of a sample from a population.

### Definition

**Sample** A portion of the population selected for study is referred to as a *sample*.



**Figure 1.1** Population and sample.

<sup>3</sup>To learn more about sampling and sampling techniques, refer to Appendix A.

The collection of information from the elements of a population or a sample is called a **survey**. A survey that includes every element of the target population is called a **census**. Often the target population is very large. Hence, in practice, a census is rarely taken because it is expensive and time-consuming. In many cases, it is even impossible to identify each element of the target population. Usually, to conduct a survey, we select a sample and collect the required information from the elements included in that sample. We then make decisions based on this sample information. Such a survey conducted on a sample is called a **sample survey**. As an example, if we collect information on the 2011 incomes of all families in Connecticut, it will be referred to as a census. On the other hand, if we collect information on the 2011 incomes of 50 families from Connecticut, it will be called a sample survey. Case Study 1–3 presents an example of a sample survey.

### Definition

**Census and Sample Survey** A survey that includes every member of the population is called a *census*. The technique of collecting information from a portion of the population is called a *sample survey*.

The purpose of conducting a sample survey is to make decisions about the corresponding population. It is important that the results obtained from a sample survey closely match the results that we would obtain by conducting a census. Otherwise, any decision based on a sample survey will not apply to the corresponding population. As an example, to find the average income of families living in New York City by conducting a sample survey, the sample must contain families who belong to different income groups in almost the same proportion as they exist in the population. Such a sample is called a **representative sample**. Inferences derived from a representative sample will be more reliable.

### Definition

**Representative Sample** A sample that represents the characteristics of the population as closely as possible is called a *representative sample*.

A sample may be random or nonrandom. In a **random sample**, each element of the population has a chance of being included in the sample. However, in a nonrandom sample this may not be the case.

### Definition

**Random Sample** A sample drawn in such a way that each element of the population has a chance of being selected is called a *random sample*. If all samples of the same size selected from a population have the same chance of being selected, we call it **simple random sampling**. Such a sample is called a **simple random sample**.

One way to select a random sample is by lottery or draw. For example, if we are to select 5 students from a class of 50, we write each of the 50 names on a separate piece of paper. Then we place all 50 slips in a box and mix them thoroughly. Finally, we randomly draw 5 slips from the box. The 5 names drawn give a random sample. On the other hand, if we arrange all 50 names alphabetically and then select the first 5 names on the list, it is a non-random sample because the students listed 6th to 50th have no chance of being included in the sample.

A sample may be selected with or without replacement. In sampling **with replacement**, each time we select an element from the population, we put it back in the population before we

## ARE WE BECOMING LESS "GREEN?"

Americans may be becoming less "green" according to a March 2012 online poll conducted by Harris Interactive (<http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/ctl/ReadCustom%20Default/mid/1508/ArticleId/1009/Default.aspx>). Conducted online between March 12 and March 19, 2012, the survey asked 2451 American adults of age 18 years and older about their "green" attitudes and behavior. Similar surveys were conducted by Harris Interactive in 2009 and 2010.

The 2012 poll found that Americans were engaged in fewer environmental-friendly activities compared to 2009 and 2010. For example, fewer survey respondents in 2012 said that they made an effort to use less water, purchase food in bulk, and buy organic and all-natural products. Specifically, 61% of the adults surveyed in 2012 said that they reused items instead of "throwing them away or buying new items," whereas this percentage was 65% in 2009.

American adults polled in 2012 also described themselves as less "environmentally conscious"

compared to respondents in 2009 and 2010. Only 34% of adults polled in 2012 said that they were concerned about the planet they were leaving behind for "future generations." In 2009, 43% said that they were concerned about the planet.

According to the Harris report, a possible explanation for this behavior may be that people were paying less attention to "green" and environmental issues in 2012 because there were other "pressing issues" such as the bad economic conditions and the upcoming election.

Some of the other findings from this 2012 poll included the following: 68% of the adults polled said that they always/often recycle, 36% always/often buy locally grown produce, 17% always/often carpool or use public transportation, 15% always/often buy organic products, 57% always/often make an effort to use less water, and 20% describe themselves as conservationists, 17% as "green," and 16% as environmentalists.

*Data Source:* <http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/ctl/ReadCustom%20Default/mid/1508/ArticleId/1009/Default.aspx>.

select the next element. Thus, in sampling with replacement, the population contains the same number of items each time a selection is made. As a result, we may select the same item more than once in such a sample. Consider a box that contains 25 marbles of different colors. Suppose we draw a marble, record its color, and put it back in the box before drawing the next marble. Every time we draw a marble from this box, the box contains 25 marbles. This is an example of sampling with replacement. The experiment of rolling a die many times is another example of sampling with replacement because every roll has the same six possible outcomes.

Sampling **without replacement** occurs when the selected element is not replaced in the population. In this case, each time we select an item, the size of the population is reduced by one element. Thus, we cannot select the same item more than once in this type of sampling. Most of the time, samples taken in statistics are without replacement. Consider an opinion poll based on a certain number of voters selected from the population of all eligible voters. In this case, the same voter is not selected more than once. Therefore, this is an example of sampling without replacement.

## EXERCISES

### CONCEPTS AND PROCEDURES

**1.3** Briefly explain the terms *population*, *sample*, *representative sample*, *random sample*, *sampling with replacement*, and *sampling without replacement*.

**1.4** Give one example each of sampling with and sampling without replacement.

**1.5** Briefly explain the difference between a census and a sample survey. Why is conducting a sample survey preferable to conducting a census?

## ■ APPLICATIONS

- 1.6** Explain whether each of the following constitutes data collected from a population or a sample.
- Opinions on a certain issue obtained from all adults living in a city.
  - The price of a gallon of regular unleaded gasoline on a given day at each of 28 gas stations in the Miami, Florida, metropolitan area.
  - Credit card debts of 100 families selected from a given city.
  - The percentage of all U.S. registered voters in each state who voted in the 2012 Presidential election.
  - The number of left-handed students in each of 50 classes selected from a given university.
- 1.7** Explain whether each of the following constitutes data collected from a population or a sample.
- The number of pizzas ordered on Fridays during 2012 at all of the pizza parlors in your town.
  - The dollar values of auto insurance claims filed in 2012 for 200 randomly selected policies.
  - The opening price of each of the 500 stocks in the S&P 500 stock index on January 3, 2012.
  - The total home attendance for each of the 18 teams in Major League Soccer during the 2012 season.
  - The living areas of 35 houses listed for sale on March 7, 2012 in Chicago, Illinois.

## 1.3 Basic Terms

It is very important to understand the meaning of some basic terms that will be used frequently in this text. This section explains the meaning of an element (or member), a variable, an observation, and a data set. An element and a data set were briefly defined in Section 1.1. This section defines these terms formally and illustrates them with the help of an example.

Table 1.1 gives information on the total revenues (in millions of U.S. dollars) for the year 2010 of the top six revenue-earning companies in the world. We can call this group of companies a sample of six companies. (Note that it is not a random sample.) Each company listed in this table is called an **element** or a **member** of the sample. Table 1.1 contains information on six elements. Note that elements are also called *observational units*.

### Definition

**Element or Member** An *element* or *member* of a sample or population is a specific subject or object (for example, a person, firm, item, state, or country) about which the information is collected.

**Table 1.1** Total Revenues for 2010 of Six Companies

Company	2010 Total Revenue (millions of dollars)	← Variable
Wal-Mart Stores	421,849	
Royal Dutch Shell	378,152	
An element } or a member → Exxon Mobil	354,674	← { An observation or measurement
BP	308,928	
Sinopec Group	273,422	
China National Petroleum	240,192	

*Source:* Fortune Magazine, July 25, 2011.

The 2010 revenue in our example is called a variable. The 2010 revenue is a characteristic of companies on which we are collecting information.

**Definition**

**Variable** A *variable* is a characteristic under study that assumes different values for different elements. In contrast to a variable, the value of a *constant* is fixed.

A few other examples of variables are household incomes, the number of houses built in a city per month during the past year, the makes of cars owned by people, the gross profits of companies, and the number of insurance policies sold by a salesperson per day during the past month.

In general, a variable assumes different values for different elements, as do the 2010 revenues for the six companies in Table 1.1. For some elements in a data set, however, the values of the variable may be the same. For example, if we collect information on incomes of households, these households are expected to have different incomes, although some of them may have the same income.

A variable is often denoted by  $x$ ,  $y$ , or  $z$ . For instance, in Table 1.1, the 2010 revenue for companies may be denoted by any one of these letters. Starting with Section 1.7, we will begin to use these letters to denote variables.

Each of the values representing the 2010 revenues of the six companies in Table 1.1 is called an **observation** or **measurement**.

**Definition**

**Observation or Measurement** The value of a variable for an element is called an *observation* or *measurement*.

From Table 1.1, the 2010 revenue of Exxon Mobil was \$354,674 million. The value \$354,674 million is an observation or a measurement. Table 1.1 contains six observations, one for each of the six companies.

The information given in Table 1.1 on the 2010 revenues of companies is called the **data** or a **data set**.

**Definition**

**Data Set** A *data set* is a collection of observations on one or more variables.

Other examples of data sets are a list of the prices of 25 recently sold homes, test scores of 15 students, opinions of 100 voters, and ages of all employees of a company.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 1.8 Explain the meaning of an element, a variable, an observation, and a data set.

### ■ APPLICATIONS

- 1.9 The following table gives the number of dog bites reported to the police last year in six cities.

City	Number of Bites
Center City	47
Elm Grove	32
Franklin	51
Bay City	44
Oakdale	12
Sand Point	3

Briefly explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

**1.10** The following table lists the number of billionaires in eight countries as of February 2011, as reported in *The New York Times* of July 27, 2011.

Country	Number of Billionaires
United States	413
China	115
Russia	101
India	55
Germany	52
Britain	32
Brazil	30
Japan	26

*Source:* *Forbes*, International Monetary Fund.

Briefly explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

**1.11** Refer to the data set in Exercise 1.9.

- a. What is the variable for this data set?
- b. How many observations are in this data set?
- c. How many elements does this data set contain?

**1.12** Refer to the data set in Exercise 1.10.

- a. What is the variable for this data set?
- b. How many observations are in this data set?
- c. How many elements does this data set contain?

## 1.4 Types of Variables

In Section 1.3, we learned that a variable is a characteristic under investigation that assumes different values for different elements. Family income, height of a person, gross sales of a company, price of a college textbook, make of the car owned by a family, number of accidents, and status (freshman, sophomore, junior, or senior) of a student enrolled at a university are examples of variables.

A variable may be classified as quantitative or qualitative. These two types of variables are explained next.

### 1.4.1 Quantitative Variables

Some variables (such as the price of a home) can be measured numerically, whereas others (such as hair color) cannot. The first is an example of a **quantitative variable** and the second of a qualitative variable.

#### Definition

**Quantitative Variable** A variable that can be measured numerically is called a *quantitative variable*. The data collected on a quantitative variable are called *quantitative data*.

Incomes, heights, gross sales, prices of homes, number of cars owned, and number of accidents are examples of quantitative variables because each of them can be expressed numerically.

For instance, the income of a family may be \$81,520.75 per year, the gross sales for a company may be \$567 million for the past year, and so forth. Such quantitative variables may be classified as either *discrete variables* or *continuous variables*.

## Discrete Variables

The values that a certain quantitative variable can assume may be countable or noncountable. For example, we can count the number of cars owned by a family, but we cannot count the height of a family member. A variable that assumes countable values is called a **discrete variable**. Note that there are no possible intermediate values between consecutive values of a discrete variable.

### Definition

**Discrete Variable** A variable whose values are countable is called a *discrete variable*. In other words, a discrete variable can assume only certain values with no intermediate values.

For example, the number of cars sold on any given day at a car dealership is a discrete variable because the number of cars sold must be 0, 1, 2, 3, . . . and we can count it. The number of cars sold cannot be between 0 and 1, or between 1 and 2. Other examples of discrete variables are the number of people visiting a bank on any day, the number of cars in a parking lot, the number of cattle owned by a farmer, and the number of students in a class.

## Continuous Variables

Some variables cannot be counted, and they can assume any numerical value between two numbers. Such variables are called **continuous variables**.

### Definition

**Continuous Variable** A variable that can assume any numerical value over a certain interval or intervals is called a *continuous variable*.

The time taken to complete an examination is an example of a continuous variable because it can assume any value, let us say, between 30 and 60 minutes. The time taken may be 42.6 minutes, 42.67 minutes, or 42.674 minutes. (Theoretically, we can measure time as precisely as we want.) Similarly, the height of a person can be measured to the tenth of an inch or to the hundredth of an inch. However, neither time nor height can be counted in a discrete fashion. Other examples of continuous variables are weights of people, amount of soda in a 12-ounce can (note that a can does not contain exactly 12 ounces of soda), and yield of potatoes (in pounds) per acre. Note that any variable that involves money and can assume a large number of values is typically treated as a continuous variable.

## 1.4.2 Qualitative or Categorical Variables

Variables that cannot be measured numerically but can be divided into different categories are called **qualitative** or **categorical variables**.

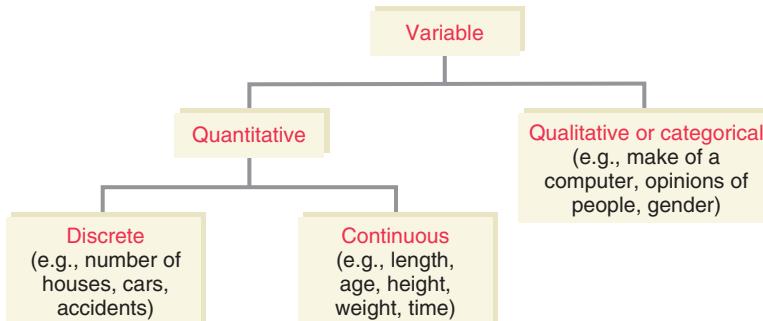
### Definition

**Qualitative or Categorical Variable** A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a *qualitative* or *categorical variable*. The data collected on such a variable are called *qualitative data*.

For example, the status of an undergraduate college student is a qualitative variable because a student can fall into any one of four categories: freshman, sophomore, junior, or senior. Other

examples of qualitative variables are the gender of a person, the brand of a computer, the opinions of people, and the make of a car.

Figure 1.2 illustrates the types of variables.



**Figure 1.2** Types of variables.

## EXERCISES

### CONCEPTS AND PROCEDURES

**1.13** Explain the meaning of the following terms.

- a. Quantitative variable
- b. Qualitative variable
- c. Discrete variable
- d. Continuous variable
- e. Quantitative data
- f. Qualitative data

### APPLICATIONS

**1.14** Indicate which of the following variables are quantitative and which are qualitative.

- a. Number of persons in a family
- b. Color of a car
- c. Marital status of a person
- d. Time to commute from home to work
- e. Number of errors in a person's credit report

**1.15** Indicate which of the following variables are quantitative and which are qualitative.

- a. The amount of time a student spent studying for an exam
- b. The amount of rain last year in 30 cities
- c. The arrival status of an airline flight (early, on time, late, canceled) at an airport
- d. A person's blood type
- e. The amount of gasoline put into a car at a gas station

**1.16** Classify the quantitative variables in Exercise 1.14 as discrete or continuous.

**1.17** Classify the quantitative variables in Exercise 1.15 as discrete or continuous.

## 1.5 Cross-Section Versus Time-Series Data

Based on the time over which they are collected, data can be classified as either cross-section or time-series data.

### 1.5.1 Cross-Section Data

**Cross-section data** contain information on different elements of a population or sample for the *same* period of time. The information on incomes of 100 families for 2012 is an example of cross-section data. All examples of data already presented in this chapter have been cross-section data.

#### Definition

**Cross-Section Data** Data collected on different elements at the same point in time or for the same period of time are called *cross-section data*.

Table 1.1 is reproduced here as Table 1.2 and shows the 2010 revenues of the six top revenue-earning companies in the world. Because this table presents data on the revenues of six companies for the same period (2010), it is an example of cross-section data.

**Table 1.2 Total Revenues for 2010 of Six Companies**

Company	2010 Total Revenue (millions of dollars)
Wal-Mart Stores	421,849
Royal Dutch Shell	378,152
Exxon Mobil	354,674
BP	308,928
Sinopec Group	273,422
China National Petroleum	240,192

Source: Fortune Magazine, July 25, 2011.

### 1.5.2 Time-Series Data

**Time-series data** contain information on the same element at *different* points in time. Information on U.S. exports for the years 1983 to 2012 is an example of time-series data.

#### Definition

**Time-Series Data** Data collected on the same element for the same variable at different points in time or for different periods of time are called *time-series data*.

The data given in Table 1.3 are an example of time-series data. This table lists the money recovered by federal agents from health care fraud judgments during the budget years 2006 to 2010 (*Source*: The U.S. Department of Health and Human Services and the U.S. Department of Justice).

**Table 1.3 Money Recovered from Health Care Fraud Judgments**

Year	Money Recovered (billions of dollars)
2006	2.2
2007	1.8
2008	1.0
2009	1.6
2010	2.5

## 1.6 Sources of Data

The availability of accurate and appropriate data is essential for deriving reliable results.<sup>4</sup> Data may be obtained from internal sources, external sources, or surveys and experiments.

Many times data come from *internal sources*, such as a company's personnel files or accounting records. For example, a company that wants to forecast the future sales of its product may use the data of past periods from its records. For most studies, however, all the data that are needed are not usually available from internal sources. In such cases, one may have to depend on outside sources to obtain data. These sources are called *external sources*. For instance, the *Statistical Abstract of the United States* (published annually), which contains various kinds of data on the United States, is an external source of data.

A large number of government and private publications can be used as external sources of data. The following is a list of some government publications.

1. *Statistical Abstract of the United States*
2. *Employment and Earnings*
3. *Handbook of Labor Statistics*
4. *Source Book of Criminal Justice Statistics*
5. *Economic Report of the President*
6. *County & City Data Book*
7. *State & Metropolitan Area Data Book*
8. *Digest of Education Statistics*
9. *Health United States*
10. *Agricultural Statistics*

Most of the data contained in these books can be accessed on Internet sites such as [www.census.gov](http://www.census.gov) (Census Bureau), [www.bls.gov](http://www.bls.gov) (Bureau of Labor Statistics), [www.ojp.usdoj.gov/bjs](http://www.ojp.usdoj.gov/bjs) (Office of Justice Program, U.S. Department of Justice, Bureau of Justice Statistics), [www.os.dhhs.gov](http://www.os.dhhs.gov) (U.S. Department of Health and Human Services), and [www.usda.gov/nass/pubs/agstats.htm](http://www.usda.gov/nass/pubs/agstats.htm) (U.S. Department of Agriculture, Agricultural Statistics).

Besides these government publications, a large number of private publications (e.g., *Standard & Poor's Security Owner's Stock Guide* and *World Almanac and Book of Facts*) and periodicals (e.g., *The Wall Street Journal*, *USA TODAY*, *Fortune*, *Forbes*, and *Bloomberg Businessweek*) can be used as external data sources.

Sometimes the needed data may not be available from either internal or external sources. In such cases, the investigator may have to conduct a survey or experiment to obtain the required data. Appendix A discusses surveys and experiments in detail.

### EXERCISES

#### ■ CONCEPTS AND PROCEDURES

**1.18** Explain the difference between cross-section and time-series data. Give an example of each of these two types of data.

**1.19** Briefly describe internal and external sources of data.

#### ■ APPLICATIONS

**1.20** Classify the following as cross-section or time-series data.

- a. Food bill of a family for each month of 2012
- b. Number of armed robberies each year in Dallas from 1998 to 2012

<sup>4</sup>Sources of data are discussed in more detail in Appendix A.

- c. Number of supermarkets in 40 cities on December 31, 2011
- d. Gross sales of 200 ice cream parlors in July 2012

**1.21** Classify the following as cross-section or time-series data.

- a. Average prices of houses in 100 cities
- b. Salaries of 50 employees
- c. Number of cars sold each year by General Motors from 1980 to 2012
- d. Number of employees employed by a company each year from 1985 to 2012

## 1.7 Summation Notation

Sometimes mathematical notation helps express a mathematical relationship concisely. This section describes the **summation notation** that is used to denote the sum of values.

Suppose a sample consists of five books, and the prices of these five books are \$175, \$80, \$165, \$97, and \$88, respectively. The variable *price of a book* can be denoted by  $x$ . The prices of the five books can be written as follows:

$$\begin{aligned} \text{Price of the first book} &= x_1 = \$175 \\ &\quad \uparrow \\ &\quad \text{Subscript of } x \text{ denotes the} \\ &\quad \text{number of the book} \end{aligned}$$

Similarly,

$$\begin{aligned} \text{Price of the second book} &= x_2 = \$80 \\ \text{Price of the third book} &= x_3 = \$165 \\ \text{Price of the fourth book} &= x_4 = \$97 \\ \text{Price of the fifth book} &= x_5 = \$88 \end{aligned}$$

In this notation,  $x$  represents the price, and the subscript denotes a particular book.

Now, suppose we want to add the prices of all five books. We obtain

$$x_1 + x_2 + x_3 + x_4 + x_5 = 175 + 80 + 165 + 97 + 88 = \$605$$

The uppercase Greek letter  $\Sigma$  (pronounced *sigma*) is used to denote the sum of all values. Using  $\Sigma$  notation, we can write the foregoing sum as follows:

$$\Sigma x = x_1 + x_2 + x_3 + x_4 + x_5 = \$605$$

The notation  $\Sigma x$  in this expression represents the sum of all values of  $x$  and is read as “sigma  $x$ ” or “sum of all values of  $x$ .”

### ■ EXAMPLE 1-1

Annual salaries (in thousands of dollars) of four workers are 75, 90, 125, and 61, respectively. Find

- (a)  $\Sigma x$
- (b)  $(\Sigma x)^2$
- (c)  $\Sigma x^2$

**Solution** Let  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  be the annual salaries (in thousands of dollars) of the first, second, third, and fourth worker, respectively. Then,

$$x_1 = 75, \quad x_2 = 90, \quad x_3 = 125, \quad \text{and} \quad x_4 = 61$$

- (a)  $\Sigma x = x_1 + x_2 + x_3 + x_4 = 75 + 90 + 125 + 61 = 351 = \$351,000$
- (b) Note that  $(\Sigma x)^2$  is the square of the sum of all  $x$  values. Thus,

$$(\Sigma x)^2 = (351)^2 = 123,201$$

Using summation notation:  
one variable.



© Troels Graugaard/iStockphoto

- (c) The expression  $\Sigma x^2$  is the sum of the squares of  $x$  values. To calculate  $\Sigma x^2$ , we first square each of the  $x$  values and then sum these squared values. Thus,

$$\begin{aligned}\Sigma x^2 &= (75)^2 + (90)^2 + (125)^2 + (61)^2 \\ &= 5625 + 8100 + 15,625 + 3721 = 33,071\end{aligned}$$



## ■ EXAMPLE 1–2

*Using summation notation:  
two variables.*

The following table lists four pairs of  $m$  and  $f$  values:

$m$	12	15	20	30
$f$	5	9	10	16

Compute the following:

- (a)  $\Sigma m$     (b)  $\Sigma f^2$     (c)  $\Sigma mf$     (d)  $\Sigma m^2f$

**Solution** We can write

$$\begin{array}{llll}m_1 = 12 & m_2 = 15 & m_3 = 20 & m_4 = 30 \\ f_1 = 5 & f_2 = 9 & f_3 = 10 & f_4 = 16\end{array}$$

- (a)  $\Sigma m = 12 + 15 + 20 + 30 = 77$   
 (b)  $\Sigma f^2 = (5)^2 + (9)^2 + (10)^2 + (16)^2 = 25 + 81 + 100 + 256 = 462$   
 (c) To compute  $\Sigma mf$ , we multiply the corresponding values of  $m$  and  $f$  and then add the products as follows:

$$\begin{aligned}\Sigma mf &= m_1f_1 + m_2f_2 + m_3f_3 + m_4f_4 \\ &= 12(5) + 15(9) + 20(10) + 30(16) = 875\end{aligned}$$

- (d) To calculate  $\Sigma m^2f$ , we square each  $m$  value, then multiply the corresponding  $m^2$  and  $f$  values, and add the products. Thus,

$$\begin{aligned}\Sigma m^2f &= (m_1)^2f_1 + (m_2)^2f_2 + (m_3)^2f_3 + (m_4)^2f_4 \\ &= (12)^2(5) + (15)^2(9) + (20)^2(10) + (30)^2(16) = 21,145\end{aligned}$$

The calculations done in parts (a) through (d) to find the values of  $\Sigma m$ ,  $\Sigma f^2$ ,  $\Sigma mf$ , and  $\Sigma m^2f$  can be performed in tabular form, as shown in Table 1.4.

**Table 1.4**

$m$	$f$	$f^2$	$mf$	$m^2f$
12	5	$5 \times 5 = 25$	$12 \times 5 = 60$	$12 \times 12 \times 5 = 720$
15	9	$9 \times 9 = 81$	$15 \times 9 = 135$	$15 \times 15 \times 9 = 2025$
20	10	$10 \times 10 = 100$	$20 \times 10 = 200$	$20 \times 20 \times 10 = 4000$
30	16	$16 \times 16 = 256$	$30 \times 16 = 480$	$30 \times 30 \times 16 = 14,400$
$\Sigma m = 77$	$\Sigma f = 40$	$\Sigma f^2 = 462$	$\Sigma mf = 875$	$\Sigma m^2f = 21,145$

The columns of Table 1.4 can be explained as follows.

1. The first column lists the values of  $m$ . The sum of these values gives  $\Sigma m = 77$ .
2. The second column lists the values of  $f$ . The sum of this column gives  $\Sigma f = 40$ .

3. The third column lists the squares of the  $f$  values. For example, the first value, 25, is the square of 5. The sum of the values in this column gives  $\sum f^2 = 462$ .
4. The fourth column records products of the corresponding  $m$  and  $f$  values. For example, the first value, 60, in this column is obtained by multiplying 12 by 5. The sum of the values in this column gives  $\sum mf = 875$ .
5. Next, the  $m$  values are squared and multiplied by the corresponding  $f$  values. The resulting products, denoted by  $m^2f$ , are recorded in the fifth column. For example, the first value, 720, is obtained by squaring 12 and multiplying this result by 5. The sum of the values in this column gives  $\sum m^2f = 21,145$ . ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**1.22** The following table lists five pairs of  $m$  and  $f$  values.

$m$	5	10	17	20	25
$f$	12	8	6	16	4

Compute the value of each of the following:

a.  $\sum m$     b.  $\sum f^2$     c.  $\sum mf$     d.  $\sum m^2f$

**1.23** The following table lists six pairs of  $m$  and  $f$  values.

$m$	3	6	25	12	15	18
$f$	16	11	16	8	4	14

Calculate the value of each of the following:

a.  $\sum f$     b.  $\sum m^2$     c.  $\sum mf$     d.  $\sum m^2f$

**1.24** The following table contains information on the NCAA Men's Basketball Championship Tournament Final Four teams for the 33-year period from 1979 to 2011. The table shows how many teams with each seeding qualified for the Final Four during these 33 years. For example, 54 of the 132 Final Four teams were seeded number 1, 28 of the 132 Final Four teams were seeded number 2, and so on.

<b>Seed</b>	1	2	3	4	5	6	7	8	9	11
<b>Number of Teams in Men's Final Four</b>	54	28	16	11	7	6	1	5	1	3

Letting  $y$  denote the seed and  $x$  denote the number of teams having that seed, calculate the following:

a.  $\sum x$     b.  $\sum y$     c.  $\sum xy$     d.  $\sum y^2$     e.  $(\sum y)^2$

**1.25** The following table contains the same kind of information as the table in Exercise 1.24 but for the NCAA Women's Basketball Championship Tournament Final Four teams for the 30-year period from 1982 to 2011.

<b>Seed</b>	1	2	3	4	5	6	7	8	9
<b>Number of Teams in Women's Final Four</b>	63	29	13	9	1	2	1	1	1

Letting  $y$  denote the seed and  $x$  denote the number of teams having that seed, calculate the following:

a.  $\sum x$     b.  $\sum y$     c.  $\sum xy$     d.  $\sum y^2$     e.  $(\sum y)^2$

## ■ APPLICATIONS

**1.26** Eight randomly selected customers at a local grocery store spent the following amounts on groceries in a single visit: \$216, \$184, \$35, \$92, \$144, \$175, \$11, and \$57, respectively. Let  $y$  denote the amount spent by a customer on groceries in a single visit. Find:

- a.  $\Sigma y$       b.  $(\Sigma y)^2$       c.  $\Sigma y^2$

**1.27** The number of pizzas delivered to a college campus on six randomly selected nights is 48, 103, 95, 188, 286, and 136, respectively. Let  $x$  denote the number of pizzas delivered to this college campus on any given night. Find:

- a.  $\Sigma x$       b.  $(\Sigma x)^2$       c.  $\Sigma x^2$

**1.28** Nine randomly selected customers at a local fast-food restaurant ordered meals having the following calorie counts: 975, 520, 1560, 872, 1105, 437, 910, 785, and 1335. Let  $y$  denote the calorie content of a meal ordered at this restaurant. Find:

- a.  $\Sigma y$       b.  $(\Sigma y)^2$       c.  $\Sigma y^2$

**1.29** A car was filled with 16 gallons of gas on seven occasions. The number of miles that the car was able to travel on each tankful was 387, 414, 404, 396, 410, 422, and 414. Let  $x$  denote the distance traveled on 16 gallons of gas. Find:

- a.  $\Sigma x$       b.  $(\Sigma x)^2$       c.  $\Sigma x^2$

## USES AND MISUSES...

## SPEAKING THE LANGUAGE OF STATISTICS

Have you ever heard a statement like "The average American family has .90 children?" What is wrong with this statement, and how can we fix it? How about, "In a representative sample of 20 American families, one can expect 18 children." The statement is wordy, but more accurate. Why do we care?

Statisticians pay close attention to definitions because, without them, calculations would be impossible to make and interpretations of the data would be meaningless. Often, when you read statistics reported in a newspaper, the journalist or editor sometimes chooses to describe the results in a way that is easier to understand but that distorts the actual statistical result.

Let us pick apart our example. The word *average* has a very specific meaning in probability (Chapters 4 and 5). The intended meaning of the word here really is *typical*. The adjective *American* helps us define the population. The Census Bureau defines *family* as "a group of two or more people (one of whom is the householder) related by birth, marriage, or adoption and residing together; all such people (including related subfamily members) are considered as members of one family." It defines *children* as "all persons under 18 years, excluding people who maintain households, families, or subfamilies as a reference person or spouse." We understand implicitly that a family cannot have a fractional number of children, so we accept that this discrete variable takes on the properties of a continuous variable when we are talking about the characteristics of a large

population. How large does the population need to be before we can derive continuous variables from discrete variables? The answer comes in the chapters that follow.

When people hear this statistic, a common reaction is, "How can this be true? Every family that has children has one or more children. How could the average be .90?" Once again, it is important to recognize what is being measured. The statistic described in the previous paragraphs includes many families that have no children, families whose children have grown up and moved out, and families whose children are still living at home but are at least 18 years old. All such families are counted in the average, and they lower the average because they either have no children or their children are 18 or older and hence are not included in the calculation of the average. Also note that this average is for families and not for households.

The moral of the story is that whenever you read a statistical result, be sure that you understand the definitions of the terms used to describe the result and relate those terms to the definitions that you already know. In some cases *year* is a categorical variable, in others it is a discrete variable, and in others it is a continuous variable. Many surveys will report that "respondents feel better, the same, or worse" about a particular subject. Although *better*, *same*, and *worse* have a natural order to them, they do not have numerical values.

## Glossary

**Census** A survey that includes all members of the population.

**Continuous variable** A (quantitative) variable that can assume any numerical value over a certain interval or intervals.

**Cross-section data** Data collected on different elements at the same point in time or for the same period of time.

**Data or data set** Collection of observations or measurements on a variable.

**Descriptive statistics** Collection of methods for organizing, displaying, and describing data using tables, graphs, and summary measures.

**Discrete variable** A (quantitative) variable whose values are countable.

**Element or member** A specific subject or object included in a sample or population.

**Inferential statistics** Collection of methods that help make decisions about a population based on sample results.

**Observation or measurement** The value of a variable for an element.

**Population or target population** The collection of all elements whose characteristics are being studied.

**Qualitative or categorical data** Data generated by a qualitative variable.

**Qualitative or categorical variable** A variable that cannot assume numerical values but is classified into two or more categories.

**Quantitative data** Data generated by a quantitative variable.

**Quantitative variable** A variable that can be measured numerically.

**Random sample** A sample drawn in such a way that each element of the population has some chance of being included in the sample.

**Representative sample** A sample that contains the same characteristics as the corresponding population.

**Sample** A portion of the population of interest.

**Sample survey** A survey that includes elements of a sample.

**Simple random sampling** If all samples of the same size selected from a population have the same chance of being selected, it is called simple random sampling. Such a sample is called a simple random sample.

**Statistics** Science of collecting, analyzing, presenting, and interpreting data and making decisions.

**Survey** Collection of data on the elements of a population or sample.

**Time-series data** Data that give the values of the same variable for the same element at different points in time or for different periods of time.

**Variable** A characteristic under study or investigation that assumes different values for different elements.

## Supplementary Exercises

**1.30** The following table lists the number of reports filed with the U.S. Department of Transportation about mishandled baggage during the first nine months of 2010, as reported in the *USA TODAY* of July 14, 2011.

Airline	Mishandled Baggage Reports
AirTran	30,801
Alaska	36,525
American	205,247
Delta	247,660
JetBlue	41,174
Hawaiian	11,987

Explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

**1.31** The following table lists the total compensations (base salary, cash bonus, perks, stock awards, and option awards) of eight CEOs with the highest total compensations for the year 2010 as reported in *The New York Times* of April 10, 2011.

CEO	Total Compensation (millions of dollars)
Philippe P. Dauman (Viacom)	84.5
Ray R. Irani (Occidental)	76.1
Lawrence J. Ellison (Oracle)	70.1
Michael D. White (DirecTV)	32.9
John F. Lundgren (Stanley Black & Decker)	32.6
Brian L. Roberts (Comcast)	28.2
Robert A. Iger (Walt Disney)	28.0
Alan Mulally (Ford Motor)	26.5
Samuel J. Palmisano (IBM)	25.2

Source: Compiled by the research firm Equilar.

Explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

**1.32** Refer to Exercises 1.30 and 1.31. Classify these data sets as either cross-section or time-series data.

- 1.33** Indicate whether each of the following constitutes data collected from a population or a sample.
- A group of 25 patients selected to test a new drug
  - Total items produced on a machine for each year from 1995 to 2012
  - Yearly expenditures on clothes for 50 persons
  - Number of houses sold by each of the 10 employees of a real estate agency during 2012
- 1.34** Indicate whether each of the following constitutes data collected from a population or a sample.
- Salaries of CEOs of all companies in New York City
  - Five hundred houses selected from a city
  - Gross sales for 2012 of four fast-food chains
  - Annual incomes of all 33 employees of a restaurant
- 1.35** State which of the following is an example of sampling with replacement and which is an example of sampling without replacement.
- Selecting 10 patients out of 100 to test a new drug
  - Selecting one professor to be a member of the university senate and then selecting one professor from the same group to be a member of the curriculum committee
- 1.36** State which of the following is an example of sampling with replacement and which is an example of sampling without replacement.
- Selecting seven cities to market a new deodorant
  - Selecting a high school teacher to drive students to a lecture in March, then selecting a teacher from the same group to chaperone a dance in April
- 1.37** The number of shoe pairs owned by six women is 8, 14, 3, 7, 10, and 5, respectively. Let  $x$  denote the number of shoe pairs owned by a woman. Find:
- $\Sigma x$
  - $(\Sigma x)^2$
  - $\Sigma x^2$
- 1.38** The number of restaurants in each of five small towns is 4, 12, 8, 10, and 5, respectively. Let  $y$  denote the number of restaurants in a small town. Find:
- $\Sigma y$
  - $(\Sigma y)^2$
  - $\Sigma y^2$
- 1.39** The following table lists five pairs of  $m$  and  $f$  values.

$m$	3	16	11	9	20
$f$	7	32	17	12	34

Compute the value of each of the following:

- $\Sigma m$
- $\Sigma f^2$
- $\Sigma mf$
- $\Sigma m^2f$
- $\Sigma m^2$

- 1.40** The following table lists six pairs of  $x$  and  $y$  values.

$x$	7	11	8	4	14	28
$y$	5	15	7	10	9	19

Compute the value of each of the following:

- $\Sigma y$
- $\Sigma x^2$
- $\Sigma xy$
- $\Sigma x^2y$
- $\Sigma y^2$

- 1.41** A sports economist conducted a study to determine the impact of a variety of variables on the salaries of rookie National Football League (NFL) players who were selected in the NFL draft. Specifically, the study included each player's draft round (1 through 7), 40-yard dash speed, and position (such as quarterback, linebacker, etc.), the drafting team's current payroll, the player's power ratio (a measure of his weight-lifting ability in pounds to his body weight), whether or not the team has a quality starter at the player's position, and the player's standing high jump (in inches). Classify each variable in this study as being quantitative or qualitative. Classify each quantitative variable as discrete or continuous.

## Self-Review Test

- A population in statistics means a collection of all
  - men and women
  - subjects or objects of interest
  - people living in a country

2. A sample in statistics means a portion of the
- people selected from the population of a country
  - people selected from the population of an area
  - population of interest
3. Indicate which of the following is an example of a sample with replacement and which is a sample without replacement.
- Five friends go to a livery stable and select five horses to ride (each friend must choose a different horse).
  - A box contains five marbles of different colors. A marble is drawn from this box, its color is recorded, and it is put back into the box before the next marble is drawn. This experiment is repeated 12 times.
4. Indicate which of the following variables are quantitative and which are qualitative. Classify the quantitative variables as discrete or continuous.
- Women's favorite TV programs
  - Salaries of football players
  - Number of pets owned by families
  - Favorite breed of dog for each of 20 persons
5. The following table gives information on the total money spent on different categories of products by all people in Canada during May 2011. The first column contains the category, and the second column contains the amount spent in billions of Canadian dollars.

Category	Amount Spent in May 2011 (in billions of Canadian dollars)
Food and beverages	8.57
Vehicles and parts	8.07
Gas stations	4.79
General merchandise	4.72
Health, personal care	2.68
Building materials, garden supplies	2.21
Clothing, accessories	2.13
Furniture, furnishings	1.25
Electronics, appliances	1.22
Sporting goods, hobby, books, music	.94
Miscellaneous	.89

Source: www.cbc.ca/news/business.

Explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

6. The number of types of cereal in the pantries of six households is 6, 11, 3, 5, 6, and 2, respectively. Let  $x$  be the number of types of cereal in the pantry of a household. Find:
- $\sum x$
  - $(\sum x)^2$
  - $\sum x^2$
7. The 17th hole at the TPC Sawgrass golf course may be the most famous golf hole in the world due to its island green and the large number of shots that go in the water. The following table contains the number of strokes various players in the 2011 Players Championship needed to complete the hole. In the table, the frequency row represents the number of players.

<b>Number of Strokes (<math>m</math>)</b>	2	3	4	5	6	7	8
<b>Frequency (<math>f</math>)</b>	82	278	43	16	6	3	1

Source: pgatour.com.

Calculate

- $\sum m$
- $\sum f$
- $\sum m^2$
- $\sum mf$
- $\sum m^2f$
- $(\sum f)^2$

## Mini-Project

### ■ MINI-PROJECT 1-1

In this mini-project, you are going to obtain a data set of interest to you that you will use for mini-projects in some of the other chapters. The data set should contain at least one qualitative variable and one quantitative variable, although having two of each will be necessary in some cases. Ask your instructor how many variables you should have. A good-size data set to work with should contain somewhere between 50 and 100 observations.

Here are some examples of the procedures to use to obtain data:

1. Take a random sample of used cars and collect data on them. You may use Web sites like Cars.com, AutoTrader.com, and so forth. Quantitative variables may include the price, mileage, and age of a car. Categorical variables may include the model, drive train (front wheel, rear wheel, and so forth), and type (compact, SUV, minivan, and so forth). You can concentrate on your favorite type of car, or look at a variety of types.
2. Examine the real estate ads in your local newspaper or online and obtain information on houses for sale that may include listed price, number of bedrooms, lot size, living space, town, type of house, number of garage spaces, and number of bathrooms.
3. Use an almanac or go to a government Web site, such as [www.census.gov](http://www.census.gov) or [www.cdc.gov](http://www.cdc.gov), to obtain information for each state. Quantitative variables may include income, birth and death rates, cancer incidence, and the proportion of people living below the poverty level. Categorical variables may include things like the region of the country where each state is located and which party won the state governorship in the last election. You can also collect this information on a worldwide level and use the continent or world region as a categorical variable.
4. Take a random sample of students and ask them questions such as:
  - How much money did you spend on books last semester?
  - How many credit hours did you take?
  - What is your major?
5. If you are a sports fan, you can use an almanac or sports Web site to obtain statistics on a random sample of athletes. You can look at sport-specific statistics such as home runs, runs batted in, position, left-handed/right-handed, and so forth in baseball, or you could collect information to compare different sports by gathering information on salary, career length, weight, and so forth.

Once you have collected the information, write a brief report that includes answers to the following tasks/questions:

- a. Describe the variables on which you have collected information.
- b. Describe a reasonable target population for the sample you used.
- c. Is your sample a random sample from this target population?
- d. Do you feel that your sample is representative of this population?
- e. Is this an example of sampling with or without replacement?
- f. For each quantitative variable, state whether it is continuous or discrete.
- g. Describe the meaning of an element, a variable, and a measurement for this data set.
- h. Describe any problems you faced in collecting these data.
- i. Were any of the data values unusable? If yes, explain why.

Your instructor will probably want to see a copy of the data you collected. If you are using statistical software in the class, enter the data into that software and submit a copy of the data file. If you are using a handheld technology calculator, such as a graphing calculator, you will probably have to print out a hard copy version of the data set. Save this data set for projects in future chapters.

### DECIDE FOR YOURSELF

### FOR THE SECOND YEAR, JOHNNY DEPP IS AMERICA'S FAVORITE ACTOR

Society is inundated with data and data summaries. The results of surveys and studies can be found in print, on television and radio, and on a plethora of Web sites. Quite often, the people who publish articles about the results of these studies write headlines in a way that will attract attention and draw in more readers.

The following are the headline and excerpts from an article based on an online survey of 2237 adults (age 18 years and older) about actors conducted by Harris Interactive between December 5 and 12, 2011. (*Source: www.harrisinteractive.com/vault/Harris Poll 8 - Movie Stars\_1.19.12.pdf.*)

**Denzel Washington remains at number two, but in a tie this year with Clint Eastwood**

New York, N.Y.—January 19, 2012—In 2011 he was the voice of Rango, he was Captain Jack Sparrow (again) and he was also a journalist. And, again this year, Johnny Depp has the distinction of being America's Favorite Actor. Next on the list are two actors who haven't actually acted in a movie this past year. Tied for number two are Denzel Washington, who was in the second spot last year, and Clint Eastwood who was number 9 on the list last year.

- Based on the headline, what is the population of interest for this study?
- Based on the remainder of the information, it is unreasonable to generalize the results to the population implied in the headline. What large group of people was not interviewed and, hence, is not in

the population? Do you think that the results would change substantially if this group had been included in the survey? Why or why not?

- The article also includes information about how the survey was conducted. Why would this method eliminate some people from having a chance to participate in the study? Using this information and your answer to question 2 above, write a brief description of the population that the study actually examined.
- Is there any reason to believe that the group eliminated by the method used to conduct the survey could have a substantial impact on the results if they had been included in the study? Why or why not?

Well-respected polling agencies such as Harris, Gallup, and the Pew Research Center include a great deal of information about the limitations of surveys and studies. If you go to the full article on the Internet, you can read more details about the limitations of the study, which are given at the end of the article.

## TECHNOLOGY INSTRUCTION

### Entering and Saving Data

Technology makes the process of data analysis much easier and faster. Therefore, you need to be able to enter the data, proofread them, and revise them. Moreover, you can save the data and retrieve them for use at a later date.

#### TI-84

The Technology Instruction feature of this text is written for TI-84 graphing calculators running the 2.55MP operating system. Some screenshots, menus, and functions will be slightly different in older operating systems.

#### Entering Data in a List

L1	L2	L3	1
-----	-----	-----	
<b>L1(0)=</b>			

Screen 1.1

L1	L2	L3	2
25	50	-----	
64	53	-----	
53	52	-----	
42	51	-----	
31	50	-----	
20	49	-----	
<b>L2(0)=50</b>			

Screen 1.2

EX1	EX2	-----	3
25	50	-----	
64	53	-----	
53	52	-----	
42	51	-----	
31	50	-----	
20	49	-----	
<b>EX2(0)=50</b>			

Screen 1.3

#### Changing List Names/Establishing Visible Lists

- The TI-84 has only six "scratch" lists. In some cases you will be using your data at a later date. You can rename a list so that you do not have to reenter the data. Select **STAT** **>EDIT** **>SetUpEditor**, and then type in the names of your variables separated by commas, and then press **ENTER**. (see Screens 1.3 and 1.4). Names can be one to five letters long, with the letters found in green on your keypad. You can use the green **ALPHA** key with each letter, or press **A-LOCK** (**2nd > ALPHA**) while you are typing the name. To turn off **A-LOCK**, press **ALPHA**.
- You can use the arrow keys to move around and go back to a cell to edit its contents. When editing values, you will need to press **ENTER** for the changes to take effect.
- SetUpEditor** determines what lists are displayed in the editor. Changing what **SetUpEditor** displays does not delete any lists. Your lists remain in storage when the calculator is turned off.

EX1	EX2	-----	3
25	50	-----	
64	53	-----	
53	52	-----	
42	51	-----	
31	50	-----	
20	49	-----	
<b>EX2(0)=50</b>			

Screen 1.4

## Numeric Operations on Lists

- To calculate the sum of the values in a list, such as L1, select **LIST (2nd > STAT) > MATH > sum()**. Enter the name of the list (e.g., **2nd > 1** for L1), then type the right parenthesis. Press **ENTER**. (See Screens 1.5 and 1.6.)
- If you need to find the sum of values and the square of the sum denoted by  $(\Sigma x)^2$ , you can use the same instructions as in item 1. However, just before you press **ENTER**, press the  **$x^2$**  button. If you wish to square each value and calculate the sum of the squared values, which is denoted by  $\Sigma x^2$ , press the  **$x^2$**  button after entering the name of the list but prior to typing the right parenthesis. Screen 1.6 shows the appearance of these two processes.

NAMES OPS MATH  
 1:min()  
 2:max()  
 3:mean()  
 4:median()  
 5:sum()  
 6:Prod()  
 7:stdDev()  
 8:  
 9:  
 0:  
 .:  
 -:  
 /:  
 ^:  
 %:  
 $\Sigma$ :  
 $\Sigma x^2$ :  
 $\Sigma x^3$ :  
 $\Sigma x^4$ :

Screen 1.5

sum(L1) 285  
 sum(L1) $^2$  81225  
 sum(L1 $^3$ ) 15666  
 sum(L1 $^4$ ) 44666

Screen 1.6

## Minitab

### Entering and Saving Data

	C1	C2	C3-T
	year	sales	employee
1	2008	35	J. Smith
2	2008	38	A. Jones
3	2009	50	J. Smith
4	2009	48	A. Jones

Screen 1.7

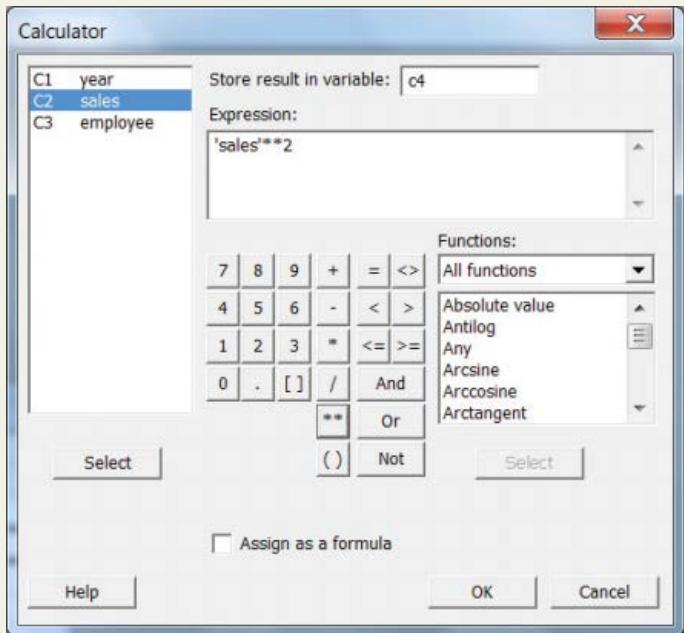
- Start Minitab. You will see the computer screen divided into two parts—**Session** window, which will contain numeric output; and a **Worksheet**, which looks similar to a spreadsheet, where you will enter your data (see Screen 1.7). You are allowed to have multiple worksheets within a project.
- Use the mouse or the arrow keys to select where you want to start entering your data in the worksheet. Each column in the worksheet corresponds to a variable, so you can enter only one kind of data into a given column. Data can be numeric, text, or date/time. The rectangles in the worksheet are called *cells*, and the cells are organized into columns such as C1, C2, and so on, each with rows 1, 2, and so forth. Note that if a column contains text data, Minitab will add “-T” to the column heading.
- The blank row between the column labels and the first row is for variable names. In these blank cells, you can type the names of variables.
- You can change whether you are typing the data across in rows or down in columns by clicking the direction arrow at the top left of the worksheet (also shown in Screen 1.7).
- Click on a cell and begin typing. Press **ENTER** when you are finished with that cell.
- If you need to revise an entry, go to that cell with the mouse or the arrow keys and begin typing. Press **ENTER** to put the revised entry into the cell.
- When you are done, select **File >Save Project As** to save your work for the first time as a file on your computer. Note that Minitab will automatically assign the file extension. *mpj* to your work after you choose the filename.
- Try entering the following data into Minitab:

January	52	.08
February	48	.06
March	49	.07

Name the columns *Month*, *Sales*, *Increase*. Save the result as the file *test.mpj*.

- To retrieve the file, select **File >Open** and select the file *test.mpj*.

- 10.** If you are already in Minitab and you want to start a new worksheet, select **File > New** and choose **Worksheet**. Whenever you save a project, Minitab will automatically save all of the worksheets in the project.



Screen 1.8

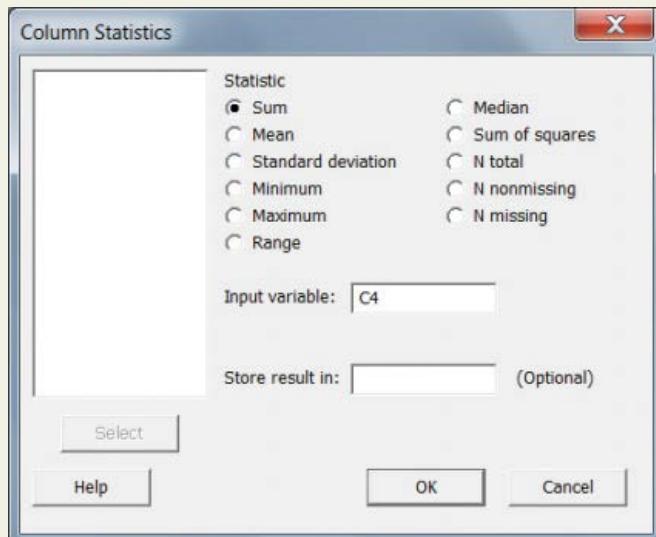
### Creating New Columns from Existing Columns

In some circumstances, such as when you need to calculate  $\Sigma x^2$  or  $\Sigma xy$ , you will need to calculate a new column of values using one or more existing columns. To calculate a column containing the squares of the values in the column *Sales* as shown in Screen 1.7,

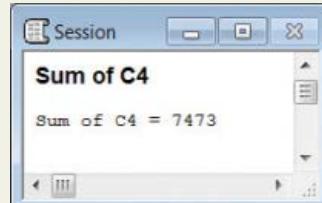
1. Select **Calc > Calculator**.
2. Type the name of the column to contain the new values (such as **C4**) in the **Store result in variable:** box.
3. Click inside the **Expression:** box, click **C2 Sales** in the column to the left of the **Expression:** box, and click **Select**. Click on the exponentiation (\*\*\*) button. Type **2** after the two asterisks in the **Expression:** box. Click **OK**. (See Screen 1.8.)
4. The numbers 1225, 1444, 2500, and 2304 should appear in column C4.

### Calculating the Sum of a Column

1. To calculate the sum of the values in a column, select **Calc > Column Statistics**, which will produce a dialog box. From the **Statistic** list, select **Sum**.
2. Click in the **Input Variable:** box. The list of variables will appear in the left portion of the dialog box. Click on the variable you wish to sum, then click **Select**. (See Screen 1.9.)
3. Click **OK**. The result will appear in the **Session** window. (See Screen 1.10.)



Screen 1.9



Screen 1.10

**Excel****Entering and Saving Data in Excel**

1. Start Excel.
2. Use the mouse or the arrow keys to select where you want to start entering your data. Data can be numeric or text. The rectangles are called *cells*, and the cells are collectively known as a *spreadsheet*.
3. You can format your data by selecting the cells that you want to format, then selecting **Format > Cells**, and then choosing whether you want to format a number, align text, and so forth. For common formatting tasks, you have icons on the toolbar, such as a dollar sign (\$) to format currency, a percent sign (%) to format numbers as percents, and icons representing left-, center-, and right-aligned text to change your alignment.
4. If you need to revise an entry, go to that cell with the mouse or the arrow keys. You can retype the entry or you can edit it. To edit it, double-click on the cell and use the arrow keys and the backspace key to help you revise the entry, then press **ENTER** to put the revised entry into the cell.
5. When you are done, select **File > Save As** to save your work for the first time as a file on your computer. Note that Excel will automatically assign the file extension.xls to your work after you choose the filename.
6. Try entering the following data into Excel:

January	52	.08
February	48	.06
March	49	.07

Save the result as the file *test.xls*.

January	\$52.00	8%
February	\$48.00	6%
March	\$49.00	7%

	A	B	C
1	year	sales	employee
2	2008	35	J. Smith
3	2008	38	A. Jones
4	2009	50	J. Smith
5	2009	48	A. Jones

Screen 1.11

**Creating New Columns from Existing Columns**

Many times, such as when you need to calculate  $\Sigma x^2$  or  $\Sigma xy$ , you will need to calculate a new column of values using one or more existing columns. To calculate the squares of the values in cells **B1** to **B3** and place them in cells **D1** to **D3**:

	A	B	C	D
1	January	\$52.00	8%	=B1^2
2	February	\$48.00	6%	
3	March	\$49.00	7%	

Screen 1.12

1. Click on cell **D1**.

2. Type **=B1^2**. Press **ENTER**. (See Screen 1.12.)

While still on cell **D1**, select **Edit > Copy**. Highlight cells **D2** and **D3**. Select **Edit > Paste**.

3. The numbers 2704, 2304, and 2401 should appear in **D1** to **D3**.

### Calculating the Sum of a Column

SUM				$\Sigma$	=SUM(D1:D3)
	A	B	C	D	
1	January	\$52.00	8%	2704	
2	February	\$48.00	6%	2304	
3	March	\$49.00	7%	2401	
4				=SUM(D1:D3)	

Screen 1.13

To calculate the sum of the values in a column, go to the empty cell below the values you wish to find the sum of. Click the sigma ( $\Sigma$ ) button in the upper-right portion of the **Home** tab. This will enter the **Sum** function into the cell along with the list of cells involved in the sum. (Note: If the list is incorrect, you can type any changes.) Press **ENTER**. (See Screen 1.13.)

## TECHNOLOGY ASSIGNMENTS

**TA1.1** The following table gives the names, hours worked, and salary for the past week for five workers.

Name	Hours Worked	Salary (\$)
John	42	1325
Shannon	33	2583
Kathy	28	3255
David	47	5090
Steve	40	1020

- Enter these data into the spreadsheet. Save the data file as WORKER. Exit the session or program. Then restart the program or software and retrieve the file WORKER.
- Print a hard copy of the spreadsheet containing data you entered.

**TA1.2** Refer to data on total revenues for 2010 of six companies given in Table 1.1. Enter those data into the spreadsheet and save this file as REVENUES.

# CHAPTER 2



Media Bakery

## Organizing and Graphing Data

### 2.1 Organizing and Graphing Qualitative Data

#### Case Study 2–1 Will Today’s Children Be Better Off Than Their Parents?

#### Case Study 2–2 Employees’ Overall Financial Stress Levels

#### 2.2 Organizing and Graphing Quantitative Data

#### Case Study 2–3 How Long Does Your Typical One-Way Commute Take?

#### Case Study 2–4 How Much Does It Cost to Insure a Car?

#### Case Study 2–5 How Many Cups of Coffee Do You Drink a Day?

#### 2.3 Cumulative Frequency Distributions

#### 2.4 Stem-and-Leaf Displays

#### 2.5 Dotplots

How would you classify your financial stress level? Is it overwhelming? Is it high? Or are you one of the lucky people who have no financial stress? In a 2011 poll of employees conducted by Financial Finesse Inc., 5% of the employees polled said that their financial stress level was overwhelming, 16% indicated that they had a high financial stress level, 65% had some financial stress, and 14% had no financial stress. (See Case Study 2–2.)

In addition to thousands of private organizations and individuals, a large number of U.S. government agencies (such as the Bureau of the Census, the Bureau of Labor Statistics, the National Agricultural Statistics Service, the National Center for Education Statistics, the National Center for Health Statistics, and the Bureau of Justice Statistics) conduct hundreds of surveys every year. The data collected from each of these surveys fill hundreds of thousands of pages. In their original form, these data sets may be so large that they do not make sense to most of us. Descriptive statistics, however, supplies the techniques that help to condense large data sets by using tables, graphs, and summary measures. We see such tables, graphs, and summary measures in newspapers and magazines every day. At a glance, these tabular and graphical displays present information on every aspect of life. Consequently, descriptive statistics is of immense importance because it provides efficient and effective methods for summarizing and analyzing information.

This chapter explains how to organize and display data using tables and graphs. We will learn how to prepare frequency distribution tables for qualitative and quantitative data; how to construct bar graphs, pie charts, histograms, and polygons for such data; and how to prepare stem-and-leaf displays.

## 2.1 Organizing and Graphing Qualitative Data

This section discusses how to organize and display qualitative (or categorical) data. Data sets are organized into tables and displayed using graphs. First we discuss the concept of raw data.

### 2.1.1 Raw Data

When data are collected, the information obtained from each member of a population or sample is recorded in the sequence in which it becomes available. This sequence of data recording is random and unranked. Such data, before they are grouped or ranked, are called **raw data**.

#### Definition

**Raw Data** Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

Suppose we collect information on the ages (in years) of 50 students selected from a university. The data values, in the order they are collected, are recorded in Table 2.1. For instance, the first student's age is 21, the second student's age is 19 (second number in the first row), and so forth. The data in Table 2.1 are quantitative raw data.

**Table 2.1** Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Suppose we ask the same 50 students about their student status. The responses of the students are recorded in Table 2.2. In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. This is an example of qualitative (or categorical) raw data.

**Table 2.2** Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

The data presented in Tables 2.1 and 2.2 are also called **ungrouped data**. An ungrouped data set contains information on each member of a sample or population individually. If we rank the data of Table 2.1 from lowest to the highest age, they will still be ungrouped data but not raw data.

### 2.1.2 Frequency Distributions

A sample of 100 students enrolled at a university were asked what they intended to do after graduation. Forty-four of them said that they wanted to work for private companies/businesses,

16 said they wanted to work for the federal government, 23 wanted to work for state or local governments, and 17 intended to start their own businesses. Table 2.3 lists the types of employment and the number of students who intend to engage in each type of employment. In this table, the variable is the *type of employment*, which is a qualitative variable. The categories (representing the type of employment) listed in the first column are mutually exclusive. In other words, each of the 100 students belongs to one and only one of these categories. The number of students who belong to a certain category is called the *frequency* of that category. A **frequency distribution** exhibits how the frequencies are distributed over various categories. Table 2.3 is called a *frequency distribution table* or simply a *frequency table*.

**Table 2.3** Type of Employment Students Intend to Engage In

Variable →	Type of Employment	Number of Students	← Frequency column
	Private companies/businesses	44	
Category →	Federal government	16	← Frequency
	State/local government	23	
	Own business	17	
		Sum = 100	

### Definition

**Frequency Distribution of a Qualitative Variable** A *frequency distribution* of a qualitative variable lists all categories and the number of elements that belong to each of the categories.

Example 2–1 illustrates how a frequency distribution table is constructed for a qualitative variable.

### ■ EXAMPLE 2–1

A sample of 30 persons who often consume donuts were asked what variety of donuts is their favorite. The responses from these 30 persons are as follows:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

Construct a frequency distribution table for these data.

**Solution** Note that the variable in this example is *favorite variety of donut*. This variable has five categories (varieties of donuts): glazed, filled, frosted, plain, and other. To prepare a frequency distribution, we record these five categories in the first column of Table 2.4. Then we read each response (each person's favorite variety of donut) from the given information and mark a *tally*, denoted by the symbol |, in the second column of Table 2.4 next to the corresponding category. For example, the first response is *glazed*. We show this in the frequency table by marking a tally in the second column next to the category *glazed*. Note that the tallies are marked in blocks of five for counting convenience. Finally, we record the total of the tallies for each category in the third column of the table. This column is called the *column of frequencies* and is usually denoted by *f*. The sum of the entries in the frequency column gives the sample size or total frequency. In Table 2.4, this total is 30, which is the sample size.

Constructing a frequency distribution table for qualitative data.



© Jack Puccio/iStockphoto

**Table 2.4** Frequency Distribution of Favorite Donut Variety

Donut Variety	Tally	Frequency ( $f$ )
Glazed		8
Filled		7
Frosted		5
Plain		3
Other		7
		Sum = 30

### 2.1.3 Relative Frequency and Percentage Distributions

The **relative frequency** of a category is obtained by dividing the frequency of that category by the sum of all frequencies. Thus, the relative frequency shows what fractional part or proportion of the total frequency belongs to the corresponding category. A *relative frequency distribution* lists the relative frequencies for all categories.

#### Calculating Relative Frequency of a Category

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

The **percentage** for a category is obtained by multiplying the relative frequency of that category by 100. A *percentage distribution* lists the percentages for all categories.

#### Calculating Percentage

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100\%$$

### ■ EXAMPLE 2–2

Determine the relative frequency and percentage distributions for the data in Table 2.4.

**Solution** The relative frequencies and percentages from Table 2.4 are calculated and listed in Table 2.5. Based on this table, we can state that 26.7% of the people in the sample said that glazed donuts are their favorite. By adding the percentages for the first two categories, we can state that 50% of the persons included in the sample said that glazed or filled donuts are their favorite. The other numbers in Table 2.5 can be interpreted in similar ways.

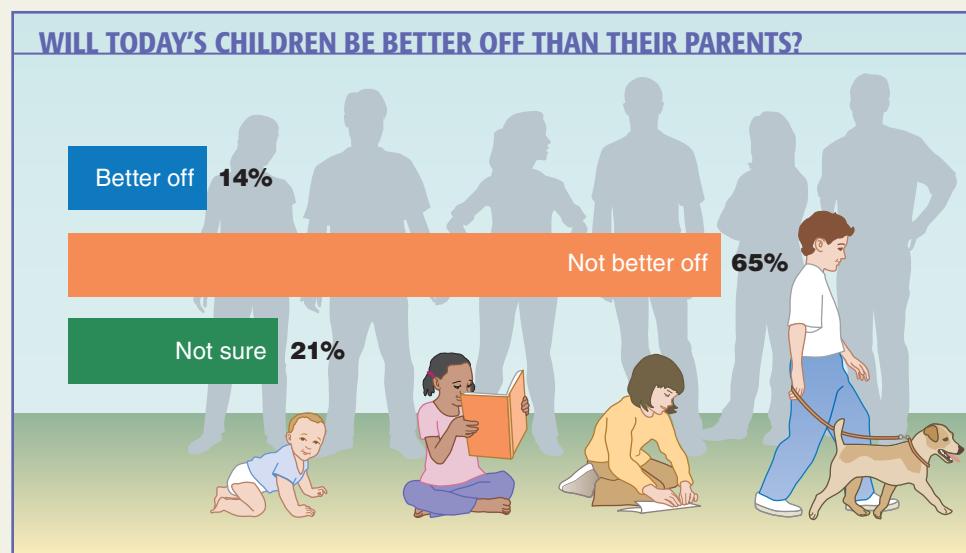
Constructing relative frequency and percentage distributions.

**Table 2.5** Relative Frequency and Percentage Distributions of Favorite Donut Variety

Donut Variety	Relative Frequency	Percentage
Glazed	$8/30 = .267$	$.267(100) = 26.7$
Filled	$7/30 = .233$	$.233(100) = 23.3$
Frosted	$5/30 = .167$	$.167(100) = 16.7$
Plain	$3/30 = .100$	$.100(100) = 10.0$
Other	$7/30 = .233$	$.233(100) = 23.3$
	Sum = 1.000	Sum = 100%

Notice that the sum of the relative frequencies is always 1.00 (or approximately 1.00 if the relative frequencies are rounded), and the sum of the percentages is always 100 (or approximately 100 if the percentages are rounded). ■

## WILL TODAY'S CHILDREN BE BETTER OFF THAN THEIR PARENTS?



Data source: Rasmussen Reports national telephone survey of American adults.

**Data Source:** [http://www.rasmussenreports.com/public\\_content/business/jobs\\_employment/july\\_2012/new\\_low\\_just\\_14\\_think\\_today\\_s\\_children\\_will\\_be\\_better\\_off\\_than\\_their\\_parents](http://www.rasmussenreports.com/public_content/business/jobs_employment/july_2012/new_low_just_14_think_today_s_children_will_be_better_off_than_their_parents).

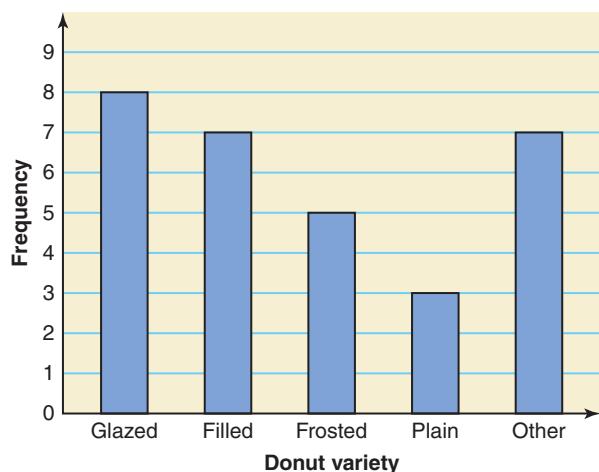
Rasmussen Reports conducted a national telephone survey on July 22–23, 2012, that included 1000 American adults. Among other questions, these adults were asked, "Will today's children be better off than their parents?" As the accompanying chart shows, 14% of the adults polled said that today's children will be better off than their parents, 65% believed they would not be better off, and 21% were not sure. As we can notice, these data are categorical, with three categories that are listed in the chart. Note that in this chart, the bars are drawn horizontally.

### 2.1.4 Graphical Presentation of Qualitative Data

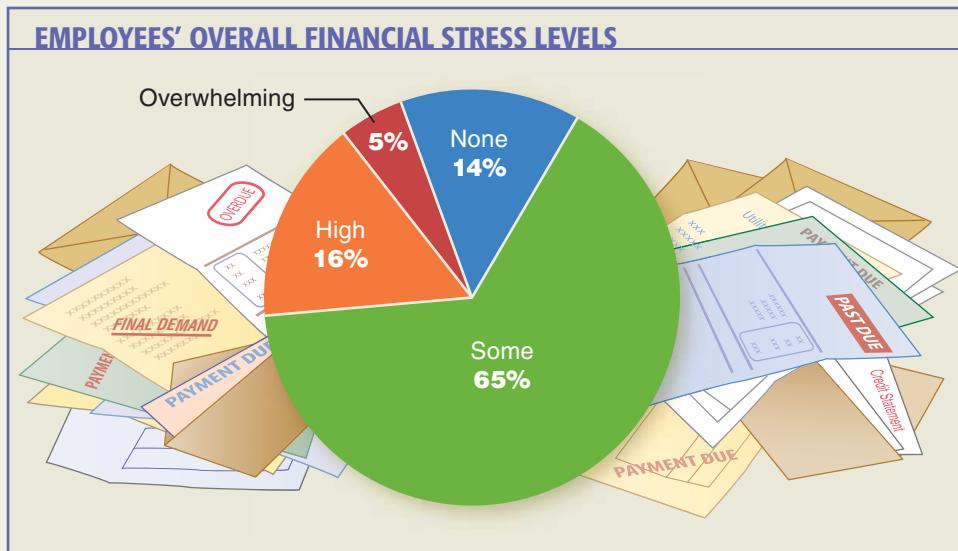
All of us have heard the adage "a picture is worth a thousand words." A graphic display can reveal at a glance the main characteristics of a data set. The *bar graph* and the *pie chart* are two types of graphs that are commonly used to display qualitative data.

#### Bar Graphs

To construct a **bar graph** (also called a *bar chart*), we mark the various categories on the horizontal axis as in Figure 2.1. Note that all categories are represented by intervals of the same width. We mark the frequencies on the vertical axis. Then we draw one bar for each category such that



**Figure 2.1** Bar graph for the frequency distribution of Table 2.4.



Data source: Financial Finesse, Inc.

In a 2011 survey conducted by Financial Finesse Inc., employees were asked about their overall financial stress levels. As shown in the accompanying pie chart, 5% of the employees surveyed said that they had overwhelming financial stress, 16% mentioned a high level of financial stress, 65% indicated some financial stress, and 14% said that they did not have any financial stress. The pie chart represents the categorical variable *overall financial stress level*.

*Data Source:* <http://www.financialfinesse.com/wp-content/uploads/2011/05/2011-Financial-Stress-Research.pdf>.

the height of the bar represents the frequency of the corresponding category. We leave a small gap between adjacent bars. Figure 2.1 gives the bar graph for the frequency distribution of Table 2.4.

### Definition

**Bar Graph** A graph made of bars whose heights represent the frequencies of respective categories is called a *bar graph*.

The bar graphs for relative frequency and percentage distributions can be drawn simply by marking the relative frequencies or percentages, instead of the frequencies, on the vertical axis.

Sometimes a bar graph is constructed by marking the categories on the vertical axis and the frequencies on the horizontal axis. Case Study 2-1 presents such an example.

### Pie Charts

A **pie chart** is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. Then we divide the pie into different portions that represent the different categories.

### Definition

**Pie Chart** A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories is called a *pie chart*.

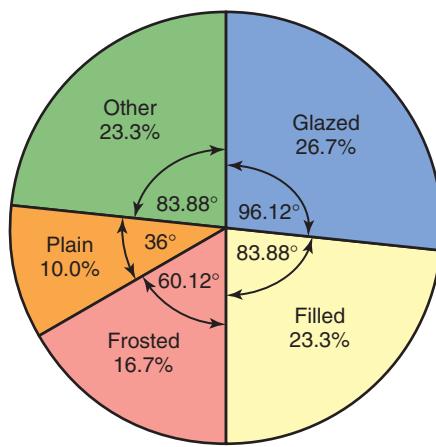
As we know, a circle contains 360 degrees. To construct a pie chart, we multiply 360 by the relative frequency of each category to obtain the degree measure or size of the angle for the corresponding category. Table 2.6 shows the calculation of angle sizes for the various categories of Table 2.5.

## EMPLOYEES' OVERALL FINANCIAL STRESS LEVELS

**Table 2.6** Calculating Angle Sizes for the Pie Chart

Donut Variety	Relative Frequency	Angle Size (degrees)
Glazed	.267	360 (.267) = 96.12
Filled	.233	360 (.233) = 83.88
Frosted	.167	360 (.167) = 60.12
Plain	.100	360 (.100) = 36.00
Other	.233	360 (.233) = 83.88
	Sum = 1.000	Sum = 360

Figure 2.2 shows the pie chart for the percentage distribution of Table 2.5, which uses the angle sizes calculated in Table 2.6.

**Figure 2.2** Pie chart for the percentage distribution of Table 2.5.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 2.1** Why do we need to group data in the form of a frequency table? Explain briefly.
- 2.2** How are the relative frequencies and percentages of categories obtained from the frequencies of categories? Illustrate with the help of an example.
- 2.3** The following data give the results of a sample survey. The letters A, B, and C represent the three categories.

A	B	B	A	C	B	C	C	C	A
C	B	C	A	C	C	B	C	C	A
A	B	C	C	B	C	B	A	C	A

- Prepare a frequency distribution table.
- Calculate the relative frequencies and percentages for all categories.
- What percentage of the elements in this sample belong to category B?
- What percentage of the elements in this sample belong to category A or C?
- Draw a bar graph for the frequency distribution.

- 2.4** The following data give the results of a sample survey. The letters Y, N, and D represent the three categories.

D	N	N	Y	Y	Y	N	Y	D	Y
Y	Y	Y	Y	N	Y	Y	N	N	Y
N	Y	Y	N	D	N	Y	Y	Y	Y
Y	Y	N	N	Y	Y	N	N	D	Y

- Prepare a frequency distribution table.
- Calculate the relative frequencies and percentages for all categories.

- c. What percentage of the elements in this sample belong to category Y?
- d. What percentage of the elements in this sample belong to category N or D?
- e. Draw a pie chart for the percentage distribution.

## ■ APPLICATIONS

**2.5** A July 2011 ESPN SportsNation poll asked, “Which is the best Fourth of July weekend sports tradition?” (<http://espn.go.com/espn/fp/flashPollResultsState?sportIndex=frontpage&pollId=116290>). The choices were Major League baseball game (B), Nathan’s Famous International Hot Dog Eating Contest (H), Breakfast at Wimbledon (W), or NASCAR race at Daytona (N). The following data represent the responses of a random sample of 45 persons who were asked the same question.

H	H	B	W	N	B	H	N	W
N	H	B	W	H	N	N	H	H
B	B	W	H	H	B	W	H	B
H	B	B	H	B	H	B	N	H
B	B	H	H	H	B	H	H	N

- a. Prepare a frequency distribution table.
- b. Calculate the relative frequencies and percentages for all categories.
- c. What percentage of the respondents mentioned *Major League baseball game* or *Breakfast at Wimbledon*?
- d. Draw a bar graph for the frequency distribution.

**2.6** Thirty adults were asked which of the following conveniences they would find most difficult to do without: television (T), refrigerator (R), air conditioning (A), public transportation (P), or microwave (M). Their responses are listed below.

R	A	R	P	P	T	R	M	P	A
A	R	R	T	P	P	T	R	A	A
R	P	A	T	R	P	R	A	P	R

- a. Prepare a frequency distribution table.
- b. Calculate the relative frequencies and percentages for all categories.
- c. What percentage of these adults named refrigerator or air conditioning as the convenience that they would find most difficult to do without?
- d. Draw a bar graph for the relative frequency distribution.

**2.7** A whatjapanthinks.com survey asked residents of Japan to name their favorite pizza topping. The possible responses included the following choices: pig-based meats, for example, bacon or ham (PI); seafood, for example, tuna, crab, or cod roe (S); vegetables and fruits (V); poultry (PO); beef (B); and cheese (C). The following data represent the responses of a random sample of 36 people.

V	PI	B	PI	V	PO	S	PI	V	S	V	S
PI	S	V	V	V	PI	S	S	V	PI	C	V
V	V	C	V	S	PO	V	PI	S	PI	PO	PI

- a. Prepare a frequency distribution table.
- b. Calculate the relative frequencies and percentages for all categories.
- c. What percentage of the respondents mentioned *vegetables and fruits*, *poultry*, or *cheese*?
- d. Draw a bar graph for the relative frequency distribution.

**2.8** The following data show the method of payment by 16 customers in a supermarket checkout line. Here, C refers to cash, CK to check, CC to credit card, and D to debit card, and O stands for other.

C	CK	CK	C	CC	D	O	C
CK	CC	D	CC	C	CK	CK	CC

- a. Construct a frequency distribution table.
- b. Calculate the relative frequencies and percentages for all categories.
- c. Draw a pie chart for the percentage distribution.

**2.9** In a May 4, 2011 Quinnipiac University poll, a random sample of New York City residents were asked, “How serious is the problem of police officers fixing tickets: very serious, somewhat serious, not too serious, or not at all serious?” (Note: In 2010 to 2011, New York City investigated the widespread problem of traffic ticket fixing by police officers. Many police officers were charged with this crime after the investigation.) The following table summarizes residents’ responses.

Response	Percentage of Responses
Very serious	38
Somewhat serious	26
Not too serious	17
Not at all serious	8

Source: www.quinnipiac.edu.

Note that these percentages add up to 89%. The remaining respondents stated that they did not know or had no opinion. Assume that 11% belong to the category *did not know*. Draw a pie chart for this percentage distribution.

**2.10** A July 7, 2011 Pew Research Center poll asked a random sample of Americans to name the current news story that they were following the most at that time. The following table summarizes their responses.

Response	Percentage of Responses
Casey Anthony verdict	37
Economy	17
Deficit and national debt	14
Last Space Shuttle launch	5
2012 Elections	4
Dominique Strauss-Kahn	1
Other	22

Source: Pew Research Center, people-press.org.

Draw a bar graph to display this percentage distribution.

## 2.2 Organizing and Graphing Quantitative Data

In the previous section we learned how to group and display qualitative data. This section explains how to group and display quantitative data.

### 2.2.1 Frequency Distributions

Table 2.7 gives the weekly earnings of 100 employees of a large company. The first column lists the *classes*, which represent the (quantitative) variable *weekly earnings*. For quantitative data, an interval that includes all the values that fall within two numbers—the lower and upper limits—is called a **class**. Note that the classes always represent a variable. As we

**Table 2.7** Weekly Earnings of 100 Employees of a Company

Variable →	Weekly Earnings (dollars)	Number of Employees <i>f</i>	← Frequency column
	801 to 1000	9	
	1001 to 1200	22	
Third class →	1201 to 1400	39	{ Frequency of the third class
	1401 to 1600	15	
	1601 to 1800	9	
	1801 to 2000	6	
Lower limit of the sixth class			
			Upper limit of the sixth class

can observe, the classes are nonoverlapping; that is, each value for earnings belongs to one and only one class. The second column in the table lists the number of employees who have earnings within each class. For example, 9 employees of this company earn \$801 to \$1000 per week. The numbers listed in the second column are called the **frequencies**, which give the number of values that belong to different classes. The frequencies are denoted by  $f$ .

For quantitative data, the frequency of a class represents the number of values in the data set that fall in that class. Table 2.7 contains six classes. Each class has a *lower limit* and an *upper limit*. The values 801, 1001, 1201, 1401, 1601, and 1801 give the lower limits, and the values 1000, 1200, 1400, 1600, 1800, and 2000 are the upper limits of the six classes, respectively. The data presented in Table 2.7 are an illustration of a **frequency distribution table** for quantitative data. Whereas the data that list individual values are called ungrouped data, the data presented in a frequency distribution table are called **grouped data**.

### Definition

**Frequency Distribution for Quantitative Data** A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called *grouped data*.

To find the midpoint of the upper limit of the first class and the lower limit of the second class in Table 2.7, we divide the sum of these two limits by 2. Thus, this midpoint is

$$\frac{1000 + 1001}{2} = 1000.5$$

The value 1000.5 is called the *upper boundary* of the first class and the *lower boundary* of the second class. By using this technique, we can convert the class limits of Table 2.7 to **class boundaries**, which are also called *real class limits*. The second column of Table 2.8 lists the boundaries for Table 2.7.

### Definition

**Class Boundary** A *class boundary* is given by the midpoint of the upper limit of one class and the lower limit of the next class.

The difference between the two boundaries of a class gives the **class width**. The class width is also called the **class size**.

### Finding Class Width

$$\text{Class width} = \text{Upper boundary} - \text{Lower boundary}$$

Thus, in Table 2.8,

$$\text{Width of the first class} = 1000.5 - 800.5 = 200$$

The class widths for the frequency distribution of Table 2.7 are listed in the third column of Table 2.8. Each class in Table 2.8 (and Table 2.7) has the same width of 200.

The **class midpoint** or **mark** is obtained by dividing the sum of the two limits (or the two boundaries) of a class by 2.

### Calculating Class Midpoint or Mark

$$\text{Class midpoint or mark} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Thus, the midpoint of the first class in Table 2.7 or Table 2.8 is calculated as follows:

$$\text{Midpoint of the first class} = \frac{801 + 1000}{2} = 900.5$$

The class midpoints for the frequency distribution of Table 2.7 are listed in the fourth column of Table 2.8.

**Table 2.8 Class Boundaries, Class Widths, and Class Midpoints for Table 2.7**

Class Limits	Class Boundaries	Class Width	Class Midpoint
801 to 1000	800.5 to less than 1000.5	200	900.5
1001 to 1200	1000.5 to less than 1200.5	200	1100.5
1201 to 1400	1200.5 to less than 1400.5	200	1300.5
1401 to 1600	1400.5 to less than 1600.5	200	1500.5
1601 to 1800	1600.5 to less than 1800.5	200	1700.5
1801 to 2000	1800.5 to less than 2000.5	200	1900.5

Note that in Table 2.8, when we write classes using class boundaries, we write *to less than* to ensure that each value belongs to one and only one class. As we can see, the upper boundary of the preceding class and the lower boundary of the succeeding class are the same.

## 2.2.2 Constructing Frequency Distribution Tables

When constructing a frequency distribution table, we need to make the following three major decisions.

### Number of Classes

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set.<sup>1</sup> It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily made by the data organizer.

### Class Width

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

#### Calculation of Class Width

$$\text{Approximate class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

Usually this approximate class width is rounded to a convenient number, which is then used as the class width. Note that rounding this number may slightly change the number of classes initially intended.

<sup>1</sup>One rule to help decide on the number of classes is Sturge's formula:

$$c = 1 + 3.3 \log n$$

where  $c$  is the number of classes and  $n$  is the number of observations in the data set. The value of  $\log n$  can be obtained by using a calculator.

## Lower Limit of the First Class or the Starting Point

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example 2–3 illustrates the procedure for constructing a frequency distribution table for quantitative data.

### ■ EXAMPLE 2–3

The following data give the total number of iPods® sold by a mail order company on each of 30 days. Construct a frequency distribution table.

8	25	11	15	29	22	10	5	17	21
22	13	26	16	18	12	9	26	20	16
23	14	19	23	20	16	27	16	21	14

*Constructing a frequency distribution table for quantitative data.*

**Solution** In these data, the minimum value is 5, and the maximum value is 29. Suppose we decide to group these data using five classes of equal width. Then,

$$\text{Approximate width of each class} = \frac{29 - 5}{5} = 4.8$$

Now we round this approximate width to a convenient number, say 5. The lower limit of the first class can be taken as 5 or any number less than 5. Suppose we take 5 as the lower limit of the first class. Then our classes will be

5–9,      10–14,      15–19,      20–24,      and      25–29

We record these five classes in the first column of Table 2.9.

Now we read each value from the given data and mark a tally in the second column of Table 2.9 next to the corresponding class. The first value in our original data set is 8, which belongs to the 5–9 class. To record it, we mark a tally in the second column next to the 5–9 class. We continue this process until all the data values have been read and entered in the tally column. Note that tallies are marked in blocks of five for counting convenience. After the tally column is completed, we count the tally marks for each class and write those numbers in the third column. This gives the column of frequencies. These frequencies represent the number of days on which the number of iPods indicated by each class were sold. For example, on 8 of these 30 days, 15 to 19 iPods were sold.

**Table 2.9 Frequency Distribution for the Data on iPods Sold**

iPods Sold	Tally	f
5–9		3
10–14		6
15–19		8
20–24		8
25–29		5
		$\Sigma f = 30$

In Table 2.9, we can denote the frequencies of the five classes by  $f_1, f_2, f_3, f_4$ , and  $f_5$ , respectively. Therefore,

$$f_1 = \text{Frequency of the first class} = 3$$

Similarly,

$$f_2 = 6, \quad f_3 = 8, \quad f_4 = 8, \quad \text{and} \quad f_5 = 5$$

Using the  $\Sigma$  notation (see Section 1.7 of Chapter 1), we can denote the sum of frequencies of all classes by  $\Sigma f$ . Hence,

$$\Sigma f = f_1 + f_2 + f_3 + f_4 + f_5 = 3 + 6 + 8 + 8 + 5 = 30$$

The number of observations in a sample is usually denoted by  $n$ . Thus, for the sample data,  $\Sigma f$  is equal to  $n$ . The number of observations in a population is denoted by  $N$ . Consequently,  $\Sigma f$  is equal to  $N$  for population data. Because the data set on the total iPods sold on 30 days in Table 2.9 is for only 30 days, it represents a sample. Therefore, in Table 2.9 we can denote the sum of frequencies by  $n$  instead of  $\Sigma f$ . ■

Note that when we present the data in the form of a frequency distribution table, as in Table 2.9, we lose the information on individual observations. We cannot know the exact number of iPods sold on any given day from Table 2.9. All we know is that for 3 days, 5 to 9 iPods were sold, and so forth.

### 2.2.3 Relative Frequency and Percentage Distributions

Using Table 2.9, we can compute the relative frequency and percentage distributions in the same way as we did for qualitative data in Section 2.1.3. The relative frequencies and percentages for a quantitative data set are obtained as follows. Note that relative frequency is the same as proportion.

#### Calculating Relative Frequency and Percentage

$$\text{Relative frequency of a class} = \frac{\text{Frequency of that class}}{\text{Sum of all frequencies}} = \frac{f}{\Sigma f}$$

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100\%$$

Example 2–4 illustrates how to construct relative frequency and percentage distributions.

#### ■ EXAMPLE 2–4

*Constructing relative frequency and percentage distributions.*

Calculate the relative frequencies and percentages for Table 2.9.

**Solution** The relative frequencies and percentages for the data in Table 2.9 are calculated and listed in the third and fourth columns, respectively, of Table 2.10. Note that the class boundaries are listed in the second column of Table 2.10.

**Table 2.10** Relative Frequency and Percentage Distributions for Table 2.9

iPods Sold	Class Boundaries	Relative Frequency	Percentage
5–9	4.5 to less than 9.5	$3/30 = .100$	10.0
10–14	9.5 to less than 14.5	$6/30 = .200$	20.0
15–19	14.5 to less than 19.5	$8/30 = .267$	26.7
20–24	19.5 to less than 24.5	$8/30 = .267$	26.7
25–29	24.5 to less than 29.5	$5/30 = .167$	16.7
		Sum = 1.001	Sum = 100.1

Using Table 2.10, we can make statements about the percentage of days with the number of iPods sold within a certain interval. For example, on 20% of the days, 10 to 14 iPods were sold. By adding the percentages for the first two classes, we can state that 5 to 14 iPods were sold on 30% of the days. Similarly, by adding the percentages of the last two classes, we can state that 20 to 29 iPods were sold on 43.4% of the days. ■

## 2.2.4 Graphing Grouped Data

Grouped (quantitative) data can be displayed in a *histogram* or a *polygon*. This section describes how to construct such graphs. We can also draw a pie chart to display the percentage distribution for a quantitative data set. The procedure to construct a pie chart is similar to the one for qualitative data explained in Section 2.1.4; it will not be repeated in this section.

### Histograms

A **histogram** can be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution. To draw a histogram, we first mark classes on the horizontal axis and frequencies (or relative frequencies or percentages) on the vertical axis. Next, we draw a bar for each class so that its height represents the frequency of that class. The bars in a histogram are drawn adjacent to each other with no gap between them. A histogram is called a **frequency histogram**, a **relative frequency histogram**, or a **percentage histogram** depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis.

#### Definition

**Histogram** A *histogram* is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentages are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other.

Figures 2.3 and 2.4 show the frequency and the relative frequency histograms, respectively, for the data of Tables 2.9 and 2.10 of Sections 2.2.2 and 2.2.3. The two histograms look alike because they represent the same data. A percentage histogram can be drawn for the percentage distribution of Table 2.10 by marking the percentages on the vertical axis.

In Figures 2.3 and 2.4, we used class limits to mark classes on the horizontal axis. However, we can show the intervals on the horizontal axis by using the class boundaries instead of the class limits.

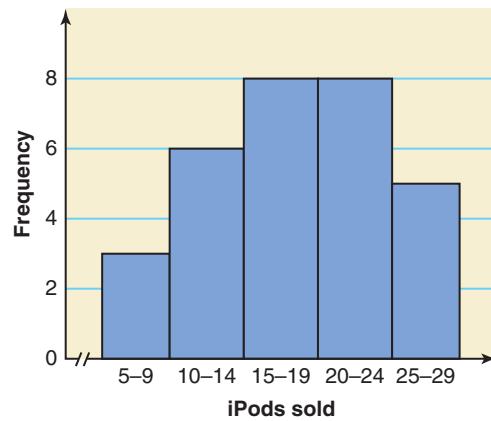


Figure 2.3 Frequency histogram for Table 2.9.

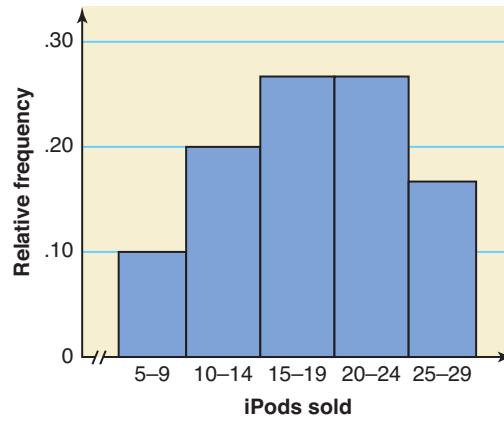


Figure 2.4 Relative frequency histogram for Table 2.10.

### Polygons

A **polygon** is another device that can be used to present quantitative data in graphic form. To draw a **frequency polygon**, we first mark a dot above the midpoint of each class at a height equal to the frequency of that class. This is the same as marking the midpoint at the top of each bar in a histogram. Next we mark two more classes, one at each end, and mark their midpoints. Note that these two classes have zero frequencies. In the last step, we join the adjacent dots with straight lines. The resulting line graph is called a frequency polygon or simply a polygon.

## CASE STUDY 2–3

### HOW LONG DOES YOUR TYPICAL ONE-WAY COMMUTE TAKE?



Data source: IBM 2011 Commuter Pain Survey of 271 adults from New York City who drive a car alone or a motorbike as their main mode of transportation to work or school.

Every year IBM conducts a survey called the Commuter Pain Survey. In this survey, people are polled in many cities around the world to collect data on many aspects related to commuting. For the 2011 IBM Commuter Pain Survey, 8042 adults of age 18 to 65 years who drove a car alone or a motorbike as their main mode of transportation to work or school were selected from 20 cities around the world, and information was collected from them on many variables. In New York City, this information was collected from a sample of 271 such adults. The accompanying graph gives the histogram for the percentage distribution of time spent by such adults to commute one way to work or school in New York City. According to the information in the graph, 23% of the adults in the sample from New York City said that they spend less than 15 minutes to commute one way to work or school, and so on. Note that the first class (less than 15 minutes) has no lower limit (although implicitly this lower limit is zero minutes) and the last class (more than 60 minutes) has no upper limit. Such classes are called *open-ended* classes.

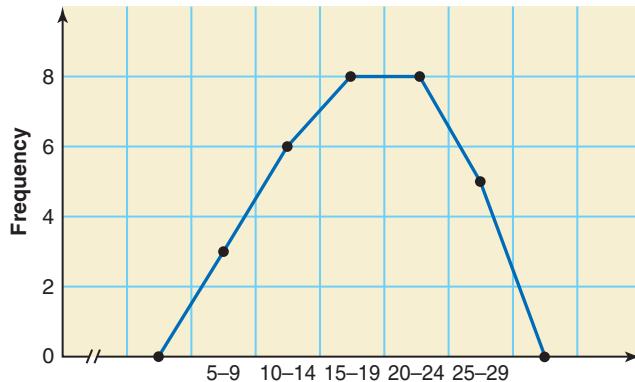
**Data Source:** IBM Commuter Pain Survey, 2011. We are thankful to IBM for providing these data to us.

A polygon with relative frequencies marked on the vertical axis is called a *relative frequency polygon*. Similarly, a polygon with percentages marked on the vertical axis is called a *percentage polygon*.

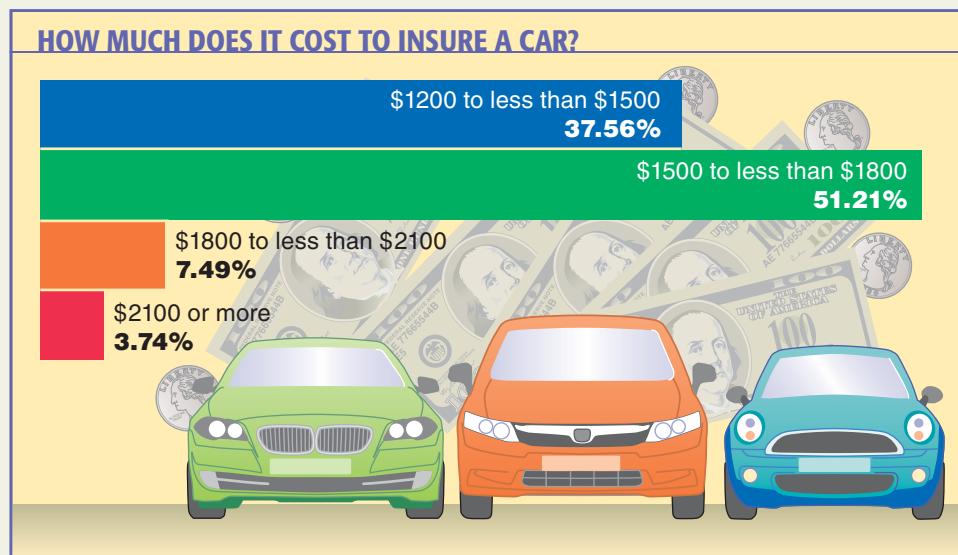
#### Definition

**Polygon** A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon*.

Figure 2.5 shows the frequency polygon for the frequency distribution of Table 2.9.



**Figure 2.5** Frequency polygon for Table 2.9.

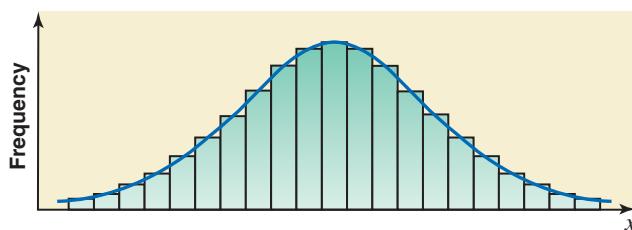


Data source: Insure.com and <http://money.msn.com/auto-insurance/auto-insurance-quotes.aspx>.

MSN Money, along with Insure.com, collects information on the average six-month car insurance premium for a large number of different types of new cars. Using this information on car insurance rates given on the Web site for 908 different cars in July 2012, the authors prepared the data that is represented by the histogram in the accompanying graph. Thus, according to the information given in the histogram, 37.56% of the cars cost \$1200 to less than \$1500 to insure for six months, and so on. Note that the first three classes given in the histogram are of the same width, which is \$300. The last class, however, is an *open-ended* class, as it has no upper limit.

*Data Source:* <http://money.msn.com/auto-insurance/auto-insurance-quotes.aspx>.

For a very large data set, as the number of classes is increased (and the width of classes is decreased), the frequency polygon eventually becomes a smooth curve. Such a curve is called a *frequency distribution curve* or simply a *frequency curve*. Figure 2.6 shows the frequency curve for a large data set with a large number of classes.



**Figure 2.6** Frequency distribution curve.

## 2.2.5 More on Classes and Frequency Distributions

This section presents two alternative methods for writing classes to construct a frequency distribution for quantitative data.

### Less-Than Method for Writing Classes

The classes in the frequency distribution given in Table 2.9 for the data on iPods sold were written as 5–9, 10–14, and so on. Alternatively, we can write the classes in a frequency distribution table using the *less-than* method. The technique for writing classes shown in Table 2.9 is more

## HOW MUCH DOES IT COST TO INSURE A CAR?

commonly used for data sets that do not contain fractional values. The *less-than* method is more appropriate when a data set contains fractional values. Example 2–5 illustrates the *less-than* method.

*Constructing a frequency distribution using the less-than method.*

### ■ EXAMPLE 2–5

The percentage of the population working in the United States peaked in 2000 but dropped to the lowest level in 30 years in 2010. Table 2.11 shows the percentage of the population working in each of the 50 states in 2010. These percentages exclude military personnel and self-employed persons. (*Source:* USA TODAY, April 14, 2011. Based on data from the U.S. Census Bureau and U.S. Bureau of Labor Statistics.)

**Table 2.11** Percentage of Population Working in 2010

State	Percentage	State	Percentage
AL	39.1	MT	43.3
AK	45.7	NE	51.4
AZ	37.2	NV	41.3
AR	39.9	NH	47.3
CA	37.3	NJ	43.8
CO	44.1	NM	38.9
CT	45.0	NY	44.1
DE	46.0	NC	40.5
FL	38.2	ND	55.8
GA	39.5	OH	43.6
HI	43.1	OK	40.7
ID	38.5	OR	41.8
IL	43.7	PA	44.2
IN	43.1	RI	43.6
IA	48.2	SC	39.0
KS	46.4	SD	49.5
KY	40.8	TN	41.2
LA	41.6	TX	41.1
ME	44.6	UT	42.7
MD	43.5	VT	47.5
MA	48.7	VA	45.3
MI	39.1	WA	41.3
MN	49.7	WV	40.3
MS	36.7	WI	48.1
MO	44.2	WY	50.1

Construct a frequency distribution table. Calculate the relative frequencies and percentages for all classes.

**Solution** The minimum value in the data set of Table 2.11 is 36.7%, and the maximum value is 55.8%. Suppose we decide to group these data using six classes of equal width. Then,

$$\text{Approximate width of a class} = \frac{55.8 - 36.7}{6} = 3.18$$

We round this number to a more convenient number—say 3—and take 3 as the width of each class. We can take the lower limit of the first class equal to 36.7 or any number lower than 36.7. If we start the first class at 36, the classes will be written as 36 to less than 39, 39 to less than 42, and so on. Note that when we round the width to a convenient number, we end up with seven classes. These seven classes, which cover all the data values of Table 2.11, are recorded in the first column of Table 2.12. The second column in Table 2.12 lists the frequencies of these classes. A value in the data set that is 36 or larger but less than 39 belongs to the first class, a value that is 39 or larger but less than 42 falls into the second class, and so on. The relative frequencies and percentages for classes are recorded in the third and fourth columns, respectively, of Table 2.12. Note that this table does not contain a column of tallies.

**Table 2.12** Frequency, Relative Frequency, and Percentage Distributions of the Percentage of Population Working

Percentage of Population Working	<i>f</i>	Relative Frequency	Percentage
36 to less than 39	6	.12	12
39 to less than 42	15	.30	30
42 to less than 45	14	.28	28
45 to less than 48	7	.14	14
48 to less than 51	6	.12	12
51 to less than 54	1	.02	2
54 to less than 57	1	.02	2
$\Sigma f = 50$		Sum = 1.00	Sum = 100

A histogram and a polygon for the data of Table 2.12 can be drawn the same way as for the data of Tables 2.9 and 2.10.

### Single-Valued Classes

If the observations in a data set assume only a few distinct (integer) values, it may be appropriate to prepare a frequency distribution table using *single-valued classes*—that is, classes that are made of single values and not of intervals. This technique is especially useful in cases of discrete data with only a few possible values. Example 2–6 exhibits such a situation.

### ■ EXAMPLE 2–6

The administration in a large city wanted to know the distribution of the number of vehicles owned by households in that city. A sample of 40 randomly selected households from this city produced the following data on the number of vehicles owned.

5	1	1	2	0	1	1	2	1	1
1	3	3	0	2	5	1	2	3	4
2	1	2	2	1	2	2	1	1	1
4	2	1	1	2	1	1	4	1	3

Construct a frequency distribution table for these data using single-valued classes.

**Solution** The observations in this data set assume only six distinct values: 0, 1, 2, 3, 4, and 5. Each of these six values is used as a class in the frequency distribution in Table 2.13, and these six classes are listed in the first column of that table. To obtain the frequencies of these classes, the observations in the data that belong to each class are counted, and the

Constructing a frequency distribution using single-valued classes.



© Jorge Salcedo/iStockphoto

## HOW MANY CUPS OF COFFEE DO YOU DRINK A DAY?



Data source: Gallup poll of U.S. adults aged 18 and older conducted July 9–12, 2012.

In a Gallup poll conducted by telephone interviews on July 9–12, 2012, U.S. adults of age 18 years and older were asked, "How many cups of coffee, if any, do you drink on an average day?" According to the results of the poll, shown in the accompanying pie chart, 36% of these adults said that they drink no coffee (represented by zero cups in the chart), 26% said that they drink one cup of coffee per day, and so on. The last class is open-ended class that indicates that 10% of these adults drink four or more cups of coffee a day. This class has no upper limit. Since the values of the variable (cups of coffee) are discrete and the variable assumes only a few possible values, the first four classes are single-valued classes.

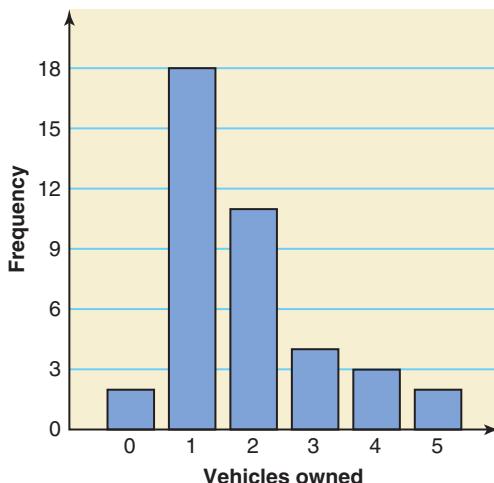
*Data Source:* <http://www.gallup.com/poll/156116/Nearly-Half-Americans-Drink-Soda-Daily.aspx>.

results are recorded in the second column of Table 2.13. Thus, in these data, 2 households own no vehicle, 18 own one vehicle each, 11 own two vehicles each, and so on.

**Table 2.13** Frequency Distribution of the Number of Vehicles Owned

Vehicles Owned	Number of Households ( $f$ )
0	2
1	18
2	11
3	4
4	3
5	2
$\Sigma f = 40$	

The data of Table 2.13 can also be displayed in a bar graph, as shown in Figure 2.7. To construct a bar graph, we mark the classes, as intervals, on the horizontal axis with a little gap between consecutive intervals. The bars represent the frequencies of respective classes.



**Figure 2.7** Bar graph for Table 2.13.

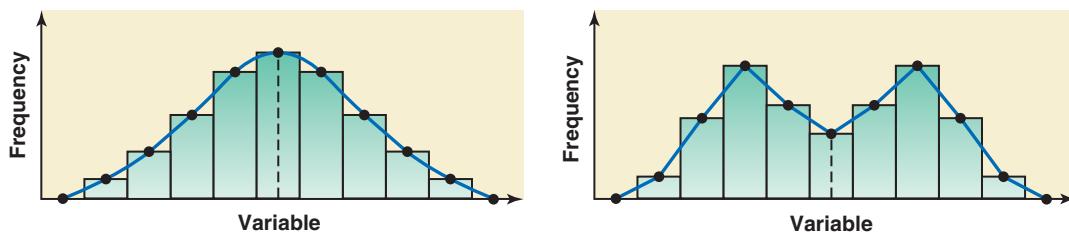
The frequencies of Table 2.13 can be converted to relative frequencies and percentages the same way as in Table 2.10. Then, a bar graph can be constructed to display the relative frequency or percentage distribution by marking the relative frequencies or percentages, respectively, on the vertical axis.

## 2.2.6 Shapes of Histograms

A histogram can assume any one of a large number of shapes. The most common of these shapes are

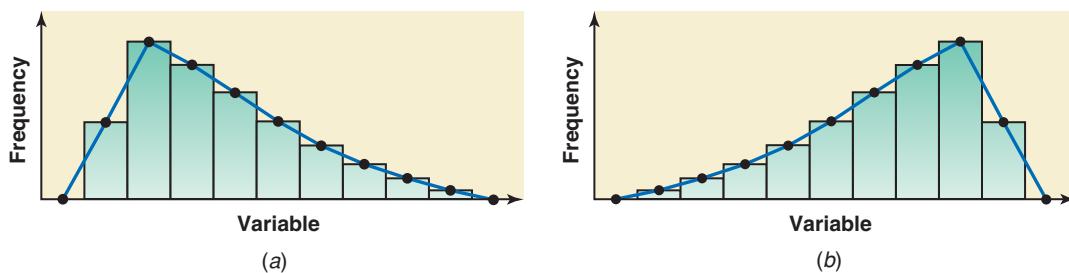
1. Symmetric
2. Skewed
3. Uniform or rectangular

A **symmetric histogram** is identical on both sides of its central point. The histograms shown in Figure 2.8 are symmetric around the dashed lines that represent their central points.



**Figure 2.8** Symmetric histograms.

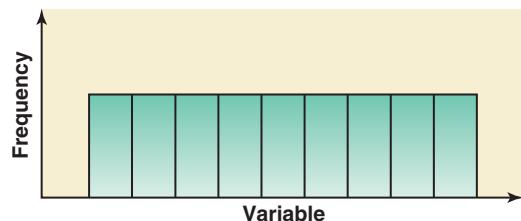
A **skewed histogram** is nonsymmetric. For a skewed histogram, the tail on one side is longer than the tail on the other side. A **skewed-to-the-right histogram** has a longer tail on the right side (see Figure 2.9a). A **skewed-to-the-left histogram** has a longer tail on the left side (see Figure 2.9b).



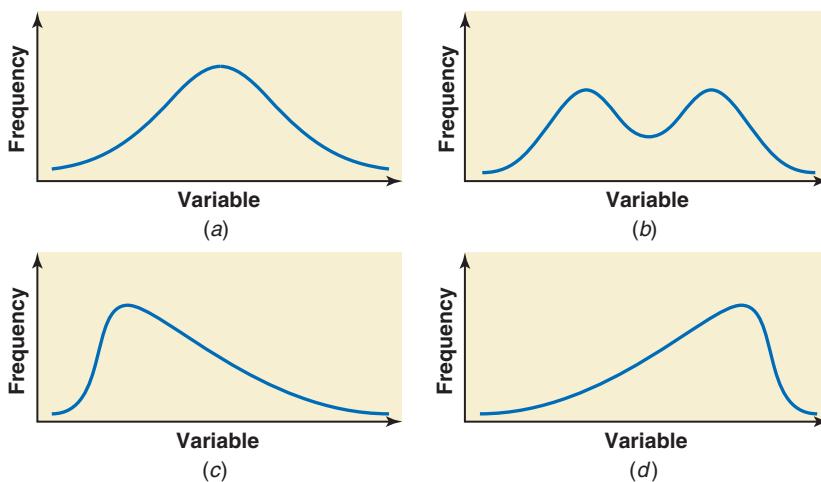
**Figure 2.9** (a) A histogram skewed to the right. (b) A histogram skewed to the left.

A **uniform** or **rectangular histogram** has the same frequency for each class. Figure 2.10 is an illustration of such a case.

**Figure 2.10** A histogram with uniform distribution.



Figures 2.11a and 2.11b display symmetric frequency curves. Figures 2.11c and 2.11d show frequency curves skewed to the right and to the left, respectively.



**Figure 2.11** (a), (b) Symmetric frequency curves. (c) Frequency curve skewed to the right. (d) Frequency curve skewed to the left.

**Warning ▶** Describing data using graphs give us insights into the main characteristics of the data. But graphs, unfortunately, can also be used, intentionally or unintentionally, to distort the facts and deceive the reader. The following are two ways to manipulate graphs to convey a particular opinion or impression.

1. *Changing the scale* either on one or on both axes—that is, shortening or stretching one or both of the axes.
2. *Truncating the frequency axis*—that is, starting the frequency axis at a number greater than zero.

When interpreting a graph, we should be very cautious. We should observe carefully whether the frequency axis has been truncated or whether any axis has been unnecessarily shortened or stretched. See the Uses and Misuses section of this chapter for such an example.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**2.11** Briefly explain the three decisions that have to be made to group a data set in the form of a frequency distribution table.

**2.12** How are the relative frequencies and percentages of classes obtained from the frequencies of classes? Illustrate with the help of an example.

**2.13** Three methods—writing classes using limits, using the *less-than* method, and grouping data using single-valued classes—were discussed to group quantitative data into classes. Explain these three methods and give one example of each.

## ■ APPLICATIONS

**2.14** A local gas station collected data from the day's receipts, recording the gallons of gasoline each customer purchased. The following table lists the frequency distribution of the gallons of gas purchased by all customers on this one day at this gas station.

Gallons of Gas	Number of Customers
0 to less than 4	31
4 to less than 8	78
8 to less than 12	49
12 to less than 16	81
16 to less than 20	117
20 to less than 24	13

- a. How many customers were served on this day at this gas station?
- b. Find the class midpoints. Do all of the classes have the same width? If so, what is this width? If not, what are the different class widths?
- c. Prepare the relative frequency and percentage distribution columns.
- d. What percentage of the customers purchased 12 gallons or more?
- e. Explain why you cannot determine exactly how many customers purchased 10 gallons or less.

**2.15** A staff member at a local grocery store was assigned the job of inspecting all containers of yogurt in the store to determine the number of days to expiry date for each container. Containers that had already expired but were still on the shelves were given a value of 0 for number of days to expiry. The following table gives the frequency distribution of the number of days to expiry date.

Number of Days	Number of Containers
0 to 5	32
6 to 11	67
12 to 17	44
18 to 23	20
24 to 29	11

- a. How many containers of yogurt were inspected?
- b. Find the class midpoints. Do all of the classes have the same width? If so, what is this width? If not, what are the different class widths?
- c. Prepare the relative frequency and percentage distribution columns.
- d. What percentage of the containers will expire in less than 18 days?
- e. Explain why you cannot determine exactly how many containers have already expired.
- f. What is the largest number of containers that may have already expired?

**2.16** A data set on money spent on lottery tickets during the past year by 200 households has a lowest value of \$1 and a highest value of \$1167. Suppose we want to group these data into six classes of equal widths.

- a. Assuming that we take the lower limit of the first class as \$1 and the width of each class equal to \$200, write the class limits for all six classes.
- b. What are the class boundaries and class midpoints?

**2.17** A data set on monthly expenditures (rounded to the nearest dollar) incurred on fast food by a sample of 500 households has a minimum value of \$3 and a maximum value of \$147. Suppose we want to group these data into six classes of equal widths.

- a. Assuming that we take the lower limit of the first class as \$1 and the upper limit of the sixth class as \$150, write the class limits for all six classes.
- b. Determine the class boundaries and class widths.
- c. Find the class midpoints.

**2.18** The accompanying table lists the 2010 median household incomes, rounded to the nearest dollar, for all 50 states and the District of Columbia.

State	2010 Median Household Income (dollars)	State	2010 Median Household Income (dollars)
AL	40,976	MT	41,467
AK	58,198	NE	52,728
AZ	47,279	NV	51,525
AR	38,571	NH	66,707
CA	54,459	NJ	63,540
CO	60,442	NM	45,098
CT	66,452	NY	49,826
DE	55,269	NC	43,753
DC	55,528	ND	51,380
FL	44,243	OH	46,093
GA	44,108	OK	43,400
HI	58,507	OR	50,526
ID	47,014	PA	48,460
IL	50,761	RI	51,914
IN	46,322	SC	41,709
IA	49,177	SD	45,669
KS	46,229	TN	38,686
KY	41,236	TX	47,464
LA	39,443	UT	56,787
ME	48,133	VT	55,942
MD	64,025	VA	60,363
MA	61,333	WA	56,253
MI	46,441	WV	42,839
MN	52,554	WI	50,522
MS	37,985	WY	52,359
MO	46,184		

Source: U.S. Census Bureau.

- Construct a frequency distribution table. Use the following classes: \$37,000–41,999, \$42,000–46,999, \$47,000–51,999, \$52,000–56,999, \$57,000–61,999, and \$62,000–66,999.
- Calculate the relative frequency and percentage for each class.
- Based on the frequency distribution, do the data appear to be symmetric or skewed?
- What percentage of these states had an estimated median household income of \$52,000 or more?

**2.19** Each state collects information on every birth that occurs within its borders. The following data give the 2008 birth rates (number of births per 1000 people) for all of the 56 counties in the state of Montana (<http://www.dphhs.mt.gov/statisticalinformation/vitalstats/index.shtml>).

10.1	22.2	15.8	12.2	7.7	3.1	14.5	7.8
13.6	8.8	10.9	8.9	14.7	9.6	14.2	14.9
18.3	22.8	5.4	5.6	19.6	8.2	9.9	14.7
13.7	10.3	9.7	9.8	8.6	9.4	14.1	12.3
10.5	11.4	2.2	9.8	10.9	4.6	6.6	8.5
10.2	14.4	20.4	18.5	10.8	6.5	11.6	12.1
10.5	9.3	8.1	7.4	10.2	9.7	5.6	14.5

- Construct a frequency distribution table using the classes 2 to less than 5, 5 to less than 8, 8 to less than 11, 11 to less than 14, 14 to less than 17, 17 to less than 20, and 20 to less than 23.
- Calculate the relative frequency and percentage for each class.
- Construct a histogram and a polygon for the birth-rate percentage distribution.
- What percentage of the counties had a birth rate of less than 11 births per 1000 people?

**2.20** The National Highway Traffic Safety Administration collects information on fatal accidents that occur on roads in the United States. Following are the number of fatal motorcycle accidents that occurred in each of South Carolina's 46 counties during the year 2009 (<http://www-fars.nhtsa.dot.gov>).

3	28	3	35	3	7	13	38	6	44	11	14
12	18	17	17	6	20	3	7	29	17	51	12
5	60	12	18	17	21	14	34	3	12	8	5
11	29	20	40	3	30	23	5	10	23		

- a. Construct a frequency distribution table using the classes 1–10, 11–20, 21–30, 31–40, 41–50, and 51–60.
- b. Calculate the relative frequency and percentage for each class.
- c. Construct a histogram and a polygon for the relative frequency distribution of part b.
- d. What percentage of the counties had between 21 and 40 fatal motorcycle accidents during 2009?

**2.21** Since 1996, Slate.com has published the *Slate 60* (compiled by the Chronicle of Philanthropy), which is a list of U.S. individuals making the largest charitable contributions each year. The accompanying table gives the names of the top 40 persons in the 2010 *Slate 60* and the money they donated (in millions) during that year (<http://www.slate.com/id/2283787/>).

Donor	Donation (millions of dollars)	Donor	Donation (millions of dollars)
George Soros	332.0	Paul Ichiro Terasaki	50.0
Michael R. Bloomberg	279.2	P. Roy and Diana T. Vagelos	50.0
T. Denny Sanford	162.5	Bennett S. LeBow	49.0
Irwin M. and Joan K. Jacobs	119.5	Lawrence J. Ellison	45.1
Eli and Edythe L. Broad	118.3	Lee G. and Jane H. Seidman	42.0
Leonard Blavatnik	117.2	Violet L. Patton	41.3
Frances Lasker Brody	110.0	Lonnie C. Jr. and Carol Johnson Poole	40.2
T. Boone Pickens	101.0	Lin Arison	39.0
Meyer and Renee Luskin	100.5	Herman Ostrow	35.0
Marc R. and Lynne Benioff	100.0	Jon L. Stryker	32.8
Mark Zuckerberg	100.0	Paul G. Allen	32.3
Terrence M. and Kim Pegula	88.0	Norton Herrick	32.0
Juanita Kious Waugh	83.7	Edward H. and Vivian Merrin	30.2
David R. and Patricia D. Atkinson	80.0	William P. and Lou W. Kennedy	30.1
Henry C. Jr. and Jane C. Woods	67.0	John C. Malone	30.0
Pierre and Pam Omidyar	61.5	Alvin S. and Terese Lane	30.0
William A. and Karen Ackman	59.3	Tamsen Ann Ziff	30.0
Charles E. Kaufman	53.3	Theodore and Vada Stanley	29.1
Ming Hsieh	50.0	Stephen and Nancy Grand	28.1
Edward P. (Ned) Evans	50.0	David M. Rubenstein	26.6

- a. Construct a frequency distribution table using the classes 25 to less than 65, 65 to less than 105, 105 to less than 145, and so on.
- b. Calculate the relative frequency and percentage for each class.
- c. Construct a histogram for the relative frequency distribution of part b.
- d. Are there any donation amounts that stand out in the histogram? If so, how do they compare to the rest of the donation amounts?

#### **Exercises 2.22 through 2.26 are based on the following data**

The following table gives the age-adjusted cancer incidence rates (new cases) per 100,000 people for three of the most common types of cancer contracted by both females and males: colon and rectum cancer, lung and bronchus cancer, and non-Hodgkin lymphoma. The rates given are for the District of Columbia and 26 states east of the Mississippi River for the years 2003 to 2007, which are the most recent data available

from the American Cancer Society. Age-adjusted rates take into account the percentage of people in different age groups within each state's population.

State	Colon and Rectum (Males)	Colon and Rectum (Females)	Lung and Bronchus (Males)	Lung and Bronchus (Females)	Non-Hodgkin Lymphoma (Males)	Non-Hodgkin Lymphoma (Females)
Alabama	60.8	41.6	106.2	53.4	20.5	13.8
Connecticut	59.4	44.4	80.5	60.3	26.0	18.1
Delaware	61.4	44.0	98.0	70.7	23.9	16.6
D.C.	58.1	47.9	79.4	46.3	22.9	13.4
Florida	53.1	40.4	86.7	59.4	21.5	15.2
Georgia	56.9	41.2	98.8	53.9	21.1	14.3
Illinois	65.6	47.3	91.2	59.4	24.2	16.2
Indiana	61.3	45.2	102.4	63.9	22.9	17.0
Kentucky	67.6	48.9	131.3	78.2	23.5	17.1
Maine	61.6	47.2	99.1	66.6	24.6	18.8
Maryland	54.4	41.3	81.5	57.9	20.9	14.4
Massachusetts	60.5	43.9	82.2	63.1	24.5	16.9
Michigan	57.1	43.4	91.9	62.5	25.7	18.7
Mississippi	63.5	45.9	113.3	55.0	20.4	14.0
New Hampshire	56.0	43.1	82.5	62.4	23.5	18.1
New Jersey	62.6	46.0	78.3	56.3	25.6	17.7
New York	58.4	44.3	78.2	54.3	25.0	17.5
North Carolina	56.0	40.9	101.0	57.6	21.9	15.4
Ohio	60.0	44.5	96.1	59.7	23.1	16.4
Pennsylvania	63.9	47.4	90.0	57.1	25.0	17.5
Rhode Island	61.8	45.7	92.6	61.9	24.9	17.4
South Carolina	58.5	42.8	100.2	53.7	20.8	14.4
Tennessee	57.9	40.8	93.6	54.5	22.8	16.3
Vermont	49.4	42.9	84.5	61.1	23.8	18.3
Virginia	54.2	41.0	88.5	53.8	20.8	13.9
West Virginia	68.0	48.7	116.3	71.3	24.0	17.3
Wisconsin	54.6	42.2	76.8	53.8	25.5	18.7

Source: American Cancer Society, [www.cancer.org/downloads/STT/2008CAFFfinalsecured.pdf](http://www.cancer.org/downloads/STT/2008CAFFfinalsecured.pdf).

- 2.22** a. Prepare a frequency distribution table for colon and rectum cancer rates for women using six classes of equal width.  
 b. Construct the relative frequency and percentage distribution columns.
- 2.23** a. Prepare a frequency distribution table for colon and rectum cancer rates for men using six classes of equal width.  
 b. Construct the relative frequency and percentage distribution columns.
- 2.24** a. Prepare a frequency distribution table for lung and bronchus cancer rates for women.  
 b. Construct the relative frequency and percentage distribution columns.  
 c. Draw a histogram and polygon for the relative frequency distribution.
- 2.25** a. Prepare a frequency distribution table for lung and bronchus cancer rates for men.  
 b. Construct the relative frequency and percentage distribution columns.  
 c. Draw a histogram and polygon for the relative frequency distribution.
- 2.26** a. Prepare a frequency distribution table for non-Hodgkin lymphoma rates for women.  
 b. Construct the relative frequency and percentage distribution columns.  
 c. Draw a histogram and polygon for the relative frequency distribution.

**2.27** The following table lists the number of strikeouts per game (K/game) for each of the 30 Major League baseball teams during the 2010 regular season. Note that Florida Marlins are now Miami Marlins.

Team	K/game	Team	K/game	Team	K/game
Arizona Diamondbacks	9.44	Florida Marlins	8.49	Philadelphia Phillies	6.57
Atlanta Braves	7.04	Houston Astros	6.33	Pittsburgh Pirates	7.45
Baltimore Orioles	6.52	Kansas City Royals	5.59	San Diego Padres	7.30
Boston Red Sox	7.04	Los Angeles Angels	6.60	San Francisco Giants	6.78
Chicago Cubs	7.63	Los Angeles Dodgers	7.31	Seattle Mariners	7.31
Chicago White Sox	5.69	Milwaukee Brewers	7.51	St. Louis Cardinals	6.34
Cincinnati Reds	7.52	Minnesota Twins	5.97	Tampa Bay Rays	7.98
Cleveland Indians	7.31	New York Mets	6.76	Texas Rangers	6.09
Colorado Rockies	7.86	New York Yankees	7.01	Toronto Blue Jays	7.19
Detroit Tigers	7.08	Oakland Athletics	6.55	Washington Nationals	7.53

Source: MLB.com.

- a. Construct a frequency distribution table. Take 5.50 as the lower boundary of the first class and .8 as the width of each class.
- b. Prepare the relative frequency and percentage distribution columns for the frequency distribution table of part a.

**2.28** The following data give the number of turnovers (fumbles and interceptions) made by both teams in each of the football games played by North Carolina State University during the 2009 and 2010 seasons.

2    3    1    1    6    5    3    5    5    1    5    2    1  
5    3    4    4    5    8    4    5    2    2    2    6

- a. Construct a frequency distribution table for these data using single-valued classes.
- b. Calculate the relative frequency and percentage for each class.
- c. What is the relative frequency of games in which there were 4 or 5 turnovers?
- d. Draw a bar graph for the frequency distribution of part a.

**2.29** Twenty-four patrons at a baseball game were observed in order to determine how many hot dogs each of them ate during the game. The following table contains the data.

4    2    1    2    1    0    2    2    2    3    0    3  
3    4    1    4    6    1    5    0    0    2    3    2

- a. Construct a frequency distribution table for these data using single-valued classes.
- b. Calculate the relative frequency and percentage for each class.
- c. What is the relative frequency of patrons who ate fewer than 4 hot dogs?
- d. Draw a bar graph for the frequency distribution of part a.

**2.30** The following table gives the frequency distribution for the numbers of parking tickets received on the campus of a university during the past week by 200 students.

Number of Tickets	Number of Students
0	59
1	44
2	37
3	32
4	28

Draw two bar graphs for these data, the first without truncating the frequency axis and the second by truncating the frequency axis. In the second case, mark the frequencies on the vertical axis starting with 25. Briefly comment on the two bar graphs.

**2.31** Eighty adults were asked to watch a 30-minute infomercial until the presentation ended or until boredom became intolerable. The following table lists the frequency distribution of the times that these adults were able to watch the infomercial.

Time (minutes)	Number of Adults
0 to less than 6	16
6 to less than 12	21
12 to less than 18	18
18 to less than 24	11
24 to less than 30	14

Draw two histograms for these data, the first without truncating the frequency axis. In the second case, mark the frequencies on the vertical axis starting with 10. Briefly comment on the two histograms.

## 2.3 Cumulative Frequency Distributions

Consider again Example 2–3 of Section 2.2.2 about the total number of iPods sold by a company. Suppose we want to know on how many days the company sold 19 or fewer iPods. Such a question can be answered by using a **cumulative frequency distribution**. Each class in a cumulative frequency distribution table gives the total number of values that fall below a certain value. A cumulative frequency distribution is constructed for quantitative data only.

### Definition

**Cumulative Frequency Distribution** A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class.

In a cumulative frequency distribution table, each class has the same lower limit but a different upper limit. Example 2–7 illustrates the procedure for preparing a cumulative frequency distribution.

### ■ EXAMPLE 2–7

Constructing a cumulative frequency distribution table.

Using the frequency distribution of Table 2.9, reproduced here, prepare a cumulative frequency distribution for the number of iPods sold by that company.

iPods Sold	<i>f</i>
5–9	3
10–14	6
15–19	8
20–24	8
25–29	5

**Solution** Table 2.14 gives the cumulative frequency distribution for the number of iPods sold. As we can observe, 5 (which is the lower limit of the first class in Table 2.9) is taken as the lower limit of each class in Table 2.14. The upper limits of all classes in Table 2.14 are the same as those in Table 2.9. To obtain the cumulative frequency of a class, we add the frequency of that class in Table 2.9 to the frequencies of all preceding classes. The cumulative frequencies are recorded in the third column of Table 2.14. The second column of this table lists the class boundaries.

**Table 2.14** Cumulative Frequency Distribution of iPods Sold

Class Limits	Class Boundaries	Cumulative Frequency
5–9	4.5 to less than 9.5	3
5–14	4.5 to less than 14.5	3 + 6 = 9
5–19	4.5 to less than 19.5	3 + 6 + 8 = 17
5–24	4.5 to less than 24.5	3 + 6 + 8 + 8 = 25
5–29	4.5 to less than 29.5	3 + 6 + 8 + 8 + 5 = 30

From Table 2.14, we can determine the number of observations that fall below the upper limit or boundary of each class. For example, 19 or fewer iPods were sold on 17 days. ■

The **cumulative relative frequencies** are obtained by dividing the cumulative frequencies by the total number of observations in the data set. The **cumulative percentages** are obtained by multiplying the cumulative relative frequencies by 100.

### Calculating Cumulative Relative Frequency and Cumulative Percentage

$$\text{Cumulative relative frequency} = \frac{\text{Cumulative frequency of a class}}{\text{Total observations in the data set}}$$

$$\text{Cumulative percentage} = (\text{Cumulative relative frequency}) \cdot 100\%$$

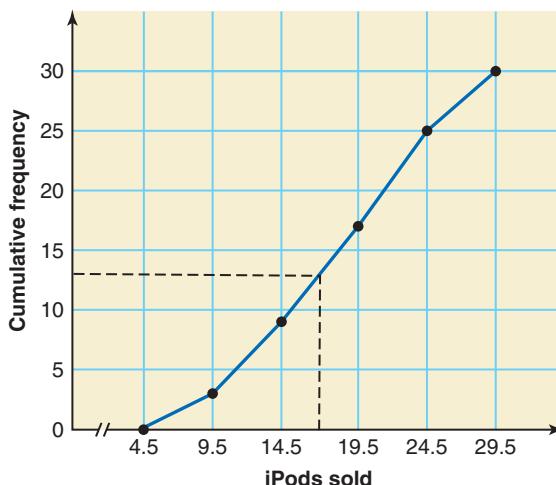
Table 2.15 contains both the cumulative relative frequencies and the cumulative percentages for Table 2.14. We can observe, for example, that 19 or fewer iPods were sold on 56.7% of the days.

**Table 2.15** Cumulative Relative Frequency and Cumulative Percentage Distributions for iPods Sold

Class Limits	Cumulative Relative Frequency	Cumulative Percentage
5–9	3/30 = .100	10.0
5–14	9/30 = .300	30.0
5–19	17/30 = .567	56.7
5–24	25/30 = .833	83.3
5–29	30/30 = 1.000	100.0

### Ogives

When plotted on a diagram, the cumulative frequencies give a curve that is called an **ogive** (pronounced *o-jive*). Figure 2.12 gives an ogive for the cumulative frequency distribution of Table 2.14. To draw the ogive in Figure 2.12, the variable, which is total iPods sold, is marked on the horizontal axis and the cumulative frequencies on the vertical axis. Then the dots are marked above the upper boundaries of various classes at the heights equal to the corresponding cumulative frequencies. The ogive is obtained by joining consecutive points with straight lines. Note that the ogive starts at the lower boundary of the first class and ends at the upper boundary of the last class.



**Figure 2.12** Ogive for the cumulative frequency distribution of Table 2.14.

### Definition

**Ogive** An *ogive* is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at heights equal to the cumulative frequencies of respective classes.

One advantage of an ogive is that it can be used to approximate the cumulative frequency for any interval. For example, we can use Figure 2.12 to estimate the number of days for which 17 or fewer iPods were sold. First, draw a vertical line from 17 on the horizontal axis up to the ogive. Then draw a horizontal line from the point where this line intersects the ogive to the vertical axis. This point gives the estimated cumulative frequency of the class 5 to 17. In Figure 2.12, this cumulative frequency is (approximately) 13 as shown by the dashed line. Therefore, 17 or fewer iPods were sold on (approximately) 13 days.

We can draw an ogive for cumulative relative frequency and cumulative percentage distributions the same way as we did for the cumulative frequency distribution.

## EXERCISES

### CONCEPTS AND PROCEDURES

**2.32** Briefly explain the concept of cumulative frequency distribution. How are the cumulative relative frequencies and cumulative percentages calculated?

**2.33** Explain for what kind of frequency distribution an ogive is drawn. Can you think of any use for an ogive? Explain.

### APPLICATIONS

**2.34** The following table, reproduced from Exercise 2.14, gives the frequency distribution of the gallons of gasoline purchased by all customers on one day at a certain gas station.

Number of Gallons	Number of Customers
0 to less than 4	31
4 to less than 8	78
8 to less than 12	49
12 to less than 16	81
16 to less than 20	117
20 to less than 24	13

- Prepare a cumulative frequency distribution.
- Calculate the cumulative relative frequency and cumulative percentage for each class.
- Find the percentage of customers who purchased less than 16 gallons.
- Draw an ogive for the cumulative percentage distribution.
- Using the ogive, estimate the percentage of customers who purchased less than 10 gallons.

**2.35** The following table, reproduced from Exercise 2.15, gives the frequency distribution of the number of days to expiry date for all containers of yogurt in stock at a local grocery store. Containers that had already expired but were still on the shelves were given a value of 0 for number of days to expiry date.

Number of Days	Number of Containers
0 to 5	32
6 to 11	67
12 to 17	44
18 to 23	20
24 to 29	11

- Prepare a cumulative frequency distribution.
- Calculate the cumulative relative frequency and cumulative percentage for each class.
- Find the percentage of the containers that will expire in 12 or more days.
- Draw an ogive for the cumulative percentage distribution.
- Using the ogive, estimate the percentage of containers that will expire in fewer than 20 days.

**2.36** Using the frequency distribution table constructed in Exercise 2.18, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

**2.37** Using the frequency distribution table constructed in Exercise 2.19, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

**2.38** Using the frequency distribution table constructed in Exercise 2.20, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

**2.39** Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution constructed in Exercise 2.23.

**2.40** Using the frequency distribution table constructed for the data of Exercise 2.25, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

**2.41** Refer to the frequency distribution table constructed in Exercise 2.26. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using that table.

**2.42** Using the frequency distribution table constructed for the data of Exercise 2.21, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions. Draw an ogive for the cumulative frequency distribution. Using the ogive, find the (approximate) number of individuals who made charitable contributions of \$85 million or less.

**2.43** Refer to the frequency distribution table constructed in Exercise 2.27. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions. Draw an ogive for the cumulative frequency distribution. Using the ogive, find the (approximate) number of teams with 6.8 or fewer strikeouts per game.

## 2.4 Stem-and-Leaf Displays

Another technique that is used to present quantitative data in condensed form is the **stem-and-leaf display**. An advantage of a stem-and-leaf display over a frequency distribution is that by preparing a stem-and-leaf display we do not lose information on individual observations. A stem-and-leaf display is constructed only for quantitative data.

### Definition

**Stem-and-Leaf Display** In a *stem-and-leaf display* of quantitative data, each value is divided into two portions—a stem and a leaf. The leaves for each stem are shown separately in a display.

Example 2–8 describes the procedure for constructing a stem-and-leaf display.

### ■ EXAMPLE 2–8

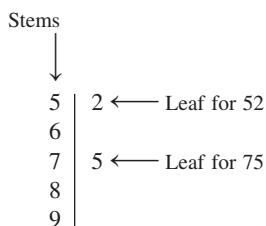
*Constructing a stem-and-leaf display for two-digit numbers.*

The following are the scores of 30 college students on a statistics test.

75	52	80	96	65	79	71	87	93	95
69	72	81	61	76	86	79	68	50	92
83	84	77	64	71	87	72	92	57	98

Construct a stem-and-leaf display.

**Solution** To construct a stem-and-leaf display for these scores, we split each score into two parts. The first part contains the first digit of a score, which is called the *stem*. The second part contains the second digit of a score, which is called the *leaf*. Thus, for the score of the first student, which is 75, 7 is the stem and 5 is the leaf. For the score of the second student, which is 52, the stem is 5 and the leaf is 2. We observe from the data that the stems for all scores are 5, 6, 7, 8, and 9 because all these scores lie in the range 50 to 98. To create a stem-and-leaf display, we draw a vertical line and write the stems on the left side of it, arranged in increasing order, as shown in Figure 2.13.



**Figure 2.13** Stem-and-leaf display.

After we have listed the stems, we read the leaves for all scores and record them next to the corresponding stems on the right side of the vertical line. For example, for the first score we write the leaf 5 next to the stem 7; for the second score we write the leaf 2 next to the stem 5. The recording of these two scores in a stem-and-leaf display is shown in Figure 2.13.

Now, we read all the scores and write the leaves on the right side of the vertical line in the rows of corresponding stems. The complete stem-and-leaf display for scores is shown in Figure 2.14.

5	2	0	7
6	5	9	1
7	8	4	
7	5	9	1
7	2	6	9
7	9	7	1
8	2	1	2
8	0	3	4
8	7	4	7
9	6	3	5
9	2	2	8

**Figure 2.14** Stem-and-leaf display of test scores.

By looking at the stem-and-leaf display of Figure 2.14, we can observe how the data values are distributed. For example, the stem 7 has the highest frequency, followed by stems 8, 9, 6, and 5.

The leaves for each stem of the stem-and-leaf display of Figure 2.14 are *ranked* (in increasing order) and presented in Figure 2.15.

5	0	2	7
6	1	4	5
6	8	9	
7	1	1	2
7	2	5	6
7	9	7	9
8	0	1	3
8	4	6	7
9	2	2	3
9	5	6	8

**Figure 2.15** Ranked stem-and-leaf display of test scores.

As already mentioned, one advantage of a stem-and-leaf display is that we do not lose information on individual observations. We can rewrite the individual scores of the 30 college students from the stem-and-leaf display of Figure 2.14 or Figure 2.15. By contrast, the information on individual observations is lost when data are grouped into a frequency table.

## ■ EXAMPLE 2–9

The following data give the monthly rents paid by a sample of 30 households selected from a small town.

880	1081	721	1075	1023	775	1235	750	965	960
1210	985	1231	932	850	825	1000	915	1191	1035
1151	630	1175	952	1100	1140	750	1140	1370	1280

Construct a stem-and-leaf display for these data.

Constructing a stem-and-leaf display for three- and four-digit numbers.

**Solution** Each of the values in the data set contains either three or four digits. We will take the first digit for three-digit numbers and the first two digits for four-digit numbers as stems. Then we will use the last two digits of each number as a leaf. Thus for the first value, which is 880, the stem is 8 and the leaf is 80. The stems for the entire data set are 6, 7, 8, 9, 10, 11, 12, and 13. They are recorded on the left side of the vertical line in Figure 2.16. The leaves for the numbers are recorded on the right side.

6	30
7	21 75 50 50
8	80 50 25
9	65 60 85 32 15 52
10	81 75 23 00 35
11	91 51 75 00 40 40
12	35 10 31 80
13	70

**Figure 2.16** Stem-and-leaf display of rents.

Sometimes a data set may contain too many stems, with each stem containing only a few leaves. In such cases, we may want to condense the stem-and-leaf display by *grouping the stems*. Example 2–10 describes this procedure.

## ■ EXAMPLE 2–10

The following stem-and-leaf display is prepared for the number of hours that 25 students spent working on computers during the past month.

Preparing a grouped stem-and-leaf display.

0	6
1	1 7 9
2	2 6
3	2 4 7 8
4	1 5 6 9 9
5	3 6 8
6	2 4 4 5 7
7	
8	5 6



Prepare a new stem-and-leaf display by grouping the stems.

Mark Harmel/Stone/Getty Images

**Solution** To condense the given stem-and-leaf display, we can combine the first three rows, the middle three rows, and the last three rows, thus getting the stems 0–2, 3–5, and 6–8. The leaves for each stem of a group are separated by an asterisk (\*), as shown in Figure 2.17. Thus, the leaf 6 in the first row corresponds to stem 0; the leaves 1, 7, and 9 correspond to stem 1; and leaves 2 and 6 belong to stem 2.

0–2	6 * 1 7 9 * 2 6
3–5	2 4 7 8 * 1 5 6 9 9 * 3 6 8
6–8	2 4 4 5 7 * * 5 6

**Figure 2.17** Grouped stem-and-leaf display.

If a stem does not contain a leaf, this is indicated in the grouped stem-and-leaf display by two consecutive asterisks. For example, in the stem-and-leaf display of Figure 2.17, there is no leaf for 7; that is, there is no number in the 70s. Hence, in Figure 2.17, we have two asterisks after the leaves for 6 and before the leaves for 8.

Some data sets produce stem-and-leaf displays that have a small number of stems relative to the number of observations in the data set and have too many leaves for each stem. In such cases, it is very difficult to determine if the distribution is symmetric or skewed, as well as other characteristics of the distribution that will be introduced in later chapters. In such a situation, we can create a stem-and-leaf display with *split stems*. To do this, each stem is split into two or five parts. Whenever the stems are split into two parts, any observation having a leaf with a value of 0, 1, 2, 3, or 4 is placed in the first split stem, while the leaves 5, 6, 7, 8, and 9 are placed in the second split stem. Sometimes we can split a stem into five parts if there are too many leaves for one stem. Whenever a stem is split into five parts, leaves with values of 0 and 1 are placed next to the first part of the split stem, leaves with values of 2 and 3 are placed next to the second part of the split stem, and so on. The stem-and-leaf display of Example 2–11 shows this procedure.

### ■ EXAMPLE 2–11

*Stem-and-leaf display with split stems.*

Consider the following stem-and-leaf display, which has only two stems. Using the split stem procedure, rewrite this stem-and-leaf display.

3	1 1 2 3 3 3 4 4 7 8 9 9 9
4	0 0 0 1 1 1 1 1 2 2 2 2 3 3 6 6 7

**Solution** To prepare a split stem-and-leaf display, let us split the two stems, 3 and 4, into two parts each as shown in Figure 2.18. The first part of each stem contains leaves from 0 to 4, and the second part of each stem contains leaves from 5 to 9.

3	1 1 2 3 3 3 4 4
3	7 8 9 9 9
4	0 0 0 1 1 1 1 1 2 2 2 2 3 3
4	6 6 7

**Figure 2.18** Split stem-and-leaf display.

In the stem-and-leaf display of Figure 2.18, the first part of stem 4 has a substantial number of leaves. So, if we decide to split stems into five parts, the new stem-and-leaf display will look as shown in Figure 2.19.

3	1 1
3	2 3 3 3
3	4 4
3	7
3	8 9 9 9
4	0 0 0 1 1 1 1 1
4	2 2 2 2 3 3
4	
4	6 6 7

**Figure 2.19** Split stem-and-leaf display.

There are two important properties to note in the split stem-and-leaf display of Figure 2.19. The third part of split stem 4 does not have any leaves. This implies that there are no observations in the data set having a value of 44 or 45. Since there are observations with values larger than 45, we need to leave an empty part of split stem 4 that corresponds to 44 and 45. Also, there are no observations with values of 48 or 49. However, since there are no values larger than 47 in the data, we do not have to write an empty split stem 4 after the largest value.

## EXERCISES

### CONCEPTS AND PROCEDURES

**2.44** Briefly explain how to prepare a stem-and-leaf display for a data set. You may use an example to illustrate.

**2.45** What advantage does preparing a stem-and-leaf display have over grouping a data set using a frequency distribution? Give one example.

**2.46** Consider the following stem-and-leaf display.

4	3 6
5	0 1 4 5
6	3 4 6 7 7 7 8 9
7	2 2 3 5 6 6 9
8	0 7 8 9

Write the data set that is represented by this display.

**2.47** Consider the following stem-and-leaf display.

2–3	18 45 56 * 29 67 83 97
4–5	04 27 33 71 * 23 37 51 63 81 92
6–8	22 36 47 55 78 89 * * 10 41

Write the data set that is represented by this display.

### APPLICATIONS

**2.48** The following data give the time (in minutes) that each of 20 students waited in line at their book-store to pay for their textbooks in the beginning of Spring 2012 semester. (*Note:* To prepare a stem-and-leaf display, each number in this data set can be written as a two-digit number. For example, 8 can be written as 08, for which the stem is 0 and the leaf is 8.)

15	8	23	21	5	17	31	22	34	6
5	10	14	17	16	25	30	3	31	19

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

**2.49** Following are the total yards gained rushing during the 2012 season by 14 running backs of 14 college football teams.

745	921	1133	1024	848	775	800
1009	1275	857	933	1145	967	995

Prepare a stem-and-leaf display. Arrange the leaves for each stem in increasing order.

**2.50** Refer to Exercise 2.20, which contains data on the number of fatal motorcycle accidents in each of South Carolina's 46 counties for the year 2009. Prepare a stem-and-leaf display for those data. Arrange the leaves for each stem in increasing order.

**2.51** Refer to Exercise 2.19, which contains data on the birth rates for all 56 counties of Montana for the year 2008. Here are the data rounded to the nearest unit:

10	22	16	12	8	3	15	8
14	9	11	9	15	10	14	15
18	23	5	6	20	8	10	15
14	10	10	10	9	9	14	12
11	11	2	10	11	5	7	9
10	14	20	19	11	7	12	12
11	9	8	7	10	10	6	15

- Prepare a stem-and-leaf display for the data. Arrange the leaves for each stem in increasing order.
- Prepare a split stem-and-leaf display for the data. Split each stem into two parts. The first part should contain the leaves 0, 1, 2, 3, and 4, and the second part should contain the leaves 5, 6, 7, 8, and 9.

- c. Which display (the one in part a or the one in part b) provides a better representation of the features of the distribution? Explain why you believe this.

**2.52** Refer to Exercise 2.21, which contains data on the amount of money donated to charity by the top 40 donors in the 2010 *Slate 60*. Here are the amounts rounded to the nearest million dollars.

332	279	163	120	118	117	110	101	101	100
100	88	84	80	67	62	59	53	50	50
50	50	49	45	42	41	40	39	35	33
32	32	30	30	30	30	30	29	28	27

- a. Prepare a stem-and-leaf display for the data. The stems should consist of the hundreds digits, and the leaves should consist of the tens and ones digits (e.g., for the number 117, the stem would be 1 and the leaf would be 17, while for the number 41, the stem would be 0 and the leaf would be 41). Arrange the leaves for each stem in increasing order.
- b. Prepare a split stem-and-leaf display for the data. Split each stem into two parts. The first part should contain the leaves with 0, 1, 2, 3, and 4 in the tens place, and the second part should contain the leaves with 5, 6, 7, 8, and 9 in the tens place.
- c. Which display (the one in part a or the one in part b) provides a better representation of the features of the distribution? Explain why you believe this.

**2.53** These data give the times (in minutes) taken to commute from home to work for 20 workers.

10	50	65	33	48	5	11	23	39	26
26	32	17	7	15	19	29	43	21	22

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

**2.54** The following data give the times served (in months) by 35 prison inmates who were released recently.

37	6	20	5	25	30	24	10	12	20
24	8	26	15	13	22	72	80	96	33
84	86	70	40	92	36	28	90	36	32
72	45	38	18	9					

- a. Prepare a stem-and-leaf display for these data.  
 b. Condense the stem-and-leaf display by grouping the stems as 0–2, 3–5, and 6–9.

**2.55** The following data give the money (in dollars) spent on textbooks by 35 students during the 2011–12 academic year.

565	728	870	620	345	868	610	765	550
845	530	705	490	258	320	505	957	787
617	721	635	438	575	702	538	720	460
840	890	560	570	706	430	968	638	

- a. Prepare a stem-and-leaf display for these data using the last two digits as leaves.  
 b. Condense the stem-and-leaf display by grouping the stems as 2–4, 5–6, and 7–9.

## 2.5 Dotplots

One of the simplest methods for graphing and understanding quantitative data is to create a dotplot. As with most graphs, statistical software should be used to make a dotplot for large data sets. However, Example 2–12 demonstrates how to create a dotplot by hand.

Dotplots can help us detect **outliers** (also called **extreme values**) in a data set. Outliers are the values that are extremely large or extremely small with respect to the rest of the data values.

### Definition

**Outliers or Extreme Values** Values that are very small or very large relative to the majority of the values in a data set are called outliers or extreme values.

## ■ EXAMPLE 2-12

Table 2.16 lists the number of minutes for which each player of the Boston Bruins hockey team was penalized during the 2011 Stanley Cup championship playoffs. Create a dotplot for these data.

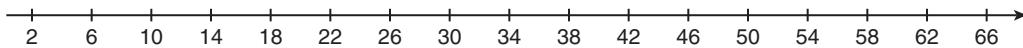
*Creating a dotplot.*

**Table 2.16** Penalty Minutes for Players of the Boston Bruins Hockey Team During the 2011 Stanley Cup Playoffs

Name	Penalty Minutes	Name	Penalty Minutes
Adam McQuaid	14	Michael Ryder	8
Andrew Ference	37	Milan Lucic	63
Brad Marchand	40	Nathan Horton	35
Chris Kelly	6	Patrice Bergeron	28
Daniel Paille	4	Rich Peverley	17
David Krejci	10	Shane Hnidy	7
Dennis Seidenberg	31	Shawn Thornton	24
Gregory Campbell	4	Tomas Kaberle	4
Johnny Boychuk	12	Tyler Seguin	2
Mark Recchi	8	Zdeno Chara	34

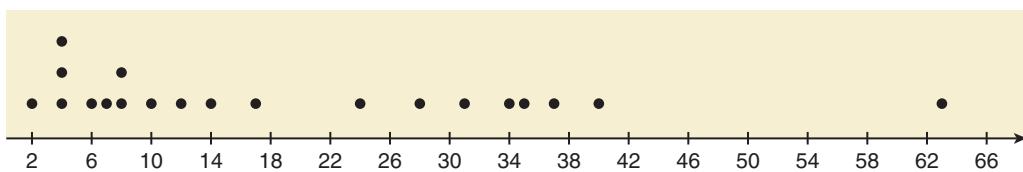
**Solution** Here we show how to make a dotplot for these data on penalty minutes.

**Step 1.** The minimum and maximum values in this data set are 2 and 63 minutes, respectively. First, we draw a horizontal line (let us call this the *numbers line*) with numbers that cover the given data as shown in Figure 2.20. Note that the numbers line in Figure 2.20 shows the values from 2 to 66.



**Figure 2.20** Numbers line.

**Step 2.** Next we place a dot above the value on the numbers line that represents each of the penalty minutes listed in the table. For example, Adam McQuaid had 14 penalty minutes. So, we place a dot above 14 on the numbers line as shown in Figure 2.21. If there are two or more observations with the same value, we stack dots above each other to represent those values. For example, as shown in Table 2.16, three members of the Boston Bruins team each had 4 penalty minutes. We stack three dots (one for each player) above 4 on the numbers line, as shown in Figure 2.21. After all the dots are placed, Figure 2.21 gives the complete dotplot.



**Figure 2.21** Dotplot for playoff penalty minutes.

As we examine the dotplot of Figure 2.21, we notice that there are two clusters (groups) of data. Sixty percent of the players had 17 or fewer penalty minutes during the playoffs, while the other 40% had 24 or more penalty minutes. Moreover, one player, Milan Lucic, amassed 63 penalty minutes, which is at least 23 minutes more than any other player. When this occurs, we suspect that such a data value could be an outlier. (In the box-and-whisker section of Chapter 3, we will learn a numerical method to determine whether a data point should be classified as an outlier.)

Dotplots are also very useful for comparing two or more data sets. To do so, we create a dotplot for each data set with numbers line and these numbers lines for all data sets should be on the same scale. We place these data sets on top of each other, resulting in what are called **stacked dotplots**. Example 2–13 shows this procedure.

### ■ EXAMPLE 2–13

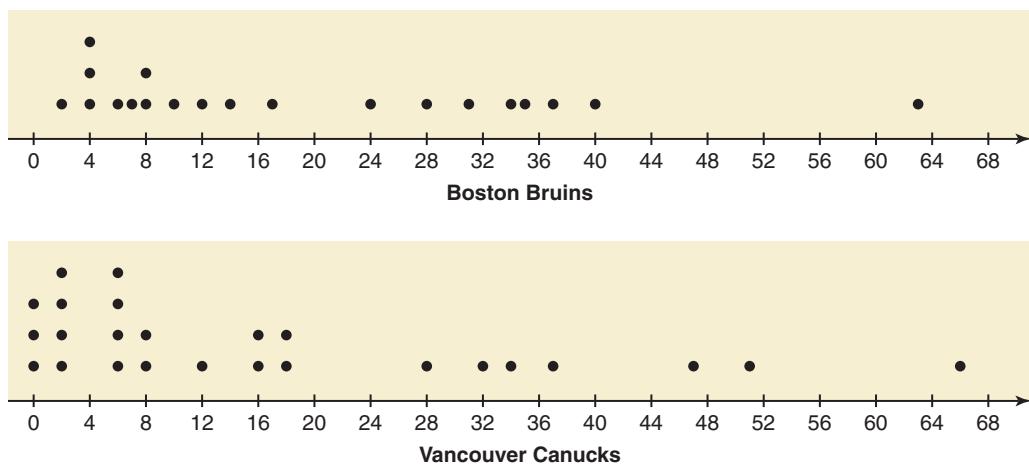
*Comparing two data sets using dotplots.*

Refer to Table 2.16 in Example 2–12, which lists the number of minutes for which each player of the 2011 Stanley Cup champion Boston Bruins hockey team was penalized during the playoffs. Table 2.17 provides the same information for the Vancouver Canucks, who lost in the finals to the Bruins in the 2011 Stanley Cup playoffs. Make dotplots for both sets of data and compare them.

**Table 2.17** Penalty Minutes for Players of the Vancouver Canucks Hockey Team During the 2011 Stanley Cup Playoffs

Name	Penalty Minutes	Name	Penalty Minutes
Aaron Rome	37	Jeff Tambellini	2
Alexander Edler	8	Keith Ballard	6
Alexandre Bolduc	0	Kevin Bieksa	51
Alexandre Burrows	34	Manny Malhotra	0
Andrew Alberts	6	Mason Raymond	6
Chris Higgins	2	Maxim Lapierre	66
Christian Ehrhoff	16	Mikael Samuelsson	8
Christopher Tanev	0	Raffi Torres	28
Cody Hodgson	2	Ryan Kesler	47
Dan Hamhuis	6	Sami Salo	2
Daniel Sedin	32	Tanner Glass	18
Henrik Sedin	16	Victor Oreskovich	12
Jannik Hansen	18		

**Solution** Figure 2.22 shows the dotplots for the given data for all players of both teams, the Boston Bruins and the Vancouver Canucks.



**Figure 2.22** Stacked dotplot of penalty minutes for the Boston Bruins and the Vancouver Canucks.

Looking at the stacked dotplot, we see that the majority of players on both teams had fewer than 20 penalty minutes each throughout the playoffs. Both teams have one outlier each, at 63 and 66 minutes, respectively. The two distributions of penalty minutes are similar in shape.

In practice, dotplots and other statistical graphs will be created using statistical software. The Technology Instruction section at the end of this chapter shows how we can do so.

## EXERCISES

### CONCEPTS AND PROCEDURES

**2.56** Briefly explain how to prepare a dotplot for a data set. You may use an example to illustrate.

**2.57** What is a stacked dotplot, and how is it used? Explain.

**2.58** Create a dotplot for the following data set.

1	2	0	5	1	1	3	2	0	5
2	1	2	1	2	0	1	3	1	2

### APPLICATIONS

**2.59** Refer to data given in Exercise 2.20 on the number of fatal motorcycle accidents in each of South Carolina's 46 counties during the year 2009. Create a dotplot for those data.

**2.60** Refer to data given in Exercise 2.28 on the number of turnovers (fumbles and interceptions) that occurred in each of North Carolina State University's football games during the 2009 and 2010 seasons. Create a dotplot for those data.

**2.61** Refer to data given in Exercise 2.29 on the number of hot dogs consumed by 24 patrons at a baseball game. Create a dotplot for those data.

**2.62** The following data give the number of times each of the 30 randomly selected account holders at a bank used that bank's ATM during a 60-day period.

3	2	3	2	2	5	0	4	1	3
2	3	3	5	9	0	3	2	2	15
1	3	2	7	9	3	0	4	2	2

Create a dotplot for these data and point out any clusters or outliers.

**2.63** The following data give the number of times each of the 20 randomly selected male students from a state university ate at fast-food restaurants during a 7-day period.

5	8	10	3	5	5	10	7	2	1
10	4	5	0	10	1	2	8	3	5

Create a dotplot for these data and point out any clusters or outliers.

**2.64** Reconsider Exercise 2.63. The following data give the number of times each of the 20 randomly selected female students from the same state university ate at fast-food restaurants during the same 7-day period.

0	0	4	2	4	10	2	5	0	5
6	1	1	4	6	2	4	5	6	0

a. Create a dotplot for these data.

b. Use the dotplots for male and female students to compare the two data sets.

**2.65** In basketball, a *double-double* occurs when a player accumulates double-digit numbers in any two of five statistical categories—points, rebounds, assists, steals, and blocked shots—in one game. The following table gives the number of times each player of the Miami Heat basketball team scored a double-double (DD) during the 2010–2011 regular season.

Player	DD	Player	DD
Carlos Arroyo	0	Juwan Howard	0
Chris Bosh	28	LeBron James	31
Dwyane Wade	10	Mario Chalmers	0
Eddie House	0	Mike Bibby	0
Erick Dampier	1	Mike Miller	3
James Jones	0	Udonis Haslem	4
Joel Anthony	0	Zydrunas Ilgauskas	3

Source: espn.com.

Create a dotplot for these data. Mention any clusters (groups) and/or outliers you observe.

## USES AND MISUSES... TRUNCATING THE AXES

Graphical analyses are an important part of statistics. However, statistical and spreadsheet software packages can allow users to change the appearance of any graph. In many cases, people will add extra features to their graphs in order to make them more exciting and eye-catching. Quite often, these changes will make the graphs misleading. As an example, consider the following data on the high temperature for each day during the week of January 22–28, 2012, at the summit of Mount Washington, New Hampshire (which happens to be known as the “Home of the World’s Worst Weather”):

Date in 2012	High Temperature (°F)
January 22	19
January 23	34
January 24	32
January 25	13
January 26	23
January 27	35
January 28	25

Source: Mount Washington Observatory, [www.mountainwashington.org/weather/f6/2012/01.pdf](http://www.mountainwashington.org/weather/f6/2012/01.pdf).

Quite often, for these kind of data, a graph that is produced for a newspaper article or weather report looks like the one given in Figure 2.23.

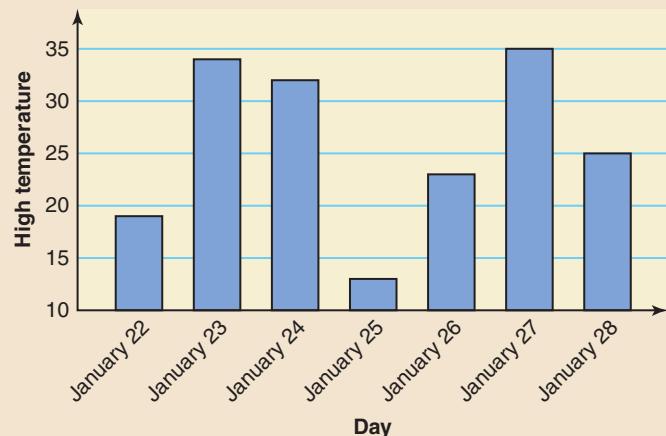


Figure 2.23 Bar chart for high temperatures.

This graph was created using typical bar chart commands. Therefore, if it is a proper bar chart, the heights of the bars should represent frequencies or relative frequencies, but they do not. Hence, it is inappropriate to use a bar chart in this situation. Moreover, there is another issue that would be a problem if this was a proper bar chart. If you look at the heights of bars for January 25 and 26, it appears that the temperature on January 26 is four times as high as it was on January 25. This is due to the fact that the vertical axis begins at 10, not zero. If you look at the data in the table, you will observe that the temperature increased from 13°F on January 25 to 23°F on January 26, which means that the temperature increased by less than 100%. Setting the minimum for the vertical axis at zero will prevent the reader from drawing the wrong conclusion due to a visual *trick*.

As you learn about the various types of graphs, it is important to learn which types of variables should be depicted in each type of graph. For the previous data, the best graph for displaying the data is called a scatterplot, as shown in Figure 2.24.

In this graph, the data are shown in chronological order, but the vertical axis now represents the temperature. The line segments that connect the dots are optional but can be quite useful when a scatterplot contains more data. In addition, the vertical axis begins at zero in this graph, so the visual interpretation of the temperature changes is appropriate. Scatterplots and their uses will be discussed in greater detail in Chapter 13.

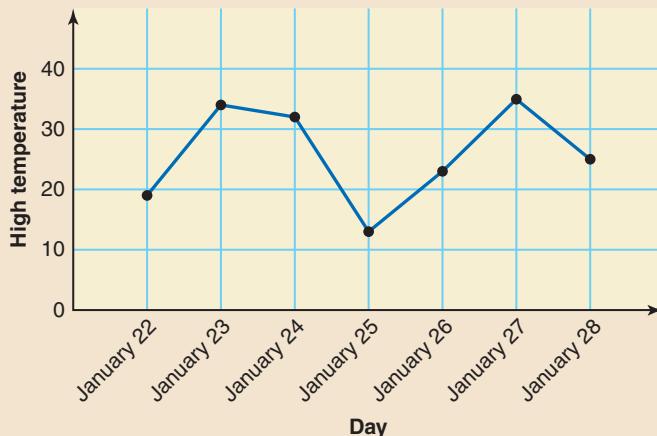


Figure 2.24 Scatterplot of high temperature.

## Glossary

**Bar graph** A graph made of bars whose heights represent the frequencies of respective categories.

**Class** An interval that includes all the values in a (quantitative) data set that fall within two numbers, the lower and upper limits of the class.

**Class boundary** The midpoint of the upper limit of one class and the lower limit of the next class.

**Class frequency** The number of values in a data set that belong to a certain class.

**Class midpoint or mark** The class midpoint or mark is obtained by dividing the sum of the lower and upper limits (or boundaries) of a class by 2.

**Class width or size** The difference between the two boundaries of a class.

**Cumulative frequency** The frequency of a class that includes all values in a data set that fall below the upper boundary of that class.

**Cumulative frequency distribution** A table that lists the total number of values that fall below the upper boundary of each class.

**Cumulative percentage** The cumulative relative frequency multiplied by 100.

**Cumulative relative frequency** The cumulative frequency of a class divided by the total number of observations.

**Frequency distribution** A table that lists all the categories or classes and the number of values that belong to each of these categories or classes.

**Grouped data** A data set presented in the form of a frequency distribution.

**Histogram** A graph in which classes are marked on the horizontal axis and frequencies, relative frequencies, or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentages of various classes are represented by the heights of bars that are drawn adjacent to each other.

**Ogive** A curve drawn for a cumulative frequency distribution.

**Outliers or Extreme values** Values that are very small or very large relative to the majority of the values in a data set.

**Percentage** The percentage for a class or category is obtained by multiplying the relative frequency of that class or category by 100.

**Pie chart** A circle divided into portions that represent the relative frequencies or percentages of different categories or classes.

**Polygon** A graph formed by joining the midpoints of the tops of successive bars in a histogram by straight lines.

**Raw data** Data recorded in the sequence in which they are collected and before they are processed.

**Relative frequency** The frequency of a class or category divided by the sum of all frequencies.

**Skewed-to-the-left histogram** A histogram with a longer tail on the left side.

**Skewed-to-the-right histogram** A histogram with a longer tail on the right side.

**Stem-and-leaf display** A display of data in which each value is divided into two portions—a stem and a leaf.

**Symmetric histogram** A histogram that is identical on both sides of its central point.

**Ungrouped data** Data containing information on each member of a sample or population individually.

**Uniform or rectangular histogram** A histogram with the same frequency for all classes.

## Supplementary Exercises

**2.66** The following data give the political party of each of the first 30 U.S. presidents. In the data, D stands for Democrat, DR for Democratic Republican, F for Federalist, R for Republican, and W for Whig.

F	F	DR	DR	DR	DR	D	D	W	W
D	W	W	D	D	R	D	R	R	R
R	D	R	D	R	R	R	D	R	R

- Prepare a frequency distribution table for these data.
- Calculate the relative frequency and percentage distributions.
- Draw a bar graph for the relative frequency distribution and a pie chart for the percentage distribution.
- What percentage of these presidents were Whigs?

**2.67** In an April 18, 2010 Pew Research Center report entitled *Distrust, Discontent, Anger and Partisan Rancor—The People and Their Government*, 2505 U.S. adults were asked, “Which is the bigger problem with the Federal government?” Of the respondents, 38% said that Federal government has the wrong priorities (W), 50% said that it runs programs inefficiently (I), and 12% had no opinion or did not know (N). Recently 44 randomly selected people were asked the same question, and their responses were as follows:

I	I	W	I	W	W	W	I	I	W	W
N	I	I	W	N	N	W	I	W	W	I
I	W	N	I	I	W	W	I	N	W	W
W	W	W	I	W	I	W	W	I	N	W

- Prepare a frequency distribution for these data.
- Calculate the relative frequencies and percentages for all classes.
- Draw a bar graph for the frequency distribution and a pie chart for the percentage distribution.
- What percentage of these respondents mentioned “Federal government has the wrong priorities” as the bigger problem?

**2.68** The following data give the numbers of television sets owned by 40 randomly selected households.

1	1	2	3	2	4	1	3	2	1
3	0	2	1	2	3	2	3	2	2
1	2	1	1	1	3	1	1	1	2
2	4	2	3	1	3	1	2	2	4

- a. Prepare a frequency distribution table for these data using single-valued classes.
- b. Compute the relative frequency and percentage distributions.
- c. Draw a bar graph for the frequency distribution.
- d. What percentage of the households own two or more television sets?

**2.69** Twenty-four students from universities in Connecticut were asked to name the five current members of the U.S. House of Representatives from Connecticut. The number of correct names supplied by the students are given below.

4	2	3	5	5	4	3	1	5	4	4	3
5	3	2	3	1	3	2	5	2	1	5	0

- a. Prepare a frequency distribution for these data using single-valued classes.
- b. Compute the relative frequency and percentage distributions.
- c. What percentage of the students in this sample named fewer than two of the representatives correctly?
- d. Draw a bar graph for the relative frequency distribution.

**2.70** The following data give the number of text messages sent on 40 randomly selected days during 2012 by a high school student:

32	33	33	34	35	36	37	37	37	37
38	39	40	41	41	42	42	42	43	44
44	45	45	45	47	47	47	47	47	48
48	49	50	50	51	52	53	54	59	61

- a. Construct a frequency distribution table. Take 32 as the lower limit of the first class and 6 as the class width.
- b. Calculate the relative frequency and percentage for each class.
- c. Construct a histogram for the frequency distribution of part a.
- d. On what percentage of these 40 days did this student send more than 44 text messages?

**2.71** The following data give the numbers of orders received for a sample of 30 hours at the Timesaver Mail Order Company.

34	44	31	52	41	47	38	35	32	39
28	24	46	41	49	53	57	33	27	37
30	27	45	38	34	46	36	30	47	50

- a. Construct a frequency distribution table. Take 23 as the lower limit of the first class and 7 as the width of each class.
- b. Calculate the relative frequencies and percentages for all classes.
- c. For what percentage of the hours in this sample was the number of orders more than 36?

**2.72** The following data give the amounts (in dollars) spent on refreshments by 30 spectators randomly selected from those who patronized the concession stands at a recent Major League Baseball game.

4.95	27.99	8.00	5.80	4.50	2.99	4.85	6.00
9.00	15.75	9.50	3.05	5.65	21.00	16.60	18.00
21.77	12.35	7.75	10.45	3.85	28.45	8.35	17.70
19.50	11.65	11.45	3.00	6.55	16.50		

- a. Construct a frequency distribution table using the *less-than* method to write classes. Take \$0 as the lower boundary of the first class and \$6 as the width of each class.
- b. Calculate the relative frequencies and percentages for all classes.
- c. Draw a histogram for the frequency distribution.

**2.73** The following data give the repair costs (in dollars) for 30 cars randomly selected from a list of cars that were involved in collisions.

2300	750	2500	410	555	1576
2460	1795	2108	897	989	1866
2105	335	1344	1159	1236	1395
6108	4995	5891	2309	3950	3950
6655	4900	1320	2901	1925	6896

- Construct a frequency distribution table. Take \$1 as the lower limit of the first class and \$1400 as the width of each class.
- Compute the relative frequencies and percentages for all classes.
- Draw a histogram and a polygon for the relative frequency distribution.
- What are the class boundaries and the width of the fourth class?

**2.74** Refer to Exercise 2.70. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using the frequency distribution table of that exercise.

**2.75** Refer to Exercise 2.71. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution table constructed for the data of that exercise.

**2.76** Refer to Exercise 2.72. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution table constructed for the data of that exercise.

**2.77** Construct the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using the frequency distribution table constructed for the data of Exercise 2.73.

**2.78** Refer to Exercise 2.70. Prepare a stem-and-leaf display for the data of that exercise.

**2.79** Construct a stem-and-leaf display for the data given in Exercise 2.71.

**2.80** The following table gives the seven most common first names among girls born in the United States during 2010 along with their frequencies (in thousands).

Name	Number of girls (in thousands)
Isabella	22.7
Sophia	20.5
Emma	17.2
Olivia	16.9
Ava	15.3
Emily	14.2
Abigail	14.1

Source: U.S. Social Security Administration, [www.ssa.gov](http://www.ssa.gov).

Draw two bar graphs for these data, the first without truncating the frequency axis, and the second by truncating this axis. In the second graph, mark the number of girls on the vertical axis starting with 13.0. Briefly comment on the two bar graphs.

**2.81** The following table lists the average price (as defined by the U.S. Energy Information Administration) per gallon of unleaded regular gasoline in each of the seven regions of the United States. These averages were calculated from the weekly averages for the period June 14, 2010 through June 6, 2011.

Region	Average Price per Gallon (in dollars)
New England	3.152
Central Atlantic	3.161
Lower Atlantic	3.038
Midwest	3.085
Gulf Coast	2.973
Rocky Mountain	3.032
West Coast	3.282

Source: U.S. Energy Information Administration, [www.eia.gov](http://www.eia.gov).

Draw two bar graphs for these data, the first without truncating the axis that represents price, and the second by truncating this axis. In the second graph, mark the prices on the vertical axis starting with \$2.95. Briefly comment on the similarities and differences of the two bar graphs.

**2.82** The following data give the waiting times (in minutes) for 25 students at the Student Health Center of a university.

39	19	19	35	18	32	20	15	20	29	25	32	28
19	42	18	32	31	21	46	27	13	14	15	28	

Create a dotplot for these data.

**2.83** Reconsider the data on the numbers of orders received for a sample of 30 hours at the Timesaver Mail Order Company given in Exercise 2.71. Create a dotplot for those data.

**2.84** Refer to Exercise 2.70, which contains data on the number of text messages sent by a high school student on each of 40 randomly selected days. Create a dotplot for those data.

**2.85** The following data give the number of visitors during visiting hours on a given evening for each of the 20 randomly selected patients at a hospital.

3	0	1	4	2	0	4	1	1	3
4	2	0	2	2	2	1	1	3	0

Create a dotplot for these data.

## Advanced Exercises

**2.86** The following frequency distribution table gives the age distribution of drivers who were at fault in auto accidents that occurred during a 1-week period in a city.

Age (years)	<i>f</i>
18 to less than 20	7
20 to less than 25	12
25 to less than 30	18
30 to less than 40	14
40 to less than 50	15
50 to less than 60	16
60 and over	35

- a. Draw a relative frequency histogram for this table.
- b. In what way(s) is this histogram misleading?
- c. How can you change the frequency distribution so that the resulting histogram gives a clearer picture?

**2.87** Refer to the data presented in Exercise 2.86. Note that there were 50% more accidents in the *25 to less than 30* age group than in the *20 to less than 25* age group. Does this suggest that the older group of drivers in this city is more accident-prone than the younger group? What other explanation might account for the difference in accident rates?

**2.88** Suppose a data set contains the ages of 135 autoworkers ranging from 20 to 53 years.

- a. Using Sturge's formula given in footnote 1 in section 2.2.2, find an appropriate number of classes for a frequency distribution for this data set.
- b. Find an appropriate class width based on the number of classes in part a.

**2.89** Stem-and-leaf displays can be used to compare distributions for two groups using a back-to-back stem-and-leaf display. In such a display, one group is shown on the left side of the stems, and the other group is shown on the right side. When the leaves are ordered, the leaves increase

as one moves away from the stems. The following stem-and-leaf display shows the money earned per tournament entered for the top 30 money winners in the 2008–09 Professional Bowlers Association men’s tour and for the top 21 money winners in the 2008–09 Professional Bowlers Association women’s tour.

Women's	Men's
8	0
8871	1
65544330	2
840	3
52	4
21	5
	9
	6
5	7
	8
	9
	5

The leaf unit for this display is 100. In other words, the data used represent the earnings in hundreds of dollars. For example, for the women’s tour, the first number is 08, which is actually 800. The second number is 11, which actually is 1100.

- a. Do the top money winners, as a group, on one tour (men’s or women’s) tend to make more money per tournament played than on the other tour? Explain how you can come to this conclusion using the stem-and-leaf display.
- b. What would be a typical earnings level amount per tournament played for each of the two tours?
- c. Do the data appear to have similar spreads for the two tours? Explain how you can come to this conclusion using the stem-and-leaf display.
- d. Does either of the tours appears to have any outliers? If so, what are the earnings levels for these players?

**2.90** Statisticians often need to know the shape of a population to make inferences. Suppose that you are asked to specify the shape of the population of weights of all college students.

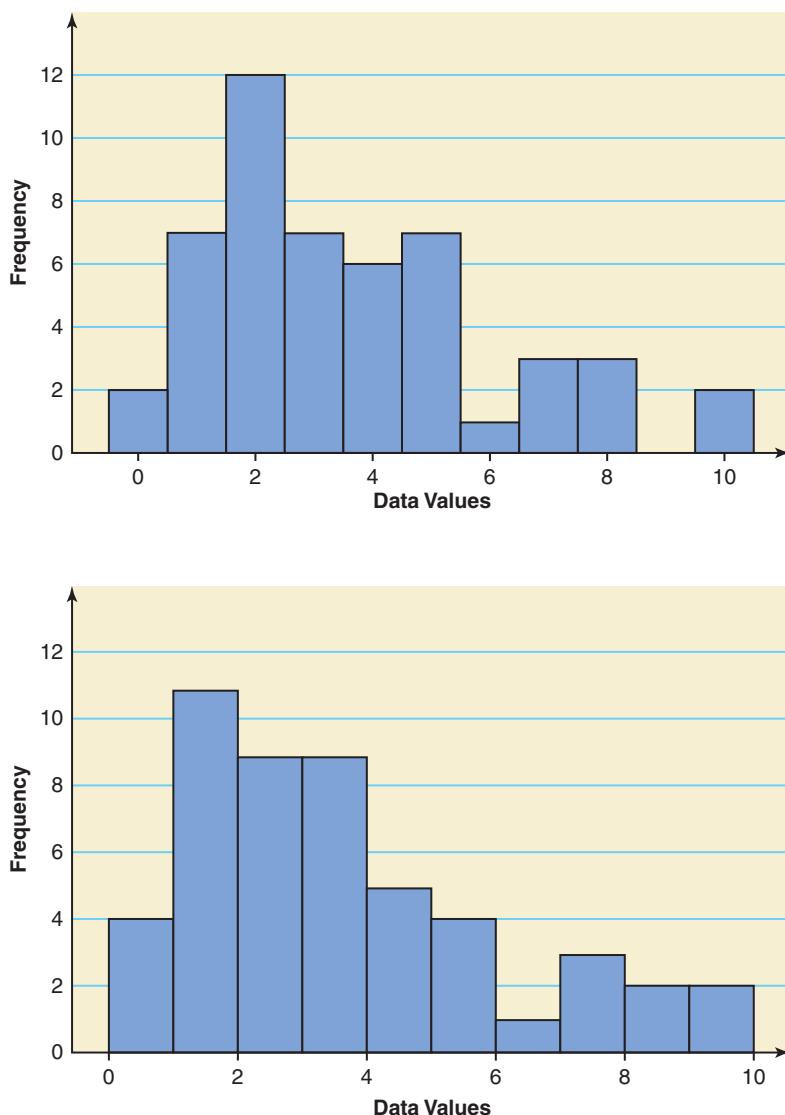
- a. Sketch a graph of what you think the weights of all college students would look like.
- b. The following data give the weights (in pounds) of a random sample of 44 college students (F and M indicate female and male, respectively).

123 F	195 M	138 M	115 F	179 M	119 F	148 F	147 F
180 M	146 F	179 M	189 M	175 M	108 F	193 M	114 F
179 M	147 M	108 F	128 F	164 F	174 M	128 F	159 M
193 M	204 M	125 F	133 F	115 F	168 M	123 F	183 M
116 F	182 M	174 M	102 F	123 F	99 F	161 M	162 M
155 F	202 M	110 F	132 M				

- i. Construct a stem-and-leaf display for these data.
- ii. Can you explain why these data appear the way they do?
- c. Construct a back-to-back stem-and-leaf display for the data on weights, placing the weights of the female students to the left of the stems and those of the male students to the right of the stems. (See Exercise 2.89 for an example of a back-to-back stem-and-leaf plot.) Does one gender tend to have higher weights than the other? Explain how you know this from the display.

**2.91** Consider the two histograms given in Figure 2.25, which are drawn for the same data set. In this data set, none of the values are integers.

- a. What are the endpoints and widths of classes in each of the two histograms?
- b. In the first histogram, of the observations that fall in the interval that is centered at 8, how many are actually between the left endpoint of that interval and 8? Note that you have to consider both histograms to answer this question.
- c. Observe the leftmost bars in both histograms. Why is the leftmost bar in the first histogram misleading?



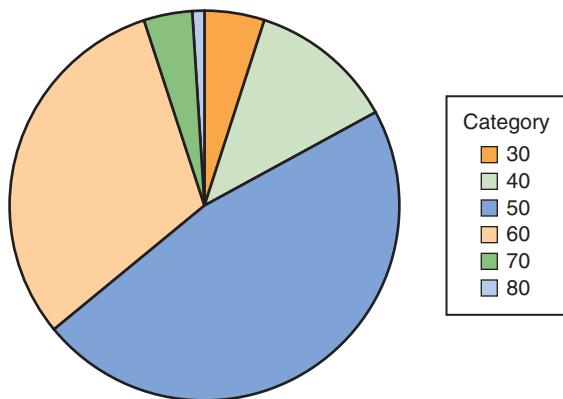
**Figure 2.25** Two histograms for the same data.

**2.92** Refer to the data on weights of 44 college students given in Exercise 2.90. Create a dotplot of all 44 weights. Then create stacked dotplots for the weights of male and female students. Describe the similarities and differences in the distributions of weights of male and female students. Using all three dotplots, explain why you cannot distinguish the lightest males from the heaviest females when you consider only the dotplot of all 44 weights.

**2.93** The pie chart in Figure 2.26 shows the percentage distribution of ages (i.e., the percentages of all prostate cancer patients falling in various age groups) for men who were recently diagnosed with prostate cancer.

- a. Are more or fewer than 50% of these patients in their 50s? How can you tell?
- b. Are more or fewer than 75% of these patients in their 50s and 60s? How can you tell?
- c. A reporter looks at this pie chart and says, “Look at all these 50-year-old men who are getting prostate cancer. This is a major concern for a man once he turns 50.” Explain why the reporter cannot necessarily conclude from this pie chart that there are a lot of 50-year-old men with prostate cancer. Can you think of any other way to present these cancer cases (both graph and variable) to determine if the reporter’s claim is valid?

**2.94** As shown in Exercise 2.89, back-to-back stem-and-leaf displays can be used to compare the distribution of a variable for two different groups. Consider the following data, which give the alcohol



**Figure 2.26** Pie chart of age groups.

content by volume (%) for different beers produced by the Flying Dog Brewery and the Sierra Nevada Brewery.

**Flying Dog Brewery:**

4.7	4.7	4.8	5.1	5.5	5.5	5.6	6.0	7.1
7.4	7.8	8.3	8.3	9.2	9.9	10.2	11.5	

**Sierra Nevada Brewery:**

4.4	5.0	5.0	5.6	5.6	5.8	5.9	5.9	6.7	6.8	6.9	7.0	9.6
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

- Create a back-to-back stem-and-leaf display of these data. Place the Flying Dog Brewery data to the left of the stems.
- What would you consider to be a typical alcohol content of the beers made by each of the two breweries?
- Does one brewery tend to have higher alcohol content in its beers than the other brewery? If so, which one? Explain how you reach this conclusion by using the stem-and-leaf display.
- Do the alcohol content distributions for the two breweries appear to have the same levels of variability? Explain how you reach this conclusion by using the stem-and-leaf display.

**2.95** The following table lists the earnings per event that were referred to in Exercise 2.89. Although the table lists earnings per event, players are listed in order of their total earnings, not their earnings per event. Note that men and women are ranked together in the table.

Name	Earnings per Event (in dollars)	Name	Earnings per Event (in dollars)
Norm Duke	9568.67	Mika Koivuniemi	3396.47
Wes Malott	8795.63	Jeff Carter	3410.94
Patrick Allen	6979.41	Michael Machuga	3455.33
Chris Barnes	5970.00	Ryan Shafer	2983.53
Walter Ray Williams Jr.	4758.82	Mike Wolfe	2902.35
Bill O'Neill	4884.38	Steve Jaros	2884.12
Rhino Page	4872.50	Chris Loschetter	3035.63
John Nolen	4801.56	Mike DeVaney	2681.76
Mike Scroggins	4307.06	Ken Simard	2412.19
Brad Angelo	4291.18	Eugene McCune	2475.88
Pete Weber	4135.29	Ronnie Russell	2540.63
Parker Bohn III	4101.47	Ritchie Allen	2340.00
Michael Fagan	3851.76	Jack Jurek	2322.94
Steve Harman	4035.63	Liz Johnson	7500.00
Tommy Jones	3715.88	Michelle Feldman	5214.29
Danny Wiseman	3648.82	Carolyn Dorin-Ballard	5185.71
Sean Rash	3399.41	Stefanie Nation	4542.86

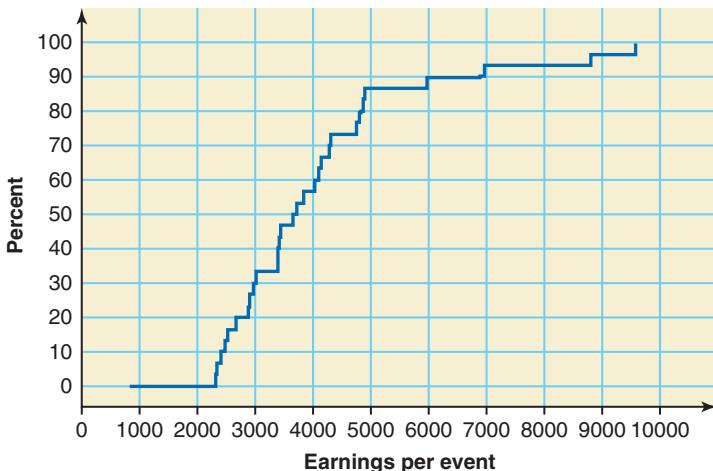
*(Continued)*

Jennifer Petrick	4285.71	Tennelle Milligan	2331.67
Jodi Woessner	3885.71	Shannon O'Keefe	2640.00
Shannon Pluhowsky	3433.33	Joy Esterson	1807.14
Missy Bellinder	2386.25	Adrienne Miller	1798.57
Diandra Asbaty	2542.86	Brenda Mack	1833.33
Trisha Reid	2400.00	Olivia Sandham	1100.00
Wendy Macpherson	3056.00	Amy Stoltz	2500.00
Clara Guerrero	2466.67	Kelly Kulick	830.00
Shalin Zulkifli	2098.57		

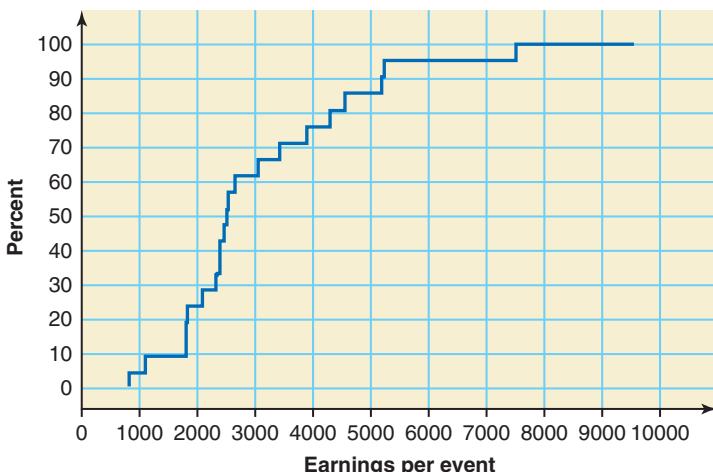
Source: Professional Bowlers Association, April 13, 2009.

A graph that is similar to an ogive is a graph of the *empirical cumulative distribution function* (CDF). The primary difference between an ogive and an empirical CDF is that the empirical CDF looks like a set of steps, as opposed to a set of slanted lines. The height of each step corresponds to the percentage of observations that occur at a specific value. Longer (not higher) steps occur when there are bigger gaps between observations.

- a. Figures 2.27(a) and (b) contain the empirical CDFs of the earnings per event for the two tours (men's and women's), in some order. In other words, one of these two figures is for the men's tour, and the other is for the women's tour, but not in that order necessarily. Match the CDFs to the respective tours. Give three reasons for your choices.
- b. Both distributions are skewed to the right. Use the information about longer steps to explain



**Figure 2.27 (a)** Empirical CDF of Earnings Per Event.



**Figure 2.27 (b)** Empirical CDF of Earnings Per Event.

why the distributions are skewed to the right.

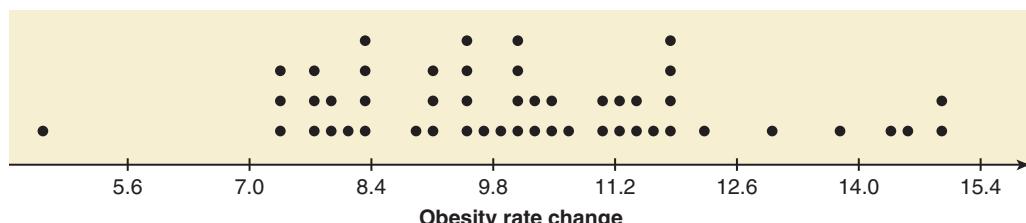
- c. What are the approximate values of the CDFs corresponding to \$3000 per tournament played and \$4000 per tournament played? Based on this information, what is the approximate percentage of bowlers who earned between \$3000 and \$4000 per tournament played?

**2.96** Table 2.18 shows the differences in obesity rates (called Rate Change in the table) for the years 2010 and 1997 for each of the 50 states and the District of Columbia. The obesity rate is the percentage of people having a body mass index (BMI) of 30 or higher. Figure 2.28 is a dotplot of these data.

**Table 2.18 Difference in 2010 and 1997 Obesity Rates By State**

State	Rate Change	State	Rate Change	State	Rate Change
Alabama	14	Kentucky	9.5	North Dakota	10.2
Alaska	4.8	Louisiana	11.4	Ohio	11.5
Arizona	11.9	Maine	10.6	Oklahoma	15.3
Arkansas	12	Maryland	9.6	Oregon	7.4
California	8	Massachusetts	8.2	Pennsylvania	11.1
Colorado	9.2	Michigan	11.6	Rhode Island	11.7
Connecticut	7.8	Minnesota	8.3	South Carolina	14.6
Delaware	9.2	Mississippi	12	South Dakota	10.3
D.C.	7.7	Missouri	11.4	Tennessee	13.1
Florida	10.5	Montana	8.4	Texas	12.3
Georgia	15.2	Nebraska	9.9	Utah	7.3
Hawaii	9.1	Nevada	8.3	Vermont	7.3
Idaho	10.2	New Hampshire	10.8	Virginia	9.6
Illinois	11.1	New Jersey	7.8	Washington	10.3
Indiana	8.4	New Mexico	10.2	West Virginia	11.9
Iowa	9	New York	7.9	Wisconsin	9.7
Kansas	14.7	North Carolina	9.5	Wyoming	10.1

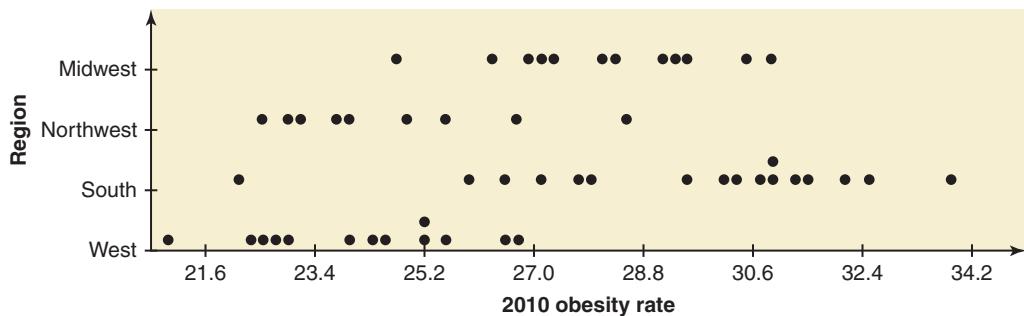
Source: www.cdc.gov.



**Figure 2.28** Dotplot of obesity rate change (year 2010 minus year 1997).

- By looking at the dotplot, what value would you provide if asked to report a *typical* obesity rate change? Why did you choose this value?
- What number do you feel most accurately represents the number of outliers in this data set: 0, 1, 3, 5, 6, or 7? Explain your reasoning, including the identification of the observations, if any, that you feel are outliers.
- Would you classify this distribution as being skewed to the left, skewed to the right, or approximately symmetric? Explain.
- The largest increase in the obesity rate during this period took place in Oklahoma (15.3), while the smallest increase took place in Alaska (4.8). Explain why this information should not lead you to conclude that Oklahoma had the highest obesity rate in 2010 and that Alaska had the lowest obesity rate in 2010. (Note: The highest and lowest obesity rates in 2010 were in Mississippi and Colorado, respectively.)

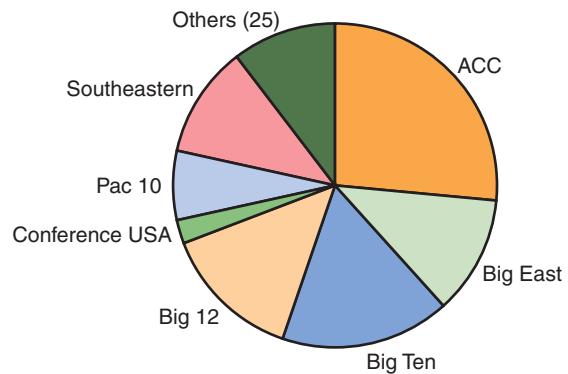
**2.97** Figure 2.29 contains stacked dotplots of 2010 state obesity rates for the regions Midwest, Northeast, South, and West.



**Figure 2.29** 2010 state obesity rates, by geographic region.

- Which region has the least variability (greatest consistency) of obesity rates? Which region has the most variability (least consistency) of obesity rates? Justify your choices.
- Which region tends to have the highest obesity rates? Which region tends to have the lowest obesity rates? Justify your choices.
- Are there any regions that have at least one obesity rate that could be considered an outlier? If so, specify the region(s) and the observation(s).

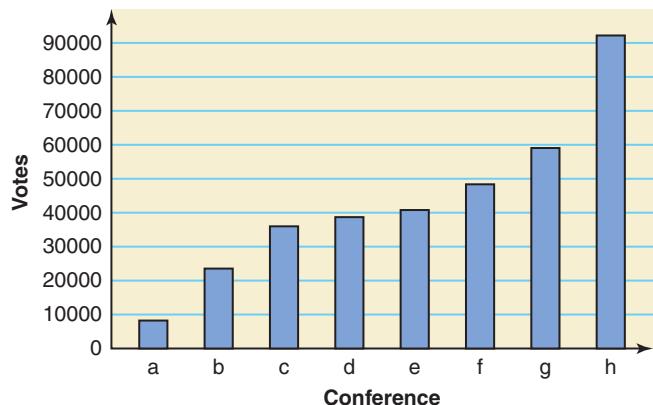
**2.98** CBS Sports had a Facebook page for the 2009 NCAA Men's Basketball Tournament including bracket contests, discussion sites, and a variety of polls. One of the polls asked users to identify their most despised teams. The pie chart in Figure 2.30 gives a breakdown of the votes by the conference of the most despised teams as of 10:53 EDT on March 16, 2009. (Note that Pac 10 is now Pac 12.)



**Figure 2.30** Pie chart of conference of the most despised NCAA men's basketball team.

- Are there any conferences that received more than 25% of the votes? If so, which conference(s)? How can you tell?
- Which two conferences appear to have the closest percentages of the votes?
- A bar chart for the same data is presented in Figure 2.31. Comparing the bar chart to the pie chart, match the conferences to the bars. In other words, explain which bar represents which conference.

**Figure 2.31** Bar chart of conference of the most despised NCAA men's basketball team.



## Self-Review Test

1. Briefly explain the difference between ungrouped and grouped data and give one example of each type.

2. The following table gives the frequency distribution of times (to the nearest hour) that 90 fans spent waiting in line to buy tickets to a rock concert.

Waiting Time (hours)	Frequency
0 to 6	5
7 to 13	27
14 to 20	30
21 to 27	20
28 to 34	8

Circle the correct answer in each of the following statements, which are based on this table.

- a. The number of classes in the table is 5, 30, 90.  
 b. The class width is 6, 7, 34.  
 c. The midpoint of the third class is 16.5, 17, 17.5.  
 d. The lower boundary of the second class is 6.5, 7, 7.5.  
 e. The upper limit of the second class is 12.5, 13, 13.5.  
 f. The sample size is 5, 90, 11.  
 g. The relative frequency of the second class is .22, .41, .30.
3. Briefly explain and illustrate with the help of graphs a symmetric histogram, a histogram skewed to the right, and a histogram skewed to the left.
4. Twenty elementary school children were asked if they live with both parents (B), father only (F), mother only (M), or someone else (S). The responses of the children follow.

M	B	B	M	F	S	B	M	F	M
B	F	B	M	M	B	B	F	B	M

- a. Construct a frequency distribution table.  
 b. Write the relative frequencies and percentages for all categories.  
 c. What percentage of the children in this sample live with their mothers only?  
 d. Draw a bar graph for the frequency distribution and a pie chart for the percentages.
5. A large Midwestern city has been chronically plagued by false fire alarms. The following data set gives the number of false alarms set off each week for a 24-week period in this city.

10	4	8	7	3	7	10	2	6	12	11	8
1	6	5	13	9	7	5	1	14	5	15	3

- a. Construct a frequency distribution table. Take 1 as the lower limit of the first class and 3 as the width of each class.  
 b. Calculate the relative frequencies and percentages for all classes.  
 c. What percentage of these weeks had 9 or fewer false alarms?  
 d. Draw the frequency histogram and polygon.
6. Refer to the frequency distribution prepared in Problem 5. Prepare the cumulative percentage distribution using that table. Draw an ogive for the cumulative percentage distribution.
7. Construct a stem-and-leaf display for the following data, which give the times (in minutes) that 24 customers spent waiting to speak to a customer service representative when they called about problems with their Internet service provider.

12	15	7	29	32	16	10	14	17	8	19	21
4	14	22	25	18	6	22	16	13	16	12	20

8. Consider this stem-and-leaf display:

3	0 3 7
4	2 4 6 7 9
5	1 3 3 6
6	0 7 7
7	1 9

Write the data set that was used to construct this display.

9. Make a dotplot for the data given in Problem 5.

## Mini-Projects

### ■ MINI-PROJECT 2-1

Using the data you gathered for the mini-project in Chapter 1, prepare a summary of that data set that includes the following.

- a. Prepare an appropriate type of frequency distribution table for one of the quantitative variables and then compute relative frequencies and cumulative relative frequencies.
- b. Create a histogram, a stem-and-leaf display (by rounding the numbers to integers), and a dotplot of the data. Comment on any symmetry or skewness and on the presence of clusters and any potential outliers.
- c. Make stacked dotplots of the same variable (as in parts a and b) based on the values of one of your categorical variables. For example, if your quantitative variable is GPAs of students, your categorical variable could be gender then you can make stacked dotplot of GPAs of males and females, respectively. Comment on the similarities and differences between the distributions for the different values of your categorical variable.

### ■ MINI-PROJECT 2-2

Choose 15 of each of two types of magazines (news, sports, fitness, entertainment, and so on) and record the percentage of pages that contain at least one advertisement. Using these percentages and the types of magazines, write a brief report that covers the following:

- a. Prepare an appropriate type of frequency distribution table for the quantitative variable and then compute relative frequencies and cumulative relative frequencies.
- b. Create a histogram, a stem-and-leaf plot (by rounding the percentages to integers), and a dotplot of all of the data. Comment on any symmetry or skewness, as well as the presence of clusters and any potential outliers.
- c. Make stacked dotplots of the same variable for each of the two types of magazines. Comment on the similarities and differences between the distributions for the two types of magazines.

### ■ MINI-PROJECT 2-3

Go to the website [www.ncaa.com](http://www.ncaa.com). From the menus Men's Sports and Women's Sports, choose one sport for which statistics are available. (Note: As you scroll over each sport's name, you will see a list of the items that are available for each sport. Make sure that you choose a sport that contains *Statistics* in this list.) After you choose a sport, click on *Statistics* in the menu below that sport. Next, look for *Custom Reporting* toward the bottom of the page. In custom reporting, choose a *Division* (I, II, or III), then select the most recent reporting week available, select *Individual* for Category, and request *All Statistics*. If you are using statistical software, choose CSV for the report format, as this is a spreadsheet file that can be opened by most statistical software packages. If not, choose any of the formats.

- a. Select a random sample of 30 athletes and record the values of three variables for each of the athletes chosen. If you are using statistical software, the software will be able to generate the sample for you. If you are using a graphing calculator, you can use the Random Integer function to choose your sample. If you are not using technology, you can use the random number generator at <http://www.randomizer.org/form.htm> to generate your sample.
- b. Prepare an appropriate type of frequency distribution table for each of the three variables. Compute relative frequencies and cumulative relative frequencies.
- c. Create a histogram, a stem-and-leaf display, and a dotplot of the data for each variable. Comment on any symmetry or skewness, as well as the presence of clusters and any potential outliers.

## DECIDE FOR YOURSELF

## DECIDING ABOUT STATISTICAL PROPERTIES

Look around you. Graphs are everywhere. Business reports, newspapers, magazines, and so forth are all loaded with graphs. Unfortunately, some people feel that the primary purpose of graphs is to provide a break from the humdrum text. Executive summaries will often contain graphs so that CEOs and executive vice presidents need only to glance at these graphs to assume that they understand everything without reading more than a paragraph or so of the report. In reality, the usefulness of graphs is somewhere between the fluff of the popular press and the quick answer of the boardroom.

Here you are asked to interpret some graphs, primarily by using them to compare distributions of a variable. As we will discuss in Chapter 3, some of our concerns have to do with the location of the center of a distribution and the variability or spread of a distribution. We can use graphs to compare the centers and variability of two or more distributions.

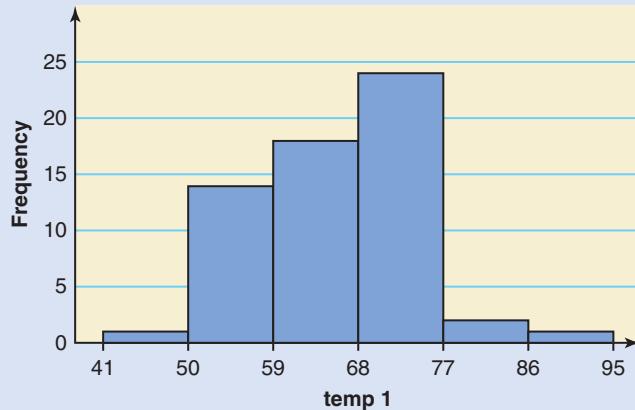
In practice, the graphs are made using statistical software, so it is important to recognize that computer software is programmed to use the same format for each graph of a specific type, unless you tell the software to do differently. For example, consider the two histograms in Figures 2.32 and 2.33 that are drawn for two different data sets.

1. Examine the two graphs of Figures 2.32 and 2.33.
2. Explain what is meant by the statement “the shapes of the two distributions are the same.”
3. Does the fact that the shapes of the two distributions are the same imply that the centers of the two distributions are the same? Why or why not? Explain.
4. Does the fact that the shapes of the two distributions are the same imply that the spreads of the two distributions are the same? Why or why not? Explain.
5. It turns out that the same variable was represented in the two graphs but with different units of measurement. Can you figure out the units?

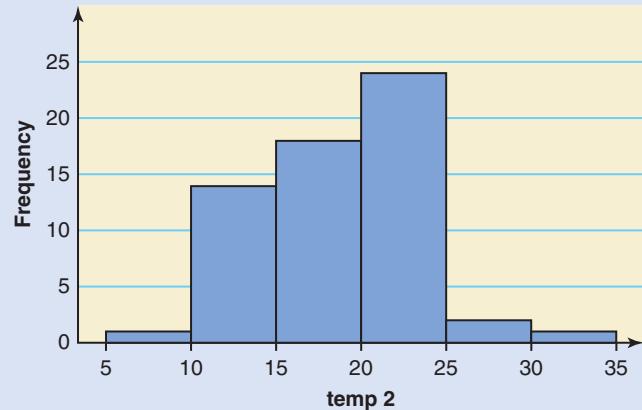
Another situation that is important to compare is when two graphs cover a similar range but have different shapes, such as the histograms in Figures 2.34 and 2.35.

1. Examine the two histograms of Figures 2.34 and 2.35.
2. These two distributions have the same center but do not have the same spread. Decide which distribution has the larger spread and explain the reasoning behind your decision.

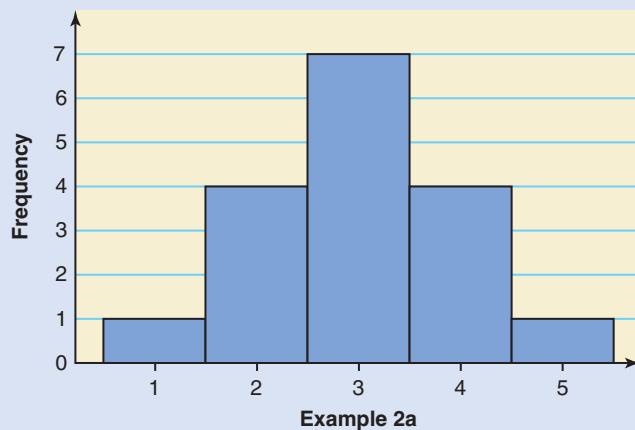
Answer all the above questions again after reading Chapter 3.



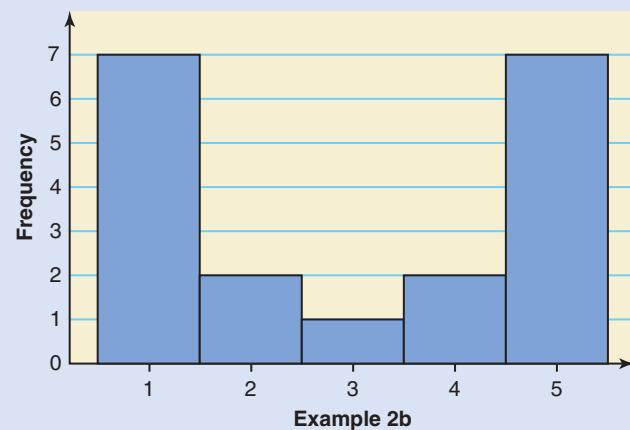
**Figure 2.32** Histogram of data temp 1.



**Figure 2.33** Histogram of data temp 2.



**Figure 2.34** Histogram of example 2a.



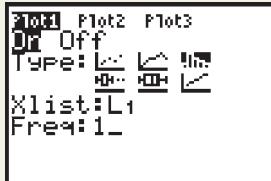
**Figure 2.35** Histogram of example 2b.

# TECHNOLOGY INSTRUCTION

## Organizing Data

### TI-84

1. To create a frequency histogram for a list of data, press **STAT PLOT**, which you access by pressing **2nd > Y =**. The **Y =** key is located at the top left of the calculator buttons.



Screen 2.1

2. Make sure that only one plot is turned on. If more than one plot is turned on, you can turn off the unwanted plots by using the following steps. Press the number corresponding to the plot you wish to turn off. A screen similar to **Screen 2.1** will appear. Use the arrow keys to move the cursor to the **Off** button, then press **ENTER**. Now use the arrow keys to move to the row with **Plot1**, **Plot2**, and **Plot3**. If there is another plot that you need to turn off, select that plot by moving the cursor to that plot, pressing **ENTER**, and repeating the previous procedure. If not, move the cursor to the plot you wish to use and press **ENTER**.



Screen 2.2

3. At the **Type** prompt use the right arrow to move to the third column in the first row that looks like a histogram, and press **ENTER**. Move to the **Xlist** prompt to enter the name of the list where the data are located. Press **2nd > Stat**, then use the up and down arrows to move through the list names until you find the list you want to use. Press **ENTER**. Enter 1 at the **Freq** prompt. (Note: if you are using one of the lists named **L1**, **L2**, **L3**, **L4**, **L5**, or **L6**, you can enter the list name by pressing **2nd** followed by one of the numbers **1** through **6**, as they correspond to the list names L1 through L6.)

4. To see the graph, select **ZOOM > 9** (the **ZOOMSTAT** function), where **ZOOM** is the third key in the top row. This sets the window settings to display your graph.
5. If you would like to change the class width and/or the starting point of the first interval, select **WINDOW** (see **Screen 2.2**). To change the class width, change the value of **Xscl** to the desired width. To change the starting point of the first interval, change the value of **Xmin** to the desired point. Press **GRAPH**, which is the fifth button in the top row. (Note: After making either or both of these changes, you may need to change the values of **Xmax** and **Ymax** to see the entire graph. The difference between **Xmax** and **Xmin** should be a multiple of **Xscl**. As an example, if **Xmin** = 5 and **Xscl** = 10 and the largest data point is 93, then **Xmax** should be set to 95 because  $95 - 5 = 90$ , which is a multiple of 10, and 95 is larger than the largest data point. The purpose of changing **Ymax** is to be able to see the tops of the bars of the histogram. If the bars run off the top of the calculator screen, increase **Ymax**, and press **GRAPH**.)
6. If you would like to see the interval endpoints and the number of observations in each class (which is given by the height of the corresponding bar), press **TRACE**, then use the left and right arrows to move from one bar to the next. When you are done, press **CLEAR**.

### Minitab

The functions for creating many common graphs are listed in the pulldown menu **Graph**. The following instructions will demonstrate how to use Minitab to create two types of graphs for categorical variables—a bar chart and a pie chart—and three types of graphs for quantitative variables—a frequency histogram, a stem-and-leaf display, and a dotplot.

#### Bar Chart

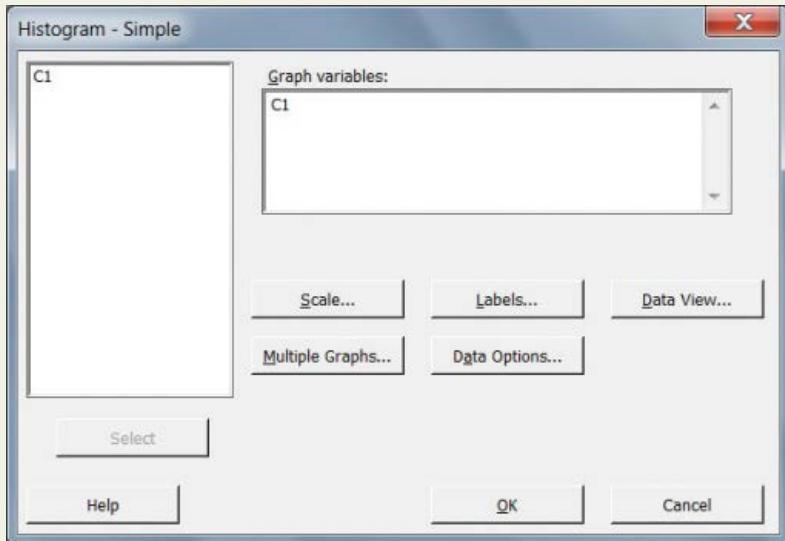
- If you have raw (or ungrouped) categorical data entered in a column (such as **C1**), select **Graph > Bar Chart**. In the resulting dialog box, select **Bars Represent: Counts of unique values** and **Simple**. Click **OK**. In the new dialog box, type **C1** in the box below **Categorical Variables** and click **OK**.
- If you have categorical data in a frequency table, with the categories entered in **C1** and the frequencies in **C2**, select **Graph > Bar Chart**. In the resulting dialog box, select **Bars Represent: Values from a table** and **Simple**. Click **OK**. In the new dialog box, type **C2** in the box below **Graph variables** and **C1** in the box below **Categorical Variable** and click **OK**.

## Pie Chart

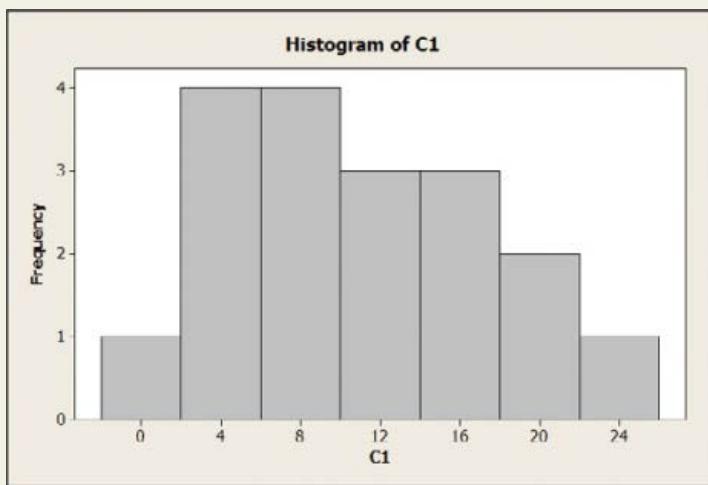
1. If you have raw categorical data entered in C1, select **Graph > Pie Chart**. In the resulting dialog box, select **Chart raw data**, type C1 in the box below **Categorical Variables**, and click **OK**.
- 2 If you have categorical data in a frequency table, with the categories entered in C1 and the frequencies in C2, select **Graph > Pie Chart**. In the resulting dialog box, select **Chart values from a table**, type C2 in the box below **Summary variables** and C1 in the box below **Categorical Variable**, and click **OK**.

## Frequency Histogram

For a quantitative data set entered in C1, select **Graph > Histogram**, select **Simple**, and click **OK**. In the resulting dialog box, type C1 in the box below **Graph Variables** (see **Screen 2.3**) and click **OK**. Minitab will produce a separate window that contains the histogram (see **Screen 2.4**).



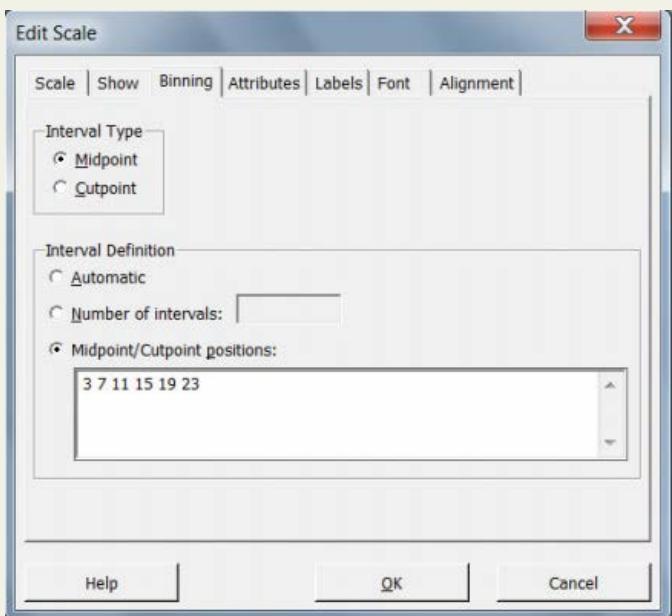
Screen 2.3



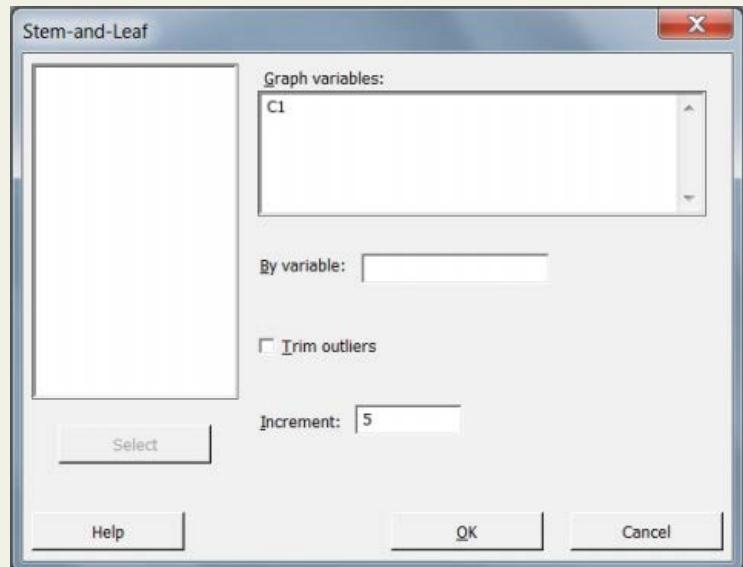
Screen 2.4

Once you have created a histogram, you can change the interval widths and starting points to whatever you desire. To do this, double-click on any of the bars in the histogram and this will produce the **Edit Bars** window. Click on the **Binning** tab. If you know the interval endpoints (boundaries) you wish to use, select **Cutpoint** under **Interval Type** in this box, then select **Midpoint/Cutpoint positions** under **Interval Definition**. Enter the endpoints for all

intervals that you wish to use, including the right endpoint of the last interval. **Screen 2.5** shows the entries necessary to have intervals of width four, with the first interval beginning at 1. Here the intervals will be 1 to less than 5, 5 to less than 9, and so on. The last interval is 21 to less than 25. If you know the midpoints of intervals that you want to use, then click next to **Midpoint** under **Interval Type** in the **Edit Bars** box. Then click next to **Midpoint/Cutpoint positions** under **Interval Definition**. Then enter the midpoints of all intervals in the box. For example, in this case, the midpoints that you will enter will be 3 7 11 15 19 23. Finally, click **OK** to obtain the desired histogram.



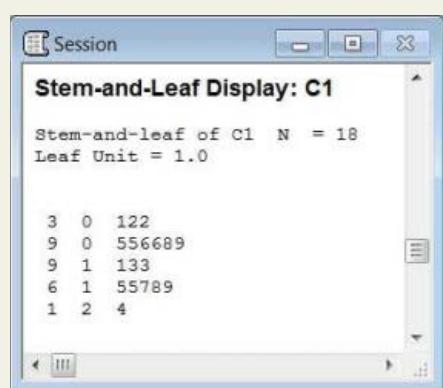
Screen 2.5



Screen 2.6

### Stem-and-Leaf Display

For a quantitative data set entered in column **C1**, select **Graph > Stem-and-Leaf**, then type **C1** in the dialog box shown under **Graph Variables**, and click **OK** (see **Screen 2.6**). The display will appear in the Session window. If there are too many stems, you can specify an **Interval** for each branch of the stem-and leaf display in the box next to **Increment** of the above dialog box. For example, the stem-and-leaf display shown in **Screen 2.7** has an interval of size 5.



Screen 2.7

## Dotplot

For a quantitative data set entered in C1, select **Graph > Dotplot**, select the appropriate dotplot from the choices, and click **OK**. In the resulting dialog box, type **C1** in the box below **Graph Variables** and click **OK**. The dotplot will appear in a new window.

### Excel

- To create a frequency distribution for a range of numerical data in Excel, decide how many categories you will have. Choose class boundaries between the categories so that you have one more boundary than classes. Type the class boundaries into Excel.
- Select where you want the class frequencies to appear, and select a range of cells equal to the number of boundaries you have.
- Type **=frequency**.
- Select the range of cells of numerical data, and then type a comma.
- Select the range of class boundaries, and type a right parenthesis (see **Screen 2.8**).
- Highlight the cells to the right of the class boundaries, including the cell that contains the **FREQUENCY** function. Press **F2**, which will make the **FREQUENCY** function appear. Press **Ctrl + Shift + Enter**. Excel will fill in the rest of the group frequencies (see **Screen 2.9**).

The screenshot shows a Microsoft Excel spreadsheet. The formula bar at the top displays the formula `=FREQUENCY(A3:A20,B3:B8)`. The worksheet has three columns: 'Data' (A), 'Boundaries' (B), and 'Frequencies' (C). The 'Frequencies' column contains the following data: 0, 5, 10, 15, 20, 25, 6.6, 8.4, 9.3, 11, 13, 13.4, 15.1, 15.7, 17, 18.2, 19.3, and 24. The cell containing the formula `=FREQUENCY(A3:A20,B3:B8)` is highlighted with a green border.

Screen 2.8

The screenshot shows the same Microsoft Excel spreadsheet as Screen 2.8, but with a different selection. The entire row of frequencies (from cell C3 to C20) is selected and highlighted with a blue background. The formula bar now displays the array formula `{=FREQUENCY(A3:A20,B3:B8)}`. The data in columns A and B remains the same as in Screen 2.8.

Screen 2.9

## TECHNOLOGY ASSIGNMENTS

- TA2.1** Construct a bar graph and a pie chart for the frequency distribution prepared in Exercise 2.5.
- TA2.2** Construct a bar graph and a pie chart for the frequency distribution prepared in Exercise 2.6.
- TA2.3** Refer to the Data Set IV, which contains results on the 5875 runners who finished the 2010 Beach to Beacon 10k Road Race in Cape Elizabeth, Maine. Take a random sample of 200 runners and complete the following tasks/questions.
- Create a bar graph of the variable Maine, which identifies whether a runner is from Maine or from somewhere else (as stated in Maine and Away). Are there more runners in your sample who are from Maine or who are from Away?
  - Create two histograms of the runners' times (given in seconds)—one for the Maine group and the second for the Away group. Make sure that the histograms are on the same scale. Does one group of runners tend to be faster than the other? Explain.
  - Create stacked dotplots of the runners' ages (given in years) for male and female runners. Write a note on the comparison of the age distributions for the male and female runners in your sample.
- TA2.4** Refer to Data Set I that accompanies this text on the prices of various products in different cities across the country. Select a subsample of 60 from the column that contains information on pizza prices and then construct a histogram for these data.
- TA2.5** Construct a histogram for the data given in Exercise 2.21 on charitable contributions of the top 40 individuals according to the 2010 *Slate 60*. Let your technology choose the interval widths. Construct two more histograms. In the first new histogram, cut the original interval width in half. In the second new histogram, double the original interval width. Discuss the similarities and differences of the three histograms. State which version you feel provides the best picture of the data and why you believe this to be the case.
- TA2.6** Prepare a stem-and-leaf display for the data given in Exercise 2.48.
- TA2.7** Prepare a stem-and-leaf display for the data of Exercise 2.53.
- TA2.8** Prepare a bar graph for the frequency distribution obtained in Exercise 2.28.
- TA2.9** Prepare a bar graph for the frequency distribution obtained in Exercise 2.29.
- TA2.10** Make a pie chart for the frequency distribution obtained in Exercise 2.19.
- TA2.11** Make a pie chart for the frequency distribution obtained in Exercise 2.29.
- TA2.12** Make a dotplot for the data of Exercise 2.64.
- TA2.13** Make a dotplot for the data of Exercise 2.65.
- TA2.14** Using the data in the file *Kickers2010*, create a stacked dotplot of the percentage of field goals made in the National Football League (NFL) and the Canadian Football League (CFL) during the 2010 season. The stacked dotplot should have three groups corresponding to the kickers in (1) the American Football Conference (AFC) and (2) the National Football Conference (NFC) of the NFL and (3) the CFL. Discuss the similarities and differences among the three groups.
- TA2.15** Using the data set *Billboard*, create a histogram of the number of weeks spent on the charts for the songs in the Billboard Hot 100 for the week of July 9, 2011. Discuss the features of the graph. Now create separate histograms of the number of weeks spent on the charts for the top 50 songs and for the songs from 51 through 100. Explain the differences and similarities between the two groups.



Mark Cornelson/Photoshot Holdings Ltd.

## Numerical Descriptive Measures

Is the *average* over? You may ask, "How can the *average* be over?" Well, it seems like a strange question unless you are Thomas L. Friedman, a columnist for *The New York Times*. You may ask, "Why does Mr. Friedman think that the *average* is over?" Read his article, which appears as Case Study 3–2.

In Chapter 2 we discussed how to summarize data using different methods and to display data using graphs. Graphs are one important component of statistics; however, it is also important to numerically describe the main characteristics of a data set. The numerical summary measures, such as the ones that identify the center and spread of a distribution, identify many important features of a distribution. For example, the techniques learned in Chapter 2 can help us graph data on family incomes. However, if we want to know the income of a "typical" family (given by the center of the distribution), the spread of the distribution of incomes, or the relative position of a family with a particular income, the numerical summary measures can provide more detailed information (see Figure 3.1). The measures that we discuss in this chapter include measures of (1) central tendency, (2) dispersion (or spread), and (3) position.

### 3.1 Measures of Central Tendency for Ungrouped Data

**Case Study 3–1 Average NFL Ticket Prices in the Secondary Market**

**Case Study 3–2 Average Is Over**

**Case Study 3–3 Education Pays**

### 3.2 Measures of Dispersion for Ungrouped Data

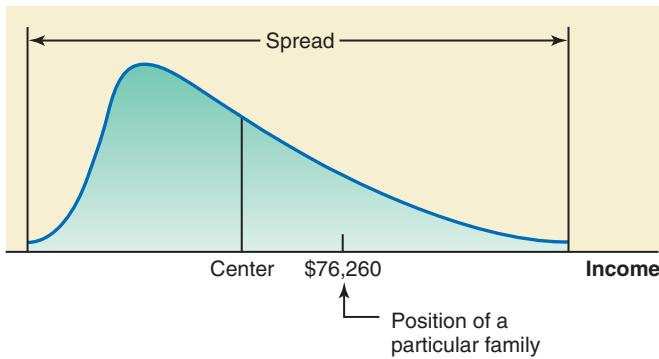
**3.3 Mean, Variance, and Standard Deviation for Grouped Data**

**3.4 Use of Standard Deviation**

**Case Study 3–4 Does Spread Mean the Same as Variability and Dispersion?**

### 3.5 Measures of Position

### 3.6 Box-and-Whisker Plot

**Figure 3.1****3.1****Measures of Central Tendency for Ungrouped Data**

We often represent a data set by numerical summary measures, usually called the *typical values*. A **measure of central tendency** gives the center of a histogram or a frequency distribution curve. This section discusses three different measures of central tendency: the mean, the median, and the mode; however, a few other measures of central tendency, such as the trimmed mean, the weighted mean, and the geometric mean, are explained in exercises following this section. We will learn how to calculate each of these measures for ungrouped data. Recall from Chapter 2 that the data that give information on each member of the population or sample individually are called *ungrouped data*, whereas *grouped data* are presented in the form of a frequency distribution table.

**3.1.1 Mean**

The **mean**, also called the *arithmetic mean*, is the most frequently used measure of central tendency. This book will use the words *mean* and *average* synonymously. For ungrouped data, the mean is obtained by dividing the sum of all values by the number of values in the data set:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

The mean calculated for sample data is denoted by  $\bar{x}$  (read as “ $x$  bar”), and the mean calculated for population data is denoted by  $\mu$  (Greek letter *mu*). We know from the discussion in Chapter 2 that the number of values in a data set is denoted by  $n$  for a sample and by  $N$  for a population. In Chapter 1, we learned that a variable is denoted by  $x$ , and the sum of all values of  $x$  is denoted by  $\Sigma x$ . Using these notations, we can write the following formulas for the mean.

**Calculating Mean for Ungrouped Data** The *mean for ungrouped data* is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data: } \mu = \frac{\Sigma x}{N}$$

$$\text{Mean for sample data: } \bar{x} = \frac{\Sigma x}{n}$$

where  $\Sigma x$  is the sum of all values,  $N$  is the population size,  $n$  is the sample size,  $\mu$  is the population mean, and  $\bar{x}$  is the sample mean.

## ■ EXAMPLE 3–1

Table 3.1 lists the total cash donations (rounded to millions of dollars) given by eight U.S. companies during the year 2010 (*Source:* Based on U.S. Internal Revenue Service data analyzed by *The Chronicle of Philanthropy* and *USA TODAY*).

*Calculating the sample mean for ungrouped data.*

**Table 3.1** Cash Donations in 2010 by Eight U.S. Companies

Company	Cash Donations (millions of dollars)
Wal-Mart	319
Exxon Mobil	199
Citigroup	110
Home Depot	63
Best Buy	21
Goldman Sachs	315
American Express	26
Nike	63

Find the mean of cash donations made by these eight companies.

**Solution** The variable in this example is the 2010 cash donations by a company. Let us denote this variable by  $x$ . Then, the eight values of  $x$  are

$$\begin{aligned}x_1 &= 319, & x_2 &= 199, & x_3 &= 110, & x_4 &= 63, \\x_5 &= 21, & x_6 &= 315, & x_7 &= 26, & \text{and} & x_8 = 63\end{aligned}$$

where  $x_1 = 319$  represents the 2010 cash donations (in millions of dollars) by Wal-Mart,  $x_2 = 199$  represents the 2010 cash donations by Exxon Mobil, and so on. The sum of the 2010 cash donations by these eight companies is

$$\begin{aligned}\Sigma x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 \\&= 319 + 199 + 110 + 63 + 21 + 315 + 26 + 63 = 1116\end{aligned}$$

Note that the given data include only eight companies. Hence, it represents a sample. Because the given data set contains eight companies,  $n = 8$ . Substituting the values of  $\Sigma x$  and  $n$  in the sample formula, we obtain the mean of 2010 cash donations of the eight companies as follows:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1116}{8} = 139.5 = \$139.5 \text{ million}$$

Thus, these eight companies donated an average of \$139.5 million in 2010 for charitable purposes. ■

## ■ EXAMPLE 3–2

The following are the ages (in years) of all eight employees of a small company:

53    32    61    27    39    44    49    57

Find the mean age of these employees.

*Calculating the population mean for ungrouped data.*

**Solution** Because the given data set includes *all* eight employees of the company, it represents the population. Hence,  $N = 8$ . We have

$$\Sigma x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

The population mean is

$$\mu = \frac{\sum x}{N} = \frac{362}{8} = 45.25 \text{ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months. ■

Reconsider Example 3–2. If we take a sample of three employees from this company and calculate the mean age of those three employees, this mean will be denoted by  $\bar{x}$ . Suppose the three values included in the sample are 32, 39, and 57. Then, the mean age for this sample is

$$\bar{x} = \frac{32 + 39 + 57}{3} = 42.67 \text{ years}$$

If we take a second sample of three employees of this company, the value of  $\bar{x}$  will (most likely) be different. Suppose the second sample includes the values 53, 27, and 44. Then, the mean age for this sample is

$$\bar{x} = \frac{53 + 27 + 44}{3} = 41.33 \text{ years}$$

Consequently, we can state that the value of the population mean  $\mu$  is constant. However, the value of the sample mean  $\bar{x}$  varies from sample to sample. The value of  $\bar{x}$  for a particular sample depends on what values of the population are included in that sample.

Sometimes a data set may contain a few very small or a few very large values. As mentioned in Chapter 2, such values are called *outliers* or *extreme values*.

A major shortcoming of the mean as a measure of central tendency is that it is very sensitive to outliers. Example 3–3 illustrates this point.

### ■ EXAMPLE 3–3

*Illustrating the effect of an outlier on the mean.*

Table 3.2 lists the total number of homes lost to foreclosure in seven states during 2010.

**Table 3.2 Number of Homes Foreclosed in 2010**

State	Number of Homes Foreclosed
California	173,175
Illinois	49,723
Minnesota	20,352
New Jersey	10,824
Ohio	40,911
Pennsylvania	18,038
Texas	61,848

Note that the number of homes foreclosed in California is very large compared to those in the other six states. Hence, it is an outlier. Show how the inclusion of this outlier affects the value of the mean.

**Solution** If we do not include the number of homes foreclosed in California (the outlier), the mean of the number of foreclosed homes in six states is

$$\begin{aligned} \text{Mean without the outlier} &= \frac{49,723 + 20,352 + 10,824 + 40,911 + 18,038 + 61,848}{6} \\ &= \frac{201,696}{6} = 33,616 \text{ homes} \end{aligned}$$



Data source: seatgeek.com.

The accompanying chart, based on data from seatgeek.com, shows the 2011–2012 average secondary market ticket price for all NFL teams as well as for a selected number of NFL teams. (Note that the secondary market tickets are the tickets that are resold via ticket Web sites such as Seatgeek.com. These tickets are not purchased directly from NFL franchises.) According to the data on seatgeek.com, the New York Giants had the highest secondary market average ticket price at \$332.82 and the San Francisco 49ers had the lowest secondary market average ticket price at \$27.99 for the 2011–2012 season. As we can see from the chart, there is a huge variation in the secondary market average ticket prices for these 10 teams, and this is true for all NFL teams. The secondary market average ticket price for all NFL teams was \$113.17 during the 2011–2012 season.

Source: <http://seatgeek.com/blog/nfl-average-ticket-prices-nfl>.

Now, to see the impact of the outlier on the value of the mean, we include the number of homes foreclosed in California and find the mean number of homes foreclosed in the seven states. This mean is

$$\text{Mean with the outlier} = \frac{173,175 + 49,723 + 20,352 + 10,824 + 40,911 + 18,038 + 61,848}{7}$$

$$= \frac{374,871}{7} = 53,553$$

Thus, including the number of homes foreclosed in California increases the mean by about 60% from 33,616 to 53,553. ■

The preceding example should encourage us to be cautious. We should remember that the mean is not always the best measure of central tendency because it is heavily influenced by outliers. Sometimes other measures of central tendency give a more accurate impression of a data set. For example, when a data set has outliers, instead of using the mean, we can use either the trimmed mean (defined in Exercise 3.33) or the median (to be discussed next) as a measure of central tendency.

### 3.1.2 Median

Another important measure of central tendency is the **median**. It is defined as follows.

#### Definition

**Median** The *median* is the value of the middle term in a data set that has been ranked in increasing order.

## CASE STUDY 3–2

### AVERAGE IS OVER

By Thomas L. Friedman. The following article was originally published in *The New York Times*.

In an essay, entitled "Making It in America," in the latest issue of the *Atlantic*, the author Adam Davidson relates a joke from cotton country about just how much a modern textile mill has been automated: The average mill has only two employees today: "a man and a dog. The man is there to feed the dog, and the dog is there to keep the man away from the machines."

Davidson's article is one of a number of pieces that have recently appeared making the point that the reason we have such stubbornly high unemployment and sagging middle-class incomes today is largely because of the big drop in demand because of the Great Recession, but it is also because of the quantum advances in both globalization and the information technology revolution, which are more rapidly than ever replacing labor with machines or foreign workers.

In the past, workers with average skills, doing an average job, could earn an average lifestyle. But, today, average is officially over. Being average just won't earn you what it used to. It can't when so many more employers have so much more access to so much more above average cheap foreign labor, cheap robotics, cheap software, cheap automation and cheap genius. Therefore, everyone needs to find their extra—their unique value contribution that makes them stand out in whatever is their field of employment. *Average is over.*

Yes, new technology has been eating jobs forever, and always will. As they say, if horses could have voted, there never would have been cars. But there's been an acceleration. As Davidson notes, "In the 10 years ending in 2009, [U.S.] factories shed workers so fast that they erased almost all the gains of the previous 70 years; roughly one out of every three manufacturing jobs—about 6 million in total—disappeared."

And you ain't seen nothin' yet. Last April, Annie Lowrey of *Slate* wrote about a start-up called "E la Carte" that is out to shrink the need for waiters and waitresses: The company "has produced a kind of souped-up iPad that lets you order and pay right at your table. The brainchild of a bunch of M.I.T. engineers, the nifty invention, known as the *Presto*, might be found at a restaurant near you soon . . . You select what you want to eat and add items to a cart. Depending on the restaurant's preferences, the console could show you nutritional information, ingredients lists and photographs. You can make special requests, like 'dressing on the side' or 'quintuple bacon.' When you're done, the order zings over to the kitchen, and the *Presto* tells you how long it will take for your items

to come out. . . . Bored with your companions? Play games on the machine. When you're through with your meal, you pay on the console, splitting the bill item by item if you wish and paying however you want. And you can have your receipt emailed to you . . . Each console goes for \$100 per month. If a restaurant serves meals eight hours a day, seven days a week, it works out to 42 cents per hour per table—making the *Presto* cheaper than even the very cheapest waiter.

What the iPad won't do in an above average way a Chinese worker will. Consider this paragraph from Sunday's terrific article in *The Times* by Charles Duhigg and Keith Bradsher about why Apple does so much of its manufacturing in China: "Apple had redesigned the iPhone's screen at the last minute, forcing an assembly-line overhaul. New screens began arriving at the [Chinese] plant near midnight. A foreman immediately roused 8,000 workers inside the company's dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day. 'The speed and flexibility is breathtaking,' the executive said. 'There's no American plant that can match that.'

And automation is not just coming to manufacturing, explains Curtis Carlson, the chief executive of SRI International, a Silicon Valley idea lab that invented the Apple iPhone program known as Siri, the digital personal assistant. "Siri is the beginning of a huge transformation in how we interact with banks, insurance companies, retail stores, health care providers, information retrieval services and product services."

There will always be change—new jobs, new products, new services. But the one thing we know for sure is that with each advance in globalization and the I.T. revolution, the best jobs will require workers to have more and better education to make themselves above average. Here are the latest unemployment rates from the Bureau of Labor Statistics for Americans over 25 years old: those with less than a high school degree, 13.8 percent; those with a high school degree and no college, 8.7 percent; those with some college or associate degree, 7.7 percent; and those with bachelor's degree or higher, 4.1 percent.

In a world where average is officially over, there are many things we need to do to buttress employment, but nothing would be more important than passing some kind of G.I. Bill for the 21st century that ensures that every American has access to post-high school education.

**Source:** Thomas L. Friedman, *The New York Times*, January 25, 2012. Copyright © 2012, *The New York Times*. All rights reserved. Used with permission and protected by the Copyright Laws of the United States. The printing, copying, redistribution, or retransmission of this Content without express written permission is prohibited.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the data set in increasing order.
2. Find the middle term. The value of this term is the median.<sup>1</sup>

Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

### ■ EXAMPLE 3–4

Refer to the data on the number of homes foreclosed in seven states given in Table 3.2 of Example 3–3. Those values are listed below.

173,175    49,723    20,352    10,824    40,911    18,038    61,848

*Calculating the median for ungrouped data: odd number of data values.*

Find the median for these data.

**Solution** First, we rank the given data in increasing order as follows:

10,824    18,038    20,352    40,911    49,723    61,848    173,175

Since there are seven homes in this data set and the middle term is the fourth term, the median is given by the value of the fourth term in the ranked data as shown below.

10,824    18,038    20,352    **40,911**    49,723    61,848    173,175  
 ↑  
 Median

Thus, the median number of homes foreclosed in these seven states was 40,911 in 2010. ■

### ■ EXAMPLE 3–5

Table 3.3 gives the total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies.

**Table 3.3 Total Compensations of 12 Highest-Paid CEOs for the Year 2010**

*Calculating the median for ungrouped data: even number of data values.*

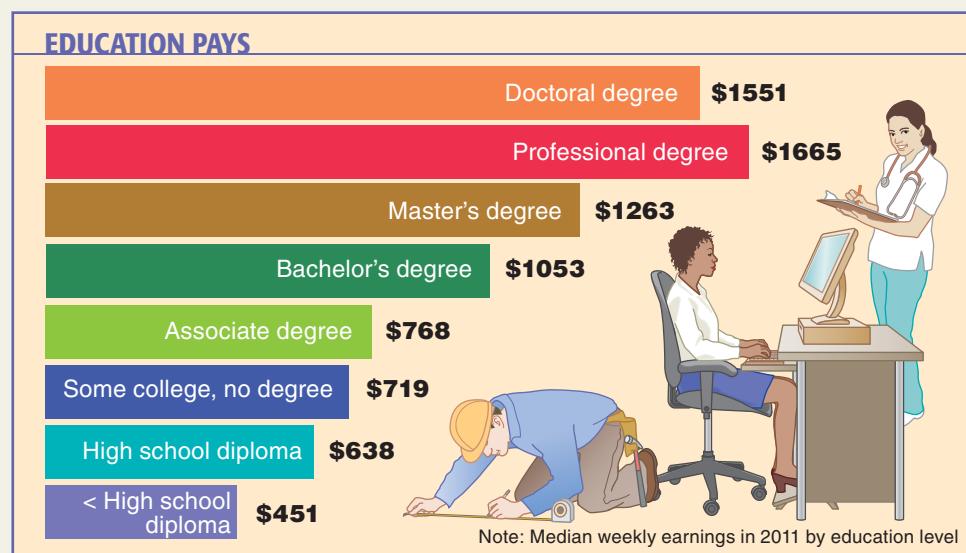
CEO and Company	2010 Total Compensation (millions of dollars)
Michael D. White (DirecTV)	32.9
David N. Farr (Emerson Electric)	22.9
Brian L. Roberts (Comcast)	28.2
Philippe P. Dauman (Viacom)	84.5
William C. Weldon (Johnson & Johnson)	21.6
Robert A. Iger (Walt Disney)	28.0
Ray R. Iran (Occidental Petroleum)	76.1
Samuel J. Palmisano (IBM)	25.2
John F. Lundgren (Stanley Black & Decker)	32.6
Lawrence J. Ellison (Oracle)	70.1
Alan Mulally (Ford Motor)	26.5
Howard Schultz (Starbucks)	21.7

Find the median for these data.

<sup>1</sup>The value of the middle term in a data set ranked in *decreasing* order will also give the value of the median.

## CASE STUDY 3–3

### EDUCATION PAYS



Data source: U.S. Bureau of Labor Statistics.

The accompanying chart shows the 2011 median weekly salaries by education level for persons of age 25 years and older who held full-time jobs. These salaries are based on the Current Population Survey conducted by the Bureau of Labor Statistics. Although this survey is called the Current Population Survey, it is actually based on a sample. Usually the samples taken by the Bureau of Labor Statistics for these surveys are very large. As shown in the chart, the highest median weekly earning (of \$1665) was for workers with a professional degree and the lowest (of \$451) was for workers with less than a high school diploma.

Data source: [http://www.bls.gov/emp/ep\\_chart\\_001.htm/](http://www.bls.gov/emp/ep_chart_001.htm/).

**Solution** First we rank the given total compensations of the 12 CEOs as follows:

21.6    21.7    22.9    25.2    26.5    28.0    28.2    32.6    32.9    70.1    76.1    84.5

There are 12 values in this data set. Because there is an even number of values in the data set, the median will be given by the average of the two middle values. The two middle values are the sixth and seventh in the arranged data, and these two values are 28.0 and 28.2. The median, which is given by the average of these two values, is calculated as follows:

21.6    21.7    22.9    25.2    26.5    28.0    28.2    32.6    32.9    70.1    76.1    84.5  
↑  
Median = 28.1

$$\text{Median} = \frac{28.0 + 28.2}{2} = \frac{56.2}{2} = 28.1 = \$28.1 \text{ million}$$

Thus, the median for the 2010 compensations of these 12 CEOs is \$28.1 million. ■

The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median. The advantage of using the median as a measure of central tendency is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers.

### 3.1.3 Mode

**Mode** is a French word that means *fashion*—an item that is most popular or common. In statistics, the mode represents the most common value in a data set.

#### Definition

**Mode** The *mode* is the value that occurs with the highest frequency in a data set.

## ■ EXAMPLE 3–6

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77    82    74    81    79    84    74    78

*Calculating the mode for ungrouped data.*

Find the mode.

**Solution** In this data set, 74 occurs twice, and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode. Therefore,

$$\text{Mode} = \mathbf{74 \text{ miles per hour}}$$



A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median. For instance, a data set with each value occurring only once has no mode. A data set with only one value occurring with the highest frequency has only one mode. The data set in this case is called **unimodal**. A data set with two values that occur with the same (highest) frequency has two modes. The distribution, in this case, is said to be **bimodal**. If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be **multimodal**.

## ■ EXAMPLE 3–7

Last year's incomes of five randomly selected families were \$76,150, \$95,750, \$124,985, \$87,490, and \$53,740. Find the mode.

*Data set with no mode.*

**Solution** Because each value in this data set occurs only once, this data set contains **no mode**.



## ■ EXAMPLE 3–8

A small company has 12 employees. Their commuting times (rounded to the nearest minute) from home to work are 23, 36, 12, 23, 47, 32, 8, 12, 26, 31, 18, and 28, respectively. Find the mode for these data.

*Data set with two modes.*

**Solution** In the given data on the commuting times of these 12 employees, each of the values 12 and 23 occurs twice, and each of the remaining values occurs only once. Therefore, this data set has two modes: 12 and 23 minutes.



## ■ EXAMPLE 3–9

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22, and 30 years, respectively. Find the mode.

*Data set with three modes.*

**Solution** This data set has three modes: **19**, **21**, and **22**. Each of these three values occurs with a (highest) frequency of 2.



One advantage of the mode is that it can be calculated for both kinds of data—quantitative and qualitative—whereas the mean and median can be calculated for only quantitative data.

## ■ EXAMPLE 3–10

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, and senior, respectively. Find the mode.

*Finding the mode for qualitative data.*

**Solution** Because **senior** occurs more frequently than the other categories, it is the mode for this data set. We cannot calculate the mean and median for this data set.



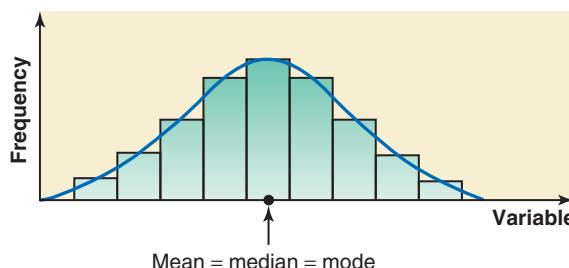
To sum up, we cannot say for sure which of the three measures of central tendency is a better measure overall. Each of them may be better under different situations. Probably the mean is the most-used measure of central tendency, followed by the median. The mean has the advantage that its calculation includes each value of the data set. The median is a better measure when a data set includes outliers. The mode is simple to locate, but it is not of much use in practical applications.

### 3.1.4 Relationships Among the Mean, Median, and Mode

As discussed in Chapter 2, two of the many shapes that a histogram or a frequency distribution curve can assume are symmetric and skewed. This section describes the relationships among the mean, median, and mode for three such histograms and frequency distribution curves. Knowing the values of the mean, median, and mode can give us some idea about the shape of a frequency distribution curve.

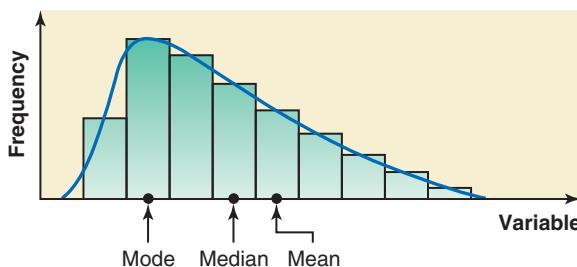
1. For a symmetric histogram and frequency distribution curve with one peak (see Figure 3.2), the values of the mean, median, and mode are identical, and they lie at the center of the distribution.

**Figure 3.2** Mean, median, and mode for a symmetric histogram and frequency distribution curve.



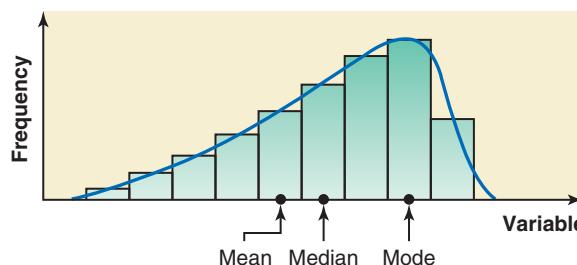
2. For a histogram and a frequency distribution curve skewed to the right (see Figure 3.3), the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two. (Notice that the mode always occurs at the peak point.) The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail. These outliers pull the mean to the right.

**Figure 3.3** Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.



3. If a histogram and a frequency distribution curve are skewed to the left (see Figure 3.4), the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two. In this case, the outliers in the left tail pull the mean to the left.

**Figure 3.4** Mean, median, and mode for a histogram and frequency distribution curve skewed to the left.



## EXERCISES

### CONCEPTS AND PROCEDURES

- 3.1** Explain how the value of the median is determined for a data set that contains an odd number of observations and for a data set that contains an even number of observations.
- 3.2** Briefly explain the meaning of an outlier. Is the mean or the median a better measure of central tendency for a data set that contains outliers? Illustrate with the help of an example.
- 3.3** Using an example, show how outliers can affect the value of the mean.
- 3.4** Which of the three measures of central tendency (the mean, the median, and the mode) can be calculated for quantitative data only, and which can be calculated for both quantitative and qualitative data? Illustrate with examples.
- 3.5** Which of the three measures of central tendency (the mean, the median, and the mode) can assume more than one value for a data set? Give an example of a data set for which this summary measure assumes more than one value.
- 3.6** Is it possible for a (quantitative) data set to have no mean, no median, or no mode? Give an example of a data set for which this summary measure does not exist.
- 3.7** Explain the relationships among the mean, median, and mode for symmetric and skewed histograms. Illustrate these relationships with graphs.
- 3.8** Prices of cars have a distribution that is skewed to the right with outliers in the right tail. Which of the measures of central tendency is the best to summarize this data set? Explain.

- 3.9** The following data set belongs to a population:

5      -7      2      0      -9      16      10      7

Calculate the mean, median, and mode.

- 3.10** The following data set belongs to a sample:

14      18      -10      8      8      -16

Calculate the mean, median, and mode.

### APPLICATIONS

- 3.11** The following table gives the standard deductions and personal exemptions for persons filing with “single” status on their 2011 state income taxes in a random sample of 9 states. Calculate the mean and median for the data on standard deductions for these states.

State	Standard Deduction (in dollars)	Personal Exemption (in dollars)
Delaware	3250	110
Hawaii	2000	1040
Kentucky	2190	20
Minnesota	5450	3500
North Dakota	5700	3650
Oregon	1945	176
Rhode Island	5700	3650
Vermont	5700	3650
Virginia	3000	930

Source: [www.taxfoundation.org](http://www.taxfoundation.org).

- 3.12** Refer to the data table in Exercise 3.11. Calculate the mean and median for the data on personal exemptions for these states.

**3.13** The following data give the 2010 gross domestic product (in billions of dollars) for all 50 states. The data are entered in alphabetical order by state (*Source*: Bureau of Economic Analysis).

173	49	254	103	1901	258	237	62	748	403
67	55	652	276	143	127	163	219	52	295
379	384	270	97	244	36	90	126	60	487
80	1160	425	35	478	148	174	570	49	164
40	255	1207	115	26	424	340	65	248	39

- a. Calculate the mean and median for these data. Are these values of the mean and the median sample statistics or population parameters? Explain.
- b. Do these data have a mode? Explain.

**3.14** The following data give the 2010 revenues (in millions of dollars) of the six Maryland-based companies listed in the 2010 *Fortune 500* ([www.money.cnn.com/magazines/fortune/fortune500/2010/states/MD.html](http://www.money.cnn.com/magazines/fortune/fortune500/2010/states/MD.html)). The data refer to the following companies, respectively: Lockheed Martin, Constellation Energy, Coventry Health Care, Marriott International, Black & Decker, and Host Hotels & Resorts.

45,189.0    15,598.8    13,993.3    10,908.0    4775.1    4216.0

Find the mean and median for these data. Do these data have a mode? Assume that these six companies constitute the population of companies from Maryland in the 2010 *Fortune 500*.

**3.15** The following table lists the 2009 total profits (rounded to millions of dollars) of the seven *Fortune 500* companies in the Computers, Office Equipment category (*Source*: [www.money.cnn.com/magazines/fortune/fortune500/2010/industries/8/index.html](http://www.money.cnn.com/magazines/fortune/fortune500/2010/industries/8/index.html)).

Company	2009 Profit (millions of dollars)
Hewlett-Packard	7660
Dell	1433
Apple	5704
Xerox	485
Sun Microsystems	-2234
Pitney Bowes	423
NCR	-33

Find the mean and median for these data. (Note: The negative values for Sun Microsystems and NCR imply that both companies lost money in 2009.) Assume that these seven companies constitute the population of such companies in the 2009 *Fortune 500*.

**3.16** The following data give the numbers of car thefts that occurred in a city during the past 12 days.

6    3    7    11    4    3    8    7    2    6    9    15

Find the mean, median, and mode.

**3.17** The following data give the amount of money (in dollars) that each of six Canadian social service charities spent to raise \$100 in donations during 2010 ([www.moneysense.ca](http://www.moneysense.ca)). The values, listed in that order, are for the Calgary Inter-Faith Food Bank Society, Covenant House Toronto, The Salvation Army Territorial Headquarters for Canada and Bermuda, Second Harvest Food Support Committee, Teen Challenge, and Toronto Windfall Clothing Support Service.

.20    29.30    11.30    5.30    9.90    .50

Compute the mean and median. Do these data have a mode? Why or why not?

**3.18** The following table gives the number of major penalties for each of the 15 teams in the Eastern Conference of the National Hockey League during the 2010–11 season ([www.nhl.com](http://www.nhl.com)). A major penalty is subject to 5 minutes in the penalty box for a player.

Team	Number of Major Penalties
Pittsburgh	74
Boston	73
New York Islanders	71
New York Rangers	62
Columbus	59
Toronto	53
Ottawa	51
Philadelphia	49
Washington	46
New Jersey	39
Montreal	35
Atlanta	34
Buffalo	30
Florida	26
Tampa Bay	23

Compute the mean and median for the data on major penalties. Do these data have a mode? Why or why not?

- 3.19** Due to antiquated equipment and frequent windstorms, the town of Oak City often suffers power outages. The following data give the numbers of power outages for each of the past 12 months.

4      5      7      3      2      0      2      3      2      1      2      4

Compute the mean, median, and mode for these data.

- 3.20** Standard milk chocolate M&Ms™ come in six colors. The Fun Size bags typically contain between 16 and 20 candies, so it is common for a Fun Size bag to have some of the six colors missing. Each of the 14 students in a summer statistics class was given a Fun Size bag and asked to count the number of colors present in the bag. The following data are the number of colors found in these 14 bags:

3      6      5      4      6      3      2      5      5      4      5      6      3      4

Find the mean, median, and mode for these data. Are the values of these summary measures population parameters or sample statistics? Explain why.

- 3.21** Nixon Corporation manufactures computer monitors. The following data are the numbers of computer monitors produced at the company for a sample of 10 days.

24      32      27      23      35      33      29      40      23      28

Calculate the mean, median, and mode for these data.

- 3.22** Grand Jury indictment data for Gloucester County, New Jersey, are published every week in the *Gloucester County Times* newspaper ([www.nj.com/gloucester](http://www.nj.com/gloucester)). The following data are the number of indictments for a sample of 11 weeks selected from July 2010 through June 2011:

35      13      17      21      21      29      20      26      24      13      23

Find the mean, median, and mode for these data.

- 3.23** The following data represent the numbers of tornadoes that touched down during 1950 to 1994 in the 12 states that had the most tornadoes during this period. The data for these states are given in the following order: CO, FL, IA, IL, KS, LA, MO, MS, NE, OK, SD, TX.

1113    2009    1374    1137    2110    1086    1166    1039    1673    2300    1139    5490

- a. Calculate the mean and median for these data.
- b. Identify the outlier in this data set. Drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
- c. Which is the better summary measure for these data, the mean or the median? Explain.

- 3.24** The following data set lists the number of women from each of 12 countries who were on the Rolex Women's World Golf Rankings Top 50 list as of July 18, 2011. The data, listed in that order, are for the

following countries: Australia, Chinese Taipei, England, Germany, Japan, Netherlands, Norway, Scotland, South Korea, Spain, Sweden, and the United States.

3      1      1      1      10      1      1      1      18      2      3      8

- Calculate the mean and median for these data.
- Identify the outlier in this data set. Drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
- Which is the better summary measure for these data, the mean or the median? Explain.

**\*3.25** One property of the mean is that if we know the means and sample sizes of two (or more) data sets, we can calculate the **combined mean** of both (or all) data sets. The combined mean for two data sets is calculated by using the formula

$$\text{Combined mean} = \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

where  $n_1$  and  $n_2$  are the sample sizes of the two data sets and  $\bar{x}_1$  and  $\bar{x}_2$  are the means of the two data sets, respectively. Suppose a sample of 10 statistics books gave a mean price of \$140 and a sample of 8 mathematics books gave a mean price of \$160. Find the combined mean. (*Hint:* For this example:  $n_1 = 10$ ,  $n_2 = 8$ ,  $\bar{x}_1 = \$140$ ,  $\bar{x}_2 = \$160$ .)

**\*3.26** Twenty business majors and 18 economics majors go bowling. Each student bowls one game. The scorekeeper announces that the mean score for the 18 economics majors is 144 and the mean score for the entire group of 38 students is 150. Find the mean score for the 20 business majors.

**\*3.27** For any data, the sum of all values is equal to the product of the sample size and mean; that is,  $\Sigma x = n\bar{x}$ . Suppose the average amount of money spent on shopping by 10 persons during a given week is \$105.50. Find the total amount of money spent on shopping by these 10 persons.

**\*3.28** The mean 2011 income for five families was \$99,520. What was the total 2011 income of these five families?

**\*3.29** The mean age of six persons is 46 years. The ages of five of these six persons are 57, 39, 44, 51, and 37 years, respectively. Find the age of the sixth person.

**\*3.30** Seven airline passengers in economy class on the same flight paid an average of \$361 per ticket. Because the tickets were purchased at different times and from different sources, the prices varied. The first five passengers paid \$420, \$210, \$333, \$695, and \$485. The sixth and seventh tickets were purchased by a couple who paid identical fares. What price did each of them pay?

**\*3.31** Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Notice that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

**\*3.32** Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	8	16	30	18	22

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

**\*3.33** The **trimmed mean** is calculated by dropping a certain percentage of values from each end of a ranked data set. The trimmed mean is especially useful as a measure of central tendency when a data set contains a few outliers. Suppose the following data give the ages (in years) of 10 employees of a company:

47      53      38      26      39      49      19      67      31      23

To calculate the 10% trimmed mean, first rank these data values in increasing order; then drop 10% of the smallest values and 10% of the largest values. The mean of the remaining 80% of the values will give the 10% trimmed mean. Note that this data set contains 10 values, and 10% of 10 is 1. Thus, if we drop

the smallest value and the largest value from this data set, the mean of the remaining 8 values will be called the 10% trimmed mean. Calculate the 10% trimmed mean for this data set.

- \*3.34 The following data give the prices (in thousands of dollars) of 20 houses sold recently in a city.

184	297	365	309	245	387	369	438	195	390
323	578	410	679	307	271	457	795	259	590

Find the 20% trimmed mean for this data set.

\*3.35 In some applications, certain values in a data set may be considered more important than others. For example, to determine students' grades in a course, an instructor may assign a weight to the final exam that is twice as much as that to each of the other exams. In such cases, it is more appropriate to use the **weighted mean**. In general, for a sequence of  $n$  data values  $x_1, x_2, \dots, x_n$  that are assigned weights  $w_1, w_2, \dots, w_n$ , respectively, the **weighted mean** is found by the formula

$$\text{Weighted mean} = \frac{\sum xw}{\sum w}$$

where  $\sum xw$  is obtained by multiplying each data value by its weight and then adding the products. Suppose an instructor gives two exams and a final, assigning the final exam a weight twice that of each of the other exams. Find the weighted mean for a student who scores 73 and 67 on the first two exams and 85 on the final. (Hint: Here,  $x_1 = 73$ ,  $x_2 = 67$ ,  $x_3 = 85$ ,  $w_1 = w_2 = 1$ , and  $w_3 = 2$ .)

\*3.36 When studying phenomena such as inflation or population changes that involve periodic increases or decreases, the **geometric mean** is used to find the average change over the entire period under study. To calculate the geometric mean of a sequence of  $n$  values  $x_1, x_2, \dots, x_n$ , we multiply them together and then find the  $n$ th root of this product. Thus

$$\text{Geometric mean} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n}$$

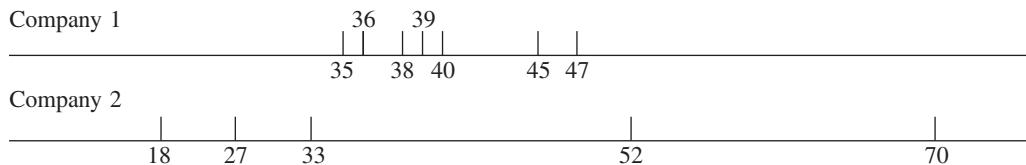
Suppose that the inflation rates for the last five years are 4%, 3%, 5%, 6%, and 8%, respectively. Thus at the end of the first year, the price index will be 1.04 times the price index at the beginning of the year, and so on. Find the mean rate of inflation over the 5-year period by finding the geometric mean of the data set 1.04, 1.03, 1.05, 1.06, and 1.08. (Hint: Here,  $n = 5$ ,  $x_1 = 1.04$ ,  $x_2 = 1.03$ , and so on. Use the  $x^{1/n}$  key on your calculator to find the fifth root. Note that the mean inflation rate will be obtained by subtracting 1 from the geometric mean.)

## 3.2 Measures of Dispersion for Ungrouped Data

The measures of central tendency, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than for the other data set. (Note that the words *dispersion*, *spread*, and *variation* have similar meanings.) Consider the following two data sets on the ages (in years) of all workers working for each of two small companies.

Company 1:	47	38	35	40	36	45	39
Company 2:		70	33	18	52	27	

The mean age of workers in both these companies is the same, 40 years. If we do not know the ages of individual workers at these two companies and are told only that the mean age of the workers at both companies is the same, we may deduce that the workers at these two companies have a similar age distribution. As we can observe, however, the variation in the workers' ages for each of these two companies is very different. As illustrated in the diagram, the ages of the workers at the second company have a much larger variation than the ages of the workers at the first company.



Thus, the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us learn about the spread of a data set are called the **measures of dispersion**. The measures of central tendency and dispersion taken together give a better picture of a data set than the measures of central tendency alone. This section discusses three measures of dispersion: range, variance, and standard deviation. Another measure of spread, called the coefficient of variation (CV), is explained in Exercise 3.57.

### 3.2.1 Range

The **range** is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

#### Finding the Range for Ungrouped Data

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

#### ■ EXAMPLE 3-11

*Calculating the range for ungrouped data.*

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

**Solution** The maximum total area for a state in this data set is 267,277 square miles, and the smallest area is 49,651 square miles. Therefore,

$$\begin{aligned}\text{Range} &= \text{Largest value} - \text{Smallest value} \\ &= 267,277 - 49,651 = \mathbf{217,626 \text{ square miles}}\end{aligned}$$

Thus, the total areas of these four states are spread over a range of 217,626 square miles. ■

The range, like the mean, has the disadvantage of being influenced by outliers. In Example 3-11, if the state of Texas with a total area of 267,277 square miles is dropped, the range decreases from 217,626 square miles to 20,252 square miles. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.

Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

### 3.2.2 Variance and Standard Deviation

The **standard deviation** is the most-used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The *standard deviation is obtained by taking the positive square root of the variance*. The variance calculated for population data is denoted by  $\sigma^2$  (read as *sigma squared*)<sup>2</sup>, and the variance calculated for sample data is denoted by  $s^2$ . Consequently, the standard deviation

<sup>2</sup>Note that  $\Sigma$  is uppercase sigma and  $\sigma$  is lowercase sigma of the Greek alphabet.

calculated for population data is denoted by  $\sigma$ , and the standard deviation calculated for sample data is denoted by  $s$ . Following are what we will call the *basic formulas* that are used to calculate the variance and standard deviation.<sup>3</sup>

$$\begin{aligned}\sigma^2 &= \frac{\sum(x - \mu)^2}{N} & \text{and} & \quad s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \\ \sigma &= \sqrt{\frac{\sum(x - \mu)^2}{N}} & \text{and} & \quad s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}\end{aligned}$$

where  $\sigma^2$  is the population variance,  $s^2$  is the sample variance,  $\sigma$  is the population standard deviation, and  $s$  is the sample standard deviation.

The quantity  $x - \mu$  or  $x - \bar{x}$  in the above formulas is called the *deviation* of the  $x$  value from the mean. The sum of the deviations of the  $x$  values from the mean is always zero; that is,  $\sum(x - \mu) = 0$  and  $\sum(x - \bar{x}) = 0$ .

For example, suppose the midterm scores of a sample of four students are 82, 95, 67, and 92, respectively. Then, the mean score for these four students is

$$\bar{x} = \frac{82 + 95 + 67 + 92}{4} = 84$$

The deviations of the four scores from the mean are calculated in Table 3.5. As we can observe from the table, the sum of the deviations of the  $x$  values from the mean is zero; that is,  $\sum(x - \bar{x}) = 0$ . For this reason we square the deviations to calculate the variance and standard deviation.

**Table 3.5**

$x$	$x - \bar{x}$
82	$82 - 84 = -2$
95	$95 - 84 = +11$
67	$67 - 84 = -17$
92	$92 - 84 = +8$
	$\sum(x - \bar{x}) = 0$

From the computational point of view, it is easier and more efficient to use *short-cut formulas* to calculate the variance and standard deviation. By using the short-cut formulas, we reduce the computation time and round-off errors. Use of the basic formulas for ungrouped data is illustrated in Section A3.1.1 of Appendix 3.1 of this chapter. The short-cut formulas for calculating the variance and standard deviation are as follows.

#### Short-Cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

where  $\sigma^2$  is the population variance and  $s^2$  is the sample variance.

The standard deviation is obtained by taking the positive square root of the variance.

$$\begin{aligned}\text{Population standard deviation:} \quad \sigma &= \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}} \\ \text{Sample standard deviation:} \quad s &= \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}\end{aligned}$$

<sup>3</sup>From the formula for  $\sigma^2$ , it can be stated that the population variance is the mean of the squared deviations of  $x$  values from the mean. However, this is not true for the variance calculated for a sample data set.

Note that the denominator in the formula for the population variance is  $N$ , but that in the formula for the sample variance it is  $n - 1$ .<sup>4</sup>

### ■ EXAMPLE 3–12

*Calculating the sample variance and standard deviation for ungrouped data.*



© Hazlan Abdul Hakim/Stockphoto

Until about 2009, airline passengers were not charged for checked baggage. Around 2009, however, many U.S. airlines started charging a fee for bags. According to the Bureau of Transportation Statistics, U.S. airlines collected more than \$3 billion in baggage fee revenue in 2010. The following table lists the baggage fee revenues of six U.S. airlines for the year 2010. (Note that Delta's revenue reflects a merger with Northwest. Also note that since then United and Continental have merged; and American filed for bankruptcy and may merge with another airline.)

Airline	Baggage Fee Revenue (millions of dollars)
United	313
Continental	342
American	581
Delta	952
US Airways	514
AirTran	152

Find the variance and standard deviation for these data.

**Solution** Let  $x$  denote the 2010 baggage fee revenue (in millions of dollars) of an airline. The values of  $\Sigma x$  and  $\Sigma x^2$  are calculated in Table 3.6.

Table 3.6

$x$	$x^2$
313	97,969
342	116,964
581	337,561
952	906,304
514	264,196
152	23,104
$\Sigma x = 2854$	$\Sigma x^2 = 1,746,098$

Calculation of the variance and standard deviation involves the following four steps.

**Step 1. Calculate  $\Sigma x$ .**

The sum of the values in the first column of Table 3.6 gives the value of  $\Sigma x$ , which is 2854.

**Step 2. Find  $\Sigma x^2$ .**

The value of  $\Sigma x^2$  is obtained by squaring each value of  $x$  and then adding the squared values. The results of this step are shown in the second column of Table 3.6. Notice that  $\Sigma x^2 = 1,746,098$ .

<sup>4</sup>The reason that the denominator in the sample formula is  $n - 1$  and not  $n$  follows: The sample variance underestimates the population variance when the denominator in the sample formula for variance is  $n$ . However, the sample variance does not underestimate the population variance if the denominator in the sample formula for variance is  $n - 1$ . In Chapter 8 we will learn that  $n - 1$  is called the degrees of freedom.

**Step 3.** Determine the variance.

Substitute all the values in the variance formula and simplify. Because the given data are for the baggage fee revenues of only six airlines, we use the formula for the sample variance:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{1,746,098 - \frac{(2854)^2}{6}}{6-1} = \frac{1,746,098 - 1,357,552.667}{5} = 77,709.06666$$

**Step 4.** Obtain the standard deviation.

The standard deviation is obtained by taking the (positive) square root of the variance:

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{77,709.06666} = 278.7634601 = \$278.76 \text{ million}$$

Thus, the standard deviation of the 2010 baggage fee revenues of these six airlines is \$278.76 million. ■

- The values of the variance and the standard deviation are never negative.** That is, the numerator in the formula for the variance should never produce a negative value. Usually the values of the variance and standard deviation are positive, but if a data set has no variation, then the variance and standard deviation are both zero. For example, if four persons in a group are the same age—say, 35 years—then the four values in the data set are

35      35      35      35

If we calculate the variance and standard deviation for these data, their values are zero. This is because there is no variation in the values of this data set.

- The measurement units of the variance are always the square of the measurement units of the original data.** This is so because the original values are squared to calculate the variance. In Example 3–12, the measurement units of the original data are millions of dollars. However, the measurement units of the variance are squared millions of dollars, which, of course, does not make any sense. Thus, the variance of the 2010 baggage fee revenue of the six airlines in Example 3–12 is 77,709.06666 squared million dollars. But the measurement units of the standard deviation are the same as the measurement units of the original data because the standard deviation is obtained by taking the square root of the variance.

**Two Observations****EXAMPLE 3–13**

Following are the 2011 earnings (in thousands of dollars) before taxes for all six employees of a small company.

88.50      108.40      65.50      52.50      79.80      54.60

Calculate the variance and standard deviation for these data.

**Solution** Let  $x$  denote the 2011 earnings before taxes of an employee of this company. The values of  $\sum x$  and  $\sum x^2$  are calculated in Table 3.7.

Calculating the population variance and standard deviation for ungrouped data.

**Table 3.7**

$x$	$x^2$
88.50	7832.25
108.40	11,750.56
65.50	4290.25
52.50	2756.25
79.80	6368.04
54.60	2981.16
$\sum x = 449.30$	$\sum x^2 = 35,978.51$

Because the data in this example are on earnings of *all* employees of this company, we use the population formula to compute the variance. Thus, the variance is

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} = \frac{35,978.51 - \frac{(449.30)^2}{6}}{6} = \mathbf{388.90}$$

The standard deviation is obtained by taking the (positive) square root of the variance:

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}} = \sqrt{388.90} = \mathbf{19.721 \text{ thousand} = \$19,721}$$

Thus, the standard deviation of the 2011 earnings of all six employees of this company is \$19,721. ■

**Warning ▶** Note that  $\sum x^2$  is not the same as  $(\sum x)^2$ . The value of  $\sum x^2$  is obtained by squaring the  $x$  values and then adding them. The value of  $(\sum x)^2$  is obtained by squaring the value of  $\sum x$ .

The uses of the standard deviation are discussed in Section 3.4. Later chapters explain how the mean and the standard deviation taken together can help in making inferences about the population.

### 3.2.3 Population Parameters and Sample Statistics

A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a *population parameter*, or simply a **parameter**. A summary measure calculated for a sample data set is called a *sample statistic*, or simply a **statistic**. Thus,  $\mu$  and  $\sigma$  are population parameters, and  $\bar{x}$  and  $s$  are sample statistics. As an illustration,  $\bar{x} = \$139.5$  million in Example 3–1 is a sample statistic, and  $\mu = 45.25$  years in Example 3–2 is a population parameter. Similarly,  $s = \$278.76$  million in Example 3–12 is a sample statistic, whereas  $\sigma = \$19,721$  in Example 3–13 is a population parameter.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**3.37** The range, as a measure of spread, has the disadvantage of being influenced by outliers. Illustrate this with an example.

**3.38** Can the standard deviation have a negative value? Explain.

**3.39** When is the value of the standard deviation for a data set zero? Give one example. Calculate the standard deviation for the example and show that its value is zero.

**3.40** Briefly explain the difference between a population parameter and a sample statistic. Give one example of each.

**3.41** The following data set belongs to a population:

5      -7      2      0      -9      16      10      7

Calculate the range, variance, and standard deviation.

**3.42** The following data set belongs to a sample:

14      18      -10      8      8      -16

Calculate the range, variance, and standard deviation.

## ■ APPLICATIONS

**3.43** The following data give the number of shoplifters apprehended during each of the past 8 weeks at a large department store.

7      10      8      3      15      12      6      11

- Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
- Calculate the range, variance, and standard deviation.

**3.44** The following data give the prices of seven textbooks randomly selected from a university bookstore.

\$89      \$170      \$104      \$113      \$56      \$161      \$147

- Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
- Calculate the range, variance, and standard deviation.

**3.45** Refer to Exercise 3.20, which listed the number of colors of M&Ms that each of the 14 Fun Size bags contained. Those data are reproduced here:

3      6      5      4      6      3      2      5      5      4      5      6      3      4

Calculate the range, variance, and standard deviation.

**3.46** Refer to the data in Exercise 3.23, which contained the numbers of tornadoes that touched down in 12 states that had the most tornadoes during the period 1950 to 1994. The data are reproduced here.

1113    2009    1374    1137    2110    1086    1166    1039    1673    2300    1139    5490

Find the range, variance, and standard deviation for these data.

**3.47** Refer to Exercise 3.22, which listed the number of indictments handed out by the Gloucester County, New Jersey, Grand Jury during 11 randomly selected weeks from July 2010 to June 2011. The data are reproduced here:

35      13      17      21      21      29      20      26      24      13      23

Calculate the range, variance, and standard deviation.

**3.48** The following data give the number of highway collisions with large wild animals, such as deer or moose, in one of the northeastern states during each week of a 9-week period.

7      10      3      8      2      5      7      4      9

Find the range, variance, and standard deviation.

**3.49** Refer to Exercise 3.24, which listed the number of women from each of 12 countries who were on the Rolex Women's World Golf Rankings Top 50 list as of July 18, 2011. Those data are reproduced here:

3      1      1      1      10      1      1      18      2      3      8

Calculate the range, variance, and standard deviation.

**3.50** The lengths (in seconds) of the eight most recent songs played on 98.9 FM WCLZ and WCLZ.com (Portland, ME) at 1:28 p.m. on Wednesday July 20, 2011 were as follows:

251      252      213      182      244      259      262      216

Calculate the range, variance, and standard deviation.

**3.51** Following are the temperatures (in degrees Fahrenheit) observed during eight wintry days in a mid-western city:

23      14      6      -7      -2      11      16      19

Compute the range, variance, and standard deviation.

**3.52** Refer to Exercise 2.94, which listed the alcohol content by volume for each of the 13 varieties of beer produced by Sierra Nevada Brewery. Those data are reproduced here:

4.4      5.0      5.0      5.6      5.6      5.8      5.9      5.9      6.7      6.8      6.9      7.0      9.6

Calculate the range, variance, and standard deviation.

**3.53** The following data represent the total points scored in each of the NFL Super Bowl games played from 2001 through 2012, in that order:

41    37    69    61    45    31    46    31    50    48    56    38

Compute the range, variance, and standard deviation for these data.

**3.54** The following data represent the 2011 guaranteed salaries (in thousands of dollars) of the head coaches of the final eight teams in the 2011 NCAA Men's Basketball Championship. The data represent the 2011 salaries of basketball coaches of the following universities, entered in that order: Arizona, Butler, Connecticut, Florida, Kansas, Kentucky, North Carolina, and Virginia Commonwealth. (Source: www.usatoday.com).

1950    434    2300    3575    3376    3800    1655    418

Compute the range, variance, and standard deviation for these data.

**3.55** The following data give the hourly wage rates of eight employees of a company.

\$22    22    22    22    22    22    22    22

Calculate the standard deviation. Is its value zero? If yes, why?

**3.56** The following data are the ages (in years) of six students.

19    19    19    19    19    19

Calculate the standard deviation. Is its value zero? If yes, why?

**\*3.57** One disadvantage of the standard deviation as a measure of dispersion is that it is a measure of absolute variability and not of relative variability. Sometimes we may need to compare the variability of two different data sets that have different units of measurement. The **coefficient of variation** is one such measure. The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean and is computed as follows:

$$\text{For population data: } CV = \frac{\sigma}{\mu} \times 100\%$$

$$\text{For sample data: } CV = \frac{s}{\bar{x}} \times 100\%$$

The yearly salaries of all employees who work for a company have a mean of \$62,350 and a standard deviation of \$6820. The years of experience for the same employees have a mean of 15 years and a standard deviation of 2 years. Is the relative variation in the salaries larger or smaller than that in years of experience for these employees?

**\*3.58** The SAT scores of 100 students have a mean of 975 and a standard deviation of 105. The GPAs of the same 100 students have a mean of 3.16 and a standard deviation of .22. Is the relative variation in SAT scores larger or smaller than that in GPAs?

**\*3.59** Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Note that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the standard deviation for each of these two data sets using the formula for sample data. Comment on the relationship between the two standard deviations.

**\*3.60** Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	8	16	30	18	22

Note that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the standard deviation for each of these two data sets using the formula for population data. Comment on the relationship between the two standard deviations.

### 3.3 Mean, Variance, and Standard Deviation for Grouped Data

In Sections 3.1.1 and 3.2.2, we learned how to calculate the mean, variance, and standard deviation for ungrouped data. In this section, we will learn how to calculate the mean, variance, and standard deviation for grouped data.

### 3.3.1 Mean for Grouped Data

We learned in Section 3.1.1 that the mean is obtained by dividing the sum of all values by the number of values in a data set. However, if the data are given in the form of a frequency table, we no longer know the values of individual observations. Consequently, in such cases, we cannot obtain the sum of individual values. We find an approximation for the sum of these values using the procedure explained in the next paragraph and example. The formulas used to calculate the mean for grouped data follow.

#### Calculating Mean for Grouped Data

$$\text{Mean for population data: } \mu = \frac{\sum mf}{N}$$

$$\text{Mean for sample data: } \bar{x} = \frac{\sum mf}{n}$$

where  $m$  is the midpoint and  $f$  is the frequency of a class.

To calculate the mean for grouped data, first find the midpoint of each class and then multiply the midpoints by the frequencies of the corresponding classes. The sum of these products, denoted by  $\sum mf$ , gives an approximation for the sum of all values. To find the value of the mean, divide this sum by the total number of observations in the data.

#### ■ EXAMPLE 3-14

Table 3.8 gives the frequency distribution of the daily commuting times (in minutes) from home to work for *all* 25 employees of a company.

*Calculating the population mean for grouped data.*

Table 3.8

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of the daily commuting times.

**Solution** Note that because the data set includes *all* 25 employees of the company, it represents the population. Table 3.9 shows the calculation of  $\sum mf$ . Note that in Table 3.9,  $m$  denotes the midpoints of the classes.

Table 3.9

Daily Commuting Time (minutes)	$f$	$m$	$mf$
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
$N = 25$		$\sum mf = 535$	

To calculate the mean, we first find the midpoint of each class. The class midpoints are recorded in the third column of Table 3.9. The products of the midpoints and the corresponding frequencies are listed in the fourth column. The sum of the fourth column values, denoted by  $\Sigma mf$ , gives the approximate total daily commuting time (in minutes) for all 25 employees. The mean is obtained by dividing this sum by the total frequency. Therefore,

$$\mu = \frac{\Sigma mf}{N} = \frac{535}{25} = 21.40 \text{ minutes}$$

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work. ■

What do the numbers 20, 135, 150, 140, and 90 in the column labeled *mf* in Table 3.9 represent? We know from this table that 4 employees spend 0 to less than 10 minutes commuting per day. If we assume that the time spent commuting by these 4 employees is evenly spread in the interval 0 to less than 10, then the midpoint of this class (which is 5) gives the mean time spent commuting by these 4 employees. Hence,  $4 \times 5 = 20$  is the approximate total time (in minutes) spent commuting per day by these 4 employees. Similarly, 9 employees spend 10 to less than 20 minutes commuting per day, and the total time spent commuting by these 9 employees is approximately 135 minutes a day. The other numbers in this column can be interpreted in the same way. Note that these numbers give the approximate commuting times for these employees based on the assumption of an even spread within classes. The total commuting time for all 25 employees is approximately 535 minutes. Consequently, 21.40 minutes is an approximate and not the exact value of the mean. We can find the exact value of the mean only if we know the exact commuting time for each of the 25 employees of the company.

### ■ EXAMPLE 3-15

*Calculating the sample mean for grouped data.*

Table 3.10 gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

**Table 3.10**

Number of Orders	Number of Days
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the mean.

**Solution** Because the data set includes only 50 days, it represents a sample. The value of  $\Sigma mf$  is calculated in Table 3.11.

**Table 3.11**

Number of Orders	f	m	mf
10–12	4	11	44
13–15	12	14	168
16–18	20	17	340
19–21	14	20	280
	$n = 50$		$\Sigma mf = 832$

The value of the sample mean is

$$\bar{x} = \frac{\Sigma mf}{n} = \frac{832}{50} = 16.64 \text{ orders}$$

Thus, this mail-order company received an average of 16.64 orders per day during these 50 days.

### 3.3.2 Variance and Standard Deviation for Grouped Data

Following are what we will call the *basic formulas* that are used to calculate the population and sample variances for grouped data:

$$\sigma^2 = \frac{\sum f(m - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\sum f(m - \bar{x})^2}{n - 1}$$

where  $\sigma^2$  is the population variance,  $s^2$  is the sample variance, and  $m$  is the midpoint of a class.

In either case, the standard deviation is obtained by taking the positive square root of the variance.

Again, the *short-cut formulas* are more efficient for calculating the variance and standard deviation. Section A3.1.2 of Appendix 3.1 at the end of this chapter shows how to use the basic formulas to calculate the variance and standard deviation for grouped data.

#### Short-Cut Formulas for the Variance and Standard Deviation for Grouped Data

$$\sigma^2 = \frac{\sum m^2f - \frac{(\sum mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum m^2f - \frac{(\sum mf)^2}{n}}{n - 1}$$

where  $\sigma^2$  is the population variance,  $s^2$  is the sample variance, and  $m$  is the midpoint of a class.

The standard deviation is obtained by taking the positive square root of the variance.

$$\text{Population standard deviation: } \sigma = \sqrt{\frac{\sum m^2f - \frac{(\sum mf)^2}{N}}{N}}$$

$$\text{Sample standard deviation: } s = \sqrt{\frac{\sum m^2f - \frac{(\sum mf)^2}{n}}{n - 1}}$$

Examples 3–16 and 3–17 illustrate the use of these formulas to calculate the variance and standard deviation.

#### EXAMPLE 3–16

The following data, reproduced from Table 3.8 of Example 3–14, give the frequency distribution of the daily commuting times (in minutes) from home to work for all 25 employees of a company.

*Calculating the population variance and standard deviation for grouped data.*

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the variance and standard deviation.

**Solution** All four steps needed to calculate the variance and standard deviation for grouped data are shown after Table 3.12.

**Table 3.12**

Daily Commuting Time (minutes)	<i>f</i>	<i>m</i>	<i>mf</i>	<i>m</i> <sup>2</sup> <i>f</i>
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	$N = 25$		$\Sigma mf = 535$	$\Sigma m^2f = 14,825$

**Step 1.** Calculate the value of  $\Sigma mf$ .

To calculate the value of  $\Sigma mf$ , first find the midpoint *m* of each class (see the third column in Table 3.12) and then multiply the corresponding class midpoints and class frequencies (see the fourth column). The value of  $\Sigma mf$  is obtained by adding these products. Thus,

$$\Sigma mf = 535$$

**Step 2.** Find the value of  $\Sigma m^2f$ .

To find the value of  $\Sigma m^2f$ , square each *m* value and multiply this squared value of *m* by the corresponding frequency (see the fifth column in Table 3.12). The sum of these products (that is, the sum of the fifth column) gives  $\Sigma m^2f$ . Hence,

$$\Sigma m^2f = 14,825$$

**Step 3.** Calculate the variance.

Because the data set includes all 25 employees of the company, it represents the population. Therefore, we use the formula for the population variance:

$$\sigma^2 = \frac{\Sigma m^2f - \frac{(\Sigma mf)^2}{N}}{N} = \frac{14,825 - \frac{(535)^2}{25}}{25} = \frac{3376}{25} = 135.04$$

**Step 4.** Calculate the standard deviation.

To obtain the standard deviation, take the (positive) square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{135.04} = 11.62 \text{ minutes}$$

Thus, the standard deviation of the daily commuting times for these employees is 11.62 minutes. ■

Note that the values of the variance and standard deviation calculated in Example 3–16 for grouped data are approximations. The exact values of the variance and standard deviation can be obtained only by using the ungrouped data on the daily commuting times of these 25 employees.

### ■ EXAMPLE 3–17

*Calculating the sample variance and standard deviation for grouped data.*

The following data, reproduced from Table 3.10 of Example 3–15, give the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Number of Orders	<i>f</i>
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the variance and standard deviation.

**Solution** All the information required for the calculation of the variance and standard deviation appears in Table 3.13.

**Table 3.13**

Number of Orders	<i>f</i>	<i>m</i>	<i>mf</i>	<i>m</i> <sup>2</sup> <i>f</i>
10–12	4	11	44	484
13–15	12	14	168	2352
16–18	20	17	340	5780
19–21	14	20	280	5600
	$n = 50$		$\Sigma mf = 832$	$\Sigma m^2f = 14,216$

Because the data set includes only 50 days, it represents a sample. Hence, we use the sample formulas to calculate the variance and standard deviation. By substituting the values into the formula for the sample variance, we obtain

$$s^2 = \frac{\sum m^2f - \frac{(\sum mf)^2}{n}}{n-1} = \frac{14,216 - \frac{(832)^2}{50}}{50-1} = 7.5820$$

Hence, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{7.5820} = 2.75 \text{ orders}$$

Thus, the standard deviation of the number of orders received at the office of this mail-order company during the past 50 days is 2.75. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**3.61** Are the values of the mean and standard deviation that are calculated using grouped data exact or approximate values of the mean and standard deviation, respectively? Explain.

**3.62** Using the population formulas, calculate the mean, variance, and standard deviation for the following grouped data.

<i>x</i>	2–4	5–7	8–10	11–13	14–16
<i>f</i>	5	9	14	7	5

**3.63** Using the sample formulas, find the mean, variance, and standard deviation for the grouped data displayed in the following table.

<i>x</i>	<i>f</i>
0 to less than 4	17
4 to less than 8	23
8 to less than 12	15
12 to less than 16	11
16 to less than 20	8
20 to less than 24	6

### ■ APPLICATIONS

**3.64** The following table gives the frequency distribution of the amounts of telephone bills for August 2012 for a sample of 50 families.

Amount of Telephone Bill (dollars)	Number of Families
40 to less than 70	9
70 to less than 100	11
100 to less than 130	16
130 to less than 160	10
160 to less than 190	4

Calculate the mean, variance, and standard deviation.

- 3.65** The following table gives the frequency distribution of the number of hours spent last week on cell phones (making phone calls and texting) by all 100 students of the tenth grade at a school.

Hours per Week	Number of Students
0 to less than 4	14
4 to less than 8	18
8 to less than 12	25
12 to less than 16	18
16 to less than 20	16
20 to less than 24	9

Find the mean, variance, and standard deviation.

- 3.66** The following table gives the grouped data on the ounces of milk dispensed by a machine into 1-gallon jugs for a sample of 250 jugs of milk selected from a day's production. Note that 1 gallon is equal to 128 ounces.

Ounces of Milk	Number of Jugs
121 to less than 123	5
123 to less than 125	13
125 to less than 127	42
127 to less than 129	129
129 to less than 131	61

Find the mean, variance, and standard deviation.

- 3.67** The following table gives the frequency distribution of the total miles driven during 2012 by 300 car owners.

Miles Driven in 2012 (in thousands)	Number of Car Owners
0 to less than 5	7
5 to less than 10	26
10 to less than 15	59
15 to less than 20	71
20 to less than 25	62
25 to less than 30	39
30 to less than 35	22
35 to less than 40	14

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled  $mf$  in your table of calculations. What does  $\Sigma mf$  represent?

- 3.68** The following table gives information on the amounts (in dollars) of electric bills for August 2012 for a sample of 50 families.

Amount of Electric Bill (dollars)	Number of Families
0 to less than 60	5
60 to less than 120	16
120 to less than 180	11
180 to less than 240	10
240 to less than 300	8

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled  $mf$  in your table of calculations. What does  $\Sigma mf$  represent?

- 3.69** For 50 airplanes that arrived late at an airport during a week, the time by which they were late was observed. In the following table,  $x$  denotes the time (in minutes) by which an airplane was late, and  $f$  denotes the number of airplanes.

$x$	$f$
0 to less than 20	14
20 to less than 40	18
40 to less than 60	9
60 to less than 80	5
80 to less than 100	4

Find the mean, variance, and standard deviation.

- 3.70** The following table gives the frequency distribution of the number of errors committed by a college baseball team in all of the 45 games that it played during the 2011–12 season.

Number of Errors	Number of Games
0	11
1	14
2	9
3	7
4	3
5	1

Find the mean, variance, and standard deviation. (*Hint:* The classes in this example are single valued. These values of classes will be used as values of  $m$  in the formulas for the mean, variance, and standard deviation.)

- 3.71** The following table gives the frequency distribution of the number of hours spent per week on activities that involve sports and/or exercise by a sample of 400 Americans. The numbers are consistent with the summary results from the Bureau of Labor Statistics' American Time Use Survey ([www.bls.gov/tus](http://www.bls.gov/tus)).

Hours per Week	Number of People
0 to less than 3.5	34
3.5 to less than 7.0	92
7.0 to less than 10.5	55
10.5 to less than 14.0	83
14.0 to less than 28.0	121
28.0 to less than 56.0	15

Find the mean, variance, and standard deviation.

## 3.4 Use of Standard Deviation

By using the mean and standard deviation, we can find the proportion or percentage of the total observations that fall within a given interval about the mean. This section briefly discusses Chebyshev's theorem and the empirical rule, both of which demonstrate this use of the standard deviation.

### 3.4.1 Chebyshev's Theorem

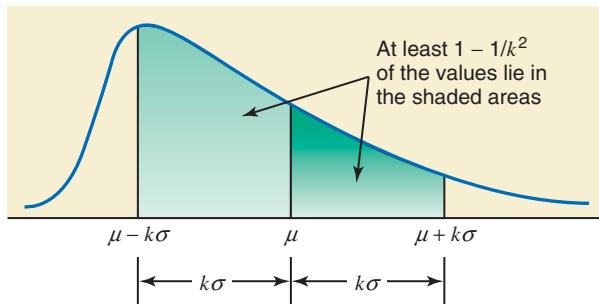
**Chebyshev's theorem** gives a lower bound for the area under a curve between two points that are on opposite sides of the mean and at the same distance from the mean.

#### Definition

**Chebyshev's Theorem** For any number  $k$  greater than 1, at least  $(1 - 1/k^2)$  of the data values lie within  $k$  standard deviations of the mean.

Figure 3.5 illustrates Chebyshev's theorem.

**Figure 3.5** Chebyshev's theorem.

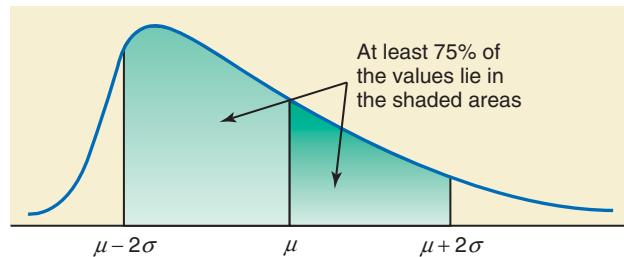


Thus, for example, if  $k = 2$ , then

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } 75\%$$

Therefore, according to Chebyshev's theorem, at least .75, or 75%, of the values of a data set lie within two standard deviations of the mean. This is shown in Figure 3.6.

**Figure 3.6** Percentage of values within two standard deviations of the mean for Chebyshev's theorem.

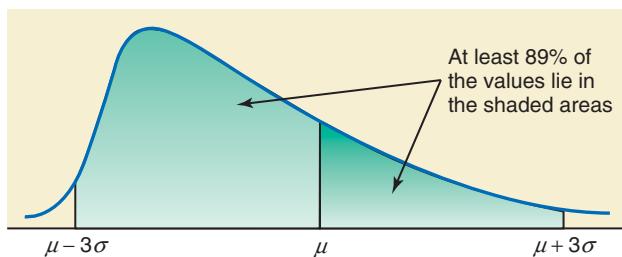


If  $k = 3$ , then,

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = 1 - .11 = .89 \text{ or } 89\% \text{ approximately}$$

According to Chebyshev's theorem, at least .89, or 89%, of the values fall within three standard deviations of the mean. This is shown in Figure 3.7.

**Figure 3.7** Percentage of values within three standard deviations of the mean for Chebyshev's theorem.



Although in Figures 3.5 through 3.7 we have used the population notation for the mean and standard deviation, the theorem applies to both sample and population data. Note that Chebyshev's theorem is applicable to a distribution of any shape. However, Chebyshev's theorem can be used only for  $k > 1$ . This is so because when  $k = 1$ , the value of  $1 - 1/k^2$  is zero, and when  $k < 1$ , the value of  $1 - 1/k^2$  is negative.

### ■ EXAMPLE 3-18

The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 mm Hg with a standard deviation of 22. Using Chebyshev's theorem, find at least what percentage of women in this group have a systolic blood pressure between 143 and 231 mm Hg.

*Applying Chebyshev's theorem.*

**Solution** Let  $\mu$  and  $\sigma$  be the mean and the standard deviation, respectively, of the systolic blood pressures of these women. Then, from the given information,

$$\mu = 187 \quad \text{and} \quad \sigma = 22$$

To find the percentage of women whose systolic blood pressures are between 143 and 231 mm Hg, the first step is to determine  $k$ . As shown below, each of the two points, 143 and 231, is 44 units away from the mean.

$$\begin{array}{ccc} |\leftarrow 143 - 187 = -44 \rightarrow| & |\leftarrow 231 - 187 = 44 \rightarrow| \\ 143 & \mu = 187 & 231 \end{array}$$

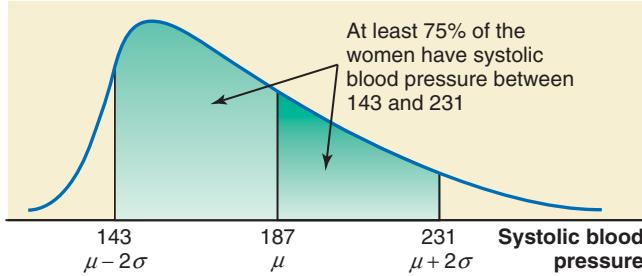
The value of  $k$  is obtained by dividing the distance between the mean and each point by the standard deviation. Thus,

$$k = 44/22 = 2$$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } 75\%$$



PhotoDisc, Inc./Getty Images



**Figure 3.8** Percentage of women with systolic blood pressure between 143 and 231.

Hence, according to Chebyshev's theorem, at least 75% of the women have systolic blood pressure between 143 and 231 mm Hg. This percentage is shown in Figure 3.8. ■

### 3.4.2 Empirical Rule

Whereas Chebyshev's theorem is applicable to any kind of distribution, the **empirical rule** applies only to a specific type of distribution called a *bell-shaped distribution*, as shown in Figure 3.9. More will be said about such a distribution in Chapter 6, where it is called a *normal curve*. In this section, only the following three rules for the curve are given.

## DOES SPREAD MEAN THE SAME AS VARIABILITY AND DISPERSION?

In any discipline, there is terminology that one needs to learn in order to become fluent. Accounting majors need to learn the difference between credits and debits, chemists need to know how an ion differs from an atom, and physical therapists need to know the difference between abduction and adduction. Statistics is no different. Failing to learn the difference between the mean and the median makes much of the remainder of this book very difficult to understand.

Another issue with terminology is the use of words other than the terminology to describe a specific concept or scenario. Sometimes the words one chooses to use can be vague or ambiguous, resulting in confusion. One debate in the statistics community involves the use of the word "spread" in place of the words "dispersion" and "variability." In a 2012 article, "Lexical ambiguity: making a case against spread," authors Jennifer Kaplan, Neal Rogness, and Diane Fisher point out that the Oxford English Dictionary has more than 25 definitions for the word spread, many of which students know coming into a statistics class. As a result of knowing some of the meanings of spread, students who use the word spread in place of variability or dispersion "do not demonstrate strong statistical meanings of the word spread at the end of a one-semester statistics course."

In order to examine the extent of this issue, the authors of the article designed a study in which they selected 160 undergraduate students taking an introductory statistics course from 14 different professors at three different universities and in the first week of the semester asked them to write sentences and definitions for spread using its primary meaning. Then, at the end of the semester, the same students were asked to write sentences and definitions for spread using its primary meaning in statistics. The authors found that responses of only one-third of the students related spread to the concept of variability, which has to do with how the data vary around the center of a distribution. A slightly larger percentage of students gave responses that "defined spread as 'to cover evenly or in a thin layer,'" while approximately one in eight responded with a definition that was synonymous with the notion of range. Seven other definitions were given by at least three students in the study.

Although more of the definitions and sentences provided at the end of the course had something to do with statistics, the authors did not see an increase in the percentage of definitions that associated spread with the concept of variability. Hence, they suggested that the ambiguity of the term spread is sufficient enough to stop using it in place of the terms variability and dispersion.

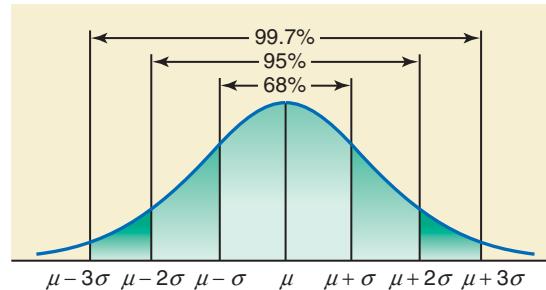
*Source:* Kaplan, J. J., Rogness, N. T., and Fisher, D. G., "Lexical ambiguity: making a case against spread," *Teaching Statistics*, 2011, 34, (2), pp. 56–60. © 2011 Teaching Statistics Trust.

**Empirical Rule** For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean.

Figure 3.9 illustrates the empirical rule. Again, the empirical rule applies to both population data and sample data.

**Figure 3.9** Illustration of the empirical rule.



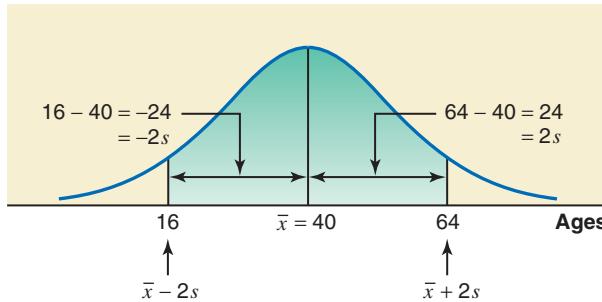
### ■ EXAMPLE 3-19

The age distribution of a sample of 5000 persons is bell shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

*Applying the empirical rule.*

**Solution** We use the empirical rule to find the required percentage because the distribution of ages follows a bell-shaped curve. From the given information, for this distribution,

$$\bar{x} = 40 \text{ years} \quad \text{and} \quad s = 12 \text{ years}$$



**Figure 3.10** Percentage of people who are 16 to 64 years old.

Each of the two points, 16 and 64, is 24 units away from the mean. Dividing 24 by 12, we convert the distance between each of the two points and the mean in terms of standard deviations. Thus, the distance between 16 and 40 and that between 40 and 64 is each equal to  $2s$ . Consequently, as shown in Figure 3.10, the area from 16 to 64 is the area from  $\bar{x} - 2s$  to  $\bar{x} + 2s$ .

Because the area within two standard deviations of the mean is approximately 95% for a bell-shaped curve, approximately **95%** of the people in the sample are 16 to 64 years old. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 3.72** Briefly explain Chebyshev's theorem and its applications.
- 3.73** Briefly explain the empirical rule. To what kind of distribution is it applied?
- 3.74** A sample of 2000 observations has a mean of 74 and a standard deviation of 12. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals  $\bar{x} \pm 2s$ ,  $\bar{x} \pm 2.5s$ , and  $\bar{x} \pm 3s$ . Note that here  $\bar{x} \pm 2s$  represents the interval  $\bar{x} - 2s$  to  $\bar{x} + 2s$ , and so on.
- 3.75** A large population has a mean of 230 and a standard deviation of 41. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals  $\mu \pm 2\sigma$ ,  $\mu \pm 2.5\sigma$ , and  $\mu \pm 3\sigma$ .
- 3.76** A large population has a bell-shaped distribution with a mean of 310 and a standard deviation of 37. Using the empirical rule, find what percentage of the observations fall in the intervals  $\mu \pm 1\sigma$ ,  $\mu \pm 2\sigma$ , and  $\mu \pm 3\sigma$ .
- 3.77** A sample of 3000 observations has a bell-shaped distribution with a mean of 82 and a standard deviation of 16. Using the empirical rule, find what percentage of the observations fall in the intervals  $\bar{x} \pm 1s$ ,  $\bar{x} \pm 2s$ , and  $\bar{x} \pm 3s$ .

### ■ APPLICATIONS

- 3.78** The mean time taken by all participants to run a road race was found to be 220 minutes with a standard deviation of 20 minutes. Using Chebyshev's theorem, find at least what percentage of runners who ran this road race completed it in
- a. 180 to 260 minutes      b. 160 to 280 minutes      c. 170 to 270 minutes
- 3.79** The 2011 gross sales of all companies in a large city have a mean of \$2.3 million and a standard deviation of \$.6 million. Using Chebyshev's theorem, find at least what percentage of companies in this city had 2011 gross sales of
- a. \$1.1 to \$3.5 million      b. \$.8 to \$3.8 million      c. \$.5 to \$4.1 million

**3.80** According to the National Center for Education Statistics ([www.nces.ed.gov](http://www.nces.ed.gov)), the amounts of all loans, including Federal Parent PLUS loans, granted to students during the 2007–2008 academic year had a distribution with a mean of \$8109.65. Suppose that the standard deviation of this distribution is \$2412.

- a. Using Chebyshev's theorem, find at least what percentage of students had 2007–2008 such loans between

i. \$2079.65 and \$14,139.65      ii. \$3285.65 and \$12,933.65

- \*b. Using Chebyshev's theorem, find the interval that contains the amounts of 2007–2008 such loans for at least 89% of all students.

**3.81** The mean monthly mortgage paid by all home owners in a town is \$2365 with a standard deviation of \$340.

- a. Using Chebyshev's theorem, find at least what percentage of all home owners in this town pay a monthly mortgage of

i. \$1685 to \$3045      ii. \$1345 to \$3385

- \*b. Using Chebyshev's theorem, find the interval that contains the monthly mortgage payments of at least 84% of all home owners in this town.

**3.82** The mean life of a certain brand of auto batteries is 44 months with a standard deviation of 3 months. Assume that the lives of all auto batteries of this brand have a bell-shaped distribution. Using the empirical rule, find the percentage of auto batteries of this brand that have a life of

- a. 41 to 47 months      b. 38 to 50 months      c. 35 to 53 months

**3.83** According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of \$4129 for family coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose that the premiums for such family coverage paid this year by all such workers have a bell-shaped distribution with a mean of \$4129 and a standard deviation of \$600. Using the empirical rule, find the approximate percentage of such workers who pay premiums for such family coverage between

- a. \$2329 and \$5929      b. \$3529 and \$4729      c. \$2929 and \$5329

**3.84** The prices of all college textbooks follow a bell-shaped distribution with a mean of \$180 and a standard deviation of \$30.

- a. Using the empirical rule, find the percentage of all college textbooks with their prices between

i. \$150 and \$210      ii. \$120 and \$240

- \*b. Using the empirical rule, find the interval that contains the prices of 99.7% of college textbooks.

**3.85** Suppose that on a certain section of I-95 with a posted speed limit of 65 mph, the speeds of all vehicles have a bell-shaped distribution with a mean of 72 mph and a standard deviation of 3 mph.

- a. Using the empirical rule, find the percentage of vehicles with the following speeds on this section of I-95.

i. 63 to 81 mph      ii. 69 to 75 mph

- \*b. Using the empirical rule, find the interval that contains the speeds of 95% of vehicles traveling on this section of I-95.

## 3.5 Measures of Position

A **measure of position** determines the position of a single value in relation to other values in a sample or a population data set. There are many measures of position; however, only quartiles, percentiles, and percentile rank are discussed in this section.

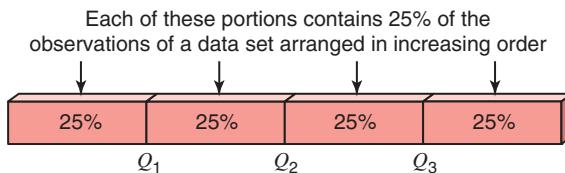
### 3.5.1 Quartiles and Interquartile Range

**Quartiles** are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts. These three measures are the **first quartile** (denoted by  $Q_1$ ), the **second quartile** (denoted by  $Q_2$ ), and the **third quartile** (denoted by  $Q_3$ ). The data should be ranked in increasing order before the quartiles are determined. The quartiles are defined as follows. Note that  $Q_1$  and  $Q_3$  are also called the lower and the upper quartiles, respectively.

#### Definition

**Quartiles** *Quartiles* are three summary measures that divide a ranked data set into four equal parts. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median.

Figure 3.11 describes the positions of the three quartiles.



**Figure 3.11** Quartiles.

Approximately 25% of the values in a ranked data set are less than  $Q_1$  and about 75% are greater than  $Q_1$ . The second quartile,  $Q_2$ , divides a ranked data set into two equal parts; hence, the second quartile and the median are the same. Approximately 75% of the data values are less than  $Q_3$  and about 25% are greater than  $Q_3$ .

The difference between the third quartile and the first quartile for a data set is called the **interquartile range (IQR)**, which is a measure of dispersion.

**Calculating Interquartile Range** The difference between the third and the first quartiles gives the *interquartile range*; that is,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1$$

Examples 3–20 and 3–21 show the calculation of the quartiles and the interquartile range.

### ■ EXAMPLE 3–20

Table 3.3 in Example 3–5 gave the total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies. That table is reproduced here:

Finding quartiles and the interquartile range.

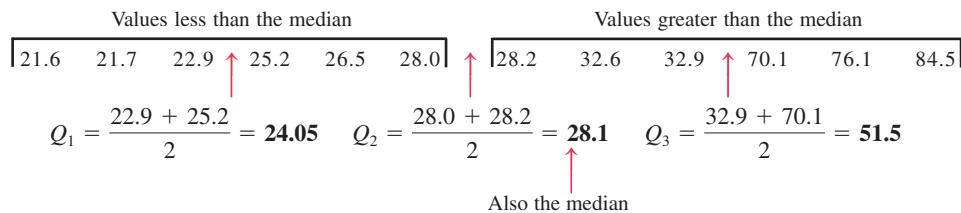
CEO and Company	2010 Total Compensation (millions of dollars)
Michael D. White (DirecTV)	32.9
David N. Farr (Emerson Electric)	22.9
Brian L. Roberts (Comcast)	28.2
Philippe P. Dauman (Viacom)	84.5
William C. Weldon (Johnson & Johnson)	21.6
Robert A. Iger (Walt Disney)	28.0
Ray R. Iran (Occidental Petroleum)	76.1
Samuel J. Palmisano (IBM)	25.2
John F. Lundgren (Stanley Black & Decker)	32.6
Lawrence J. Ellison (Oracle)	70.1
Alan Mulally (Ford Motor)	26.5
Howard Schultz (Starbucks)	21.7

- (a) Find the values of the three quartiles. Where does the total compensation of Michael D. White (CEO of DirecTV) fall in relation to these quartiles?
- (b) Find the interquartile range.

#### Solution

- (a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5



Finding quartiles for an even number of data values.

The value of  $Q_2$ , which is also the median, is given by the value of the middle term in the ranked data set. For the data of this example, this value is the average of the sixth and seventh terms. Consequently,  $Q_2$  is \$28.1 million. The value of  $Q_1$  is given by the value of the middle term of the six values that fall below the median (or  $Q_2$ ). Thus, it is obtained by taking the average of the third and fourth terms. So,  $Q_1$  is \$24.05 million. The value of  $Q_3$  is given by the value of the middle term of the six values that fall above the median. For the data of this example,  $Q_3$  is obtained by taking the average of the ninth and tenth terms, and it is \$51.5 million.

The value of  $Q_1 = \$24.05$  million indicates that 25% of the CEOs in this sample had 2010 total compensations less than \$24.05 million and 75% of them had 2010 total compensations higher than \$24.05 million. Similarly, we can state that half of these CEOs had 2010 total compensations less than \$28.1 million and the other half had higher than \$28.1 million, since the second quartile is \$28.1 million. The value of  $Q_3 = \$51.5$  million indicates that 75% of these CEOs had 2010 total compensations less than \$51.5 million and 25% had higher than this value.

By looking at the position of \$32.9 million (total compensation of Michael D. White, CEO of DirecTV), we can state that this value lies in the **bottom 75%** of these 2010 total compensations and is just below  $Q_3$ . This value falls between the second and third quartiles.

- (b) The interquartile range is given by the difference between the values of the third and first quartiles. Thus,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 51.5 - 24.05 = \$27.45 \text{ million}$$

## ■ EXAMPLE 3–21

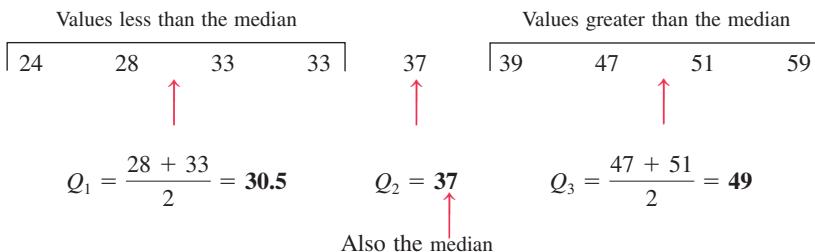
The following are the ages (in years) of nine employees of an insurance company:

47      28      39      51      33      37      59      24      33

- (a) Find the values of the three quartiles. Where does the age of 28 years fall in relation to the ages of these employees?  
 (b) Find the interquartile range.

### Solution

- (a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:



Thus the values of the three quartiles are

$$Q_1 = 30.5 \text{ years}, \quad Q_2 = 37 \text{ years}, \quad \text{and} \quad Q_3 = 49 \text{ years}$$

The age of 28 falls in the **lowest 25%** of the ages.

- (b) The interquartile range is

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 49 - 30.5 = 18.5 \text{ years}$$

### 3.5.2 Percentiles and Percentile Rank

**Percentiles** are the summary measures that divide a ranked data set into 100 equal parts. Each (ranked) data set has 99 percentiles that divide it into 100 equal parts. The data should be ranked in increasing order to compute percentiles. The  $k$ th percentile is denoted by  $P_k$ , where  $k$  is an integer in the range 1 to 99. For instance, the 25th percentile is denoted by  $P_{25}$ . Figure 3.12 shows the positions of the 99 percentiles.

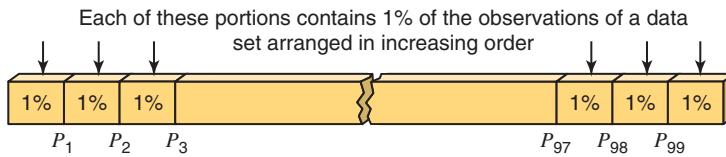


Figure 3.12 Percentiles.

Thus, the  $k$ th percentile,  $P_k$ , can be defined as a value in a data set such that about  $k\%$  of the measurements are smaller than the value of  $P_k$  and about  $(100 - k)\%$  of the measurements are greater than the value of  $P_k$ .

The approximate value of the  $k$ th percentile is determined as explained next.

**Calculating Percentiles** The (approximate) value of the  $k$ th percentile, denoted by  $P_k$ , is

$$P_k = \text{Value of the } \left( \frac{kn}{100} \right) \text{th term in a ranked data set}$$

where  $k$  denotes the number of the percentile and  $n$  represents the sample size.

Example 3–22 describes the procedure to calculate the percentiles. For convenience, we will round  $kn/100$  to the nearest whole number to find the value of  $P_k$ .

#### ■ EXAMPLE 3–22

Refer to the data on total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies given in Exercise 3–20. Find the value of the 60th percentile. Give a brief interpretation of the 60th percentile.

Finding the percentile for a data set.

**Solution** From Example 3–20, the data arranged in increasing order are as follows:

21.6    21.7    22.9    25.2    26.5    28.0    28.2    32.6    32.9    70.1    76.1    84.5

The position of the 60th percentile is

$$\frac{kn}{100} = \frac{60(12)}{100} = 7.20 \text{th term} \approx 7 \text{th term}$$

The value of the 7.20th term can be approximated by the value of the 7th term in the ranked data. Therefore,

$$P_{60} = 60 \text{th percentile} = 28.2 = \$28.2 \text{ million}$$

Thus, approximately 60% of these 12 CEOs had 2010 total compensations less than \$28.2 million. ■

We can also calculate the **percentile rank** for a particular value  $x_i$  of a data set by using the formula given below. The percentile rank of  $x_i$  gives the percentage of values in the data set that are less than  $x_i$ .

#### Finding Percentile Rank of a Value

$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100\%$$

Example 3–23 shows how the percentile rank is calculated for a data value.

*Finding the percentile rank  
for a data value.*

### ■ EXAMPLE 3–23

Refer to the data on total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies given in Exercise 3–20. Find the percentile rank for \$26.5 million (2010 total compensation of Alan Mulally, CEO of Ford Motor). Give a brief interpretation of this percentile rank.

**Solution** From Example 3–20, the data arranged in increasing order are as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

In this data set, 4 of the 12 values are less than \$26.5 million. Hence,

$$\text{Percentile rank of } 26.5 = \frac{4}{12} \times 100 = 33.33\%$$

Rounding this answer to the nearest integral value, we can state that about 33% of these 12 CEOs had 2010 total compensations less than \$26.5 million. Hence, 67% of these 12 CEOs had \$26.5 million or higher total compensations in 2010. ■

Most statistical software packages use slightly different methods to calculate quartiles and percentiles. Those methods, while more precise, are beyond the scope of this text.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**3.86** Briefly describe how the three quartiles are calculated for a data set. Illustrate by calculating the three quartiles for two examples, the first with an odd number of observations and the second with an even number of observations.

**3.87** Explain how the interquartile range is calculated. Give one example.

**3.88** Briefly describe how the percentiles are calculated for a data set.

**3.89** Explain the concept of the percentile rank for an observation of a data set.

### ■ APPLICATIONS

**3.90** The following data give the weights (in pounds) lost by 15 members of a health club at the end of 2 months after joining the club.

5	10	8	7	25	12	5	14
11	10	21	9	8	11	18	

a. Compute the values of the three quartiles and the interquartile range.

b. Calculate the (approximate) value of the 82nd percentile.

c. Find the percentile rank of 10.

**3.91** The following data give the speeds of 13 cars (in mph) measured by radar, traveling on I-84.

73	75	69	68	78	69	74
76	72	79	68	77	71	

a. Find the values of the three quartiles and the interquartile range.

b. Calculate the (approximate) value of the 35th percentile.

c. Compute the percentile rank of 71.

**3.92** The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53
51	53	51	48	46	43	52	50	54	47
44	47	50	49	52					

a. Calculate the values of the three quartiles and the interquartile range.

b. Determine the (approximate) value of the 53rd percentile.

c. Find the percentile rank of 50.

**3.93** The following data give the numbers of minor penalties accrued by each of the 30 National Hockey League franchises during the 2010–11 regular season.

249	265	269	287	287	292	299	300	300	301
302	304	311	312	320	325	330	331	335	337
344	347	347	348	352	353	354	355	363	374

- Calculate the values of the three quartiles and the interquartile range.
- Find the approximate value of the 57th percentile.
- Calculate the percentile rank of 311.

**3.94** The following data give the numbers of text messages sent by a high school student on 40 randomly selected days during 2012:

32	33	33	34	35	36	37	37	37	37
38	39	40	41	41	42	42	42	43	44
44	45	45	45	47	47	47	47	47	48
48	49	50	50	51	52	53	54	59	61

- Calculate the values of the three quartiles and the interquartile range. Where does the value 49 fall in relation to these quartiles?
- Determine the approximate value of the 91st percentile. Give a brief interpretation of this percentile.
- For what percentage of the days was the number of text messages sent 40 or higher? Answer by finding the percentile rank of 40.

**3.95** Nixon Corporation manufactures computer monitors. The following data give the numbers of computer monitors produced at the company for a sample of 30 days.

24	32	27	23	33	33	29	25	23	36
26	26	31	20	27	33	27	23	28	29
31	35	34	22	37	28	23	35	31	43

- Calculate the values of the three quartiles and the interquartile range. Where does the value of 31 lie in relation to these quartiles?
- Find the (approximate) value of the 65th percentile. Give a brief interpretation of this percentile.
- For what percentage of the days was the number of computer monitors produced 32 or higher? Answer by finding the percentile rank of 32.

**3.96** The following data give the numbers of new cars sold at a dealership during a 20-day period.

8	5	12	3	9	10	6	12	8	8
4	16	10	11	7	7	3	5	9	11

- Calculate the values of the three quartiles and the interquartile range. Where does the value of 4 lie in relation to these quartiles?
- Find the (approximate) value of the 25th percentile. Give a brief interpretation of this percentile.
- Find the percentile rank of 10. Give a brief interpretation of this percentile rank.

**3.97** According to [www.money-zine.com](http://www.money-zine.com), the average FICO score in the United States was around 692 in December 2011. Suppose the following data represent the credit scores of 22 randomly selected loan applicants.

494	728	468	533	747	639	430	690	604	422	356
805	749	600	797	702	628	625	617	647	772	572

- Calculate the values of the three quartiles and the interquartile range. Where does the value 617 fall in relation to these quartiles?
- Find the approximate value of the 30th percentile. Give a brief interpretation of this percentile.
- Calculate the percentile rank of 533. Give a brief interpretation of this percentile rank.

## 3.6 Box-and-Whisker Plot

A **box-and-whisker plot** gives a graphic presentation of data using five measures: the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences. (The inner fences are explained in Example 3–24.) A box-and-whisker plot can help us visualize the center, the spread, and the skewness of a data set. It also helps detect outliers. We can compare different distributions by making box-and-whisker plots for each of them.

**Definition**

**Box-and-Whisker Plot** A plot that shows the center, spread, and skewness of a data set. It is constructed by drawing a box and two whiskers that use the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Example 3–24 explains all the steps needed to make a box-and-whisker plot.

### ■ EXAMPLE 3–24

*Constructing a box-and-whisker plot.*

The following data are the incomes (in thousands of dollars) for a sample of 12 households.

75    69    84    112    74    104    81    90    94    144    79    98

Construct a box-and-whisker plot for these data.

**Solution** The following five steps are performed to construct a box-and-whisker plot.

**Step 1.** First, rank the data in increasing order and calculate the values of the median, the first quartile, the third quartile, and the interquartile range. The ranked data are

69    74    75    79    81    84    90    94    98    104    112    144

For these data,

$$\text{Median} = (84 + 90)/2 = 87$$

$$Q_1 = (75 + 79)/2 = 77$$

$$Q_3 = (98 + 104)/2 = 101$$

$$\text{IQR} = Q_3 - Q_1 = 101 - 77 = 24$$

**Step 2.** Find the points that are  $1.5 \times \text{IQR}$  below  $Q_1$  and  $1.5 \times \text{IQR}$  above  $Q_3$ . These two points are called the **lower** and the **upper inner fences**, respectively.

$$1.5 \times \text{IQR} = 1.5 \times 24 = 36$$

$$\text{Lower inner fence} = Q_1 - 36 = 77 - 36 = 41$$

$$\text{Upper inner fence} = Q_3 + 36 = 101 + 36 = 137$$

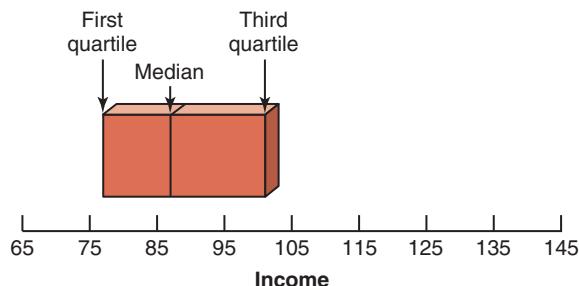
**Step 3.** Determine the smallest and the largest values in the given data set within the two inner fences. These two values for our example are as follows:

$$\text{Smallest value within the two inner fences} = 69$$

$$\text{Largest value within the two inner fences} = 112$$

**Step 4.** Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a box with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median. The result of this step is shown in Figure 3.13.

**Figure 3.13**



**Step 5.** By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 69 and 112 in this example as listed in Step 3. The two lines that join the box to these two values are called **whiskers**. A value that falls outside the two inner fences is shown by marking an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in Figure 3.14.

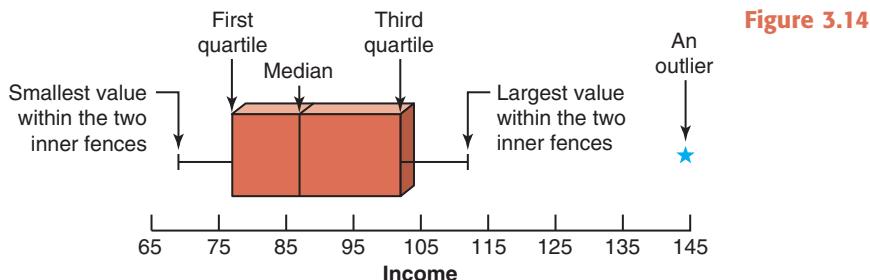


Figure 3.14

In Figure 3.14, about 50% of the data values fall within the box, about 25% of the values fall on the left side of the box, and about 25% fall on the right side of the box. Also, 50% of the values fall on the left side of the median and 50% lie on the right side of the median. The data of this example are skewed to the right because the lower 50% of the values are spread over a smaller range than the upper 50% of the values. ■

The observations that fall outside the two inner fences are called outliers. These outliers can be classified into two kinds of outliers—mild and extreme outliers. To do so, we define two outer fences—a **lower outer fence** at  $3.0 \times \text{IQR}$  below the first quartile and an **upper outer fence** at  $3.0 \times \text{IQR}$  above the third quartile. If an observation is outside either of the two inner fences but within the two outer fences, it is called a *mild outlier*. An observation that is outside either of the two outer fences is called an *extreme outlier*. For the previous example, the outer fences are at 5 and 173. Because 144 is outside the upper inner fence but inside the upper outer fence, it is a mild outlier.

For a symmetric data set, the line representing the median will be in the middle of the box and the spread of the values will be over almost the same range on both sides of the box.

## EXERCISES

### CONCEPTS AND PROCEDURES

**3.98** Briefly explain what summary measures are used to construct a box-and-whisker plot.

**3.99** Prepare a box-and-whisker plot for the following data:

36	43	28	52	41	59	47	61
24	55	63	73	32	25	35	49
31	22	61	42	58	65	98	34

Does this data set contain any outliers?

**3.100** Prepare a box-and-whisker plot for the following data:

11	8	26	31	62	19	7	3	14	75
33	30	42	15	18	23	29	13	16	6

Does this data set contain any outliers?



### APPLICATIONS

**3.101** The following data give the time (in minutes) that each of 20 students selected from a university waited in line at their bookstore to pay for their textbooks in the beginning of the Fall 2012 semester.

15	8	23	21	5	17	31	22	34	6
5	10	14	17	16	25	30	3	31	19

Prepare a box-and-whisker plot. Comment on the skewness of these data.

**3.102** Refer to Exercise 3.97. The following data represent the credit scores of 22 randomly selected loan applicants.

494	728	468	533	747	639	430	690	604	422	356
805	749	600	797	702	628	625	617	647	772	572

Prepare a box-and-whisker plot. Are these data skewed in any direction?

**3.103** The following data give the 2009 estimates of crude oil reserves (in billions of barrels) of Saudi Arabia, Iran, Iraq, Kuwait, Venezuela, the United Arab Emirates, Russia, Libya, Nigeria, Canada, the United States, China, Brazil, and Mexico (*source*: www.eia.gov).

266.7	136.2	115.0	107.0	99.4	97.8	60.0
43.7	36.2	27.7	21.3	16.0	12.6	10.5

Prepare a box-and-whisker plot. Is the distribution of these data symmetric or skewed? Are there any outliers? If so, classify them as mild or extreme.

**3.104** The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53
51	53	51	48	46	43	52	50	54	47
44	47	50	49	52					

Prepare a box-and-whisker plot. Comment on the skewness of these data.

**3.105** Refer to Exercise 3.93. The following data represent the numbers of minor penalties accrued by each of the 30 National Hockey League franchises during the 2010–11 regular season.

249	265	269	287	287	292	299	300	300	301
302	304	311	312	320	325	330	331	335	337
344	347	347	348	352	353	354	355	363	374

Prepare a box-and-whisker plot. Are these data skewed in any direction?

**3.106** Refer to Exercise 3.22. The following data represent the number of Grand Jury indictments for Gloucester County, New Jersey, for a sample of 11 weeks selected from July 2010 through June 2011 as reproduced from that exercise:

35	13	17	21	21	29	20	26	24	13	23
----	----	----	----	----	----	----	----	----	----	----

Make a box-and-whisker plot. Comment on the skewness of these data.

**3.107** Nixon Corporation manufactures computer monitors. The following are the numbers of computer monitors produced at the company for a sample of 30 days:

24	32	27	23	33	33	29	25	23	28
21	26	31	20	27	33	27	23	28	29
31	35	34	22	26	28	23	35	31	27

Prepare a box-and-whisker plot. Comment on the skewness of these data.

**3.108** The following data give the numbers of new cars sold at a dealership during a 20-day period.

8	5	12	3	9	10	6	12	8	8
4	16	10	11	7	7	3	5	9	11

Make a box-and-whisker plot. Comment on the skewness of these data.

## USES AND MISUSES... TAKING THINGS TO THE EXTREME

The first numeric summaries that students in a statistics class tend to learn are measures of center, especially the mean and the median. Many things in society enforce the notion that the most important statistics are the mean and the median, as they provide a notion of what *typically* occurs. Meteorologists report average temperatures, investment firms report average returns on mutual funds, and insurance companies examine the average claim payment. There will also be some analysis of variability, but, quite often, very little is said about *extreme* events despite the fact that extremes, while highly unlikely, can cause the most damage to a society, an economy, or a financial bottom line.

As an example, consider insurance. As a policy holder, you typically think about such things as the payment of a claim due to a car

accident, a tree that falls on and damages your house, or a theft. If you do not have insurance, the amount you would have to pay to repair a damaged car, fix your house, or replace stolen items would be an *extreme* (large) amount to you, although it would be a relatively small amount for an insurance company. You purchase insurance to protect yourself in the case of an *extreme* event. For an insurance company, an *extreme* event is one that results in simultaneous damage to the properties (cars, houses, etc.) of many policy holders by events like an earthquake, a hurricane, or a tornado. If an insurance company has to pay a few thousand dollars, or even \$200,000 to \$300,000, to one policy holder, this is relatively painless. However, if an insurance company has to pay thousands of dollars

to each of many hundreds of thousands of policy holders, this is an extreme that can affect the insurance company's assets, which help it to pay out future claims.

In recent years, several catastrophic events resulted in massive damages and losses. The earthquake and tsunami that struck Japan in 2011 had estimated insured losses of \$35 billion ([uk.reuters.com/article/2011/03/13/uk-air-worldwide-japan-idUKTRE72C1LH20110313](http://uk.reuters.com/article/2011/03/13/uk-air-worldwide-japan-idUKTRE72C1LH20110313)). The tornadoes that devastated Joplin, Missouri, and Huntsville, Alabama, in 2011 had combined insured losses of almost \$17 billion ([www.ncdc.noaa.gov/oa/reports/billionz.html#chron](http://www.ncdc.noaa.gov/oa/reports/billionz.html#chron)). Although both of these events increased the average pay out by insurance companies to their clients, insurance companies do not pay out the average

amount each day or week. The payments are made when the events occur, and the aforementioned events serve to deplete a company's reserves.

Statisticians who have to deal with these types of events use *Extreme Value Theory* to model the extremes of the distribution, namely those that are extremely large or small relative to the mean or median of a distribution. In many cases, modelers are interested in examining values that exceed a specific threshold, such as a cost that would deplete a large percentage of a company's reserves or a temperature or rainfall amount that would result in damage due to floods or drought or have the potential to result in frozen pipes or wildfires.

## Glossary

**Bimodal distribution** A distribution that has two modes.

**Box-and-whisker plot** A plot that shows the center, spread, and skewness of a data set with a box and two whiskers using the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

**Chebyshev's theorem** For any number  $k$  greater than 1, at least  $(1 - 1/k^2)$  of the values for any distribution lie within  $k$  standard deviations of the mean.

**Coefficient of variation** A measure of relative variability that expresses standard deviation as a percentage of the mean.

**Empirical rule** For a specific bell-shaped distribution, about 68% of the observations fall in the interval  $(\mu - \sigma)$  to  $(\mu + \sigma)$ , about 95% fall in the interval  $(\mu - 2\sigma)$  to  $(\mu + 2\sigma)$ , and about 99.7% fall in the interval  $(\mu - 3\sigma)$  to  $(\mu + 3\sigma)$ .

**First quartile** The value in a ranked data set such that about 25% of the measurements are smaller than this value and about 75% are larger. It is the median of the values that are smaller than the median of the whole data set.

**Geometric mean** Calculated by taking the  $n$ th root of the product of all values in a data set.

**Interquartile range (IQR)** The difference between the third and the first quartiles.

**Lower inner fence** The value in a data set that is  $1.5 \times \text{IQR}$  below the first quartile.

**Lower outer fence** The value in a data set that is  $3.0 \times \text{IQR}$  below the first quartile.

**Mean** A measure of central tendency calculated by dividing the sum of all values by the number of values in the data set.

**Measures of central tendency** Measures that describe the center of a distribution. The mean, median, and mode are three of the measures of central tendency.

**Measures of dispersion** Measures that give the spread of a distribution. The range, variance, and standard deviation are three such measures.

**Measures of position** Measures that determine the position of a single value in relation to other values in a data set. Quartiles, percentiles, and percentile rank are examples of measures of position.

**Median** The value of the middle term in a ranked data set. The median divides a ranked data set into two equal parts.

**Mode** The value (or values) that occurs with highest frequency in a data set.

**Multimodal distribution** A distribution that has more than two modes.

**Parameter** A summary measure calculated for population data.

**Percentile rank** The percentile rank of a value gives the percentage of values in the data set that are smaller than this value.

**Percentiles** Ninety-nine values that divide a ranked data set into 100 equal parts.

**Quartiles** Three summary measures that divide a ranked data set into four equal parts.

**Range** A measure of spread obtained by taking the difference between the largest and the smallest values in a data set.

**Second quartile** Middle or second of the three quartiles that divide a ranked data set into four equal parts. About 50% of the values in the data set are smaller and about 50% are larger than the second quartile. The second quartile is the same as the median.

**Standard deviation** A measure of spread that is given by the positive square root of the variance.

**Statistic** A summary measure calculated for sample data.

**Third quartile** Third of the three quartiles that divide a ranked data set into four equal parts. About 75% of the values in a data set are smaller than the value of the third quartile and about 25% are larger. It is the median of the values that are greater than the median of the whole data set.

**Trimmed mean** The  $k\%$  trimmed mean is obtained by dropping  $k\%$  of the smallest values and  $k\%$  of the largest values from the given data and then calculating the mean of the remaining  $(100 - 2k)\%$  of the values.

**Unimodal distribution** A distribution that has only one mode.

**Upper inner fence** The value in a data set that is  $1.5 \times \text{IQR}$  above the third quartile.

**Upper outer fence** The value in a data set that is  $3.0 \times \text{IQR}$  above the third quartile.

**Variance** A measure of spread.

**Weighted mean** Mean of a data set whose values are assigned different weights before the mean is calculated.

## Supplementary Exercises

**3.109** Each year the faculty at Metro Business College chooses 10 members from the current graduating class that they feel are most likely to succeed. The data below give the current annual incomes (in thousands of dollars) of the 10 members of the class of 2004 who were voted most likely to succeed.

59      68      84      78      107      382      56      74      97      60

- Calculate the mean and median.
- Does this data set contain any outlier(s)? If yes, drop the outlier(s) and recalculate the mean and median. Which of these measures changes by a greater amount when you drop the outlier(s)?
- Is the mean or the median a better summary measure for these data? Explain.

**3.110** The Belmont Stakes is the final race in the annual Triple Crown of thoroughbred horse racing. The race is 1.5 miles in length, and the record for the fastest time of 2 minutes, 24 seconds is held by Secretariat, the 1973 winner. We compared Secretariat's time from 1973 with the time of each winner of the Belmont Stakes for the years 1999–2011. The following data represent the differences (in seconds) between each winner's time for the years 1999–2011 and Secretariat's time in 1973. For example, the 1999 winner took 3.80 seconds longer than Secretariat to finish the race.

3.80    7.20    2.80    5.71    4.26    3.50    4.75    3.81    4.74    5.65    3.54    7.57    6.88

- Calculate the mean and median. Do these data have a mode? Why or why not?
- Compute the range, variance, and standard deviation for these data.

**3.111** The following table gives the total points scored by each of the top 16 National Basketball Association (NBA) scorers during the 2010–11 regular season (*source*: [www.nba.com](http://www.nba.com)).

Name	Points Scored	Name	Points Scored
Kevin Durant	2161	Kevin Martin	1876
LeBron James	2111	Blake Griffin	1845
Kobe Bryant	2078	Russell Westbrook	1793
Derrick Rose	2026	Dwight Howard	1784
Amare Stoudemire	1971	LaMarcus Aldridge	1769
Carmelo Anthony	1970	Dirk Nowitzki	1681
Dwyane Wade	1941	Brook Lopez	1673
Monta Ellis	1929	Danny Granger	1622

- Calculate the mean and median. Do these data have a mode? Why or why not?
- Compute the range, variance, and standard deviation for these data.

**3.112** The following data give the numbers of driving citations received during the last three years by 12 drivers.

4      8      0      3      11      7      4      14      8      13      7      9

- Find the mean, median, and mode for these data.
- Calculate the range, variance, and standard deviation.
- Are the values of the summary measures in parts a and b population parameters or sample statistics?

**3.113** The following table gives the distribution of the amounts of rainfall (in inches) for July 2012 for 50 cities.

Rainfall	Number of Cities
0 to less than 2	6
2 to less than 4	10
4 to less than 6	20
6 to less than 8	7
8 to less than 10	4
10 to less than 12	3

Find the mean, variance, and standard deviation. Are the values of these summary measures population parameters or sample statistics?

- 3.114** The following table gives the frequency distribution of the times (in minutes) that 50 commuter students at a large university spent looking for parking spaces on the first day of classes in the Fall semester of 2012.

Time	Number of Students
0 to less than 4	1
4 to less than 8	7
8 to less than 12	15
12 to less than 16	18
16 to less than 20	6
20 to less than 24	3

Find the mean, variance, and standard deviation. Are the values of these summary measures population parameters or sample statistics?

- 3.115** The mean time taken to learn the basics of a software program by all students is 200 minutes with a standard deviation of 20 minutes.

- a. Using Chebyshev's theorem, find at least what percentage of students will learn the basics of this software program in
  - i. 160 to 240 minutes
  - ii. 140 to 260 minutes
- \*b. Using Chebyshev's theorem, find the interval that contains the times taken by at least 75% of all students to learn this software program.

- 3.116** According to the American Time Use Survey conducted by the Bureau of Labor Statistics ([www.bls.gov/atus/](http://www.bls.gov/atus/)), Americans spent an average of 985.50 hours watching television in 2010. Suppose that the standard deviation of the distribution of times that Americans spent watching television in 2010 is 285.20 hours.

- a. Using Chebyshev's theorem, find at least what percentage of Americans watched television in 2010 for
  - i. 272.50 to 1698.50 hours
  - ii. 129.90 to 1841.10 hours
- \*b. Using Chebyshev's theorem, find the interval that contains the time (in hours) that at least 75% of Americans spent watching television in 2010.

- 3.117** Refer to Exercise 3.115. Suppose the times taken to learn the basics of this software program by all students have a bell-shaped distribution with a mean of 200 minutes and a standard deviation of 20 minutes.

- a. Using the empirical rule, find the percentage of students who will learn the basics of this software program in
  - i. 180 to 220 minutes
  - ii. 160 to 240 minutes
- \*b. Using the empirical rule, find the interval that contains the times taken by 99.7% of all students to learn this software program.

- 3.118** The annual earnings of all employees with CPA certification and 6 years of experience and working for large firms have a bell-shaped distribution with a mean of \$134,000 and a standard deviation of \$12,000.

- a. Using the empirical rule, find the percentage of all such employees whose annual earnings are between
  - i. \$98,000 and \$170,000
  - ii. \$110,000 and \$158,000
- \*b. Using the empirical rule, find the interval that contains the annual earnings of 68% of all such employees.

**3.119** Refer to the data of Exercise 3.109 on the current annual incomes (in thousands of dollars) of the 10 members of the class of 2004 of the Metro Business College who were voted most likely to succeed.

59      68      84      78      107      382      56      74      97      60

- Determine the values of the three quartiles and the interquartile range. Where does the value of 74 fall in relation to these quartiles?
- Calculate the (approximate) value of the 70th percentile. Give a brief interpretation of this percentile.
- Find the percentile rank of 97. Give a brief interpretation of this percentile rank.

**3.120** Refer to the data given in Exercise 3.111 on the total points scored by each of the top 16 NBA scorers during the 2010–11 regular season.

- Calculate the values of the three quartiles and the interquartile range. Where does the number 1681 fall in relation to these quartiles?
- Find the approximate value of the 18th percentile. Give a brief interpretation of this percentile.
- Calculate the percentile rank of 1793. Give a brief interpretation of this percentile rank.

**3.121** A student washes her clothes at a laundromat once a week. The data below give the time (in minutes) she spent in the laundromat for each of 15 randomly selected weeks. Here, time spent in the laundromat includes the time spent waiting for a machine to become available.

75      62      84      73      107      81      93      72  
135      77      85      67      90      83      112

Prepare a box-and-whisker plot. Is the data set skewed in any direction? If yes, is it skewed to the right or to the left? Does this data set contain any outliers?

**3.122** The following data give the lengths of time (in weeks) taken to find a full-time job by 18 computer science majors who graduated in 2011 from a small college.

30      43      32      21      65      8      4      18      16  
38      9      44      33      23      24      81      42      55

Make a box-and-whisker plot. Comment on the skewness of this data set. Does this data set contain any outliers?

## Advanced Exercises

**3.123** Melissa's grade in her math class is determined by three 100-point tests and a 200-point final exam. To determine the grade for a student in this class, the instructor will add the four scores together and divide this sum by 5 to obtain a percentage. This percentage must be at least 80 for a grade of B. If Melissa's three test scores are 75, 69, and 87, what is the minimum score she needs on the final exam to obtain a B grade?

**3.124** Jeffrey is serving on a six-person jury for a personal-injury lawsuit. All six jurors want to award damages to the plaintiff but cannot agree on the amount of the award. The jurors have decided that each of them will suggest an amount that he or she thinks should be awarded; then they will use the mean of these six numbers as the award to recommend to the plaintiff.

- Jeffrey thinks the plaintiff should receive \$20,000, but he thinks the mean of the other five jurors' recommendations will be about \$12,000. He decides to suggest an inflated amount so that the mean for all six jurors is \$20,000. What amount would Jeffrey have to suggest?
- How might this jury revise its procedure to prevent a juror like Jeffrey from having an undue influence on the amount of damages to be awarded to the plaintiff?

**3.125** The heights of five starting players on a basketball team have a mean of 76 inches, a median of 78 inches, and a range of 11 inches.

- If the tallest of these five players is replaced by a substitute who is 2 inches taller, find the new mean, median, and range.
- If the tallest player is replaced by a substitute who is 4 inches shorter, which of the new values (mean, median, range) could you determine, and what would their new values be?

**3.126** On a 300-mile auto trip, Lisa averaged 52 mph for the first 100 miles, 65 mph for the second 100 miles, and 58 mph for the last 100 miles.

- How long did the 300-mile trip take?
- Could you find Lisa's average speed for the 300-mile trip by calculating  $(52 + 65 + 58)/3$ ? If not, find the correct average speed for the trip.

- 3.127** A small country bought oil from three different sources in one week, as shown in the following table.

Source	Barrels Purchased	Price per Barrel (\$)
Mexico	1000	95
Kuwait	200	92
Spot Market	100	99

Find the mean price per barrel for all 1300 barrels of oil purchased in that week.

- 3.128** During the 2011–12 winter season, a homeowner received four deliveries of heating oil, as shown in the following table.

Gallons Purchased	Price per Gallon (\$)
209	2.60
182	2.40
157	2.78
149	2.74

The homeowner claimed that the mean price he paid for oil during the season was  $(2.60 + 2.40 + 2.78 + 2.74)/4 = \$2.63$  per gallon. Do you agree with this claim? If not, explain why this method of calculating the mean is not appropriate in this case and find the correct value of the mean price.

- 3.129** In the Olympic Games, when events require a subjective judgment of an athlete's performance, the highest and lowest of the judges' scores may be dropped. Consider a gymnast whose performance is judged by seven judges and the highest and the lowest of the seven scores are dropped.

- a. Gymnast A's scores in this event are 9.4, 9.7, 9.5, 9.5, 9.4, 9.6, and 9.5. Find this gymnast's mean score after dropping the highest and the lowest scores.
- b. The answer to part a is an example of (approximately) what percentage of trimmed mean?
- c. Write another set of scores for a gymnast B so that gymnast A has a higher mean score than gymnast B based on the trimmed mean, but gymnast B would win if all seven scores were counted. Do not use any scores lower than 9.0.

- 3.130** A survey of young people's shopping habits in a small city during the summer months of 2012 showed the following: Shoppers aged 12 to 14 years took an average of 8 shopping trips per month and spent an average of \$14 per trip. Shoppers aged 15 to 17 years took an average of 11 trips per month and spent an average of \$18 per trip. Assume that this city has 1100 shoppers aged 12 to 14 years and 900 shoppers aged 15 to 17 years.

- a. Find the total amount spent per month by all these 2000 shoppers in both age groups.
- b. Find the mean number of shopping trips per person per month for these 2000 shoppers.
- c. Find the mean amount spent per person per month by shoppers aged 12 to 17 years in this city.

- 3.131** The following table shows the total population and the number of deaths (in thousands) due to heart attack for two age groups (in years) in Countries A and B for 2011.

	Age 30 and Under		Age 31 and Over	
	A	B	A	B
Population	40,000	25,000	20,000	35,000
Deaths due to heart attack	1000	500	2000	3000

- a. Calculate the death rate due to heart attack per 1000 population for the 30 years and under age group for each of the two countries. Which country has the lower death rate in this age group?
- b. Calculate the death rates due to heart attack for the two countries for the 31 years and over age group. Which country has the lower death rate in this age group?
- c. Calculate the death rate due to heart attack for the entire population of Country A; then do the same for Country B. Which country has the lower overall death rate?
- d. How can the country with lower death rate in both age groups have the higher overall death rate? (This phenomenon is known as Simpson's paradox.)

**3.132** In a study of distances traveled to a college by commuting students, data from 100 commuters yielded a mean of 8.73 miles. After the mean was calculated, data came in late from three students, with respective distances of 11.5, 7.6, and 10.0 miles. Calculate the mean distance for all 103 students.

**3.133** The test scores for a large statistics class have an unknown distribution with a mean of 70 and a standard deviation of 10.

- Find  $k$  so that at least 50% of the scores are within  $k$  standard deviations of the mean.
- Find  $k$  so that at most 10% of the scores are more than  $k$  standard deviations above the mean.

**3.134** The test scores for a very large statistics class have a bell-shaped distribution with a mean of 70 points.

- If 16% of all students in the class scored above 85, what is the standard deviation of the scores?
- If 95% of the scores are between 60 and 80, what is the standard deviation?

**3.135** How much does the typical American family spend to go away on vacation each year? Twenty-five randomly selected households reported the following vacation expenditures (rounded to the nearest hundred dollars) during the past year:

2500	500	800	0	100
0	200	2200	0	200
0	1000	900	321,500	400
500	100	0	8200	900
0	1700	1100	600	3400

- Using both graphical and numerical methods, organize and interpret these data.
- What measure of central tendency best answers the original question?

**3.136** Actuaries at an insurance company must determine a premium for a new type of insurance. A random sample of 40 potential purchasers of this type of insurance were found to have suffered the following values of losses (in dollars) during the past year. These losses would have been covered by the insurance if it were available.

100	32	0	0	470	50	0	14,589	212	93
0	0	1127	421	0	87	135	420	0	250
12	0	309	0	177	295	501	0	143	0
167	398	54	0	141	0	3709	122	0	0

- Find the mean, median, and mode of these 40 losses.
- Which of the mean, median, or mode is largest?
- Draw a box-and-whisker plot for these data, and describe the skewness, if any.
- Which measure of central tendency should the actuaries use to determine the premium for this insurance?

**3.137** A local golf club has men's and women's summer leagues. The following data give the scores for a round of 18 holes of golf for 17 men and 15 women randomly selected from their respective leagues.

<b>Men</b>	87	68	92	79	83	67	71	92	112
	75	77	102	79	78	85	75	72	
<b>Women</b>	101	100	87	95	98	81	117	107	103
	97	90	100	99	94	94			

- Make a box-and-whisker plot for each of the data sets and use them to discuss the similarities and differences between the scores of the men and women golfers.
- Compute the various descriptive measures you have learned for each sample. How do they compare?

**3.138** Answer the following questions.

- The total weight of all pieces of luggage loaded onto an airplane is 12,372 pounds, which works out to be an average of 51.55 pounds per piece. How many pieces of luggage are on the plane?
- A group of seven friends, having just gotten back a chemistry exam, discuss their scores. Six of the students reveal that they received grades of 81, 75, 93, 88, 82, and 85, respectively, but the

seventh student is reluctant to say what grade she received. After some calculation she announces that the group averaged 81 on the exam. What is her score?

- 3.139** Suppose that there are 150 freshmen engineering majors at a college and each of them will take the same five courses next semester. Four of these courses will be taught in small sections of 25 students each, whereas the fifth course will be taught in one section containing all 150 freshmen. To accommodate all 150 students, there must be six sections of each of the four courses taught in 25-student sections. Thus, there are 24 classes of 25 students each and one class of 150 students.

- Find the mean size of these 25 classes.
- Find the mean class size from a student's point of view, noting that each student has five classes containing 25, 25, 25, 25, and 150 students, respectively.

Are the means in parts a and b equal? If not, why not?

- 3.140** The following data give the weights (in pounds) of a random sample of 44 college students. (Here F and M indicate female and male, respectively.)

123 F	195 M	138 M	115 F	179 M	119 F
148 F	147 F	180 M	146 F	179 M	189 M
175 M	108 F	193 M	114 F	179 M	147 M
108 F	128 F	164 F	174 M	128 F	159 M
193 M	204 M	125 F	133 F	115 F	168 M
123 F	183 M	116 F	182 M	174 M	102 F
123 F	99 F	161 M	162 M	155 F	202 M
110 F	132 M				

Compute the mean, median, and standard deviation for the weights of all students, of men only, and of women only. Of the mean and median, which is the more informative measure of central tendency? Write a brief note comparing the three measures for all students, men only, and women only.

- 3.141** The distribution of the lengths of fish in a certain lake is not known, but it is definitely not bell shaped. It is estimated that the mean length is 6 inches with a standard deviation of 2 inches.

- At least what proportion of fish in the lake are between 3 inches and 9 inches long?
- What is the smallest interval that will contain the lengths of at least 84% of the fish?
- Find an interval so that fewer than 36% of the fish have lengths outside this interval.

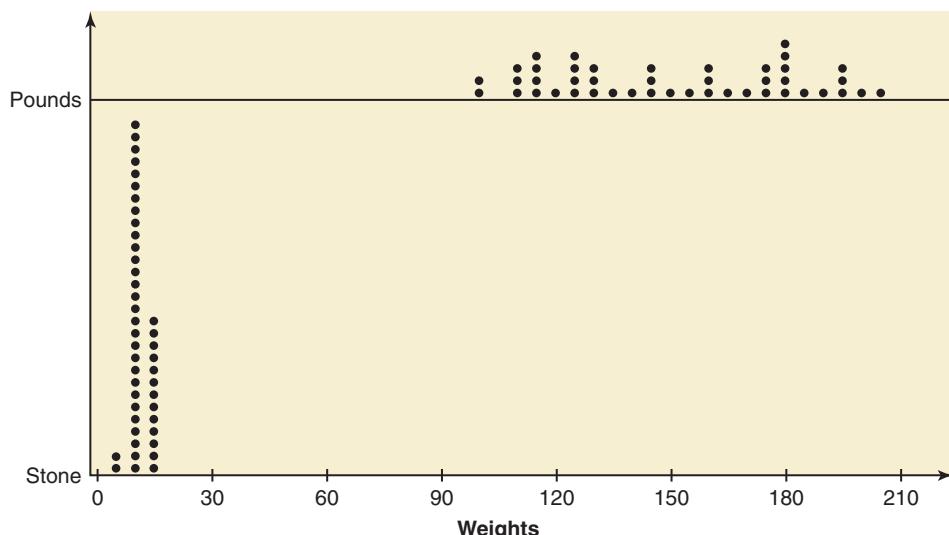
- 3.142** The following stem-and-leaf diagram gives the distances (in thousands of miles) driven during the past year by a sample of drivers in a city.

0	3 6 9
1	2 8 5 1 0 5
2	5 1 6
3	8
4	1
5	
6	2

- Compute the sample mean, median, and mode for the data on distances driven.
- Compute the range, variance, and standard deviation for these data.
- Compute the first and third quartiles.
- Compute the interquartile range. Describe what properties the interquartile range has. When would the IQR be preferable to using the standard deviation when measuring variation?

- 3.143** Refer to the data in Problem 3.140. Two individuals, one from Canada and one from England, are interested in your analysis of these data but they need your results in different units. The Canadian individual wants the results in grams (1 pound = 435.59 grams), while the English individual wants the results in stones (1 stone = 14 pounds).

- Convert the data on weights from pounds to grams, and then recalculate the mean, median, and standard deviation of weight for males and females separately. Repeat the procedure, changing the unit from pounds to stones.
- Convert your answers from Problem 3.140 to grams and stones. What do you notice about these answers and your answers from part a?



**Figure 3.15** Stacked dotplot of weights in stones and pounds.

- c. What happens to the values of the mean, median, and standard deviation when you convert from a larger unit to a smaller unit (e.g., from pounds to grams)? Does the same thing happen if you convert from a smaller unit (e.g., pounds) to a larger unit (e.g., stones)?
- d. Figure 3.15 gives a stacked dotplot of these weights in pounds and stones. Which of these two distributions has more variability? Use your results from parts a to c to explain why this is the case.
- e. Now consider the weights in pounds and grams. Make a stacked dotplot for these data and answer part d.

**3.144** Although the standard workweek is 40 hours a week, many people work a lot more than 40 hours a week. The following data give the numbers of hours worked last week by 50 people.

40.5	41.3	41.4	41.5	42.0	42.2	42.4	42.4	42.6	43.3
43.7	43.9	45.0	45.0	45.2	45.8	45.9	46.2	47.2	47.5
47.8	48.2	48.3	48.8	49.0	49.2	49.9	50.1	50.6	50.6
50.8	51.5	51.5	52.3	52.3	52.6	52.7	52.7	53.4	53.9
54.4	54.8	55.0	55.4	55.4	55.4	56.2	56.3	57.8	58.7

- a. The sample mean and sample standard deviation for this data set are 49.012 and 5.080, respectively. Using Chebyshev's theorem, calculate the intervals that contain at least 75%, 88.89%, and 93.75% of the data.
- b. Determine the actual percentages of the given data values that fall in each of the intervals that you calculated in part a. Also calculate the percentage of the data values that fall within one standard deviation of the mean.
- c. Do you think the lower endpoints provided by Chebyshev's theorem in part a are useful for this problem? Explain your answer.
- d. Suppose that the individual with the first number (54.4) in the fifth row of the data is a workaholic who actually worked 84.4 hours last week and not 54.4 hours. With this change now  $\bar{x} = 49.61$  and  $s = 7.10$ . Recalculate the intervals for part a and the actual percentages for part b. Did your percentages change a lot or a little?
- e. How many standard deviations above the mean would you have to go to capture all 50 data values? What is the lower bound for the percentage of the data that should fall in the interval, according to the Chebyshev theorem?

**3.145** Refer to the women's golf scores in Exercise 3.137. It turns out that 117 was mistakenly entered. Although this person still had the highest score among the 15 women, her score was not a mild or extreme outlier according to the box-and-whisker plot, nor was she tied for the highest score. What are the possible scores that she could have shot?

## APPENDIX 3.1

### A3.1.1 BASIC FORMULAS FOR THE VARIANCE AND STANDARD DEVIATION FOR UNGROUPED DATA

Example 3–25 below illustrates how to use the basic formulas to calculate the variance and standard deviation for ungrouped data. From Section 3.2.2, the basic formulas for variance for ungrouped data are

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

where  $\sigma^2$  is the population variance and  $s^2$  is the sample variance.

In either case, the standard deviation is obtained by taking the square root of the variance.

**EXAMPLE 3–25** Refer to Example 3–12, where we used the short-cut formulas to compute the variance and standard deviation for the data on the baggage fee revenues collected by six airlines in 2010. Calculate the variance and standard deviation for those data using the basic formula.

*Calculating the variance and standard deviation for ungrouped data using basic formulas.*

**Solution** Let  $x$  denote the baggage fee revenue (in millions of dollars) collected by an airline in 2010. Table 3.14 shows all the required calculations to find the variance and standard deviation.

**Table 3.14**

$x$	$(x - \bar{x})$	$(x - \bar{x})^2$
313	$313 - 475.67 = -162.67$	26,461.5289
342	$342 - 475.67 = -133.67$	17,867.6689
581	$581 - 475.67 = 105.33$	11,094.4089
952	$952 - 475.67 = 476.33$	226,890.2689
514	$514 - 475.67 = 38.33$	1469.1889
152	$152 - 475.67 = -323.67$	104,762.2689
$\Sigma x = 2854$		$\Sigma(x - \bar{x})^2 = 388,545.3334$

The following steps are performed to compute the variance and standard deviation.

**Step 1.** Find the mean as follows:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{2854}{6} = 475.67$$

**Step 2.** Calculate  $x - \bar{x}$ , the deviation of each value of  $x$  from the mean. The results are shown in the second column of Table 3.14.

**Step 3.** Square each of the deviations of  $x$  from  $\bar{x}$ ; that is, calculate each of the  $(x - \bar{x})^2$  values. These values are called the *squared deviations*, and they are recorded in the third column.

**Step 4.** Add all the squared deviations to obtain  $\Sigma(x - \bar{x})^2$ ; that is, sum all the values given in the third column of Table 3.14. This gives

$$\Sigma(x - \bar{x})^2 = 388,545.3334$$

**Step 5.** Obtain the sample variance by dividing the sum of the squared deviations by  $n - 1$ . Thus

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{388,545.3334}{6 - 1} = 77,709.06668$$

**Step 6.** Obtain the sample standard deviation by taking the positive square root of the variance. Hence,

$$s = \sqrt{77,709.06668} = \mathbf{278.76} = \$278.76 \text{ million}$$

### A3.1.2 BASIC FORMULAS FOR THE VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA

Example 3–26 demonstrates how to use the basic formulas to calculate the variance and standard deviation for grouped data. The basic formulas for these calculations are

$$\sigma^2 = \frac{\sum f(m - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\sum f(m - \bar{x})^2}{n - 1}$$

where  $\sigma^2$  is the population variance,  $s^2$  is the sample variance,  $m$  is the midpoint of a class, and  $f$  is the frequency of a class.

In either case, the standard deviation is obtained by taking the square root of the variance.

*Calculating the variance and standard deviation for grouped data using basic formulas.*

**EXAMPLE 3–26** In Example 3–17, we used the short-cut formula to compute the variance and standard deviation for the data on the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the variance and standard deviation for those data using the basic formula.

**Solution** All the required calculations to find the variance and standard deviation appear in Table 3.15.

**Table 3.15**

Number of Orders		$f$	$m$	$mf$	$m - \bar{x}$	$(m - \bar{x})^2$	$f(m - \bar{x})^2$
10–12	4	11		44	-5.64	31.8096	127.2384
13–15	12	14		168	-2.64	6.9696	83.6352
16–18	20	17		340	.36	.1296	2.5920
19–21	14	20		280	3.36	11.2896	158.0544
		$n = 50$		$\sum mf = 832$		$\sum f(m - \bar{x})^2 = 371.5200$	

The following steps are performed to compute the variance and standard deviation using the basic formula.

**Step 1.** Find the midpoint of each class. Multiply the corresponding values of  $m$  and  $f$ . Find  $\sum mf$ . From Table 3.15,  $\sum mf = 832$ .

**Step 2.** Find the mean as follows:

$$\bar{x} = \sum mf/n = 832/50 = 16.64$$

**Step 3.** Calculate  $m - \bar{x}$ , the deviation of each value of  $m$  from the mean. These calculations are done in the fifth column of Table 3.15.

**Step 4.** Square each of the deviations  $m - \bar{x}$ ; that is, calculate each of the  $(m - \bar{x})^2$  values. These are called *squared deviations*, and they are recorded in the sixth column.

**Step 5.** Multiply the squared deviations by the corresponding frequencies (see the seventh column of Table 3.15). Adding the values of the seventh column, we obtain

$$\sum f(m - \bar{x})^2 = 371.5200$$

**Step 6.** Obtain the sample variance by dividing  $\sum f(m - \bar{x})^2$  by  $n - 1$ . Thus,

$$s^2 = \frac{\sum f(m - \bar{x})^2}{n - 1} = \frac{371.5200}{50 - 1} = 7.5820$$

**Step 7.** Obtain the standard deviation by taking the positive square root of the variance.

$$s = \sqrt{s^2} = \sqrt{7.5820} = 2.75 \text{ orders}$$

### Self-Review Test

- The value of the middle term in a ranked data set is called the
  - mean
  - median
  - mode
- Which of the following summary measures is/are influenced by extreme values?
  - mean
  - median
  - mode
  - range

3. Which of the following summary measures can be calculated for qualitative data?  
 a. mean      b. median      c. mode
4. Which of the following can have more than one value?  
 a. mean      b. median      c. mode
5. Which of the following is obtained by taking the difference between the largest and the smallest values of a data set?  
 a. variance    b. range      c. mean
6. Which of the following is the mean of the squared deviations of  $x$  values from the mean?  
 a. standard deviation    b. population variance    c. sample variance
7. The values of the variance and standard deviation are  
 a. never negative      b. always positive      c. never zero
8. A summary measure calculated for the population data is called  
 a. a population parameter    b. a sample statistic    c. an outlier
9. A summary measure calculated for the sample data is called a  
 a. population parameter    b. sample statistic    c. box-and-whisker plot
10. Chebyshev's theorem can be applied to  
 a. any distribution      b. bell-shaped distributions only      c. skewed distributions only
11. The empirical rule can be applied to  
 a. any distribution      b. bell-shaped distributions only      c. skewed distributions only
12. The first quartile is a value in a ranked data set such that about  
 a. 75% of the values are smaller and about 25% are larger than this value  
 b. 50% of the values are smaller and about 50% are larger than this value  
 c. 25% of the values are smaller and about 75% are larger than this value
13. The third quartile is a value in a ranked data set such that about  
 a. 75% of the values are smaller and about 25% are larger than this value  
 b. 50% of the values are smaller and about 50% are larger than this value  
 c. 25% of the values are smaller and about 75% are larger than this value
14. The 75th percentile is a value in a ranked data set such that about  
 a. 75% of the values are smaller and about 25% are larger than this value  
 b. 25% of the values are smaller and about 75% are larger than this value
15. The following data give the number of items purchased by each of 14 customers who shopped on a certain day at a supermarket.

18	14	22	7	9	13	19
25	13	4	16	22	6	10

Calculate the mean, median, mode, range, variance, and standard deviation.

16. The mean, as a measure of central tendency, has the disadvantage of being influenced by extreme values. Illustrate this point with an example.
17. The range, as a measure of spread, has the disadvantage of being influenced by extreme values. Illustrate this point with an example.
18. When is the value of the standard deviation for a data set zero? Give one example of such a data set. Calculate the standard deviation for that data set to show that it is zero.
19. The following table gives the frequency distribution of the numbers of computers sold during the past 25 weeks at an electronics store.

Computers Sold	Frequency
4 to 9	2
10 to 15	4
16 to 21	10
22 to 27	6
28 to 33	3

- a. What does the frequency column in the table represent?  
 b. Calculate the mean, variance, and standard deviation.

**20.** The members of a very large health club were observed on a randomly selected day. The distribution of times they spent that day at the health club was found to have a mean of 91.8 minutes and a standard deviation of 16.2 minutes. Suppose these values of the mean and standard deviation hold true for all members of this club.

- Using Chebyshev's theorem, find at least what percentage of this health club's members spend times at this health club between
  - 59.4 and 124.2 minutes
  - 51.3 and 132.3 minutes
- Using Chebyshev's theorem, find the interval that contains the times spent at this health club by at least 89% of members.

**21.** The ages of cars owned by all people living in a city have a bell-shaped distribution with a mean of 7.3 years and a standard deviation of 2.2 years.

- Using the empirical rule, find the percentage of cars in this city that are
  - 5.1 to 9.5 years old
  - .7 to 13.9 years old
- Using the empirical rule, find the interval that contains the ages of 95% of the cars owned by all people in this city.

**22.** The following data give the number of times the metal detector was set off by passengers at a small airport during 15 consecutive half-hour periods on February 1, 2012.

7	2	12	13	0	8	10
15	3	5	14	20	1	11

- Calculate the three quartiles and the interquartile range. Where does the value of 4 lie in relation to these quartiles?
- Find the (approximate) value of the 60th percentile. Give a brief interpretation of this value.
- Calculate the percentile rank of 12. Give a brief interpretation of this value.

**23.** Make a box-and-whisker plot for the data on the number of times passengers set off the airport metal detector given in Problem 22. Comment on the skewness of this data set.

**\*24.** The mean weekly wages of a sample of 15 employees of a company are \$1035. The mean weekly wages of a sample of 20 employees of another company are \$1090. Find the combined mean for these 35 employees.

**\*25.** The mean GPA of five students is 3.21. The GPAs of four of these five students are, respectively, 3.85, 2.67, 3.45, and 2.91. Find the GPA of the fifth student.

**\*26.** The following are the prices (in thousands of dollars) of 10 houses sold recently in a city:

479	366	238	207	287	349	293	2534	463	538
-----	-----	-----	-----	-----	-----	-----	------	-----	-----

Calculate the 10% trimmed mean for this data set. Do you think the 10% trimmed mean is a better summary measure than the (simple) mean (i.e., the mean of all 10 values) for these data? Briefly explain why or why not.

**\*27.** Consider the following two data sets.

Data Set I:	8	16	20	35
Data Set II:	5	13	17	32

Note that each value of the second data set is obtained by subtracting 3 from the corresponding value of the first data set.

- Calculate the mean for each of these two data sets. Comment on the relationship between the two means.
- Calculate the standard deviation for each of these two data sets. Comment on the relationship between the two standard deviations.

## Mini-Projects

### ■ MINI-PROJECT 3-1

Refer to the data you collected for Mini-Project 1-1 of Chapter 1 and analyzed graphically in Mini-Project 2-1 of Chapter 2. Write a report summarizing those data. This report should include answers to at least the following questions.

- Calculate the summary measures mean, standard deviation, five-number summary (minimum value, first quartile, median, third quartile, and maximum value), and interquartile range for the variable you graphed in Mini-Project 2-1. Do this for the entire data set, as well as for the different groups formed by the categorical variable that you used to divide the data set in Mini-Project 2-1.
- Are the summary measures for the various groups similar to those for the entire data set? If not, which ones differ and how do they differ? Make the same comparisons among the summary

measures for various groups. Do the groups have similar levels of variability? Explain how you can determine this from the graphs that you created in Mini-Project 2–1.

- c. Draw a box-and-whisker plot for the entire data set. Also draw side-by-side box-and-whisker plots for the various groups. Are there any outliers? If so, are there any values that are outliers in any of the groups but not in the entire data set? Does the plot show any skewness?
- d. Discuss which measures for the center and spread would be more appropriate to use to describe your data set. Also, discuss your reasons for using those measures.

### MINI-PROJECT 3-2

You are employed as a statistician for a company that makes household products, which are sold by part-time salespersons who work during their spare time. The company has four salespersons employed in a small town. Let us denote these salespersons by A, B, C, and D. The sales records (in dollars) for the past 6 weeks for these four salespersons are shown in the following table.

Week	A	B	C	D
1	1774	2205	1330	1402
2	1808	1507	1295	1665
3	1890	2352	1502	1530
4	1932	1939	1104	1826
5	1855	2052	1189	1703
6	1726	1630	1441	1498

Your supervisor has asked you to prepare a brief report comparing the sales volumes and the consistency of sales of these four salespersons. Use the mean sales for each salesperson to compare the sales volumes, and then choose an appropriate statistical measure to compare the consistency of sales. Make the calculations and write a report.

### MINI-PROJECT 3-3

Refer to the data you collected and analyzed graphically for Mini-Project 2–3 of Chapter 2. Write a report summarizing those data. This report should include answers to at least the following questions.

- a. Calculate the summary measures (mean, standard deviation, five-number summary, and interquartile range) for each of the three variables you chose.
- b. Which of the three variables has the largest measures of variability? Which has the smallest? Explain why.
- c. Draw a box-and-whisker plot for each of the three variables. Are these plots consistent with your answer in part b? Are there any outliers?
- d. Discuss which measures for the center and variability would be most appropriate to use to describe the variables.

## DECIDE FOR YOURSELF

### DECIDING WHERE TO LIVE

By the time you get to college, you must have heard it over and over again: “A picture is worth a thousand words.” Now we have pictures and numbers discussed in Chapters 2 and 3, respectively. Why both? Well, although each one of them acts as a summary of a data set, it is a combination of the pictures and numbers that tells a big part of the story without having to look at the entire data set. Suppose that you ask a realtor for information on the prices of homes in two different but comparable midwestern suburbs. Let us call these Suburbs A and B. The realtor provides you with the following information that is obtained from a random sample of 40 houses in each suburb:

- a. The average price of homes in each of the two suburbs
- b. The five-number summary of prices of homes in each neighborhood
- c. The histogram of the distribution of home prices for each suburb

All the information provided by the realtor is given in the following two tables and two histograms shown in Figures 3.16 and 3.17. Note that the second table gives the minimum and maximum prices of homes (in thousands of dollars) for each suburb along with the values of  $Q_1$ , median, and  $Q_3$  (in thousands of dollars).

Suburb	A	B
Average Price (in thousands of dollars)	221.9	220.03

	Minimum	$Q_1$	Median	$Q_3$	Maximum
Suburb A	151.0	175.5	188.0	199.5	587.0
Suburb B	187.0	210.0	222.5	228.0	250.0

Before you decide which suburb you should buy the house in, answer the following questions:

1. Examine the summary statistics and graphs given here.
2. Explain how the information given here can help you to make a decision about the suburb where you should look for a house to buy.

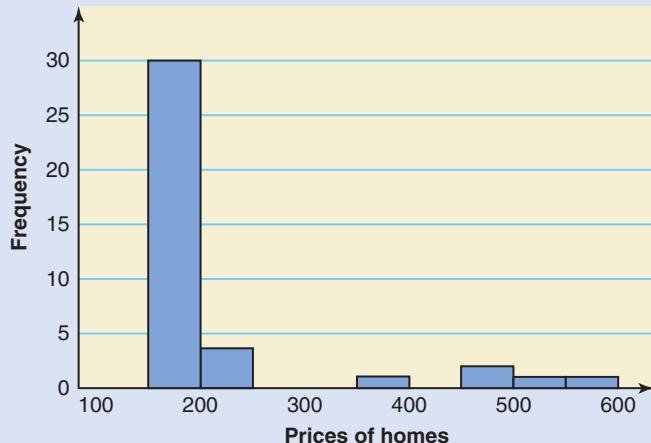


Figure 3.16 Histogram of prices of homes in Suburb A.

3. Explain how and why you might be misled by simply looking at the average prices if you are looking to spend less money to buy a house.

4. Is there any information about the suburbs not given here that you will like to obtain before making a decision about the suburb where you should buy a house?



Figure 3.17 Histogram of prices of homes in Suburb B.

## TECHNOLOGY INSTRUCTION

### Numerical Descriptive Measures

#### TI-84

```
1-Var Stats
x̄=6.8333333333
Σx=41
Σx²=377
Sx=.400757511
σx=.017323598
n=6
```

Screen 3.1

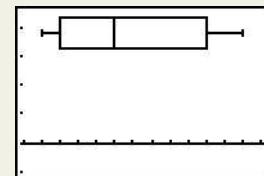
```
1-Var Stats
n=6
minX=2
Q1=3
Med=6
Q3=11
maxX=13
```

Screen 3.2

1. To calculate the **sample statistics** (e.g., mean, standard deviation, and five-number summary), first enter your data into a list such as L1, then select **STAT >CALC >1-Var Stats**, and press **ENTER**. At the List prompt, access the name of your list by pressing **2nd >STAT** and scrolling through the list of names until you get to your list name. At the FreqList prompt, enter the name of the list that contains the frequencies. If there are no frequencies, leave this field blank. Highlight Calculate and press **ENTER**. You will obtain the output shown in Screens 3.1 and 3.2.

Screen 3.1 shows, in this order, the sample mean, the sum of the data values, the sum of the squared data values, the sample standard deviation, the value of the population standard deviation (you will use this only when your data constitute a census instead of a sample), and the number of data values (e.g., the sample or population size). Pressing the downward arrow key will show the five-number summary, which is shown in Screen 3.2.

2. Constructing a box-and-whisker plot is similar to constructing a histogram. First enter your data into a list such as L1, then select **STAT PLOT** and go into one of the three plots. Make sure the plot is turned on. For the type, select the second row, first column (this boxplot will display outliers, if there are any). Enter the name of your list for **XList**. Select **ZOOM>9** to display the plot as shown in Screen 3.3.

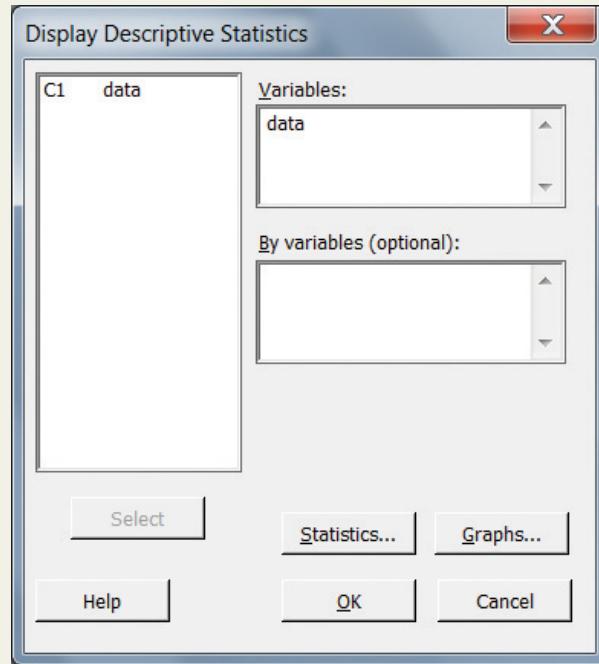


Screen 3.3

#### Minitab

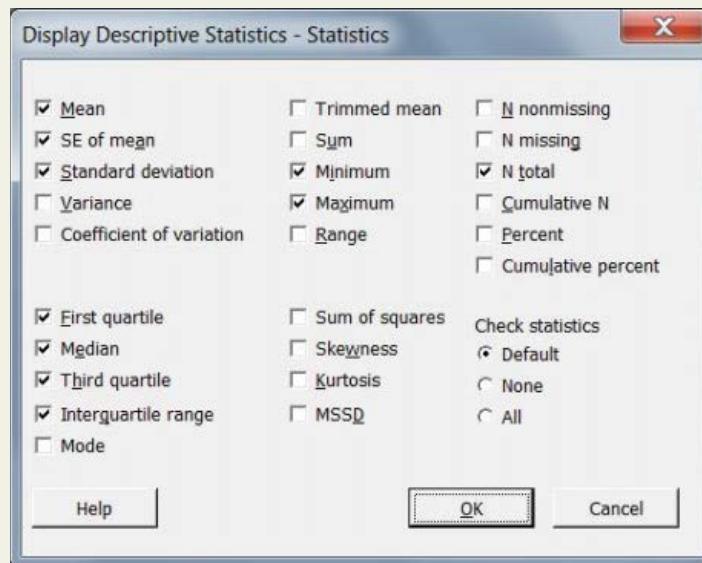
1. To find the sample statistics (e.g., the mean, standard deviation, and five-number summary), first enter the given data in a column such as C1, and then select **Stat >Basic Statistics > Display Descriptive Statistics**. In the dialog box you obtain, enter the name of the column

Screen 3.4



where your data are stored in the **Variables** box as shown in **Screen 3.4**. Click the **Statistics** button in this dialog box and choose the summary measures you want to calculate in the new dialog box as shown in **Screen 3.5**. Click **OK** in both dialog boxes. The output will appear in the **Session** window, which is shown in **Screen 3.6**.

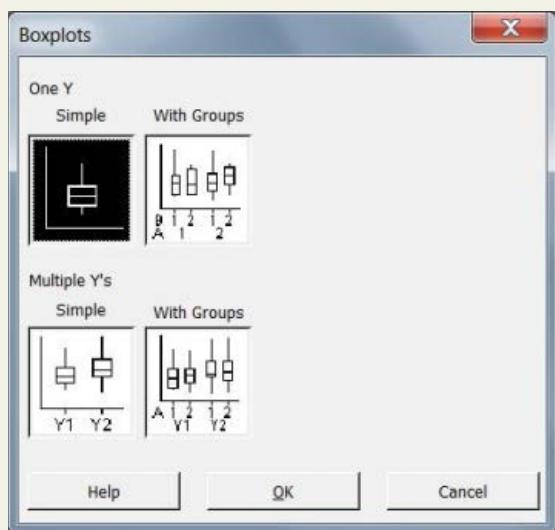
Screen 3.5



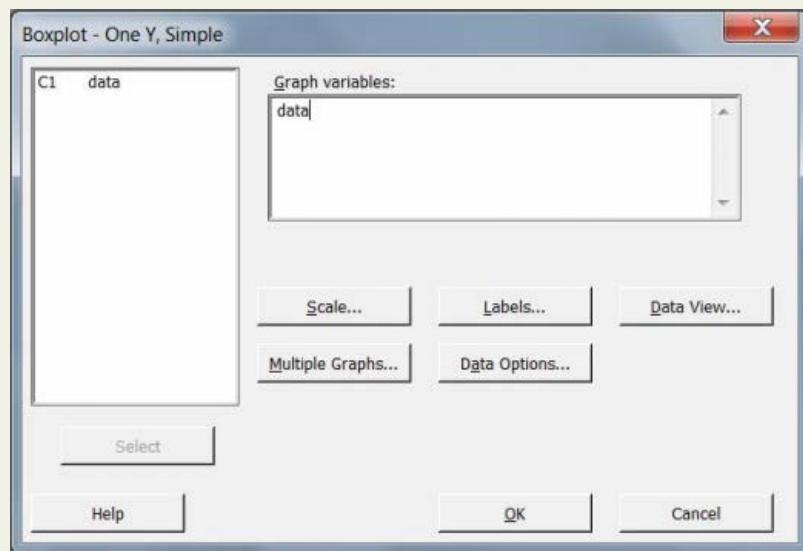
Descriptive Statistics: data											
Variable	Count	Total									
		Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR	
data	6	6.83	1.78	4.36	2.00	2.75	6.00	11.50	13.00	8.75	

Screen 3.6

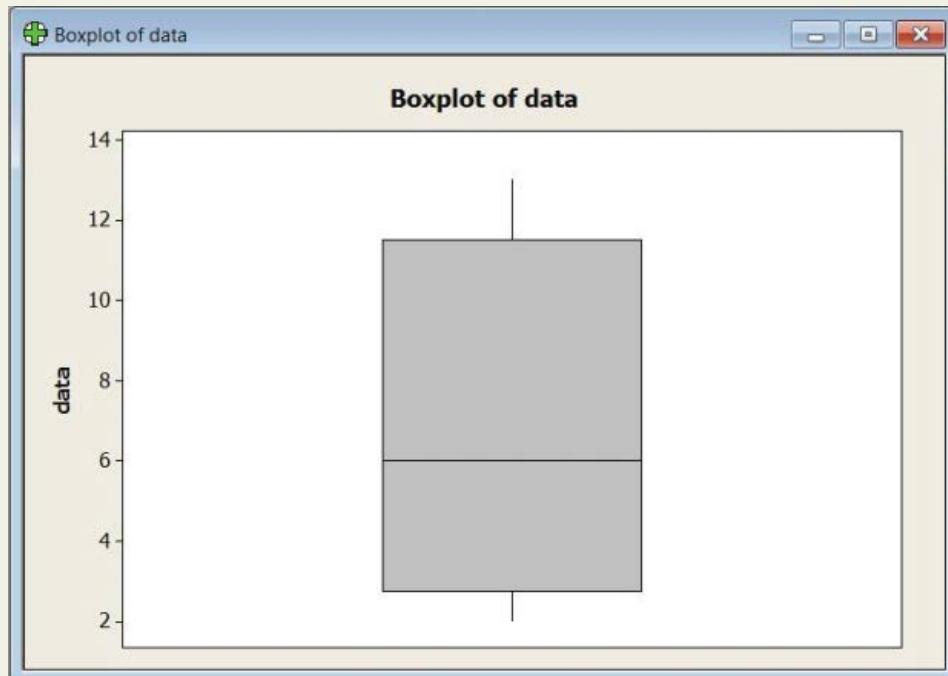
2. To create a box-and-whisker plot, enter the given data in a column such as C1, select **Graph > Boxplot > Simple**, and click **OK** (see Screen 3.7). In the dialog box that appears, enter the name of the column that contains the data in the **Graph variables** box (see Screen 3.8) and click **OK**. The boxplot shown in Screen 3.9 will appear on the screen.



Screen 3.7

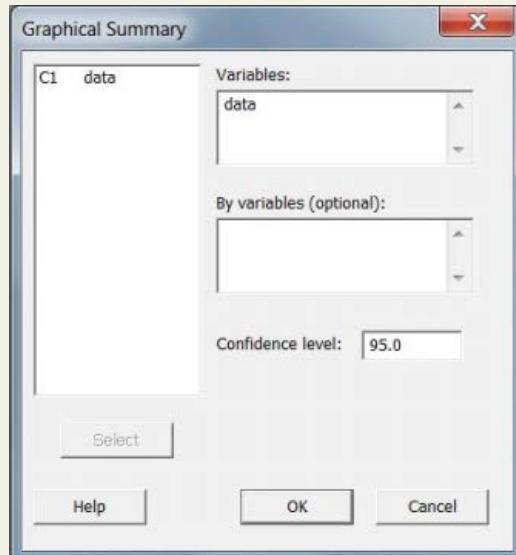
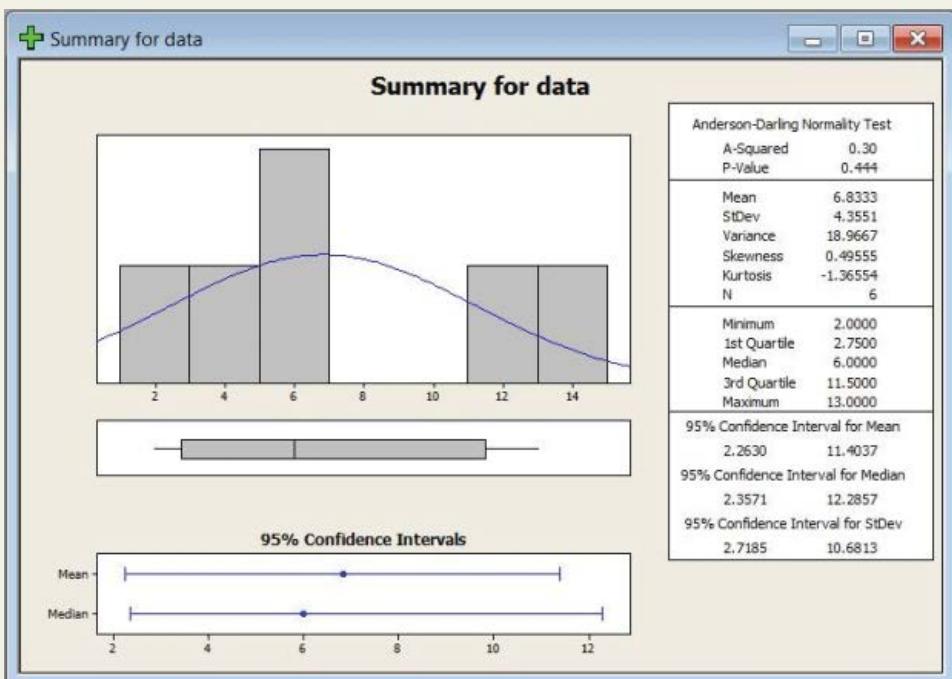


Screen 3.8

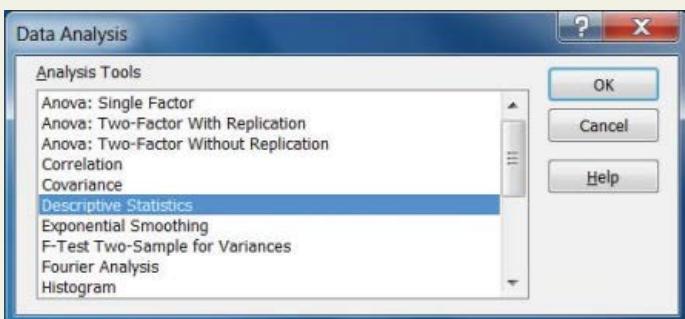


Screen 3.9

3. A nice combination of graphical and numeric summaries is available with a single command. Enter the given data in a column such as C1, select **Stat > Basic Statistics > Graphical Summary**, and click **OK**. In the dialog box that appears, enter the name of the column that contains the data in the **Variables** box (see Screen 3.10) and click **OK**. The output in Screen 3.11 will appear.

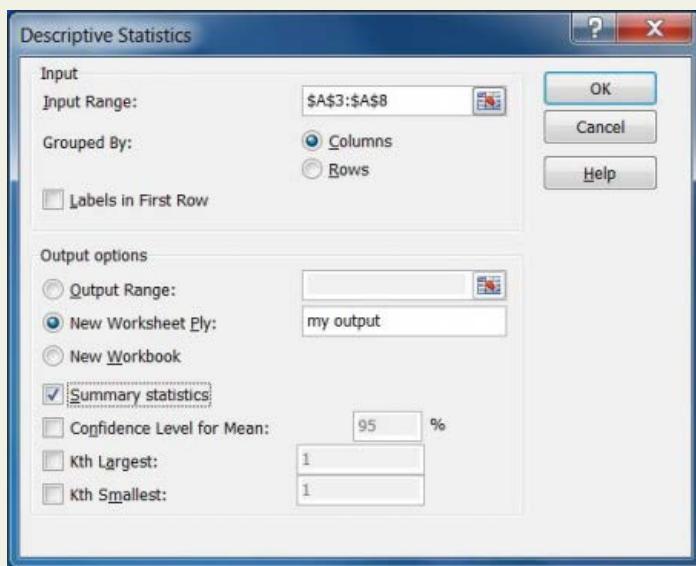
**Screen 3.10****Screen 3.11**

## Excel

**Screen 3.12**

Calculating Summary Statistics Using the Excel Analysis ToolPak Add-in:

1. Click the **Data** tab. Click **Data Analysis** in the **Analysis** group. The **Data Analysis** menu will open (see **Screen 3.12**).
2. Select **Descriptive Statistics**. Click **OK**. The **Descriptive Statistics** window will open (see **Screen 3.13**). Click in the **Input Range** box. Select the range where your data



Screen 3.13

are located. (Note: the easiest way to do this is to highlight the data with your mouse.) Select **Rows** or **Columns** to identify whether the data are grouped in rows or columns.

3. Select where you want Excel to place the output. You can select a specific range in the current spreadsheet, a new spreadsheet within the current Excel workbook, or a new Excel workbook.
4. Click **Summary Statistics**. Click **OK** (see Screen 3.14 for an example of the output).

	<i>Column1</i>
1	
2	
3	Mean 6.833333
4	Standard Error 1.777951
5	Median 6
6	Mode 6
7	Standard Deviation 4.355074
8	Sample Variance 18.96667
9	Kurtosis -1.36554
10	Skewness 0.495554
11	Range 11
12	Minimum 2
13	Maximum 13
14	Sum 41
15	Count 6

Screen 3.14

5. The **Analysis ToolPak** does not calculate the first and third quartiles. To do this, go to an empty cell in the spreadsheet. Then
  - a. Type **=quartile(**
  - b. Select the range of data and then type a comma
  - c. Type **1** for the first quartile or **3** for the third quartile
  - d. Type a right parenthesis, and then press **Enter**.
6. To find the *k*th percentile:
  - a. Type **=percentile(**
  - b. Select the range of data and then type a comma
  - c. Type the value of *k*
  - d. Type a right parenthesis, and then press **Enter**.

## TECHNOLOGY ASSIGNMENTS

**TA3.1** Refer to the Data Set IV that is on the Web site of this book, which contains results on the 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine.

- a. Calculate the mean, median, range, standard deviation, and interquartile range for the time variable.
- b. Take 10 random samples of 200 runners each, and calculate the statistics listed in part a for each of these samples. Discuss how the values of the sample statistics compare to the population parameters that you calculated in part a.

**TA3.2** Refer to the data on monthly rent given in *City Data* (Data Set I) on the Web site of this book. From that data set select the 4th value and then select every 10th value after that (i.e., select the 4th, 14th, 24th, 34th, . . . values). Such a sample taken from a population is called a *systematic random sample*. Find the mean, median, standard deviation, first quartile, and third quartile for the monthly rent for this subsample.

**TA3.3** Refer to *City Data* (Data Set I) on the prices of various products in different cities across the country. Select a subsample of the prices of regular unleaded gas for 40 cities. Find the mean, median, and standard deviation for the data of this subsample.

**TA3.4** Refer to the data of TA3.3. Make a box-and-whisker plot for those data.

**TA3.5** Refer to *City Data* (Data Set I) on the prices of various products in different cities across the country. Make a box-and-whisker plot for the data on the monthly rent.

**TA3.6** Using the data set *Billboard* on the Web site of this book, calculate the mean, median, range, standard deviation, and interquartile range, and create a boxplot of the number of weeks spent on the charts for the songs in the Billboard Hot 100 for the week of July 9, 2011. Discuss the features of the graph. Identify any outliers, and specify whether they are mild or extreme outliers. Now repeat the process for the songs ranked 1 through 50 and the songs ranked 51 through 100. Explain the differences and similarities between the two groups.

**TA3.7** Refer to Data Set VII on the stocks included in the Standard & Poor's 100 Index. Calculate the mean, median, standard deviation, and interquartile range for the data on the highest prices for the stocks in each of the market sectors. Compare the values of the various statistics for different sectors. Create a stacked dotplot of the highest prices for various sectors with each sector's data as one set of data. Explain how the results of your comparisons can be seen in the dotplot.

**TA3.8** Refer to Data Set III on the National Football League. Calculate the mean, median, standard deviation, and interquartile range for the players' ages separately for each of the position groups. Is there a position group that tends to have younger players, on average, than the other position groups? Is there a position group that tends to have less variability in the ages of the players?

**TA3.9** Calculate the five-number summaries, the values of the upper and lower inner fences, and the values of the upper and lower outer fences for the data referred to in TA3.8. Create side-by-side boxplots for the data on the position groups. Using these boxplots, compare the shapes of the age distributions for the positions. Are there any outliers? If so, classify the outliers as being mild or extreme.

**TA3.10** Using the data in the file *Motorcycle* on the Web site of this book, calculate the mean, median, range, standard deviation, and interquartile range, and create a boxplot of the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Discuss the features of the graph. Identify which counties are outliers, and specify whether they are mild or extreme outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?

**TA3.11** Using the data set *Kickers2010* on the Web site of this book, calculate the mean, median, standard deviation, and interquartile range of the longest field goals made during the 2010 NFL and Canadian Football League (CFL) seasons for the American Football Conference (AFC), the National Football Conference (NFC), and the CFL. Discuss the similarities and differences among the three groups.



© Robert Simon/iStockphoto

## Probability

### 4.1 Experiment, Outcome, and Sample Space

### 4.2 Calculating Probability

### 4.3 Marginal Probability, Conditional Probability, and Related Probability Concepts

### Case Study 4–1 Do You Worry About Your Weight?

### 4.4 Intersection of Events and the Multiplication Rule

### 4.5 Union of Events and the Addition Rule

### 4.6 Counting Rule, Factorials, Combinations, and Permutations

### Case Study 4–2 Probability of Winning a Mega Millions Lottery Jackpot

Do you worry about your weight? According to a Gallup poll, 48% of American adults worry at least some of the time (which means all or some of the time) about their weight. The poll showed that more women than men worry about their weight at least some of the time. In this poll, 55% of adult women and 41% of adult men said that they worry at least some of the time about their weight. (See Case Study 4–1.)

We often make statements about probability. For example, a weather forecaster may predict that there is an 80% chance of rain tomorrow. A health news reporter may state that a smoker has a much greater chance of getting cancer than does a nonsmoker. A college student may ask an instructor about the chances of passing a course or getting an A if he or she did not do well on the midterm examination.

Probability, which measures the likelihood that an event will occur, is an important part of statistics. It is the basis of inferential statistics, which will be introduced in later chapters. In inferential statistics, we make decisions under conditions of uncertainty. Probability theory is used to evaluate the uncertainty involved in those decisions. For example, estimating next year's sales for a company is based on many assumptions, some of which may happen to be true and others may not. Probability theory will help us make decisions under such conditions of imperfect information and uncertainty. Combining probability and probability distributions (which are discussed in Chapters 5 through 7) with descriptive statistics will help us make decisions about populations based on information obtained from samples. This chapter presents the basic concepts of probability and the rules for computing probability.

## 4.1 Experiment, Outcome, and Sample Space

Quality control inspector Jack Cook of Tennis Products Company picks up a tennis ball from the production line to check whether it is good or defective. Cook's act of inspecting a tennis ball is an example of a statistical **experiment**. The result of his inspection will be that the ball is either "good" or "defective." Each of these two observations is called an **outcome** (also called a *basic* or *final outcome*) of the experiment, and these outcomes taken together constitute the **sample space** for this experiment.

### Definition

**Experiment, Outcomes, and Sample Space** An *experiment* is a process that, when performed, results in one and only one of many observations. These observations are called the *outcomes* of the experiment. The collection of all outcomes for an experiment is called a *sample space*.

A sample space is denoted by  $S$ . The sample space for the example of inspecting a tennis ball is written as

$$S = \{\text{good, defective}\}$$

The elements of a sample space are called **sample points**.

Table 4.1 lists some examples of experiments, their outcomes, and their sample spaces.

**Table 4.1 Examples of Experiments, Outcomes, and Sample Spaces**

Experiment	Outcomes	Sample Space
Toss a coin once	Head, Tail	$S = \{\text{Head, Tail}\}$
Roll a die once	1, 2, 3, 4, 5, 6	$S = \{1, 2, 3, 4, 5, 6\}$
Toss a coin twice	$HH, HT, TH, TT$	$S = \{HH, HT, TH, TT\}$
Play lottery	Win, Lose	$S = \{\text{Win, Lose}\}$
Take a test	Pass, Fail	$S = \{\text{Pass, Fail}\}$
Select a worker	Male, Female	$S = \{\text{Male, Female}\}$

The sample space for an experiment can also be illustrated by drawing either a Venn diagram or a tree diagram. A **Venn diagram** is a picture (a closed geometric shape such as a rectangle, a square, or a circle) that depicts all the possible outcomes for an experiment. In a **tree diagram**, each outcome is represented by a branch of the tree. Venn and tree diagrams help us understand probability concepts by presenting them visually. Examples 4–1 through 4–3 describe how to draw these diagrams for statistical experiments.

### ■ EXAMPLE 4–1

Draw the Venn and tree diagrams for the experiment of tossing a coin once.

**Solution** This experiment has two possible outcomes: head and tail. Consequently, the sample space is given by

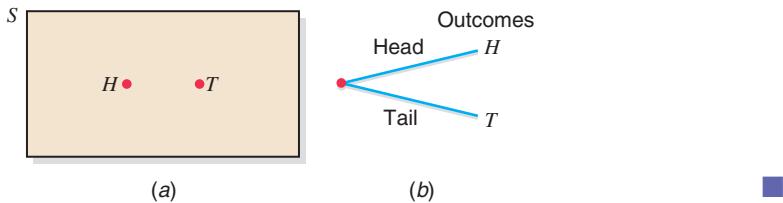
$$S = \{H, T\}, \quad \text{where } H = \text{Head} \quad \text{and} \quad T = \text{Tail}$$

To draw a Venn diagram for this example, we draw a rectangle and mark two points inside this rectangle that represent the two outcomes, Head and Tail. The rectangle is labeled  $S$  because it represents the sample space (see Figure 4.1a). To draw a tree diagram,

Drawing Venn and tree diagrams: one toss of a coin.

we draw two branches starting at the same point, one representing the head and the second representing the tail. The two final outcomes are listed at the ends of the branches (see Figure 4.1b).

**Figure 4.1** (a) Venn diagram and (b) tree diagram for one toss of a coin.



### ■ EXAMPLE 4–2

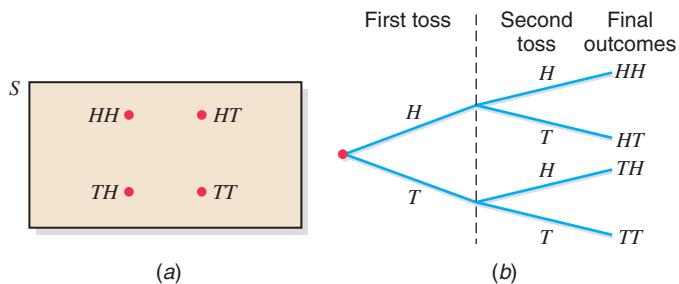
Drawing Venn and tree diagrams: two tosses of a coin.

Draw the Venn and tree diagrams for the experiment of tossing a coin twice.

**Solution** This experiment can be split into two parts: the first toss and the second toss. Suppose that the first time the coin is tossed, we obtain a head. Then, on the second toss, we can still obtain a head or a tail. This gives us two outcomes:  $HH$  (head on both tosses) and  $HT$  (head on the first toss and tail on the second toss). Now suppose that we observe a tail on the first toss. Again, either a head or a tail can occur on the second toss, giving the remaining two outcomes:  $TH$  (tail on the first toss and head on the second toss) and  $TT$  (tail on both tosses). Thus, the sample space for two tosses of a coin is

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

The Venn and tree diagrams are given in Figure 4.2. Both of these diagrams show the sample space for this experiment.



**Figure 4.2** (a) Venn diagram and (b) tree diagram for two tosses of a coin.

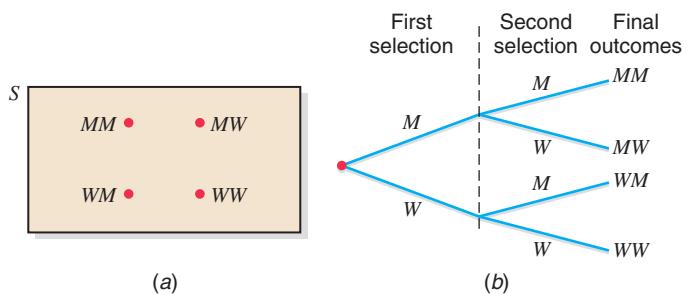
### ■ EXAMPLE 4–3

Drawing Venn and tree diagrams: two selections.

Suppose we randomly select two workers from a company and observe whether the worker selected each time is a man or a woman. Write all the outcomes for this experiment. Draw the Venn and tree diagrams for this experiment.

**Solution** Let us denote the selection of a man by  $M$  and that of a woman by  $W$ . We can compare the selection of two workers to two tosses of a coin. Just as each toss of a coin can result in one of two outcomes, head or tail, each selection from the workers of this company can result in one of two outcomes, man or woman. As we can see from the Venn and tree diagrams of Figure 4.3, there are four final outcomes:  $MM$ ,  $MW$ ,  $WM$ ,  $WW$ . Hence, the sample space is written as

$$S = \{MM, MW, WM, WW\}$$



**Figure 4.3** (a) Venn diagram and (b) tree diagram for selecting two workers.

### 4.1.1 Simple and Compound Events

An **event** consists of one or more of the outcomes of an experiment.

#### Definition

**Event** An *event* is a collection of one or more of the outcomes of an experiment.

An event may be a *simple event* or a *compound event*. A simple event is also called an *elementary event*, and a compound event is also called a *composite event*.

#### Simple Event

Each of the final outcomes for an experiment is called a **simple event**. In other words, a simple event includes one and only one outcome. Usually, simple events are denoted by  $E_1, E_2, E_3$ , and so forth. However, we can denote them by any other letter—that is, by  $A, B, C$ , and so forth. Many times we denote events by the same letter and use subscripts to distinguish them, as in  $A_1, A_2, A_3, \dots$ .

#### Definition

**Simple Event** An event that includes one and only one of the (final) outcomes for an experiment is called a *simple event* and is usually denoted by  $E_i$ .

Example 4–4 describes simple events.

### ■ EXAMPLE 4–4

Reconsider Example 4–3 on selecting two workers from a company and observing whether the worker selected each time is a man or a woman. Each of the final four outcomes ( $MM$ ,  $MW$ ,  $WM$ , and  $WW$ ) for this experiment is a simple event. These four events can be denoted by  $E_1, E_2, E_3$ , and  $E_4$ , respectively. Thus,

$$E_1 = \{MM\}, \quad E_2 = \{MW\}, \quad E_3 = \{WM\}, \quad \text{and} \quad E_4 = \{WW\}$$

Illustrating simple events.

#### Compound Event

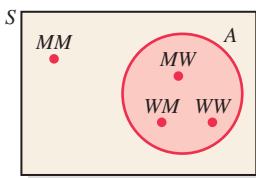
A **compound event** consists of more than one outcome.

#### Definition

**Compound Event** A *compound event* is a collection of more than one outcome for an experiment.

Compound events are denoted by  $A, B, C, D, \dots$  or by  $A_1, A_2, A_3, \dots, B_1, B_2, B_3, \dots$ , and so forth. Examples 4–5 and 4–6 describe compound events.

**Illustrating a compound event:  
two selections.**



**Figure 4.4** Venn diagram for event A.

**Illustrating simple and compound events: two selections.**

### ■ EXAMPLE 4–5

Reconsider Example 4–3 on selecting two workers from a company and observing whether the worker selected each time is a man or a woman. Let A be the event that at most one man is selected. Is event A a simple or compound event?

**Solution** Event A will occur if either no man or one man is selected. Hence, the event A is given by

$$A = \{MW, WM, WW\}$$

Because event A contains more than one outcome, it is a compound event. The Venn diagram in Figure 4.4 gives a graphic presentation of compound event A. ■

### ■ EXAMPLE 4–6

In a group of people, some are in favor of genetic engineering and others are against it. Two persons are selected at random from this group and asked whether they are in favor of or against genetic engineering. How many distinct outcomes are possible? Draw a Venn diagram and a tree diagram for this experiment. List all the outcomes included in each of the following events and state whether they are simple or compound events.

- (a) Both persons are in favor of genetic engineering.
- (b) At most one person is against genetic engineering.
- (c) Exactly one person is in favor of genetic engineering.

**Solution** Let F denote an event that a person is in favor of genetic engineering, and A denote an event that a person is against genetic engineering.

This experiment has the following four outcomes:

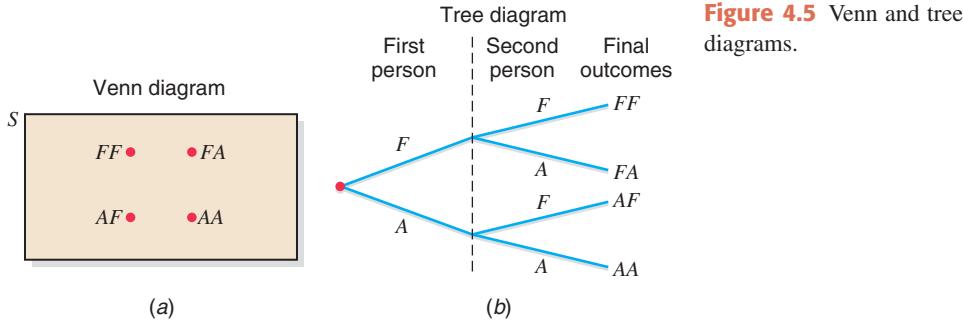
$FF$  = both persons are in favor of genetic engineering

$FA$  = the first person is in favor and the second is against

$AF$  = the first person is against and the second is in favor

$AA$  = both persons are against genetic engineering

The Venn and tree diagrams in Figure 4.5 show these four outcomes.



- (a) The event “both persons are in favor of genetic engineering” will occur if  $FF$  is obtained. Thus,

Both persons are in favor of genetic engineering =  $\{FF\}$

Because this event includes only one of the final four outcomes, it is a **simple** event.

- (b) The event “at most one person is against genetic engineering” will occur if either none or one of the persons selected is against genetic engineering. Consequently,

At most one person is against genetic engineering =  $\{FF, FA, AF\}$

Because this event includes more than one outcome, it is a **compound** event.

- (c) The event “exactly one person is in favor of genetic engineering” will occur if one of the two persons selected is in favor and the other is against genetic engineering. Hence, it includes the following two outcomes:

Exactly one person is in favor of genetic engineering = {FA, AF}

Because this event includes more than one outcome, it is a **compound** event. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

- 4.1** Define the following terms: *experiment, outcome, sample space, simple event, and compound event.*
- 4.2** List the simple events for each of the following statistical experiments in a sample space  $S$ .
- One roll of a die
  - Three tosses of a coin
  - One toss of a coin and one roll of a die
- 4.3** A box contains three items that are labeled A, B, and C. Two items are selected at random (without replacement) from this box. List all the possible outcomes for this experiment. Write the sample space  $S$ .

### APPLICATIONS

- 4.4** Two students are randomly selected from a statistics class, and it is observed whether or not they suffer from math anxiety. How many total outcomes are possible? Draw a tree diagram for this experiment. Draw a Venn diagram.
- 4.5** In a group of adults, some own iPads, and others do not. If two adults are randomly selected from this group, how many total outcomes are possible? Draw a tree diagram for this experiment.
- 4.6** An automated teller machine at a local bank is stocked with \$10 and \$20 bills. When a customer withdraws \$40 from the machine, it dispenses either two \$20 bills or four \$10 bills. If two customers withdraw \$40 each, how many outcomes are possible? Show all these outcomes in a Venn diagram, and draw a tree diagram for this experiment.
- 4.7** A box contains a certain number of computer parts, a few of which are defective. Two parts are selected at random from this box and inspected to determine if they are good or defective. How many total outcomes are possible? Draw a tree diagram for this experiment.
- 4.8** In a group of people, some are in favor of a tax increase on rich people to reduce the federal deficit and others are against it. (Assume that there is no other outcome such as “no opinion” and “do not know.”) Three persons are selected at random from this group and their opinions in favor or against raising such taxes are noted. How many total outcomes are possible? Write these outcomes in a sample space  $S$ . Draw a tree diagram for this experiment.
- 4.9** Draw a tree diagram for three tosses of a coin. List all outcomes for this experiment in a sample space  $S$ .
- 4.10** Refer to Exercise 4.4. List all the outcomes included in each of the following events. Indicate which are simple and which are compound events.
- Both students suffer from math anxiety.
  - Exactly one student suffers from math anxiety.
  - The first student does not suffer and the second suffers from math anxiety.
  - None of the students suffers from math anxiety.
- 4.11** Refer to Exercise 4.5. List all the outcomes included in each of the following events. Indicate which are simple and which are compound events.
- One person has an iPad and the other does not.
  - At least one person has an iPad.
  - Not more than one person has an iPad.
  - The first person has an iPad and the second does not.
- 4.12** Refer to Exercise 4.6. List all of the outcomes in each of the following events and mention which of these are simple and which are compound events.
- Exactly one customer receives \$20 bills.
  - Both customers receive \$10 bills.
  - At most one customer receives \$20 bills.
  - The first customer receives \$10 bills and the second receives \$20 bills.

**4.13** Refer to Exercise 4.7. List all the outcomes included in each of the following events. Indicate which are simple and which are compound events.

- At least one part is good.
- Exactly one part is defective.
- The first part is good and the second is defective.
- At most one part is good.

**4.14** Refer to Exercise 4.8. List all the outcomes included in each of the following events and mention which are simple and which are compound events.

- At most one person is against a tax increase on rich people.
- Exactly two persons are in favor of a tax increase on rich people.
- At least one person is against a tax increase on rich people.
- More than one person is against a tax increase on rich people.

## 4.2 Calculating Probability

**Probability**, which gives the likelihood of occurrence of an event, is denoted by  $P$ . The probability that a simple event  $E_i$  will occur is denoted by  $P(E_i)$ , and the probability that a compound event  $A$  will occur is denoted by  $P(A)$ .

### Definition

**Probability** *Probability* is a numerical measure of the likelihood that a specific event will occur.

#### Two Properties of Probability ►

##### 1. The probability of an event always lies in the range 0 to 1.

Whether it is a simple or a compound event, the probability of an event is never less than 0 or greater than 1. Using mathematical notation, we can write this property as follows.

#### First Property of Probability

$$0 \leq P(E_i) \leq 1$$

$$0 \leq P(A) \leq 1$$

An event that cannot occur has zero probability; such an event is called an **impossible (or null) event**. An event that is certain to occur has a probability equal to 1 and is called a **sure (or certain) event**. That is,

For an impossible event  $M$ :  $P(M) = 0$

For a sure event  $C$ :  $P(C) = 1$

##### 2. The sum of the probabilities of all simple events (or final outcomes) for an experiment, denoted by $\sum P(E_i)$ , is always 1.

#### Second Property of Probability

For an experiment,

$$\sum P(E_i) = P(E_1) + P(E_2) + P(E_3) + \dots = 1$$

From this property, for the experiment of one toss of a coin,

$$P(H) + P(T) = 1$$

For the experiment of two tosses of a coin,

$$P(HH) + P(HT) + P(TH) + P(TT) = 1$$

For one game of football by a professional team,

$$P(\text{win}) + P(\text{loss}) + P(\text{tie}) = 1$$

## 4.2.1 Three Conceptual Approaches to Probability

The three conceptual approaches to probability are (1) classical probability, (2) the relative frequency concept of probability, and (3) the subjective probability concept. These three concepts are explained next.

### Classical Probability

Many times, various outcomes for an experiment may have the same probability of occurrence. Such outcomes are called **equally likely outcomes**. The classical probability rule is applied to compute the probabilities of events for an experiment for which all outcomes are equally likely.

#### Definition

**Equally Likely Outcomes** Two or more outcomes that have the same probability of occurrence are said to be *equally likely outcomes*.

According to the **classical probability rule**, the probability of a simple event is equal to 1 divided by the total number of outcomes for the experiment. This is obvious because the sum of the probabilities of all final outcomes for an experiment is 1, and all the final outcomes are equally likely. In contrast, the probability of a compound event  $A$  is equal to the number of outcomes favorable to event  $A$  divided by the total number of outcomes for the experiment.

**Classical Probability Rule to Find Probability** Given that  $E_i$  is a simple event and  $A$  is a compound event:

$$P(E_i) = \frac{1}{\text{Total number of outcomes for the experiment}}$$

$$P(A) = \frac{\text{Number of outcomes favorable to } A}{\text{Total number of outcomes for the experiment}}$$

Examples 4–7 through 4–9 illustrate how probabilities of events are calculated using the classical probability rule.

### ■ EXAMPLE 4–7

Find the probability of obtaining a head and the probability of obtaining a tail for one toss of a coin.

*Calculating the probability of a simple event.*

**Solution** The two outcomes, head and tail, are equally likely outcomes. Therefore,<sup>1</sup>

$$P(\text{head}) = \frac{1}{\text{Total number of outcomes}} = \frac{1}{2} = .50$$

Similarly,

$$P(\text{tail}) = \frac{1}{2} = .50$$

### ■ EXAMPLE 4–8

Find the probability of obtaining an even number in one roll of a die.

*Calculating the probability of a compound event.*

<sup>1</sup>If the final answer for the probability of an event does not terminate within four decimal places, usually it is rounded to four decimal places.

**Solution** This experiment of rolling a die once has a total of six outcomes: 1, 2, 3, 4, 5, and 6. All these outcomes are equally likely. Let  $A$  be an event that an even number is observed on the die. Event  $A$  includes three outcomes: 2, 4, and 6; that is,

$$A = \{2, 4, 6\}$$

If any one of these three numbers is obtained, event  $A$  is said to occur. Hence,

$$P(A) = \frac{\text{Number of outcomes included in } A}{\text{Total number of outcomes}} = \frac{3}{6} = .50$$

### ■ EXAMPLE 4–9

*Calculating the probability of a compound event.*

In a group of 500 women, 120 have played golf at least once. Suppose one of these 500 women is randomly selected. What is the probability that she has played golf at least once?

**Solution** Because the selection is to be made randomly, each of the 500 women has the same probability of being selected. Consequently this experiment has a total of 500 equally likely outcomes. One hundred twenty of these 500 outcomes are included in the event that the selected woman has played golf at least once. Hence,

$$P(\text{selected woman has played golf at least once}) = \frac{120}{500} = .24$$

### Relative Frequency Concept of Probability

Suppose we want to calculate the following probabilities:

1. The probability that the next car that comes out of an auto factory is a “lemon”
2. The probability that a randomly selected family owns a home
3. The probability that a randomly selected woman is an excellent driver
4. The probability that an 80-year-old person will live for at least 1 more year
5. The probability that a randomly selected adult is in favor of increasing taxes to reduce the national debt
6. The probability that a randomly selected person owns a sport-utility vehicle (SUV)

These probabilities cannot be computed using the classical probability rule because the various outcomes for the corresponding experiments are not equally likely. For example, the next car manufactured at an auto factory may or may not be a lemon. The two outcomes, “it is a lemon” and “it is not a lemon,” are not equally likely. If they were, then (approximately) half the cars manufactured by this company would be lemons, and this might prove disastrous to the survival of the firm.

Although the various outcomes for each of these experiments are not equally likely, each of these experiments can be performed again and again to generate data. In such cases, to calculate probabilities, we either use past data or generate new data by performing the experiment a large number of times. The relative frequency of an event is used as an approximation for the probability of that event. This method of assigning a probability to an event is called the **relative frequency concept of probability**. Because relative frequencies are determined by performing an experiment, the probabilities calculated using relative frequencies may change almost each time an experiment is repeated. For example, every time a new sample of 500 cars is selected from the production line of an auto factory, the number of lemons in those 500 cars is expected to be different. However, the variation in the percentage of lemons will be small if the sample size is large. Note that if we are considering the population, the relative frequency will give an exact probability.

**Using Relative Frequency as an Approximation of Probability** If an experiment is repeated  $n$  times and an event  $A$  is observed  $f$  times, then, according to the relative frequency concept of probability,

$$P(A) = \frac{f}{n}$$

Examples 4–10 and 4–11 illustrate how the probabilities of events are approximated using the relative frequencies.

## ■ EXAMPLE 4–10

Ten of the 500 randomly selected cars manufactured at a certain auto factory are found to be lemons. Assuming that the lemons are manufactured randomly, what is the probability that the next car manufactured at this auto factory is a lemon?

*Approximating probability by relative frequency: sample data.*

**Solution** Let  $n$  denote the total number of cars in the sample and  $f$  the number of lemons in  $n$ . Then,

$$n = 500 \quad \text{and} \quad f = 10$$

Using the relative frequency concept of probability, we obtain

$$P(\text{next car is a lemon}) = \frac{f}{n} = \frac{10}{500} = .02$$

This probability is actually the relative frequency of lemons in 500 cars. Table 4.2 lists the frequency and relative frequency distributions for this example.

**Table 4.2 Frequency and Relative Frequency Distributions for the Sample of Cars**

Car	$f$	Relative Frequency
Good	490	490/500 = .98
Lemon	10	10/500 = .02
	$n = 500$	Sum = 1.00

The column of relative frequencies in Table 4.2 is used as the column of approximate probabilities. Thus, from the relative frequency column,

$$\begin{aligned} P(\text{next car is a lemon}) &= .02 \\ P(\text{next car is a good car}) &= .98 \end{aligned}$$

**Note that relative frequencies are not exact probabilities but are approximate probabilities unless they are based on a census.** However, if the experiment is repeated again and again, this approximate probability of an outcome obtained from the relative frequency will approach the actual probability of that outcome. This is called the **Law of Large Numbers**.

### Definition

**Law of Large Numbers** If an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability.

## ■ EXAMPLE 4–11

Allison wants to determine the probability that a randomly selected family from New York State owns a home. How can she determine this probability?

*Approximating probability by relative frequency.*

**Solution** There are two outcomes for a randomly selected family from New York State: “This family owns a home” and “This family does not own a home.” These two outcomes are not equally likely. (Note that these two outcomes will be equally likely if exactly half of the families in New York State own homes and exactly half do not own homes.) Hence, the classical probability rule cannot be applied. However, we can repeat this experiment again and again. In other words, we can select a sample of families from New York State and observe whether or not each of them owns a home. Hence, we will use the relative frequency approach to probability.

Suppose Allison selects a random sample of 1000 families from New York State and observes that 730 of them own homes and 270 do not own homes. Then,

$$n = \text{sample size} = 1000$$

$$f = \text{number of families who own homes} = 730$$

Consequently,

$$P(\text{a randomly selected family owns a home}) = \frac{f}{n} = \frac{730}{1000} = .730$$

Again, note that .730 is just an approximation of the probability that a randomly selected family from New York State owns a home. Every time Allison repeats this experiment she may obtain a different probability for this event. However, because the sample size ( $n = 1000$ ) in this example is large, the variation is expected to be relatively small. ■

### Subjective Probability

Many times we face experiments that neither have equally likely outcomes nor can be repeated to generate data. In such cases, we cannot compute the probabilities of events using the classical probability rule or the relative frequency concept. For example, consider the following probabilities of events:

1. The probability that Carol, who is taking a statistics course, will earn an A in the course
2. The probability that the Dow Jones Industrial Average will be higher at the end of the next trading day
3. The probability that the New York Giants will win the Super Bowl next season
4. The probability that Joe will lose the lawsuit he has filed against his landlord

Neither the classical probability rule nor the relative frequency concept of probability can be applied to calculate probabilities for these examples. All these examples belong to experiments that have neither equally likely outcomes nor the potential of being repeated. For example, Carol, who is taking statistics, will take the test (or tests) only once, and based on that she will either earn an A or not. The two events “she will earn an A” and “she will not earn an A” are not equally likely. The probability assigned to an event in such cases is called **subjective probability**. It is based on the individual’s judgment, experience, information, and belief. Carol may assign a high probability to the event that she will earn an A in statistics, whereas her instructor may assign a low probability to the same event.

#### Definition

**Subjective Probability** *Subjective probability* is the probability assigned to an event based on subjective judgment, experience, information, and belief.

Subjective probability is assigned arbitrarily. It is usually influenced by the biases, preferences, and experience of the person assigning the probability.

## EXERCISES

### CONCEPTS AND PROCEDURES

**4.15** Briefly explain the two properties of probability.

**4.16** Briefly describe an impossible event and a sure event. What is the probability of the occurrence of each of these two events?

**4.17** Briefly explain the three approaches to probability. Give one example of each approach.

**4.18** Briefly explain for what kind of experiments we use the classical approach to calculate probabilities of events and for what kind of experiments we use the relative frequency approach.

**4.19** Which of the following values cannot be the probability of an event and why?

2.4       $\frac{3}{8}$        $-.63$       .55       $\frac{9}{4}$        $-\frac{2}{9}$       1.0       $\frac{12}{17}$

**4.20** Which of the following values cannot be the probability of an event and why?

.67      0.0       $\frac{32}{88}$        $-.1.6$        $\frac{8}{13}$       4.8       $-.3$        $-\frac{3}{4}$

## ■ APPLICATIONS

**4.21** Suppose a randomly selected passenger is about to go through the metal detector at JFK Airport in New York City. Consider the following two outcomes: The passenger sets off the metal detector, and the passenger does not set off the metal detector. Are these two outcomes equally likely? Explain why or why not. If you are to find the probability of these two outcomes, would you use the classical approach or the relative frequency approach? Explain why.

**4.22** Fifty-six people have signed up for a karaoke contest at a local nightclub. Of them, 19 sang in a band, chorus, or choir while in high school and 37 did not. Suppose one contestant is chosen at random. Consider the following two events: The selected contestant sang in a band, chorus, or choir while in high school, and the selected contestant did not sing in a band, chorus, or choir while in high school. If you are to find the probabilities of these two events, would you use the classical approach or the relative frequency approach? Explain why.

**4.23** The president of a company has a hunch that there is a .80 probability that the company will be successful in marketing a new brand of ice cream. Is this a case of classical, relative frequency, or subjective probability? Explain why.

**4.24** A financial expert believes that the probability is .13 that the stock price of a specific technology company will double over the next year. Is this a case of classical, relative frequency, or subjective probability? Explain why.

**4.25** A hat contains 40 marbles. Of them, 18 are red and 22 are green. If one marble is randomly selected out of this hat, what is the probability that this marble is

- a. red?
- b. green?

**4.26** A die is rolled once. What is the probability that

- a. a number less than 5 is obtained?
- b. a number 3 to 6 is obtained?

**4.27** A random sample of 2000 adults showed that 1320 of them have shopped at least once on the Internet. What is the (approximate) probability that a randomly selected adult has shopped on the Internet?

**4.28** In a statistics class of 42 students, 28 have volunteered for community service in the past. Find the probability that a randomly selected student from this class has volunteered for community service in the past.

**4.29** In a group of 50 car owners, 8 own hybrid cars. If one car owner is selected at random from this group, what is the probability that this car owner owns a hybrid car?

**4.30** Out of the 3000 families who live in a given apartment complex in New York City, 600 paid no income tax last year. What is the probability that a randomly selected family from these 3000 families did pay income tax last year?

**4.31** The television game show *The Price Is Right* has a game called the Shell Game. The game has four shells, and one of these four shells has a ball under it. The contestant chooses one shell. If this shell contains the ball, the contestant wins. If a contestant chooses one shell randomly, what is the probability of each of the following outcomes?

- a. contestant wins
- b. contestant loses

Do these probabilities add up to 1.0? If yes, why?

**4.32** There are 1265 eligible voters in a town, and 972 of them are registered to vote. If one eligible voter is selected at random from this town, what is the probability that this voter is

- a. registered?
- b. not registered?

Do these two probabilities add up to 1.0? If yes, why?

**4.33** According to an article in *The Sacramento Bee* ([www.sacbee.com/2011/08/04/3816872/medicare-prescription-premiums.html](http://www.sacbee.com/2011/08/04/3816872/medicare-prescription-premiums.html)), approximately 10% of Medicare beneficiaries lack a prescription drug care plan. Suppose that a town in Florida has 2384 residents who are Medicare beneficiaries, and 216 of them do not have a prescription drug care plan. If one of the Medicare beneficiaries is chosen at random from this town, what is the probability that this person has a prescription drug care plan? What is the probability that this person does not have a prescription drug care plan? Do these probabilities add up to 1.0? If yes, why? If no, why not?

**4.34** A sample of 500 large companies showed that 120 of them offer free psychiatric help to their employees who suffer from psychological problems. If one company is selected at random from this sample, what is the probability that this company offers free psychiatric help to its employees who suffer from psychological problems? What is the probability that this company does not offer free psychiatric help to its employees who suffer from psychological problems? Do these two probabilities add up to 1.0? If yes, why?

**4.35** A sample of 400 large companies showed that 130 of them offer free health fitness centers to their employees on the company premises. If one company is selected at random from this sample, what is the probability that this company offers a free health fitness center to its employees on the company premises? What is the probability that this company does not offer a free health fitness center to its employees on the company premises? Do these two probabilities add up to 1.0? If yes, why?

**4.36** In a large city, 15,000 workers lost their jobs last year. Of them, 7400 lost their jobs because their companies closed down or moved, 4600 lost their jobs due to insufficient work, and the remainder lost their jobs because their positions were abolished. If one of these 15,000 workers is selected at random, find the probability that this worker lost his or her job

- a. because the company closed down or moved
- b. due to insufficient work
- c. because the position was abolished

Do these probabilities add up to 1.0? If so, why?

**4.37** Many colleges require students to take a placement exam to determine which math courses they are eligible to take during the first semester of their freshman year. Of the 2938 freshmen at a local state college, 214 were required to take a remedial math course, 1465 could take a nonremedial, non-calculus-based math course, and 1259 could take a calculus-based math course. If one of these freshmen is selected at random, find the probability that this student could take

- a. a calculus-based math course
- b. a nonremedial, non-calculus-based math course
- c. a remedial math course

Do these probabilities add up to 1.0? If so, why?

**4.38** In a sample of 500 families, 70 have a yearly income of less than \$40,000, 220 have a yearly income of \$40,000 to \$80,000, and the remaining families have a yearly income of more than \$80,000. Write the frequency distribution table for this problem. Calculate the relative frequencies for all classes. Suppose one family is randomly selected from these 500 families. Find the probability that this family has a yearly income of

- a. less than \$40,000
- b. more than \$80,000

**4.39** Suppose you want to find the (approximate) probability that a randomly selected family from Los Angeles earns more than \$175,000 a year. How would you find this probability? What procedure would you use? Explain briefly.

**4.40** Suppose you have a loaded die and you want to find the (approximate) probabilities of different outcomes for this die. How would you find these probabilities? What procedure would you use? Explain briefly.

## 4.3 Marginal Probability, Conditional Probability, and Related Probability Concepts

In this section first we discuss marginal and conditional probabilities, and then we discuss the concepts (in that order) of mutually exclusive events, independent and dependent events, and complementary events.

### 4.3.1 Marginal and Conditional Probabilities

Suppose all 100 employees of a company were asked whether they are in favor of or against paying high salaries to CEOs of U.S. companies. Table 4.3 gives a two-way classification of the responses of these 100 employees. Assume that every employee responds either *in favor* or *against*.

**Table 4.3** Two-Way Classification of Employee Responses

	In Favor	Against
Male	15	45
Female	4	36

Table 4.3 shows the distribution of 100 employees based on two variables or characteristics: gender (male or female) and opinion (in favor or against). Such a table is called a *contingency table*. In Table 4.3, each box that contains a number is called a *cell*. Notice that there are four cells. Each cell gives the frequency for two characteristics. For example, 15 employees in this group possess two characteristics: “male” and “in favor of paying high salaries to CEOs.” We can interpret the numbers in other cells the same way.

By adding the row totals and the column totals to Table 4.3, we write Table 4.4.

**Table 4.4** Two-Way Classification of Employee Responses with Totals

	In Favor	Against	Total
Male	15	45	60
Female	4	36	40
Total	19	81	100

Suppose one employee is selected at random from these 100 employees. This employee may be classified either on the basis of gender alone or on the basis of opinion. If only one characteristic is considered at a time, the employee selected can be a male, a female, in favor, or against. The probability of each of these four characteristics or events is called **marginal probability** or *simple probability*. These probabilities are called marginal probabilities because they are calculated by dividing the corresponding row margins (totals for the rows) or column margins (totals for the columns) by the grand total.

#### Definition

**Marginal Probability** *Marginal probability* is the probability of a single event without consideration of any other event. Marginal probability is also called *simple probability*.

For Table 4.4, the four marginal probabilities are calculated as follows:

$$P(\text{male}) = \frac{\text{Number of males}}{\text{Total number of employees}} = \frac{60}{100} = .60$$

As we can observe, the probability that a male will be selected is obtained by dividing the total of the row labeled “Male” (60) by the grand total (100). Similarly,

$$P(\text{female}) = 40/100 = .40$$

$$P(\text{in favor}) = 19/100 = .19$$

$$P(\text{against}) = 81/100 = .81$$

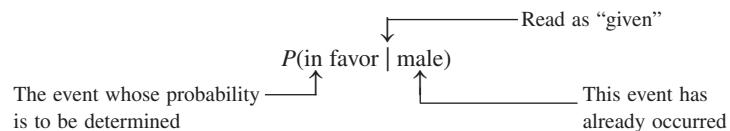
These four marginal probabilities are shown along the right side and along the bottom of Table 4.5.

**Table 4.5** Listing the Marginal Probabilities

	In Favor (A)	Against (B)	Total
Male ( $M$ )	15	45	60
Female ( $F$ )	4	36	40
Total	19	81	100

$P(A) = 19/100 = .19$        $P(B) = 81/100 = .81$

Now suppose that one employee is selected at random from these 100 employees. Furthermore, assume it is known that this (selected) employee is a male. In other words, the event that the employee selected is a male has already occurred. What is the probability that the employee selected is in favor of paying high salaries to CEOs? This probability is written as follows:



This probability,  $P(\text{in favor} \mid \text{male})$ , is called the **conditional probability** of “in favor” given that the event “male” has already happened. It is read as “the probability that the employee selected is in favor given that this employee is a male.”

## Definition

**Conditional Probability** *Conditional probability* is the probability that an event will occur given that another event has already occurred. If  $A$  and  $B$  are two events, then the conditional probability of  $A$  given  $B$  is written as

$$P(A \mid B)$$

and read as “the probability of  $A$  given that  $B$  has already occurred.”

## ■ EXAMPLE 4-12

Compute the conditional probability  $P(\text{in favor} | \text{male})$  for the data on 100 employees given in Table 4.4.

*Calculating the conditional probability: two-way table.*

**Solution** The probability  $P(\text{in favor} \mid \text{male})$  is the conditional probability that a randomly selected employee is in favor given that this employee is a male. It is known that the event “male” has already occurred. Based on the information that the employee selected is a male, we can infer that the employee selected must be one of the 60 males and, hence, must belong to the first row of Table 4.4. Therefore, we are concerned only with the first row of that table.

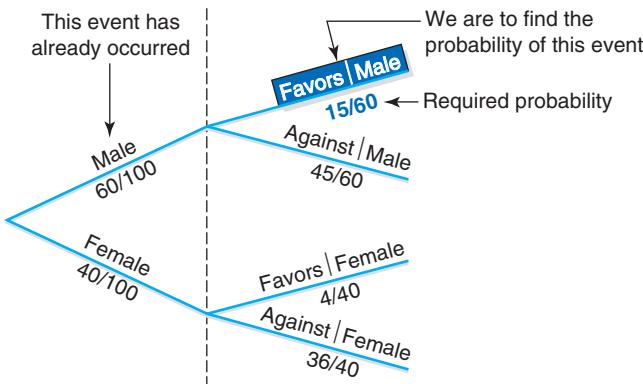
	<b>In Favor</b>	<b>Against</b>	<b>Total</b>
<b>Male</b>	15	45	60
	↑ Males who are in favor		↑ Total number of males

The required conditional probability is calculated as follows:

$$P(\text{in favor} \mid \text{male}) = \frac{\text{Number of males who are in favor}}{\text{Total number of males}} = \frac{15}{60} = .25$$

As we can observe from this computation of conditional probability, the total number of males (the event that has already occurred) is written in the denominator and the number of males who are in favor (the event whose probability we are to find) is written in the numerator. Note that we are considering the row of the event that has already occurred. The tree diagram in Figure 4.6 illustrates this example.

**Figure 4.6** Tree diagram.



*Calculating the conditional probability: two-way table.*

### ■ EXAMPLE 4-13

For the data of Table 4.4, calculate the conditional probability that a randomly selected employee is a female given that this employee is in favor of paying high salaries to CEOs.

**Solution** We are to compute the probability  $P(\text{female} \mid \text{in favor})$ . Because it is known that the employee selected is in favor of paying high salaries to CEOs, this employee must belong to the first column (the column labeled “in favor”) and must be one of the 19 employees who are in favor.

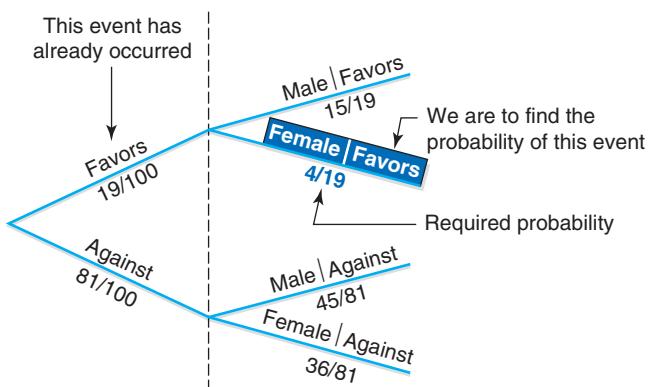
$$\begin{array}{r}
 \text{In Favor} \\
 \hline
 15 \\
 4 \leftarrow \text{Females who are in favor} \\
 \hline
 19 \leftarrow \text{Total number of employees who are in favor}
 \end{array}$$

Hence, the required probability is

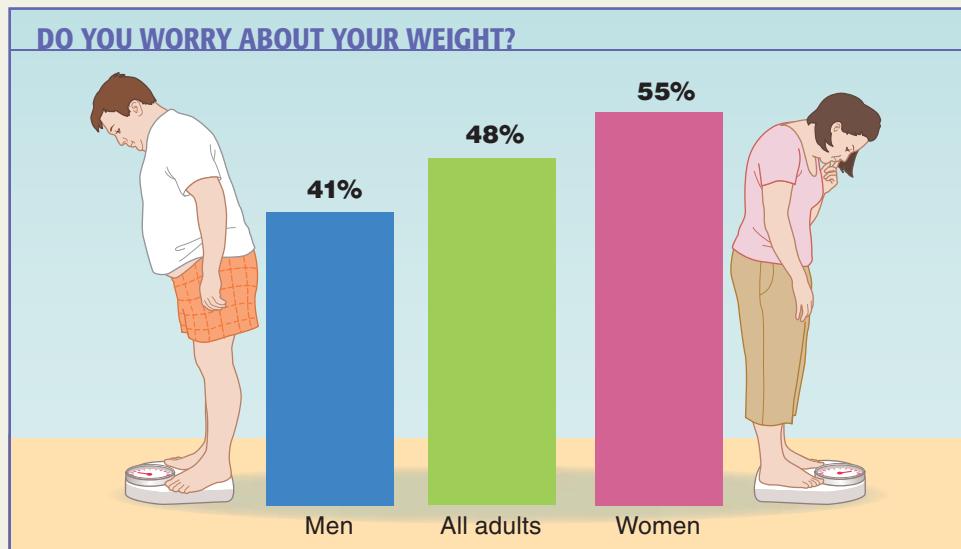
$$P(\text{female} \mid \text{in favor}) = \frac{\text{Number of females who are in favor}}{\text{Total number of employees who are in favor}} = \frac{4}{19} = .2105$$

The tree diagram in Figure 4.7 illustrates this example.

**Figure 4.7** Tree diagram.



## DO YOU WORRY ABOUT YOUR WEIGHT?



Data source: Gallup poll of 1014 adults aged 18 and older conducted July 9–12, 2012.

A Gallup poll of 1014 American adults of age 18 years and older conducted July 9–12, 2012, asked them, “How often do you worry about your weight?” The accompanying chart shows the percentage of adults included in the poll who said that they worry at least some of the time (which means all or some of the time) about their weight. According to this information, 48% of the adults in the sample said that they worry at least some of the time about their weight. When broken down based on gender, this percentage is 41% for men and 55% for women.

Assume that these percentages are true for the current population of American adults. Suppose we randomly select one American adult. Based on the overall percentage, the probability that this adult worries at least some of the time about his/her weight is

$$P(\text{a randomly selected adult worries at least some of the time about his/her weight}) = .48$$

This is a marginal probability because there is no condition imposed here.

Now suppose we randomly select one American adult. Then, given that this adult is a man, the probability is .41 that he worries at least some of the time about his weight. If the selected adult is a woman, the probability is .55 that she worries at least some of the time about her weight. These are two conditional probabilities, which can be written as follows:

$$P(\text{a randomly selected adult worries at least some of the time about his/her weight} | \text{man}) = .41$$

$$P(\text{a randomly selected adult worries at least some of the time about his/her weight} | \text{woman}) = .55$$

Note that these are approximate probabilities because the percentages given in the chart are based on a sample survey of 1014 adults.

Source: <http://www.gallup.com/poll/155903/Gender-Gap-Personal-Weight-Worries-Narrows.aspx>

### 4.3.2 Mutually Exclusive Events

Events that cannot occur together are called **mutually exclusive events**. Such events do not have any common outcomes. If two or more events are mutually exclusive, then at most one of them will occur every time we repeat the experiment. Thus the occurrence of one event excludes the occurrence of the other event or events.

**Definition**

**Mutually Exclusive Events** Events that cannot occur together are said to be *mutually exclusive events*.

For any experiment, the final outcomes are always mutually exclusive because one and only one of these outcomes is expected to occur in one repetition of the experiment. For example, consider tossing a coin twice. This experiment has four outcomes: *HH*, *HT*, *TH*, and *TT*. These outcomes are mutually exclusive because one and only one of them will occur when we toss this coin twice.

**■ EXAMPLE 4–14**

Consider the following events for one roll of a die:

$$A = \text{an even number is observed} = \{2, 4, 6\}$$

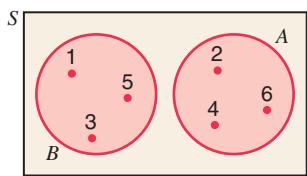
$$B = \text{an odd number is observed} = \{1, 3, 5\}$$

$$C = \text{a number less than } 5 \text{ is observed} = \{1, 2, 3, 4\}$$

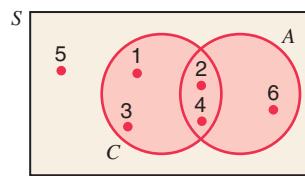
Illustrating mutually exclusive and mutually nonexclusive events.

Are events *A* and *B* mutually exclusive? Are events *A* and *C* mutually exclusive?

**Solution** Figures 4.8 and 4.9 show the Venn diagrams of events *A* and *B* and events *A* and *C*, respectively.



**Figure 4.8** Mutually exclusive events *A* and *B*.



**Figure 4.9** Mutually nonexclusive events *A* and *C*.

As we can observe from the definitions of events *A* and *B* and from Figure 4.8, events *A* and *B* have no common element. For one roll of a die, only one of the two events *A* and *B* can happen. Hence, these are two mutually exclusive events. We can observe from the definitions of events *A* and *C* and from Figure 4.9 that events *A* and *C* have two common outcomes: 2-spot and 4-spot. Thus, if we roll a die and obtain either a 2-spot or a 4-spot, then *A* and *C* happen at the same time. Hence, events *A* and *C* are not mutually exclusive. ■

**■ EXAMPLE 4–15**

Consider the following two events for a randomly selected adult:

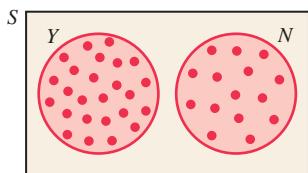
$$Y = \text{this adult has shopped on the Internet at least once}$$

$$N = \text{this adult has never shopped on the Internet}$$

Illustrating mutually exclusive events.

Are events *Y* and *N* mutually exclusive?

**Solution** Note that event *Y* consists of all adults who have shopped on the Internet at least once, and event *N* includes all adults who have never shopped on the Internet. These two events are illustrated in the Venn diagram in Figure 4.10.



**Figure 4.10** Mutually exclusive events *Y* and *N*.

As we can observe from the definitions of events  $Y$  and  $N$  and from Figure 4.10, events  $Y$  and  $N$  have no common outcome. They represent two distinct sets of adults: the ones who have shopped on the Internet at least once and the ones who have never shopped on the Internet. Hence, these two events are mutually exclusive. ■

### 4.3.3 Independent versus Dependent Events

In the case of two **independent events**, the occurrence of one event does not change the probability of the occurrence of the other event.

#### Definition

**Independent Events** Two events are said to be *independent* if the occurrence of one does not affect the probability of the occurrence of the other. In other words,  $A$  and  $B$  are *independent events* if

$$\text{either } P(A | B) = P(A) \text{ or } P(B | A) = P(B)$$

It can be shown that if one of these two conditions is true, then the second will also be true, and if one is not true, then the second will also not be true.

If the occurrence of one event affects the probability of the occurrence of the other event, then the two events are said to be **dependent events**. In probability notation, the two events are dependent if either  $P(A | B) \neq P(A)$  or  $P(B | A) \neq P(B)$ .

### ■ EXAMPLE 4–16

Illustrating two dependent events: two-way table.

Refer to the information on 100 employees given in Table 4.4 in Section 4.3.1. Are events “female ( $F$ )” and “in favor ( $A$ )” independent?

**Solution** Events  $F$  and  $A$  will be independent if

$$P(F) = P(F | A)$$

Otherwise they will be dependent.

Using the information given in Table 4.4, we compute the following two probabilities:

$$P(F) = 40/100 = .40 \quad \text{and} \quad P(F | A) = 4/19 = .2105$$

Because these two probabilities are not equal, the two events are dependent. Here, dependence of events means that the percentages of males who are in favor of and against paying high salaries to CEOs are different from the respective percentages of females who are in favor and against.

In this example, the dependence of  $A$  and  $F$  can also be proved by showing that the probabilities  $P(A)$  and  $P(A | F)$  are not equal. ■

### ■ EXAMPLE 4–17

Illustrating two independent events.

A box contains a total of 100 DVDs that were manufactured on two machines. Of them, 60 were manufactured on Machine I. Of the total DVDs, 15 are defective. Of the 60 DVDs that were manufactured on Machine I, 9 are defective. Let  $D$  be the event that a randomly selected DVD is defective, and let  $A$  be the event that a randomly selected DVD was manufactured on Machine I. Are events  $D$  and  $A$  independent?

**Solution** From the given information,

$$P(D) = 15/100 = .15 \quad \text{and} \quad P(D | A) = 9/60 = .15$$

Hence,

$$P(D) = P(D | A)$$

Consequently, the two events,  $D$  and  $A$ , are independent.

Independence, in this example, means that the probability of any DVD being defective is the same, .15, irrespective of the machine on which it is manufactured. In other words, the two machines are producing the same percentage of defective DVDs. For example, 9 of the 60 DVDs manufactured on Machine I are defective, and 6 of the 40 DVDs manufactured on Machine II are defective. Thus, for each of the two machines, 15% of the DVDs produced are defective.

Using the given information, we can prepare Table 4.6. The numbers in the shaded cells are given to us. The remaining numbers are calculated by doing some arithmetic manipulations.

**Table 4.6 Two-Way Classification Table**

	Defective (D)	Good (G)	Total
Machine I (A)	9	51	60
Machine II (B)	6	34	40
Total	15	85	100

Using this table, we can find the following probabilities:

$$P(D) = 15/100 = .15$$

$$P(D | A) = 9/60 = .15$$

Because these two probabilities are the same, the two events are independent. ■

We can make the following two important observations about mutually exclusive, independent, and dependent events.

#### ◀ Two Important Observations

1. Two events are either mutually exclusive or independent.<sup>2</sup>
  - a. Mutually exclusive events are always dependent.
  - b. Independent events are never mutually exclusive.
2. Dependent events may or may not be mutually exclusive.

### 4.3.4 Complementary Events

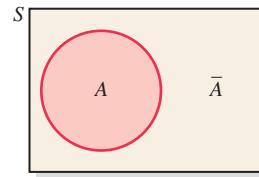
Two mutually exclusive events that taken together include all the outcomes for an experiment are called **complementary events**. Note that two complementary events are always mutually exclusive.

#### Definition

**Complementary Events** The complement of event  $A$ , denoted by  $\bar{A}$  and read as “ $A$  bar” or “ $A$  complement,” is the event that includes all the outcomes for an experiment that are not in  $A$ .

Events  $A$  and  $\bar{A}$  are complements of each other. The Venn diagram in Figure 4.11 shows the complementary events  $A$  and  $\bar{A}$ .

<sup>2</sup>The exception to this rule occurs when at least one of the two events has a zero probability.

**Figure 4.11** Venn diagram of two complementary events.

Because two complementary events, taken together, include all the outcomes for an experiment and because the sum of the probabilities of all outcomes is 1, it is obvious that

$$P(A) + P(\bar{A}) = 1$$

From this equation, we can deduce that

$$P(A) = 1 - P(\bar{A}) \quad \text{and} \quad P(\bar{A}) = 1 - P(A)$$

Thus, if we know the probability of an event, we can find the probability of its complementary event by subtracting the given probability from 1.

### ■ EXAMPLE 4-18

*Calculating probabilities of complementary events.*

In a group of 2000 taxpayers, 400 have been audited by the IRS at least once. If one taxpayer is randomly selected from this group, what are the two complementary events for this experiment, and what are their probabilities?

**Solution** The two complementary events for this experiment are

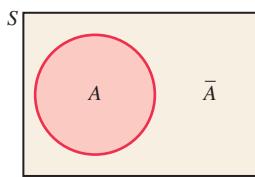
$A$  = the selected taxpayer has been audited by the IRS at least once

$\bar{A}$  = the selected taxpayer has never been audited by the IRS

Note that here event  $A$  includes the 400 taxpayers who have been audited by the IRS at least once, and  $\bar{A}$  includes the 1600 taxpayers who have never been audited by the IRS. Hence, the probabilities of events  $A$  and  $\bar{A}$  are

$$P(A) = 400/2000 = .20 \quad \text{and} \quad P(\bar{A}) = 1600/2000 = .80$$

As we can observe, the sum of these two probabilities is one. Figure 4.12 shows a Venn diagram for this example.

**Figure 4.12** Venn diagram.

### ■ EXAMPLE 4-19

*Calculating probabilities of complementary events.*

In a group of 5000 adults, 3500 are in favor of stricter gun control laws, 1200 are against such laws, and 300 have no opinion. One adult is randomly selected from this group. Let  $A$  be the event that this adult is in favor of stricter gun control laws. What is the complementary event of  $A$ ? What are the probabilities of the two events?

**Solution** The two complementary events for this experiment are

$A$  = the selected adult is in favor of stricter gun control laws

$\bar{A}$  = the selected adult is either against such laws or has no opinion

Note that here event  $\bar{A}$  includes 1500 adults who are either against stricter gun control laws or have no opinion. Also notice that events  $A$  and  $\bar{A}$  are complements of each other. Because

3500 adults in the group favor stricter gun control laws and 1500 either are against stricter gun control laws or have no opinion, the probabilities of events  $A$  and  $\bar{A}$  are

$$P(A) = 3500/5000 = .70 \quad \text{and} \quad P(\bar{A}) = 1500/5000 = .30$$

As we can observe, the sum of these two probabilities is 1. Also, once we find  $P(A)$ , we can find the probability of  $P(\bar{A})$  as

$$P(\bar{A}) = 1 - P(A) = 1 - .70 = .30$$

Figure 4.13 shows a Venn diagram for this example.

**Figure 4.13** Venn diagram.



## EXERCISES

### CONCEPTS AND PROCEDURES

**4.41** Briefly explain the difference between the marginal and conditional probabilities of events. Give one example of each.

**4.42** What is meant by two mutually exclusive events? Give one example of two mutually exclusive events and another example of two mutually nonexclusive events.

**4.43** Briefly explain the meaning of independent and dependent events. Suppose  $A$  and  $B$  are two events. What formula will you use to prove whether  $A$  and  $B$  are independent or dependent?

**4.44** What is the complement of an event? What is the sum of the probabilities of two complementary events?

**4.45** A statistical experiment has 11 equally likely outcomes that are denoted by  $a, b, c, d, e, f, g, h, i, j$ , and  $k$ . Consider three events:  $A = \{b, d, e, j\}$ ,  $B = \{a, c, f, j\}$ , and  $C = \{c, g, k\}$ .

- a. Are events  $A$  and  $B$  independent events? What about events  $A$  and  $C$ ?
- b. Are events  $A$  and  $B$  mutually exclusive events? What about  $A$  and  $C$ ? What about  $B$  and  $C$ ?
- c. What are the complements of events  $A$ ,  $B$ , and  $C$ , respectively, and what are their probabilities?

**4.46** A statistical experiment has 10 equally likely outcomes that are denoted by 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Let event  $A = \{3, 4, 6, 9\}$  and event  $B = \{1, 2, 5\}$ .

- a. Are events  $A$  and  $B$  mutually exclusive events?
- b. Are events  $A$  and  $B$  independent events?
- c. What are the complements of events  $A$  and  $B$ , respectively, and their probabilities?

### APPLICATIONS

**4.47** Two thousand randomly selected adults were asked whether or not they have ever shopped on the Internet. The following table gives a two-way classification of the responses.

	Have Shopped	Have Never Shopped
Male	500	700
Female	300	500

- a. If one adult is selected at random from these 2000 adults, find the probability that this adult
  - i. has never shopped on the Internet
  - ii. is a male
  - iii. has shopped on the Internet given that this adult is a female
  - iv. is a male given that this adult has never shopped on the Internet

- b. Are the events “male” and “female” mutually exclusive? What about the events “have shopped” and “male?” Why or why not?  
 c. Are the events “female” and “have shopped” independent? Why or why not?

**4.48** A 2010–2011 poll conducted by Gallup, ([www.gallup.com/poll/148994/Emotional-Health-Higher-Between-Older-Americans.aspx](http://www.gallup.com/poll/148994/Emotional-Health-Higher-Between-Older-Americans.aspx)) examined the emotional health of a large number of Americans. Among other things, Gallup reported on whether people had an *Emotional Health Index* score of 90 or higher, which would classify a person as being *emotionally well-off*. The report was based on a survey of 65,528 people in the age group 35–44 years and 91,802 people in the age group 65–74 years. The following table gives the results of the survey, converting percentages to frequencies.

	Emotionally Well-Off	Emotionally Not Well-Off
35–44 Age group	16,016	49,512
65–74 Age group	32,583	59,219

- a. If one person is selected at random from this sample of 157,330 Americans, find the probability that this person  
 i. is emotionally well-off  
 ii. is in the 35–44 age group  
 iii. is emotionally well-off given that this person is in the 35–44 age group  
 iv. is emotionally not well-off given that this person is in the 65–74 age group  
 b. Are the events *emotionally well-off* and *emotionally not well-off* mutually exclusive? What about the events *emotionally well-off* and *35–44 age group*? Why or why not?  
 c. Are the events *emotionally well-off* and *35–44 age group* independent? Why or why not?

**4.49** Two thousand randomly selected adults were asked if they are in favor of or against cloning. The following table gives the responses.

	In Favor	Against	No Opinion
Male	395	405	100
Female	300	680	120

- a. If one person is selected at random from these 2000 adults, find the probability that this person is  
 i. in favor of cloning  
 ii. against cloning  
 iii. in favor of cloning given the person is a female  
 iv. a male given the person has no opinion  
 b. Are the events “male” and “in favor” mutually exclusive? What about the events “in favor” and “against”? Why or why not?  
 c. Are the events “female” and “no opinion” independent? Why or why not?

**4.50** Five hundred employees were selected from a city’s large private companies, and they were asked whether or not they have any retirement benefits provided by their companies. Based on this information, the following two-way classification table was prepared.

		Have Retirement Benefits	
		Yes	No
	Men	225	75
	Women	150	50

- a. If one employee is selected at random from these 500 employees, find the probability that this employee  
 i. is a woman  
 ii. has retirement benefits  
 iii. has retirement benefits given the employee is a man  
 iv. is a woman given that she does not have retirement benefits

- b.** Are the events “man” and “yes” mutually exclusive? What about the events “yes” and “no?” Why or why not?  
**c.** Are the events “woman” and “yes” independent? Why or why not?

**4.51** A consumer agency randomly selected 1700 flights for two major airlines, A and B. The following table gives the two-way classification of these flights based on airline and arrival time. Note that “less than 30 minutes late” includes flights that arrived early or on time.

	Less Than 30 Minutes Late	30 Minutes to 1 Hour Late	More Than 1 Hour Late
Airline A	429	390	92
Airline B	393	316	80

- a.** If one flight is selected at random from these 1700 flights, find the probability that this flight is  
**i.** more than 1 hour late  
**ii.** less than 30 minutes late  
**iii.** a flight on airline A given that it is 30 minutes to 1 hour late  
**iv.** more than 1 hour late given that it is a flight on airline B  
**b.** Are the events “airline A” and “more than 1 hour late” mutually exclusive? What about the events “less than 30 minutes late” and “more than 1 hour late?” Why or why not?  
**c.** Are the events “airline B” and “30 minutes to 1 hour late” independent? Why or why not?

**4.52** A July 21, 2009 (just a reminder that July 21 is National Junk Food Day) survey on www.HuffingtonPost.com asked people to choose their favorite junk food from a list of choices. Of the 8002 people who responded to the survey, 2049 answered chocolate, 345 said sugary candy, 1271 mentioned ice cream, 775 indicated fast food, 650 said cookies, 1107 mentioned chips, 490 said cake, and 1315 indicated pizza. Although the results were not broken down by gender, suppose that the following table represents the results for the 8002 people who responded, assuming that there were 4801 females and 3201 males included in the survey.

Favorite Junk Food	Female	Male
Chocolate	1518	531
Sugary candy	218	127
Ice cream	685	586
Fast food	312	463
Cookies	431	219
Chips	458	649
Cake	387	103
Pizza	792	523

- a.** If one person is selected at random from this sample of 8002 respondents, find the probability that this person  
**i.** is a female  
**ii.** responded *chips*  
**iii.** responded *chips* given that this person is a *female*  
**iv.** responded *chocolate* given that this person is a *male*  
**b.** Are the events *chips* and *cake* mutually exclusive? What about the events *chips* and *female*? Why or why not?  
**c.** Are the events *chips* and *female* independent? Why or why not?

**4.53** There are a total of 160 practicing physicians in a city. Of them, 75 are female and 25 are pediatricians. Of the 75 females, 20 are pediatricians. Are the events “female” and “pediatrician” independent? Are they mutually exclusive? Explain why or why not.

**4.54** Of a total of 100 DVDs manufactured on two machines, 20 are defective. Sixty of the total DVDs were manufactured on Machine 1, and 10 of these 60 are defective. Are the events “Machine I” and “defective” independent? (Note: Compare this exercise with Example 4–17.)

**4.55** There are 142 people participating in a local 5K road race. Sixty-five of these runners are female. Of the female runners, 19 are participating in their first 5K road race. Of the male runners, 28 are participating in their first 5K road race. Are the events *female* and *participating in their first 5K road race* independent? Are they mutually exclusive? Explain why or why not.

**4.56** Define the following two events for two tosses of a coin:

$A$  = at least one head is obtained

$B$  = both tails are obtained

- Are  $A$  and  $B$  mutually exclusive events? Are they independent? Explain why or why not.
- Are  $A$  and  $B$  complementary events? If yes, first calculate the probability of  $B$  and then calculate the probability of  $A$  using the complementary event rule.

**4.57** Let  $A$  be the event that a number less than 3 is obtained if we roll a die once. What is the probability of  $A$ ? What is the complementary event of  $A$ , and what is its probability?

**4.58** Thirty percent of last year's graduates from a university received job offers during their last semester in school. What are the two complementary events here and what are their probabilities?

**4.59** The probability that a randomly selected college student attended at least one major league baseball game last year is .12. What is the complementary event? What is the probability of this complementary event?

## 4.4 Intersection of Events and the Multiplication Rule

This section discusses the intersection of two events and the application of the multiplication rule to compute the probability of the intersection of events.

### 4.4.1 Intersection of Events

The **intersection of two events** is given by the outcomes that are common to both events.

#### Definition

**Intersection of Events** Let  $A$  and  $B$  be two events defined in a sample space. The *intersection* of  $A$  and  $B$  represents the collection of all outcomes that are common to both  $A$  and  $B$  and is denoted by

$A \text{ and } B$

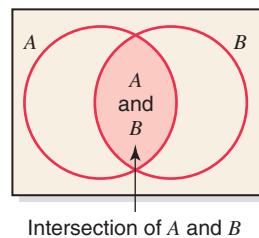
The intersection of events  $A$  and  $B$  is also denoted by either  $A \cap B$  or  $AB$ . Let

$A$  = event that a family owns a DVD player

$B$  = event that a family owns a digital camera

Figure 4.14 illustrates the intersection of events  $A$  and  $B$ . The shaded area in this figure gives the intersection of events  $A$  and  $B$ , and it includes all the families who own both a DVD player and a digital camera.

**Figure 4.14** Intersection of events  $A$  and  $B$ .



## 4.4.2 Multiplication Rule

Sometimes we may need to find the probability of two or more events happening together.

### Definition

**Joint Probability** The probability of the intersection of two events is called their *joint probability*. It is written as

$$P(A \text{ and } B)$$

The probability of the intersection of two events is obtained by multiplying the marginal probability of one event by the conditional probability of the second event. This rule is called the **multiplication rule**.

**Multiplication Rule to Find Joint Probability** The probability of the intersection of two events  $A$  and  $B$  is

$$P(A \text{ and } B) = P(A) P(B | A) = P(B) P(A | B)$$

The joint probability of events  $A$  and  $B$  can also be denoted by  $P(A \cap B)$  or  $P(AB)$ .

### ■ EXAMPLE 4-20

Table 4.7 gives the classification of all employees of a company by gender and college degree.

**Table 4.7 Classification of Employees by Gender and Education**

Calculating the joint probability of two events: two-way table.

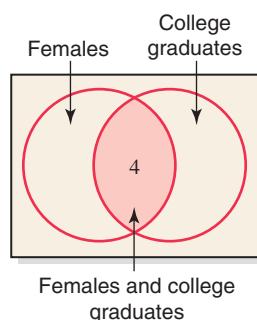
	College Graduate (G)	Not a College Graduate (N)	Total
Male (M)	7	20	27
Female (F)	4	9	13
Total	11	29	40

If one of these employees is selected at random for membership on the employee–management committee, what is the probability that this employee is a female and a college graduate?

**Solution** We are to calculate the probability of the intersection of the events “female” (denoted by  $F$ ) and “college graduate” (denoted by  $G$ ). This probability may be computed using the formula

$$P(F \text{ and } G) = P(F) P(G | F)$$

The shaded area in Figure 4.15 shows the intersection of the events “female” and “college graduate.” There are four females who are college graduates.



**Figure 4.15** Intersection of events  $F$  and  $G$ .

Notice that there are 13 females among 40 employees. Hence, the probability that a female is selected is

$$P(F) = 13/40$$

To calculate the probability  $P(G | F)$ , we know that  $F$  has already occurred. Consequently, the employee selected is one of the 13 females. In the table, there are 4 college graduates among 13 female employees. Hence, the conditional probability of  $G$  given  $F$  is

$$P(G | F) = 4/13$$

The joint probability of  $F$  and  $G$  is

$$P(F \text{ and } G) = P(F) P(G | F) = (13/40)(4/13) = .100$$

Thus, the probability is .100 that a randomly selected employee is a female and a college graduate.

The probability in this example can also be calculated without using the multiplication rule. As we can notice from Figure 4.15 and from the table, 4 employees out of a total of 40 are female and college graduates. Hence, if any of these 4 employees is selected, the events “female” and “college graduate” both happen. Therefore, the required probability is

$$P(F \text{ and } G) = 4/40 = .100$$

We can compute three other joint probabilities for the table in Example 4–20 as follows:

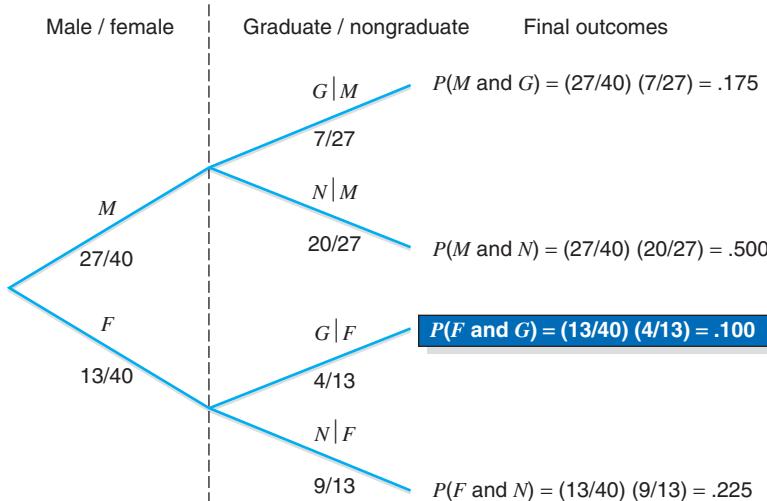
$$P(M \text{ and } G) = P(M) P(G | M) = (27/40)(7/27) = .175$$

$$P(M \text{ and } N) = P(M) P(N | M) = (27/40)(20/27) = .500$$

$$P(F \text{ and } N) = P(F) P(N | F) = (13/40)(9/13) = .225$$

The tree diagram in Figure 4.16 shows all four joint probabilities for this example. The joint probability of  $F$  and  $G$  is highlighted.

**Figure 4.16** Tree diagram for joint probabilities.



### ■ EXAMPLE 4–21

A box contains 20 DVDs, 4 of which are defective. If two DVDs are selected at random (without replacement) from this box, what is the probability that both are defective?

**Solution** Let us define the following events for this experiment:

$G_1$  = event that the first DVD selected is good

$D_1$  = event that the first DVD selected is defective

$G_2$  = event that the second DVD selected is good

$D_2$  = event that the second DVD selected is defective

*Calculating the joint probability of two events.*

We are to calculate the joint probability of  $D_1$  and  $D_2$ , which is given by

$$P(D_1 \text{ and } D_2) = P(D_1) P(D_2 | D_1)$$

As we know, there are 4 defective DVDs in 20. Consequently, the probability of selecting a defective DVD at the first selection is

$$P(D_1) = 4/20$$

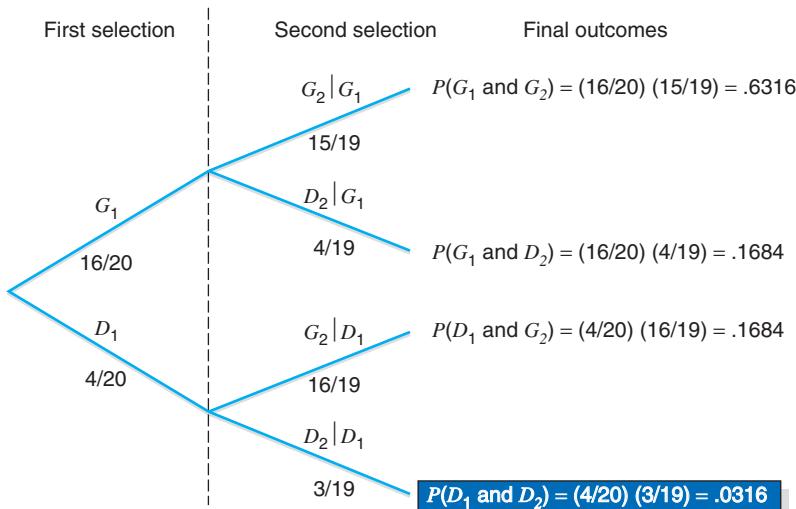
To calculate the probability  $P(D_2 | D_1)$ , we know that the first DVD selected is defective because  $D_1$  has already occurred. Because the selections are made without replacement, there are 19 total DVDs, and 3 of them are defective at the time of the second selection. Therefore,

$$P(D_2 | D_1) = 3/19$$

Hence, the required probability is

$$P(D_1 \text{ and } D_2) = P(D_1) P(D_2 | D_1) = (4/20)(3/19) = .0316$$

The tree diagram in Figure 4.17 shows the selection procedure and the final four outcomes for this experiment along with their probabilities. The joint probability of  $D_1$  and  $D_2$  is highlighted in the tree diagram.



**Figure 4.17** Selecting two DVDs. ■

Conditional probability was discussed in Section 4.3.1. It is obvious from the formula for joint probability that if we know the probability of an event  $A$  and the joint probability of events  $A$  and  $B$ , then we can calculate the conditional probability of  $B$  given  $A$ .

### Calculating Conditional Probability

If  $A$  and  $B$  are two events, then,

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} \quad \text{and} \quad P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

given that  $P(A) \neq 0$  and  $P(B) \neq 0$ .

### ■ EXAMPLE 4-22

The probability that a randomly selected student from a college is a senior is .20, and the joint probability that the student is a computer science major and a senior is .03. Find the conditional probability that a student selected at random is a computer science major given that the student is a senior.

*Calculating the conditional probability of an event.*

**Solution** Let us define the following two events:

$A$  = the student selected is a senior

$B$  = the student selected is a computer science major

From the given information,

$$P(A) = .20 \quad \text{and} \quad P(A \text{ and } B) = .03$$

Hence,

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{.03}{.20} = .15$$

Thus, the (conditional) probability is .15 that a student selected at random is a computer science major given that he or she is a senior. ■

### Multiplication Rule for Independent Events

The foregoing discussion of the multiplication rule was based on the assumption that the two events are dependent. Now suppose that events  $A$  and  $B$  are independent. Then,

$$P(A) = P(A | B) \quad \text{and} \quad P(B) = P(B | A)$$

By substituting  $P(B)$  for  $P(B | A)$  into the formula for the joint probability of  $A$  and  $B$ , we obtain

$$P(A \text{ and } B) = P(A) P(B)$$

**Multiplication Rule to Calculate the Probability of Independent Events** The probability of the intersection of two independent events  $A$  and  $B$  is

$$P(A \text{ and } B) = P(A) P(B)$$

### ■ EXAMPLE 4-23

*Calculating the joint probability of two independent events.*

An office building has two fire detectors. The probability is .02 that any fire detector of this type will fail to go off during a fire. Find the probability that both of these fire detectors will fail to go off in case of a fire.

**Solution** In this example, the two fire detectors are independent because whether or not one fire detector goes off during a fire has no effect on the second fire detector. We define the following two events:

$A$  = the first fire detector fails to go off during a fire

$B$  = the second fire detector fails to go off during a fire

Then, the joint probability of  $A$  and  $B$  is

$$P(A \text{ and } B) = P(A) P(B) = (.02)(.02) = .0004$$

The multiplication rule can be extended to calculate the joint probability of more than two events. Example 4-24 illustrates such a case for independent events.

### ■ EXAMPLE 4-24

*Calculating the joint probability of three independent events.*

The probability that a patient is allergic to penicillin is .20. Suppose this drug is administered to three patients.

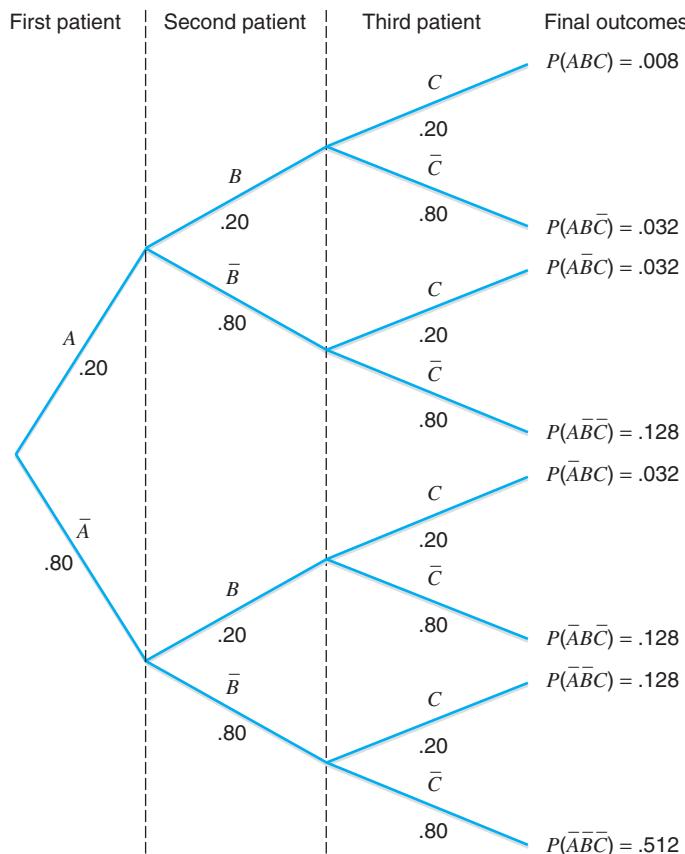
- (a) Find the probability that all three of them are allergic to it.
- (b) Find the probability that at least one of them is not allergic to it.

**Solution**

- (a) Let  $A$ ,  $B$ , and  $C$  denote the events that the first, second, and third patients, respectively, are allergic to penicillin. We are to find the joint probability of  $A$ ,  $B$ , and  $C$ . All three events are independent because whether or not one patient is allergic does not depend on whether or not any of the other patients is allergic. Hence,

$$P(A \text{ and } B \text{ and } C) = P(A) P(B) P(C) = (.20)(.20)(.20) = .008$$

The tree diagram in Figure 4.18 shows all the outcomes for this experiment. Events  $\bar{A}$ ,  $\bar{B}$ , and  $\bar{C}$  are the complementary events of  $A$ ,  $B$ , and  $C$ , respectively. They represent the events that the patients are not allergic to penicillin. Note that the intersection of events  $A$ ,  $B$ , and  $C$  is written as  $ABC$  in the tree diagram.



**Figure 4.18** Tree diagram for joint probabilities.

- (b) Let us define the following events:

$$G = \text{all three patients are allergic}$$

$$H = \text{at least one patient is not allergic}$$

Events  $G$  and  $H$  are two complementary events. Event  $G$  consists of the intersection of events  $A$ ,  $B$ , and  $C$ . Hence, from part (a),

$$P(G) = P(A \text{ and } B \text{ and } C) = .008$$

Therefore, using the complementary event rule, we obtain

$$P(H) = 1 - P(G) = 1 - .008 = .992$$



## Joint Probability of Mutually Exclusive Events

We know from an earlier discussion that two mutually exclusive events cannot happen together. Consequently, their joint probability is zero.

**Joint Probability of Mutually Exclusive Events** The joint probability of two mutually exclusive events is always zero. If  $A$  and  $B$  are two mutually exclusive events, then,

$$P(A \text{ and } B) = 0$$

*Illustrating the joint probability of two mutually exclusive events.*

### ■ EXAMPLE 4–25

Consider the following two events for an application filed by a person to obtain a car loan:

$A$  = event that the loan application is approved

$R$  = event that the loan application is rejected

What is the joint probability of  $A$  and  $R$ ?

**Solution** The two events  $A$  and  $R$  are mutually exclusive. Either the loan application will be approved or it will be rejected. Hence,

$$P(A \text{ and } R) = 0$$

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 4.60** Explain the meaning of the intersection of two events. Give one example.
- 4.61** What is meant by the joint probability of two or more events? Give one example.
- 4.62** How is the multiplication rule of probability for two dependent events different from the rule for two independent events?
- 4.63** What is the joint probability of two mutually exclusive events? Give one example.
- 4.64** Find the joint probability of  $A$  and  $B$  for the following.
- $P(A) = .36$  and  $P(B | A) = .87$
  - $P(B) = .53$  and  $P(A | B) = .22$
- 4.65** Find the joint probability of  $A$  and  $B$  for the following.
- $P(B) = .66$  and  $P(A | B) = .91$
  - $P(A) = .12$  and  $P(B | A) = .07$
- 4.66** Given that  $A$  and  $B$  are two independent events, find their joint probability for the following.
- $P(A) = .17$  and  $P(B) = .44$
  - $P(A) = .72$  and  $P(B) = .84$
- 4.67** Given that  $A$  and  $B$  are two independent events, find their joint probability for the following.
- $P(A) = .29$  and  $P(B) = .65$
  - $P(A) = .03$  and  $P(B) = .28$
- 4.68** Given that  $A$ ,  $B$ , and  $C$  are three independent events, find their joint probability for the following.
- $P(A) = .81$ ,  $P(B) = .49$ , and  $P(C) = .36$
  - $P(A) = .02$ ,  $P(B) = .03$ , and  $P(C) = .05$
- 4.69** Given that  $A$ ,  $B$ , and  $C$  are three independent events, find their joint probability for the following.
- $P(A) = .30$ ,  $P(B) = .50$ , and  $P(C) = .70$
  - $P(A) = .40$ ,  $P(B) = .50$ , and  $P(C) = .60$
- 4.70** Given that  $P(A) = .72$  and  $P(A \text{ and } B) = .38$ , find  $P(B | A)$ .
- 4.71** Given that  $P(B) = .29$  and  $P(A \text{ and } B) = .24$ , find  $P(A | B)$ .
- 4.72** Given that  $P(A | B) = .44$  and  $P(A \text{ and } B) = .33$ , find  $P(B)$ .
- 4.73** Given that  $P(B | A) = .70$  and  $P(A \text{ and } B) = .35$ , find  $P(A)$ .

## ■ APPLICATIONS

**4.74** Refer to Exercise 4.52, which contains information on a July 21, 2009 www.HuffingtonPost.com survey that asked people to choose their favorite junk food from a list of choices. The following table contains results classified by gender. (Note: There are 4801 females and 3201 males.)

Favorite Junk Food	Female	Male
Chocolate	1518	531
Sugary candy	218	127
Ice cream	685	586
Fast food	312	463
Cookies	431	219
Chips	458	649
Cake	387	103
Pizza	792	523

- a. Suppose that one person is selected at random from this sample of 8002 respondents. Find the following probabilities.
- Probability of the intersection of events *female* and *ice cream*.
  - Probability of the intersection of events *male* and *pizza*.
- b. Mention at least four other joint probabilities you can calculate for this table and then find their probabilities. You may draw a tree diagram to find these probabilities.

**4.75** The following table gives a two-way classification of all basketball players at a state university who began their college careers between 2004 and 2008, based on gender and whether or not they graduated.

	Graduated	Did Not Graduate
Male	126	55
Female	133	32

- a. If one of these players is selected at random, find the following probabilities.
- $P(\text{female and graduated})$
  - $P(\text{male and did not graduate})$
- b. Find  $P(\text{graduated and did not graduate})$ . Is this probability zero? If yes, why?
- 4.76** Five hundred employees were selected from a city's large private companies and asked whether or not they have any retirement benefits provided by their companies. Based on this information, the following two-way classification table was prepared.

	Have Retirement Benefits	
	Yes	No
Men	225	75
Women	150	50

- a. Suppose one employee is selected at random from these 500 employees. Find the following probabilities.
- Probability of the intersection of events "woman" and "yes"
  - Probability of the intersection of events "no" and "man"
- b. Mention what other joint probabilities you can calculate for this table and then find them. You may draw a tree diagram to find these probabilities.

**4.77** Two thousand randomly selected adults were asked whether or not they have ever shopped on the Internet. The following table gives a two-way classification of the responses obtained.

	Have Shopped	Have Never Shopped
Male	500	700
Female	300	500

- a. Suppose one adult is selected at random from these 2000 adults. Find the following probabilities.
- $P(\text{has never shopped on the Internet and is a male})$
  - $P(\text{has shopped on the Internet and is a female})$
- b. Mention what other joint probabilities you can calculate for this table and then find those. You may draw a tree diagram to find these probabilities.

**4.78** A consumer agency randomly selected 1700 flights for two major airlines, A and B. The following table gives the two-way classification of these flights based on airline and arrival time. Note that “less than 30 minutes late” includes flights that arrived early or on time.

	Less Than 30 Minutes Late	30 Minutes to 1 Hour Late	More Than 1 Hour Late
Airline A	429	390	92
Airline B	393	316	80

- a. Suppose one flight is selected at random from these 1700 flights. Find the following probabilities.
- $P(\text{more than 1 hour late and airline A})$
  - $P(\text{airline B and less than 30 minutes late})$
- b. Find the joint probability of events “30 minutes to 1 hour late” and “more than 1 hour late.” Is this probability zero? Explain why or why not.

**4.79** Refer to Exercise 4.48. A 2010–2011 poll conducted by Gallup ([www.gallup.com/poll/148994/Emotional-Health-Higher-Born-Older-Americans.aspx](http://www.gallup.com/poll/148994/Emotional-Health-Higher-Born-Older-Americans.aspx)) examined the emotional health of a large number of Americans. Among other things, Gallup reported on whether people had *Emotional Health Index* scores of 90 or higher, which would classify them as being *emotionally well-off*. The report was based on a survey of 65,528 people in the age group 35–44 years and 91,802 people in the age group 65–74 years. The following table gives the results of the survey, converting percentages to frequencies.

	Emotionally Well-Off	Emotionally Not Well-Off
35–44 Age group	16,016	49,512
65–74 Age group	32,583	59,219

- a. Suppose that one person is selected at random from this sample of 157,330 Americans. Find the following probabilities.
- $P(35\text{--}44 \text{ age group and emotionally not well-off})$
  - $P(\text{emotionally well-off and } 65\text{--}74 \text{ age group})$
- b. Find the joint probability of the events *35–44 age group* and *65–74 age group*. Is this probability zero? Explain why or why not.

**4.80** In a statistics class of 42 students, 28 have volunteered for community service in the past. If two students are selected at random from this class, what is the probability that both of them have volunteered for community service in the past? Draw a tree diagram for this problem.

**4.81** In a political science class of 35 students, 21 favor abolishing the electoral college and thus electing the President of the United States by popular vote. If two students are selected at random from this class, what is the probability that both of them favor abolition of the electoral college? Draw a tree diagram for this problem.

**4.82** A company is to hire two new employees. They have prepared a final list of eight candidates, all of whom are equally qualified. Of these eight candidates, five are women. If the company decides to select two persons randomly from these eight candidates, what is the probability that both of them are women? Draw a tree diagram for this problem.

**4.83** Forty-seven employees in an office wear eyeglasses. Thirty-one have single-vision correction, and 16 wear bifocals. If two employees are selected at random from this group, what is the probability that both of them wear bifocals? What is the probability that both have single-vision correction?

**4.84** Of the 35 students in a class, 22 are taking the class because it is a major requirement, and the other 13 are taking it as an elective. If two students are selected at random from this class, what is the probability that the first student is taking the class as an elective and the second is taking it because it is a major requirement? How does this probability compare to the probability that the first student is taking the class because it is a major requirement and the second is taking it as an elective?

**4.85** The probability that a student graduating from Suburban State University has student loans to pay off after graduation is .60. If two students are randomly selected from this university, what is the probability that neither of them has student loans to pay off after graduation?

**4.86** A contractor has submitted bids for two state construction projects. The probability of winning each contract is .25, and it is the same for both contracts.

- What is the probability that he will win both contracts?
- What is the probability that he will win neither contract?

Draw a tree diagram for this problem.

**4.87** Five percent of all items sold by a mail-order company are returned by customers for a refund. Find the probability that of two items sold during a given hour by this company,

- both will be returned for a refund
- neither will be returned for a refund

Draw a tree diagram for this problem.

**4.88** According to the Recording Industry Association of America, only 37% of music files downloaded from Web sites in 2009 were paid for. Suppose that this percentage holds true for such files downloaded this year. Three downloaded music files are selected at random. What is the probability that all three were paid for? What is the probability that none were paid for? Assume independence of events.

**4.89** The probability that a farmer is in debt is .80. What is the probability that three randomly selected farmers are all in debt? Assume independence of events.

**4.90** The probability that a student graduating from Suburban State University has student loans to pay off after graduation is .60. The probability that a student graduating from this university has student loans to pay off after graduation and is a male is .24. Find the conditional probability that a randomly selected student from this university is a male given that this student has student loans to pay off after graduation.

**4.91** The probability that an employee at a company is a female is .36. The probability that an employee is a female and married is .19. Find the conditional probability that a randomly selected employee from this company is married given that she is a female.

**4.92** Recent uncertain economic conditions have forced many people to change their spending habits. In a recent telephone poll of 1000 adults, 629 stated that they were cutting back on their daily spending. Suppose that 322 of the 629 people who stated that they were cutting back on their daily spending said that they were cutting back “somewhat” and 97 stated that they were cutting back “somewhat” and “delaying the purchase of a new car by at least 6 months”. If one of the 629 people who are cutting back on their spending is selected at random, what is the probability that he/she is delaying the purchase of a new car by at least 6 months given that he/she is cutting back on spending “somewhat”?

**4.93** Suppose that 20% of all adults in a small town live alone, and 8% of the adults live alone and have at least one pet. What is the probability that a randomly selected adult from this town has at least one pet given that this adult lives alone?

## 4.5 Union of Events and the Addition Rule

This section discusses the union of events and the addition rule that is applied to compute the probability of the union of events.

### 4.5.1 Union of Events

The **union of two events**  $A$  and  $B$  includes all outcomes that are either in  $A$  or in  $B$  or in both  $A$  and  $B$ .

#### Definition

**Union of Events** Let  $A$  and  $B$  be two events defined in a sample space. The *union of events*  $A$  and  $B$  is the collection of all outcomes that belong either to  $A$  or to  $B$  or to both  $A$  and  $B$  and is denoted by

$$A \text{ or } B$$

The union of events  $A$  and  $B$  is also denoted by  $A \cup B$ . Example 4–26 illustrates the union of events  $A$  and  $B$ .

### ■ EXAMPLE 4–26

*Illustrating the union of two events.*



PhotoDisc, Inc./Getty Images

**Solution** Let us define the following events:

$M$  = a senior citizen is a male

$F$  = a senior citizen is a female

$A$  = a senior citizen takes at least one medicine

$B$  = a senior citizen does not take any medicine

The union of the events “male” and “take at least one medicine” includes those senior citizens who are either male or take at least one medicine or both. The number of such senior citizens is

$$140 + 210 - 95 = 255$$

Why did we subtract 95 from the sum of 140 and 210? The reason is that 95 senior citizens (which represent the intersection of events  $M$  and  $A$ ) are common to both events  $M$  and  $A$  and, hence, are counted twice. To avoid double counting, we subtracted 95 from the sum of the other two numbers. We can observe this double counting from Table 4.8, which is constructed using the given information. The sum of the numbers in the three shaded cells gives the number of senior citizens who are either male or take at least one medicine or both. However, if we add the totals of the row labeled  $M$  and the column labeled  $A$ , we count 95 twice.

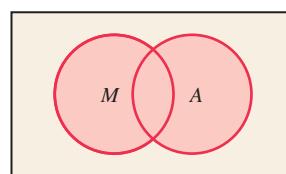
**Table 4.8**

	<i>A</i>	<i>B</i>	Total
<i>M</i>	95	45	140
<i>F</i>	115	45	160
Total	210	90	300

→Counted twice

Figure 4.19 shows the diagram for the union of the events “male” and “take at least one medicine on a permanent basis.” The union of events  $M$  and  $A$  will be written as  $(M \text{ or } A)$ .

**Figure 4.19** Union of events  $M$  and  $A$ .



Area shaded in red gives the union of events  $M$  and  $A$ , and includes 255 senior citizens

### 4.5.2 Addition Rule

The method used to calculate the probability of the union of events is called the **addition rule**. It is defined as follows.

**Addition Rule to Find the Probability of Union of Events** The probability of the union of two events  $A$  and  $B$  is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Thus, to calculate the probability of the union of two events  $A$  and  $B$ , we add their marginal probabilities and subtract their joint probability from this sum. We must subtract the joint probability of  $A$  and  $B$  from the sum of their marginal probabilities to avoid double counting because of common outcomes in  $A$  and  $B$ . This is the case where events  $A$  and  $B$  are not mutually exclusive.

### ■ EXAMPLE 4-27

A university president proposed that all students must take a course in ethics as a requirement for graduation. Three hundred faculty members and students from this university were asked about their opinions on this issue. Table 4.9 gives a two-way classification of the responses of these faculty members and students.

*Calculating the probability of the union of two events: two-way table.*

**Table 4.9** Two-Way Classification of Responses

	Favor	Oppose	Neutral	Total
Faculty	45	15	10	70
Student	90	110	30	230
Total	135	125	40	300

Find the probability that one person selected at random from these 300 persons is a faculty member or is in favor of this proposal.

**Solution** Let us define the following events:

$A$  = the person selected is a faculty member

$B$  = the person selected is in favor of the proposal

From the information given in Table 4.9,

$$P(A) = 70/300 = .2333$$

$$P(B) = 135/300 = .4500$$

$$P(A \text{ and } B) = P(A) P(B | A) = (70/300)(45/70) = .1500$$

Using the addition rule, we obtain

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = .2333 + .4500 - .1500 = .5333$$

Thus, the probability that a randomly selected person from these 300 persons is a faculty member or is in favor of this proposal is .5333.

The probability in this example can also be calculated without using the addition rule. The total number of persons in Table 4.9 who are either faculty members or in favor of this proposal is

$$45 + 15 + 10 + 90 = 160$$

Hence, the required probability is

$$P(A \text{ or } B) = 160/300 = .5333$$

### ■ EXAMPLE 4-28

In a group of 2500 persons, 1400 are female, 600 are vegetarian, and 400 are female and vegetarian. What is the probability that a randomly selected person from this group is a male or vegetarian?

*Calculating the probability of the union of two events.*

**Solution** Let us define the following events:

$F$  = the randomly selected person is a female

$M$  = the randomly selected person is a male

$V$  = the randomly selected person is a vegetarian

$N$  = the randomly selected person is a nonvegetarian

From the given information, we know that of the group, 1400 are female, 600 are vegetarian, and 400 are female and vegetarian. Hence, 1100 are male, 1900 are nonvegetarian, and 200 are male and vegetarian. We are to find the probability  $P(M \text{ or } V)$ . This probability is obtained as follows:

$$\begin{aligned} P(M \text{ or } V) &= P(M) + P(V) - P(M \text{ and } V) \\ &= \frac{1100}{2500} + \frac{600}{2500} - \frac{200}{2500} \\ &= .44 + .24 - .08 = .60 \end{aligned}$$

Actually, using the given information, we can prepare Table 4.10 for this example. In the table, the numbers in the shaded cells are given to us. The remaining numbers are calculated by doing some arithmetic manipulations.

**Table 4.10 Two-Way Classification Table**

	Vegetarian ( $V$ )	Nonvegetarian ( $N$ )	Total
Female ( $F$ )	400	1000	1400
Male ( $M$ )	200	900	1100
Total	600	1900	2500

Using Table 4.10, we find the required probability:

$$\begin{aligned} P(M \text{ or } V) &= P(M) + P(V) - P(M \text{ and } V) \\ &= \frac{1100}{2500} + \frac{600}{2500} - \frac{200}{2500} = .44 + .24 - .08 = .60 \end{aligned}$$
■

### Addition Rule for Mutually Exclusive Events

We know from an earlier discussion that the joint probability of two mutually exclusive events is zero. When  $A$  and  $B$  are mutually exclusive events, the term  $P(A \text{ and } B)$  in the addition rule becomes zero and is dropped from the formula. Thus, the probability of the union of two mutually exclusive events is given by the sum of their marginal probabilities.

**Addition Rule to Find the Probability of the Union of Mutually Exclusive Events** The probability of the union of two mutually exclusive events  $A$  and  $B$  is

$$P(A \text{ or } B) = P(A) + P(B)$$

### EXAMPLE 4–29

A university president proposed that all students must take a course in ethics as a requirement for graduation. Three hundred faculty members and students from this university were asked about their opinion on this issue. The following table, reproduced from Table 4.9 in Example 4–27, gives a two-way classification of the responses of these faculty members and students.

	Favor	Oppose	Neutral	Total
Faculty	45	15	10	70
Student	90	110	30	230
Total	135	125	40	300

*Calculating the probability of the union of two mutually exclusive events: two-way table.*

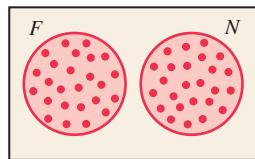
What is the probability that a randomly selected person from these 300 faculty members and students is in favor of the proposal or is neutral?

**Solution** Let us define the following events:

$F$  = the person selected is in favor of the proposal

$N$  = the person selected is neutral

As shown in Figure 4.20, events  $F$  and  $N$  are mutually exclusive because a person selected can be either in favor or neutral but not both.



**Figure 4.20** Venn diagram of mutually exclusive events  $F$  and  $N$ .

From the given information,

$$P(F) = 135/300 = .4500$$

$$P(N) = 40/300 = .1333$$

Hence,

$$P(F \text{ or } N) = P(F) + P(N) = .4500 + .1333 = .5833 \quad \blacksquare$$

The addition rule formula can easily be extended to apply to more than two events. The following example illustrates such a case.

### ■ EXAMPLE 4-30

Consider the experiment of rolling a die twice. Find the probability that the sum of the numbers obtained on two rolls is 5, 7, or 10.

**Solution** The experiment of rolling a die twice has a total of 36 outcomes, which are listed in Table 4.11. Assuming that the die is balanced, these 36 outcomes are equally likely.

*Calculating the probability of the union of three mutually exclusive events.*

**Table 4.11** Two Rolls of a Die

		Second Roll of the Die					
		1	2	3	4	5	6
First Roll of the Die	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

The events that give the sum of two numbers equal to 5 or 7 or 10 are shaded in the table. As we can observe, the three events “the sum is 5,” “the sum is 7,” and “the sum is 10” are mutually exclusive. Four outcomes give a sum of 5, six give a sum of 7, and three outcomes give a sum of 10. Thus,

$$\begin{aligned} P(\text{sum is 5 or 7 or 10}) &= P(\text{sum is 5}) + P(\text{sum is 7}) + P(\text{sum is 10}) \\ &= 4/36 + 6/36 + 3/36 = .3611 \quad \blacksquare \end{aligned}$$

*Calculating the probability of the union of three mutually exclusive events.*

### ■ EXAMPLE 4–31

The probability that a person is in favor of genetic engineering is .55 and that a person is against it is .45. Two persons are randomly selected, and it is observed whether they favor or oppose genetic engineering.

- Draw a tree diagram for this experiment.
- Find the probability that at least one of the two persons favors genetic engineering.

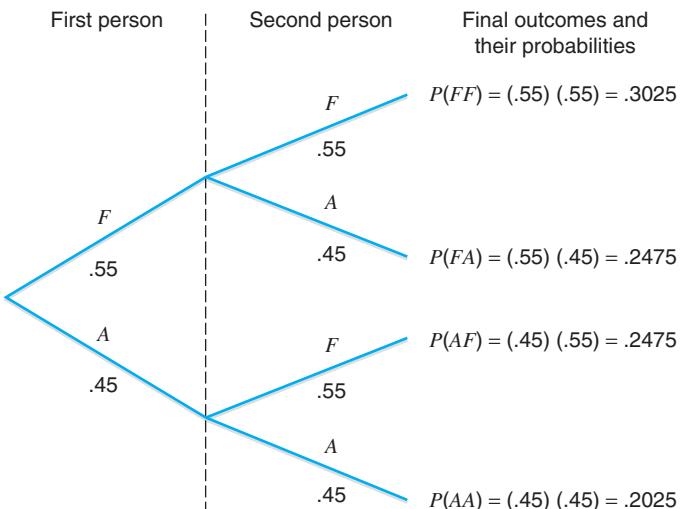
#### Solution

- Let

$$F = \text{a person is in favor of genetic engineering}$$

$$A = \text{a person is against genetic engineering}$$

This experiment has four outcomes: both persons are in favor ( $FF$ ), the first person is in favor and the second is against ( $FA$ ), the first person is against and the second is in favor ( $AF$ ), and both persons are against genetic engineering ( $AA$ ). The tree diagram in Figure 4.21 shows these four outcomes and their probabilities.



**Figure 4.21** Tree diagram.

- The probability that at least one person favors genetic engineering is given by the union of events  $FF$ ,  $FA$ , and  $AF$ . These three outcomes are mutually exclusive. Hence,

$$\begin{aligned} P(\text{at least one person favors}) &= P(FF \text{ or } FA \text{ or } AF) \\ &= P(FF) + P(FA) + P(AF) \\ &= .3025 + .2475 + .2475 = .7975 \end{aligned}$$

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- Explain the meaning of the union of two events. Give one example.
- How is the addition rule of probability for two mutually exclusive events different from the rule for two mutually nonexclusive events?
- Consider the following addition rule to find the probability of the union of two events  $A$  and  $B$ :

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

When and why is the term  $P(A \text{ and } B)$  subtracted from the sum of  $P(A)$  and  $P(B)$ ? Give one example where you might use this formula.

- 4.97** When is the following addition rule used to find the probability of the union of two events  $A$  and  $B$ ?

$$P(A \text{ or } B) = P(A) + P(B)$$

Give one example where you might use this formula.

- 4.98** Find  $P(A \text{ or } B)$  for the following.

- a.  $P(A) = .66$ ,  $P(B) = .47$ , and  $P(A \text{ and } B) = .33$
- b.  $P(A) = .84$ ,  $P(B) = .61$ , and  $P(A \text{ and } B) = .55$

- 4.99** Find  $P(A \text{ or } B)$  for the following.

- a.  $P(A) = .28$ ,  $P(B) = .39$ , and  $P(A \text{ and } B) = .08$
- b.  $P(A) = .41$ ,  $P(B) = .27$ , and  $P(A \text{ and } B) = .19$

- 4.100** Given that  $A$  and  $B$  are two mutually exclusive events, find  $P(A \text{ or } B)$  for the following.

- a.  $P(A) = .38$  and  $P(B) = .59$
- b.  $P(A) = .15$  and  $P(B) = .23$

- 4.101** Given that  $A$  and  $B$  are two mutually exclusive events, find  $P(A \text{ or } B)$  for the following.

- a.  $P(A) = .71$  and  $P(B) = .03$
- b.  $P(A) = .44$  and  $P(B) = .38$

## ■ APPLICATIONS

- 4.102** As mentioned in Exercise 4.52, a July 21 survey on [www.HuffingtonPost.com](http://www.HuffingtonPost.com) asked people to choose their favorite junk food from a list of choices. Although the results were not broken down by gender, suppose that the following table represents the results for the 8002 people who responded, assuming that there were 4801 females and 3201 males included in the survey.

Favorite Junk Food	Female	Male
Chocolate	1518	531
Sugary candy	218	127
Ice cream	685	586
Fast food	312	463
Cookies	431	219
Chips	458	649
Cake	387	103
Pizza	792	523

Suppose that one person is selected at random from this sample of 8002 respondents. Find the following probabilities.

- a. Probability of the union of events *female* and *chocolate*.
- b. Probability of the union of events *male* and *cake*.

- 4.103** The following table gives a two-way classification of all basketball players at a state university who began their college careers between 2004 and 2008, based on gender and whether or not they graduated.

	Graduated	Did Not Graduate
Male	126	55
Female	133	32

If one of these players is selected at random, find the following probabilities.

- a.  $P(\text{female or did not graduate})$
- b.  $P(\text{graduated or male})$

- 4.104** Five hundred employees were selected from a city's large private companies, and they were asked whether or not they have any retirement benefits provided by their companies. Based on this information, the following two-way classification table was prepared.

	Have Retirement Benefits	
	Yes	No
Men	225	75
Women	150	50

Suppose one employee is selected at random from these 500 employees. Find the following probabilities.

- The probability of the union of events "woman" and "yes"
- The probability of the union of events "no" and "man"

- 4.105** Two thousand randomly selected adults were asked whether or not they have ever shopped on the Internet. The following table gives a two-way classification of the responses.

	Have Shopped	Have Never Shopped
Male	500	700
Female	300	500

Suppose one adult is selected at random from these 2000 adults. Find the following probabilities.

- $P(\text{has never shopped on the Internet or is a female})$
- $P(\text{is a male or has shopped on the Internet})$
- $P(\text{has shopped on the Internet or has never shopped on the Internet})$

- 4.106** A consumer agency randomly selected 1700 flights for two major airlines, A and B. The following table gives the two-way classification of these flights based on airline and arrival time. Note that "less than 30 minutes late" includes flights that arrived early or on time.

	Less Than 30 Minutes Late	30 Minutes to 1 Hour Late	More Than 1 Hour Late
Airline A	429	390	92
Airline B	393	316	80

If one flight is selected at random from these 1700 flights, find the following probabilities.

- $P(\text{more than 1 hour late or airline A})$
- $P(\text{airline B or less than 30 minutes late})$
- $P(\text{airline A or airline B})$

- 4.107** Two thousand randomly selected adults were asked if they think they are financially better off than their parents. The following table gives the two-way classification of the responses based on the education levels of the persons included in the survey and whether they are financially better off, the same as, or worse off than their parents.

	Less Than High School	High School	More Than High School
Better off	140	450	420
Same as	60	250	110
Worse off	200	300	70

Suppose one adult is selected at random from these 2000 adults. Find the following probabilities.

- $P(\text{better off or high school})$
- $P(\text{more than high school or worse off})$
- $P(\text{better off or worse off})$

**4.108** There is an area of free (but illegal) parking near an inner-city sports arena. The probability that a car parked in this area will be ticketed by police is .35, that the car will be vandalized is .15, and that it will be ticketed and vandalized is .10. Find the probability that a car parked in this area will be ticketed or vandalized.

**4.109** Amy is trying to purchase concert tickets online for two of her favorite bands, the Leather Recliners and Double Latte No Foam. She estimates that her probability of being able to get tickets for the Leather Recliners concert is .14, the probability of being able to get tickets for the Double Latte No Foam concert is .23, and the probability of being able to get tickets for both concerts is .026. What is the probability that she will be able to get tickets for at least one of the two concerts?

**4.110** Jason and Lisa are planning an outdoor reception following their wedding. They estimate that the probability of bad weather is .25, that of a disruptive incident (a fight breaks out, the limousine is late, etc.) is .15, and that bad weather and a disruptive incident will occur is .08. Assuming these estimates are correct, find the probability that their reception will suffer bad weather or a disruptive incident.

**4.111** The probability that a randomly selected elementary or secondary school teacher from a city is a female is .68, holds a second job is .38, and is a female and holds a second job is .29. Find the probability that an elementary or secondary school teacher selected at random from this city is a female or holds a second job.

**4.112** According to the U.S. Census Bureau's most recent data on the marital status of the 242 million Americans aged 15 years and older, 124.2 million are currently married and 74.5 million have never been married. If one person from these 242 million persons is selected at random, find the probability that this person is currently married *or* has never been married. Explain why this probability is not equal to 1.0.

**4.113** According to a survey of 2000 home owners, 800 of them own homes with three bedrooms, and 600 of them own homes with four bedrooms. If one home owner is selected at random from these 2000 home owners, find the probability that this home owner owns a house that has three *or* four bedrooms. Explain why this probability is not equal to 1.0.

**4.114** According to an Automobile Association of America report, 9.6% of Americans traveled by car over the 2011 Memorial Day weekend and 88.09% stayed home. What is the probability that a randomly selected American stayed home or traveled by car over the 2011 Memorial Day weekend? Explain why this probability does not equal 1.0.

**4.115** Twenty percent of a town's voters favor letting a major discount store move into their neighborhood, 63% are against it, and 17% are indifferent. What is the probability that a randomly selected voter from this town will either be against it or be indifferent? Explain why this probability is not equal to 1.0.

**4.116** The probability that a corporation makes charitable contributions is .72. Two corporations are selected at random, and it is noted whether or not they make charitable contributions.

- Draw a tree diagram for this experiment.
- Find the probability that at most one corporation makes charitable contributions.

**4.117** The probability that an open-heart operation is successful is .84. What is the probability that in two randomly selected open-heart operations at least one will be successful? Draw a tree diagram for this experiment.

## 4.6 Counting Rule, Factorials, Combinations, and Permutations

In this section, first we discuss the counting rule that helps us calculate the total number of outcomes for experiments, and then we learn about factorials, combinations, and permutations, respectively.

### 4.6.1 Counting Rule

The experiments dealt with so far in this chapter have had only a few outcomes, which were easy to list. However, for experiments with a large number of outcomes, it may not be easy to list all outcomes. In such cases, we may use the **counting rule** to find the total number of outcomes.

**Counting Rule to Find Total Outcomes** If an experiment consists of three steps, and if the first step can result in  $m$  outcomes, the second step in  $n$  outcomes, and the third step in  $k$  outcomes, then

$$\text{Total outcomes for the experiment} = m \cdot n \cdot k$$

The counting rule can easily be extended to apply to an experiment that has fewer or more than three steps.

### ■ EXAMPLE 4–32

Applying the counting rule: 3 steps.

Consider three tosses of a coin. How many total outcomes this experiment has?

**Solution** This experiment of tossing a coin three times has three steps: the first toss, the second toss, and the third toss. Each step has two outcomes: a head and a tail. Thus,

$$\text{Total outcomes for three tosses of a coin} = 2 \times 2 \times 2 = 8$$

The eight outcomes for this experiment are  $HHH$ ,  $HHT$ ,  $HTH$ ,  $HTT$ ,  $THH$ ,  $THT$ ,  $TTH$ , and  $TTT$ . ■

### ■ EXAMPLE 4–33

Applying the counting rule: 2 steps.

A prospective car buyer can choose between a fixed and a variable interest rate and can also choose a payment period of 36 months, 48 months, or 60 months. How many total outcomes are possible?

**Solution** This experiment is made up of two steps: choosing an interest rate and selecting a loan payment period. There are two outcomes (a fixed or a variable interest rate) for the first step and three outcomes (a payment period of 36 months, 48 months, or 60 months) for the second step. Hence,

$$\text{Total outcomes} = 2 \times 3 = 6$$

### ■ EXAMPLE 4–34

Applying the counting rule: 16 steps.

A National Football League team will play 16 games during a regular season. Each game can result in one of three outcomes: a win, a loss, or a tie. How many total outcomes are possible?

**Solution** The total possible outcomes for 16 games are calculated as follows:

$$\begin{aligned}\text{Total outcomes} &= 3 \cdot 3 \\ &= 3^{16} = 43,046,721\end{aligned}$$

One of the 43,046,721 possible outcomes is all 16 wins. ■

## 4.6.2 Factorials

The symbol  $!$  (read as *factorial*) is used to denote **factorials**. The value of the factorial of a number is obtained by multiplying all the integers from that number to 1. For example,  $7!$  is read as “seven factorial” and is evaluated by multiplying all the integers from 7 to 1.

### Definition

**Factorials** The symbol  $n!$ , read as “ $n$  factorial,” represents the product of all the integers from  $n$  to 1. In other words,

$$n! = n(n - 1)(n - 2)(n - 3) \cdots 3 \cdot 2 \cdot 1$$

By definition,

$$0! = 1$$

Note that some calculators use  $r!$  instead of  $n!$  on the factorial key.



PhotoDisc, Inc./Getty Images

### ■ EXAMPLE 4–35

Evaluate 7!.

**Solution** To evaluate 7!, we multiply all the integers from 7 to 1.

$$7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = \mathbf{5040}$$

Thus, the value of 7! is 5040. ■

*Evaluating a factorial.*

### ■ EXAMPLE 4–36

Evaluate 10!.

**Solution** The value of 10! is given by the product of all the integers from 10 to 1. Thus,

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = \mathbf{3,628,800} \quad \blacksquare$$

*Evaluating a factorial.*

### ■ EXAMPLE 4–37

Evaluate  $(12 - 4)!$ .

**Solution** The value of  $(12 - 4)!$  is

$$(12 - 4)! = 8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = \mathbf{40,320} \quad \blacksquare$$

*Evaluating a factorial of the difference between two numbers.*

### ■ EXAMPLE 4–38

Evaluate  $(5 - 5)!$ .

*Evaluating a factorial of zero.*

**Solution** The value of  $(5 - 5)!$  is 1.

$$(5 - 5)! = 0! = \mathbf{1}$$

Note that 0! is always equal to 1. ■

Statistical software and most calculators can be used to find the values of factorials. Check if your calculator can evaluate factorials.

### 4.6.3 Combinations

Quite often we face the problem of selecting a few elements from a group of distinct elements. For example, a student may be required to attempt any two questions out of four in an examination. As another example, the faculty in a department may need to select 3 professors from 20 to form a committee, or a lottery player may have to pick 6 numbers from 49. The question arises: In how many ways can we make the selections in each of these examples? For instance, how many possible selections exist for the student who is to choose any two questions out of four? The answer is six. Let the four questions be denoted by the numbers 1, 2, 3, and 4. Then the six selections are

$$(1 \text{ and } 2) \quad (1 \text{ and } 3) \quad (1 \text{ and } 4) \quad (2 \text{ and } 3) \quad (2 \text{ and } 4) \quad (3 \text{ and } 4)$$

The student can choose questions 1 and 2, or 1 and 3, or 1 and 4, and so on. Note that in combinations, all selections are made without replacement.

Each of the possible selections in the above list is called a **combination**. All six combinations are distinct; that is, each combination contains a different set of questions. It is important to remember that the order in which the selections are made is not important in the

case of combinations. Thus, whether we write (1 and 2) or (2 and 1), both of these arrangements represent only one combination.

### Definition

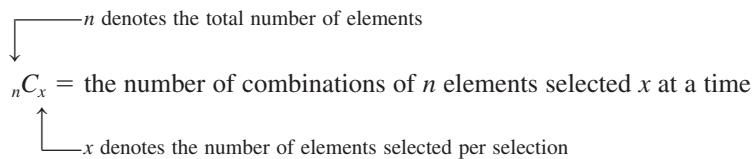
**Combinations Notation** Combinations give the number of ways  $x$  elements can be selected from  $n$  elements. The notation used to denote the total number of combinations is

$${}_nC_x$$

which is read as “the number of combinations of  $n$  elements selected  $x$  at a time.”

Note that some calculators use  $r$  instead of  $x$ , so that the combinations notation then reads  ${}_nC_r$ .

Suppose there are a total of  $n$  elements from which we want to select  $x$  elements. Then,



**Number of Combinations** The *number of combinations* for selecting  $x$  from  $n$  distinct elements is given by the formula

$${}_nC_x = \frac{n!}{x!(n-x)!}$$

where  $n!$ ,  $x!$ , and  $(n - x)!$  are read as “ $n$  factorial,” “ $x$  factorial,” and “ $n$  minus  $x$  factorial,” respectively.

In the combinations formula,

$$n! = n(n-1)(n-2)(n-3)\cdots 3 \cdot 2 \cdot 1$$

$$x! = x(x-1)(x-2)\cdots 3 \cdot 2 \cdot 1$$

$$(n-x)! = (n-x)(n-x-1)(n-x-2)\cdots 3 \cdot 2 \cdot 1$$

Note that in combinations,  $n$  is always greater than or equal to  $x$ . If  $n$  is less than  $x$ , then we cannot select  $x$  distinct elements from  $n$ .

### ■ EXAMPLE 4-39

Finding the number of combinations using the formula.

An ice cream parlor has six flavors of ice cream. Kristen wants to buy two flavors of ice cream. If she randomly selects two flavors out of six, how many combinations are possible?

**Solution** For this example,

$$n = \text{total number of ice cream flavors} = 6$$

$$x = \text{number of ice cream flavors to be selected} = 2$$

Therefore, the number of ways in which Kristen can select two flavors of ice cream out of six is

$${}_6C_2 = \frac{6!}{2!(6-2)!} = \frac{6!}{2!4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 15$$

Thus, there are 15 ways for Kristen to select two ice cream flavors out of six. ■

### ■ EXAMPLE 4-40

Three members of a jury will be randomly selected from five people. How many different combinations are possible?

*Finding the number of combinations and listing them.*

**Solution** There are a total of five persons, and we are to select three of them. Hence,

$$n = 5 \quad \text{and} \quad x = 3$$

Applying the combinations formula, we get

$${}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{120}{6 \cdot 2} = 10$$

If we assume that the five persons are A, B, C, D, and E, then the 10 possible combinations for the selection of three members of the jury are

ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE ■

### ■ EXAMPLE 4-41

Marv & Sons advertised to hire a financial analyst. The company has received applications from 10 candidates who seem to be equally qualified. The company manager has decided to call only 3 of these candidates for an interview. If she randomly selects 3 candidates from the 10, how many total selections are possible?

*Using the combinations formula.*

**Solution** The total number of ways to select 3 applicants from 10 is given by  ${}_{10}C_3$ . Here,  $n = 10$  and  $x = 3$ . We find the number of combinations as follows:

$${}_{10}C_3 = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} = \frac{3,628,800}{(6)(5040)} = 120$$

Thus, the company manager can select 3 applicants from 10 in 120 ways. ■

Statistical software and many calculators can be used to find combinations. Check to see whether your calculator can do so.

If the total number of elements and the number of elements to be selected are the same, then there is only one combination. In other words,

◀ Remember

$${}_nC_n = 1$$

Also, the number of combinations for selecting zero items from  $n$  is 1; that is,

$${}_nC_0 = 1$$

For example,

$${}_5C_5 = \frac{5!}{5!(5-5)!} = \frac{5!}{5!0!} = \frac{120}{(120)(1)} = 1$$

$${}_8C_0 = \frac{8!}{0!(8-0)!} = \frac{8!}{0!8!} = \frac{40,320}{(1)(40,320)} = 1$$

Case Study 4-2 describes the probability of winning a Mega Million Lottery jackpot.

## PROBABILITY OF WINNING A MEGA MILLIONS LOTTERY JACKPOT

Large jackpot lotteries became popular in the United States during the 1970s and 1980s. The introduction of the Powerball lottery in 1992 resulted in the growth of multi-jurisdictional lotteries, as well as the development of a second extensive multi-jurisdictional game called Mega Millions in 2002. Both of these games are offered in 44 jurisdictions (42 states, the District of Columbia, and the U.S. Virgin Islands), which has resulted in multiple jackpots of more than \$300 million.

Both games operate on a similar premise. There are two bins—one containing white balls, and one containing red (Powerball) or gold (Mega Millions) balls. When a player fills out a ticket, he or she selects five numbers from the set of white balls (1–59 for Powerball, 1–56 for Mega Millions) and one number from the set of red (1–35) or gold (1–46) balls, depending on the game. Prizes awarded to players are based on how many balls of each color are matched. If all five white ball numbers and the colored ball number are matched, the player wins the jackpot. If more than one player matches all of the numbers, then the jackpot is divided among them. The following table lists the various prizes for the Mega Millions lottery.

Number of white balls matched	Number of gold balls matched	Prize	Number of white balls matched	Number of gold balls matched	Prize
5	1	Jackpot	2	1	\$10
5	0	\$250,000	3	0	\$7
4	1	\$10,000	1	1	\$3
4	0	\$150	0	1	\$2
3	1	\$150			

The probability of winning each of the various prizes listed in the table for the Mega Millions lottery can be calculated using combinations. First, we need to calculate the number of ways to draw five white ball numbers from 56 and one gold ball number from 46. These combinations are, respectively,

$${}_{56}C_5 = 3819816 \quad \text{and} \quad {}_{46}C_1 = 46$$

To obtain the total number of ways to select six numbers (five white ball numbers and one gold ball number), we multiply the two numbers obtained above, which gives us  $3,819,816 \times 46 = 175,711,536$ . Thus, there are 175,711,536 different sets of five white ball numbers and one gold ball number that can be drawn. Then, the probability that a player with one ticket wins the jackpot is

$$P(\text{winning the jackpot}) = 1/175,711,536 = .00000000569$$

To calculate the probability of winning each of the other prizes, we calculate the number of ways any prize can be won and divide it by 175,711,536. For example, to win a prize of \$10,000, a player needs to match four white ball numbers and the gold ball number. As shown below, there are 255 ways to match four white ball numbers and one gold ball number.

$${}_5C_4 \times {}_{51}C_1 \times {}_1C_1 \times {}_{45}C_0 = 5 \times 51 \times 1 \times 1 = 255$$

Here  ${}_5C_4$  gives the number of ways to match four of the five winning white ball numbers,  ${}_{51}C_1$  gives the number of ways to match one of the 51 nonwinning white ball numbers,  ${}_1C_1$  gives the number of ways to match the winning gold ball number, and  ${}_{45}C_0$  gives the number of ways to match none of the 45 non-winning gold ball numbers. Then, the probability of winning a prize of \$10,000 is

$$P(\text{winning a } \$10,000 \text{ prize}) = 255/175,711,536 = .00000145$$

We can calculate the probabilities of winning the other prizes listed in the table in the same way.

### 4.6.4 Permutations

The concept of permutations is very similar to that of combinations but with one major difference—here the order of selection is important. Suppose there are three marbles in a jar—red, green, and purple—and we select two marbles from these three. When the order of selection is not

important, as we know from the previous section, there are three ways (combinations) to do so. Those three ways are RG, RP, and GP, where R represents that a red marble is selected, G means a green marble is selected, and P indicates a purple marble is selected. In these three combinations, the order of selection is not important, and, thus, RG and GR represent the same selection. However, if the order of selection is important, then RG and GR are not the same selections, but they are two different selections. Similarly, RP and PR are two different selections, and GP and PG are two different selections. Thus, if the order in which the marbles are selected is important, then there are six selections—RG, GR, RP, PR, GP, and PG. These are called six **permutations** or **arrangements**.

### Definition

**Permutations Notation** Permutations give the total number of selections of  $x$  elements from  $n$  (different) elements in such a way that the order of selection is important. The notation used to denote the permutations is

$${}_n P_x$$

which is read as “the number of permutations of selecting  $x$  elements from  $n$  elements.” Permutations are also called **arrangements**.

**Permutations Formula** The following formula is used to find the number of permutations or arrangements of selecting  $x$  items out of  $n$  items. Note that here, the  $n$  items should all be different.

$${}_n P_x = \frac{n!}{(n - x)!}$$

Example 4–42 shows how to apply this formula.

### ■ EXAMPLE 4–42

A club has 20 members. They are to select three office holders—president, secretary, and treasurer—for next year. They always select these office holders by drawing 3 names randomly from the names of all members. The first person selected becomes the president, the second is the secretary, and the third one takes over as treasurer. Thus, the order in which 3 names are selected from the 20 names is important. Find the total arrangements of 3 names from these 20.

*Finding the number of permutations using the formula.*

**Solution** For this example,

$$n = \text{total members of the club} = 20$$

$$x = \text{number of names to be selected} = 3$$

Since the order of selections is important, we find the number of permutations or arrangements using the following formula:

$${}_n P_x = \frac{n!}{(n - x)!} = \frac{20!}{(20 - 3)!} = \frac{20!}{17!} = 6840$$

Thus, there are 6840 permutations or arrangements for selecting 3 names out of 20. ■

Statistical software and many calculators can find permutations. Check to see whether your calculator can do it.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**4.118** How many different outcomes are possible for four rolls of a die?

**4.119** How many different outcomes are possible for 10 tosses of a coin?

**4.120** Determine the value of each of the following using the appropriate formula.

$$3! \quad (9 - 3)! \quad 9! \quad (14 - 12)! \quad {}_5C_3 \quad {}_7C_4 \quad {}_9C_3 \quad {}_4C_0 \quad {}_3C_3 \quad {}_6P_2 \quad {}_8P_4$$

**4.121** Find the value of each of the following using the appropriate formula.

$$6! \quad 11! \quad (7 - 2)! \quad (15 - 5)! \quad {}_8C_2 \quad {}_5C_0 \quad {}_5C_5 \quad {}_6C_4 \quad {}_{11}C_7 \quad {}_9P_6 \quad {}_{12}P_8$$

### ■ APPLICATIONS

**4.122** A small ice cream shop has 10 flavors of ice cream and 5 kinds of toppings for its sundaes. How many different selections of one flavor of ice cream and one kind of topping are possible?

**4.123** A man just bought 4 suits, 8 shirts, and 12 ties. All of these suits, shirts, and ties coordinate with each other. If he is to randomly select one suit, one shirt, and one tie to wear on a certain day, how many different outcomes (selections) are possible?

**4.124** A restaurant menu has four kinds of soups, eight kinds of main courses, five kinds of desserts, and six kinds of drinks. If a customer randomly selects one item from each of these four categories, how many different outcomes are possible?

**4.125** A student is to select three classes for next semester. If this student decides to randomly select one course from each of eight economics classes, six mathematics classes, and five computer classes, how many different outcomes are possible?

**4.126** A ski patrol unit has nine members available for duty, and two of them are to be sent to rescue an injured skier. In how many ways can two of these nine members be selected? Now suppose the order of selection is important. How many arrangements are possible in this case?

**4.127** An ice cream shop offers 25 flavors of ice cream. How many ways are there to select 2 different flavors from these 25 flavors? How many permutations are possible?

**4.128** A veterinarian assigned to a racetrack has received a tip that one or more of the 12 horses in the third race have been doped. She has time to test only 3 horses. How many ways are there to randomly select 3 horses from these 12 horses? How many permutations are possible?

**4.129** An environmental agency will randomly select 4 houses from a block containing 25 houses for a radon check. How many total selections are possible? How many permutations are possible?

**4.130** An investor will randomly select 6 stocks from 20 for an investment. How many total combinations are possible? If the order in which stocks are selected is important, how many permutations will there be?

**4.131** A company employs a total of 16 workers. The management has asked these employees to select 2 workers who will negotiate a new contract with management. The employees have decided to select the 2 workers randomly. How many total selections are possible? Considering that the order of selection is important, find the number of permutations.

**4.132** In how many ways can a sample (without replacement) of 9 items be selected from a population of 20 items?

**4.133** In how many ways can a sample (without replacement) of 5 items be selected from a population of 15 items?

## USES AND MISUSES...

### 1. STATISTICS VERSUS PROBABILITY

At this point, you may think that probability and statistics are basically the same things. They both use the term mean, they both report results in terms of percentages, and so on. Do not be fooled: Although they share many of the same mathematical tools, probability and statistics are very different sciences. The first three chapters of this text were very careful to specify whether a particular set of data was a population or a sample. This is because statistics takes a sample of data and, based upon the properties of that sample—mean, median, mode, standard deviation—attempts to say something about a population. Probability does exactly the opposite: In probability, we know the properties of the population based on the sample space and the probability distribution, and we want to make statements about a sample from the population.

Here's an example viewed from a statistical and a probabilistic point of view. A sequence of outcomes from 10 independent coin tosses is {H, T, H, T, H, T, H, T, H, T}. A statistician will ask the question: Based on the observed 4 heads and 6 tails, what combination of heads and tails would he or she expect from 100 or 1000 tosses, and how certain would he or she be of that answer? Someone using probability will ask: If the coin toss was fair (the probability of the event that a single coin toss be a head or tail is .5), what is the probability that the compound event of four heads and six tails will occur? These are substantially different questions.

The distinction between a statistical approach and a probabilistic approach to a problem can be surprising. Imagine that you must determine the average life of an automotive part. One approach

would be to take a sample of parts, test each of them until they fail to work, and then perform some calculations regarding the distribution of failures. However, if this particular part has outliers with long life spans (several years), you are going to be spending a lot of time in the laboratory. An approach using probabilistic techniques could develop a hypothetical life span based on the physical properties of the part, the conditions of its use, and the manufacturing characteristics. Then you can use your experimental results over a relatively short period of time—including data on those parts that did not fail—to adjust your prior understanding of what makes the part fail, saving yourself a lot of time.

### 2. ODDS AND PROBABILITY

One of the first things we learn in probability is that the sum of the probabilities of all outcomes for an experiment must equal 1.0. We also learn about the probabilities that are developed from relative frequencies and about subjective probabilities. In the latter case, many of the probabilities involve personal opinions of experts in the field. Still, both scenarios (probabilities obtained from relative frequencies and subjective probabilities) require that all probabilities must be nonnegative and the sum of the probabilities of all (simple) outcomes for an experiment must equal 1.0.

Although probabilities and probability models are all around us—in weather prediction, medicine, financial markets, and so forth—they are most obvious in the world of gaming and gambling. Sports betting agencies publish odds of each team winning a specific game or championship. The following table gives the odds, as of August 16,

Team	Odds	Team	Odds
Arizona Cardinals	1:100	Miami Dolphins	1:40
Atlanta Falcons	1:9	Minnesota Vikings	1:60
Baltimore Ravens	1:10	New York Giants	1:15
Buffalo Bills	1:125	New York Jets	1:8
Carolina Panthers	1:125	New England Patriots	1:7
Chicago Bears	1:14	New Orleans Saints	1:9
Cincinnati Bengals	1:100	Oakland Raiders	1:75
Cleveland Browns	1:100	Philadelphia Eagles	1:12
Dallas Cowboys	1:10	Pittsburgh Steelers	1:9
Denver Broncos	1:125	San Diego Chargers	1:9
Detroit Lions	1:60	San Francisco 49ers	1:40
Green Bay Packers	1:7	Seattle Seahawks	1:60
Houston Texans	1:30	St. Louis Rams	1:50
Indianapolis Colts	1:9	Tampa Bay Buccaneers	1:50
Jacksonville Jaguars	1:40	Tennessee Titans	1:60
Kansas City Chiefs	1:25	Washington Redskins	1:40

2011, of each National Football League team winning Super Bowl XLVI, held in February 2012. These odds were obtained from the Web site [www2.vegas.com/gaming/futures/superbowl.html](http://www2.vegas.com/gaming/futures/superbowl.html).

Note that the odds listed in this table are called the odds in favor of winning the Super Bowl. For example, the defending champion Green Bay Packers had 1:7 (which is read as 1 to 7) odds of winning Super Bowl XLVI. If we switch the numbers around, we can state that the odds were 7:1 (or 7 to 1) against the Packers winning Super Bowl XLVI.

How do we convert these odds into probabilities? Let us consider the Green Bay Packers. Odds of 1:7 imply that out of 8 chances, there was 1 chance that the Packers would win Super Bowl XLVI and 7 chances that the Packers would not win Super Bowl XLVI. Thus, the probability that the Packers would win Super Bowl XLVI was  $\frac{1}{1+7} = \frac{1}{8} = .1250$  and the probability that Packers would not win Super Bowl XLVI was  $\frac{7}{1+7} = \frac{7}{8} = .8750$ . Similarly, for the Chicago Bears, the probability of winning Super Bowl XLVI was  $\frac{1}{1+14} = \frac{1}{15} = .0667$  and the probability of not winning Super Bowl XLVI was  $\frac{14}{1+14} = \frac{14}{15} = .9333$ . We can calculate these probabilities for all teams listed in the table by using this procedure.

Note that here the 32 outcomes (that each team would win Super Bowl XLVI) are mutually exclusive events because it is impossible for two or more teams to win the Super Bowl during the same year. Hence, if we add the probabilities of winning Super Bowl XLVI for all teams, we should obtain a value of 1.0. However, if you calculate the probability of winning Super Bowl XLVI for each of the

32 teams using the odds given in the table and then add all these probabilities, the sum is 1.588759461. So, what happened? Did these odds makers flunk their statistics and probability courses? Probably not.

Casinos and odds makers, who are in the business of making money, are interested in encouraging people to gamble. These probabilities, which seem to violate the basic rule of probability theory, still obey the primary rule for the casinos, which is that, on average, a casino is going to make a profit.

**Note:** When casinos create odds for sports betting, they recognize that many people will bet on one of their favorite teams, such as the Dallas Cowboys or the Pittsburgh Steelers. To meet the rule that the sum of all of the probabilities is 1.0, the probabilities for the teams more likely to win would have to be lowered. Lowering a probability corresponds to lowering the odds. For example, if the odds for the Green Bay Packers had been lowered from 1:7 to 1:20, the probability for them to win would have decreased from .125 to .0476. If the Packers had remained as one of the favorites, many people would have bet on them. However, if they had won, the casino would have paid \$20, instead of \$7, for every \$1 bet. The casinos do not want to do this and, hence, they ignore the probability rule in order to make more money. However, the casinos cannot do this with their traditional games, which are bound by the standard rules. From a mathematical standpoint, it is not acceptable to ignore the rule that the probabilities of all final outcomes for an experiment must add up to 1.0. (For the information of the reader, the New York Giants won Super Bowl XLVI in February 2012.)

## Glossary

**Classical probability rule** The method of assigning probabilities to outcomes or events of an experiment with equally likely outcomes.

**Combinations** The number of ways  $x$  elements can be selected from  $n$  elements. Here order of selection is not important.

**Complementary events** Two events that taken together include all the outcomes for an experiment but do not contain any common outcome.

**Compound event** An event that contains more than one outcome of an experiment. It is also called a *composite event*.

**Conditional probability** The probability of an event subject to the condition that another event has already occurred.

**Dependent events** Two events for which the occurrence of one changes the probability of the occurrence of the other.

**Equally likely outcomes** Two (or more) outcomes or events that have the same probability of occurrence.

**Event** A collection of one or more outcomes of an experiment.

**Experiment** A process with well-defined outcomes that, when performed, results in one and only one of the outcomes per repetition.

**Factorial** Denoted by the symbol  $!$ . The product of all the integers from a given number to 1. For example,  $n!$  (read as “ $n$  factorial”) represents the product of all the integers from  $n$  to 1.

**Impossible event** An event that cannot occur.

**Independent events** Two events for which the occurrence of one does not change the probability of the occurrence of the other.

**Intersection of events** The intersection of events is given by the outcomes that are common to two (or more) events.

**Joint probability** The probability that two (or more) events occur together.

**Law of Large Numbers** If an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability.

**Marginal probability** The probability of one event or characteristic without consideration of any other event.

**Mutually exclusive events** Two or more events that do not contain any common outcome and, hence, cannot occur together.

**Outcome** The result of the performance of an experiment.

**Permutations** Number of arrangements of  $x$  items selected from  $n$  items. Here order of selection is important.

**Probability** A numerical measure of the likelihood that a specific event will occur.

**Relative frequency as an approximation of probability** Probability assigned to an event based on the results of an experiment or based on historical data.

**Sample point** An outcome of an experiment.

**Sample space** The collection of all sample points or outcomes of an experiment.

**Simple event** An event that contains one and only one outcome of an experiment. It is also called an *elementary event*.

**Subjective probability** The probability assigned to an event based on the information and judgment of a person.

**Sure event** An event that is certain to occur.

**Tree diagram** A diagram in which each outcome of an experiment is represented by a branch of a tree.

**Union of two events** Given by the outcomes that belong either to one or to both events.

**Venn diagram** A picture that represents a sample space or specific events.

## Supplementary Exercises

- 4.134** A car rental agency currently has 44 cars available, 28 of which have a GPS navigation system. One of the 44 cars is selected at random. Find the probability that this car

- has a GPS navigation system
- does not have a GPS navigation system

- 4.135** In a class of 35 students, 13 are seniors, 9 are juniors, 8 are sophomores, and 5 are freshmen. If one student is selected at random from this class, what is the probability that this student is

- a junior?
- a freshman?

- 4.136** A random sample of 250 juniors majoring in psychology or communication at a large university is selected. These students are asked whether or not they are happy with their majors. The following table gives the results of the survey. Assume that none of these 250 students is majoring in both areas.

	Happy	Unhappy
Psychology	80	20
Communication	115	35

- If one student is selected at random from this group, find the probability that this student is
  - happy with the choice of major
  - a psychology major
  - a communication major given that the student is happy with the choice of major
  - unhappy with the choice of major given that the student is a psychology major
  - a psychology major *and* is happy with that major
  - a communication major *or* is unhappy with his or her major
- Are the events “psychology major” and “happy with major” independent? Are they mutually exclusive? Explain why or why not.

- 4.137** A random sample of 250 adults was taken, and they were asked whether they prefer watching sports or opera on television. The following table gives the two-way classification of these adults.

	Prefer Watching Sports	Prefer Watching Opera
Male	96	24
Female	45	85

- If one adult is selected at random from this group, find the probability that this adult
  - prefers watching opera
  - is a male
  - prefers watching sports given that the adult is a female
  - is a male given that he prefers watching sports
  - is a female *and* prefers watching opera
  - prefers watching sports *or* is a male
- Are the events “female” and “prefers watching sports” independent? Are they mutually exclusive? Explain why or why not.

**4.138** A random sample of 80 lawyers was taken, and they were asked if they are in favor of or against capital punishment. The following table gives the two-way classification of their responses.

	Favors Capital Punishment	Opposes Capital Punishment
Male	32	24
Female	13	11

- a. If one lawyer is randomly selected from this group, find the probability that this lawyer
  - i. favors capital punishment
  - ii. is a female
  - iii. opposes capital punishment given that the lawyer is a female
  - iv. is a male given that he favors capital punishment
  - v. is a female *and* favors capital punishment
  - vi. opposes capital punishment *or* is a male
- b. Are the events “female” and “opposes capital punishment” independent? Are they mutually exclusive? Explain why or why not.

**4.139** A random sample of 400 college students was asked if college athletes should be paid. The following table gives a two-way classification of the responses.

	Should Be Paid	Should Not Be Paid
Student athlete	90	10
Student nonathlete	210	90

- a. If one student is randomly selected from these 400 students, find the probability that this student
  - i. is in favor of paying college athletes
  - ii. favors paying college athletes given that the student selected is a nonathlete
  - iii. is an athlete *and* favors paying student athletes
  - iv. is a nonathlete *or* is against paying student athletes
- b. Are the events “student athlete” and “should be paid” independent? Are they mutually exclusive? Explain why or why not.

**4.140** An appliance repair company that makes service calls to customers’ homes has found that 5% of the time there is nothing wrong with the appliance and the problem is due to customer error (appliance unplugged, controls improperly set, etc.). Two service calls are selected at random, and it is observed whether or not the problem is due to customer error. Draw a tree diagram. Find the probability that in this sample of two service calls

- a. both problems are due to customer error
- b. at least one problem is not due to customer error

**4.141** According to the National Science Foundation, during the Fall 2008 semester (the most recent data available) 30.13% of all science and engineering graduate students enrolled in doctorate-granting colleges in the United States were temporary visa holders ([www.nsf.gov/statistics/nsf11311/pdf/tab46.pdf](http://www.nsf.gov/statistics/nsf11311/pdf/tab46.pdf)). Assume that this percentage holds true for current such students. Suppose that two students from the aforementioned group are selected and their status (U.S. citizens/permanent residents or temporary visa holders) is observed. Draw a tree diagram for this problem using C to denote U.S. citizens/permanent residents and V to denote temporary visa holders. Find the probability that in this sample of two graduate students

- a. at least one student is a temporary visa holder
- b. both students are U.S. citizens/permanent residents

**4.142** Refer to Exercise 4.134. Two cars are selected at random from these 44 cars. Find the probability that both of these cars have GPS navigation systems.

**4.143** Refer to Exercise 4.135. Two students are selected at random from this class of 35 students. Find the probability that the first student selected is a junior and the second is a sophomore.

**4.144** A company has installed a generator to back up the power in case there is a power failure. The probability that there will be a power failure during a snowstorm is .30. The probability that the generator will stop working during a snowstorm is .09. What is the probability that during a snowstorm the company will lose both sources of power? Note that the two sources of power are independent.

**4.145** Terry & Sons makes bearings for autos. The production system involves two independent processing machines so that each bearing passes through these two processes. The probability that the first processing machine is not working properly at any time is .08, and the probability that the second machine is not working properly at any time is .06. Find the probability that both machines will not be working properly at any given time.

## Advanced Exercises

**4.146** A player plays a roulette game in a casino by betting on a single number each time. Because the wheel has 38 numbers, the probability that the player will win in a single play is  $1/38$ . Note that each play of the game is independent of all previous plays.

- Find the probability that the player will win for the first time on the 10th play.
- Find the probability that it takes the player more than 50 plays to win for the first time.
- A gambler claims that because he has 1 chance in 38 of winning each time he plays, he is certain to win at least once if he plays 38 times. Does this sound reasonable to you? Find the probability that he will win at least once in 38 plays.

**4.147** A certain state's auto license plates have three letters of the alphabet followed by a three-digit number.

- How many different license plates are possible if all three-letter sequences are permitted and any number from 000 to 999 is allowed?
- Arnold witnessed a hit-and-run accident. He knows that the first letter on the license plate of the offender's car was a B, that the second letter was an O or a Q, and that the last number was a 5. How many of this state's license plates fit this description?

**4.148** The median life of Brand LT5 batteries is 100 hours. What is the probability that in a set of three such batteries, exactly two will last longer than 100 hours?

**4.149** Powerball is a game of chance that has generated intense interest because of its large jackpots. To play this game, a player selects five different numbers from 1 through 59, and then picks a Powerball number from 1 through 39. The lottery organization randomly draws 5 different white balls from 59 balls numbered 1 through 59, and then randomly picks a Powerball number from 1 through 39. Note that it is possible for the Powerball number to be the same as one of the first five numbers.

- If a player's first five numbers match the numbers on the five white balls drawn by the lottery organization and the player's Powerball number matches the Powerball number drawn by the lottery organization, the player wins the jackpot. Find the probability that a player who buys one ticket will win the jackpot. (Note that the order in which the five white balls are drawn is unimportant.)
- If a player's first five numbers match the numbers on the five white balls drawn by the lottery organization, the player wins about \$200,000. Find the probability that a player who buys one ticket will win this prize.

**4.150** A trimotor plane has three engines—a central engine and an engine on each wing. The plane will crash only if the central engine fails *and* at least one of the two wing engines fails. The probability of failure during any given flight is .005 for the central engine and .008 for each of the wing engines. Assuming that the three engines operate independently, what is the probability that the plane will crash during a flight?

**4.151** A box contains 10 red marbles and 10 green marbles.

- Sampling at random from this box five times with replacement, you have drawn a red marble all five times. What is the probability of drawing a red marble the sixth time?
- Sampling at random from this box five times without replacement, you have drawn a red marble all five times. Without replacing any of the marbles, what is the probability of drawing a red marble the sixth time?
- You have tossed a fair coin five times and have obtained heads all five times. A friend argues that according to the law of averages, a tail is due to occur and, hence, the probability of obtaining a head on the sixth toss is less than .50. Is he right? Is coin tossing mathematically equivalent to the procedure mentioned in part a or the procedure mentioned in part b above? Explain.

**4.152** A gambler has four cards—two diamonds and two clubs. The gambler proposes the following game to you: You will leave the room and the gambler will put the cards face down on a table. When you return to the room, you will pick two cards at random. You will win \$10 if both cards are diamonds, you will win \$10 if both are clubs, and for any other outcome you will lose \$10. Assuming that there is no cheating, should you accept this proposition? Support your answer by calculating your probability of winning \$10.

**4.153** A thief has stolen Roger's automatic teller machine (ATM) card. The card has a four-digit personal identification number (PIN). The thief knows that the first two digits are 3 and 5, but he does not know the last two digits. Thus, the PIN could be any number from 3500 to 3599. To protect the customer, the automatic teller machine will not allow more than three unsuccessful attempts to enter the PIN. After the third wrong PIN, the machine keeps the card and allows no further attempts.

- What is the probability that the thief will find the correct PIN within three tries? (Assume that the thief will not try the same wrong PIN twice.)
- If the thief knew that the first two digits were 3 and 5 and that the third digit was either 1 or 7, what is the probability of the thief guessing the correct PIN in three attempts?

**4.154** Consider the following games with two dice.

- A gambler is going to roll a die four times. If he rolls at least one 6, you must pay him \$5. If he fails to roll a 6 in four tries, he will pay you \$5. Find the probability that you must pay the gambler. Assume that there is no cheating.
- The same gambler offers to let you roll a pair of dice 24 times. If you roll at least one double 6, he will pay you \$10. If you fail to roll a double 6 in 24 tries, you will pay him \$10. The gambler says that you have a better chance of winning because your probability of success on each of the 24 rolls is  $1/36$  and you have 24 chances. Thus, he says, your probability of winning \$10 is  $24(1/36) = 2/3$ . Do you agree with this analysis? If so, indicate why. If not, point out the fallacy in his argument, and then find the correct probability that you will win.

**4.155** A gambler has given you two jars and 20 marbles. Of these 20 marbles, 10 are red and 10 are green. You must put all 20 marbles in these two jars in such a way that each jar must have at least one marble in it. Then a friend of yours, who is blindfolded, will select one of the two jars at random and then will randomly select a marble from this jar. If the selected marble is red, you and your friend win \$100.

- If you put 5 red marbles and 5 green marbles in each jar, what is the probability that your friend selects a red marble?
- If you put 2 red marbles and 2 green marbles in one jar and the remaining marbles in the other jar, what is the probability that your friend selects a red marble?
- How should these 20 marbles be distributed among the two jars in order to give your friend the highest possible probability of selecting a red marble?

**4.156** A screening test for a certain disease is prone to giving false positives or false negatives. If a patient being tested has the disease, the probability that the test indicates a (false) negative is .13. If the patient does not have the disease, the probability that the test indicates a (false) positive is .10. Assume that 3% of the patients being tested actually have the disease. Suppose that one patient is chosen at random and tested. Find the probability that

- this patient has the disease and tests positive
- this patient does not have the disease and tests positive
- this patient tests positive
- this patient has the disease given that he or she tests positive

(Hint: A tree diagram may be helpful in part c.)

**4.157** A pizza parlor has 12 different toppings available for its pizzas, and 2 of these toppings are pepperoni and anchovies. If a customer picks 2 toppings at random, find the probability that

- neither topping is anchovies
- pepperoni is one of the toppings

**4.158** An insurance company has information that 93% of its auto policy holders carry collision coverage or uninsured motorist coverage on their policies. Eighty percent of the policy holders carry collision coverage, and 60% have uninsured motorist coverage.

- What percentage of these policy holders carry both collision and uninsured motorist coverage?
- What percentage of these policy holders carry neither collision nor uninsured motorist coverage?
- What percentage of these policy holders carry collision but not uninsured motorist coverage?

**4.159** Many states have a lottery game, usually called a Pick-4, in which you pick a four-digit number such as 7359. During the lottery drawing, there are four bins, each containing balls numbered 0 through 9. One ball is drawn from each bin to form the four-digit winning number.

- You purchase one ticket with one four-digit number. What is the probability that you will win this lottery game?
- There are many variations of this game. The primary variation allows you to win if the four digits in your number are selected in any order as long as they are the same four digits as obtained

by the lottery agency. For example, if you pick four digits making the number 1265, then you will win if 1265, 2615, 5216, 6521, and so forth, are drawn. The variations of the lottery game depend on how many unique digits are in your number. Consider the following four different versions of this game.

- i. All four digits are unique (e.g., 1234)
- ii. Exactly one of the digits appears twice (e.g., 1223 or 9095)
- iii. Two digits each appear twice (e.g., 2121 or 5588)
- iv. One digit appears three times (e.g., 3335 or 2722)

Find the probability that you will win this lottery in each of these four situations.

**4.160** A restaurant chain is planning to purchase 100 ovens from a manufacturer, provided that these ovens pass a detailed inspection. Because of high inspection costs, 5 ovens are selected at random for inspection. These 100 ovens will be purchased if at most 1 of the 5 selected ovens fails inspection. Suppose that there are 8 defective ovens in this batch of 100 ovens. Find the probability that this batch of ovens is purchased. (Note: In Chapter 5 you will learn another method to solve this problem.)

**4.161** A production system has two production lines; each production line performs a two-part process, and each process is completed by a different machine. Thus, there are four machines, which we can identify as two first-level machines and two second-level machines. Each of the first-level machines works properly 98% of the time, and each of the second-level machines works properly 96% of the time. All four machines are independent in regard to working properly or breaking down. Two products enter this production system, one in each production line.

- a. Find the probability that both products successfully complete the two-part process (i.e., all four machines are working properly).
- b. Find the probability that neither product successfully completes the two-part process (i.e., at least one of the machines in each production line is not working properly).

**4.162** The Big Six Wheel (or Wheel of Fortune) is a casino and carnival game that is well known for being a big money maker for the casinos. The wheel has 54 equally likely slots (outcomes) on it. The slot that pays the largest amount of money is called the *joker*. If a player bets on the *joker*, the probability of winning is  $1/54$ . The outcome of any given play of this game (a spin of the wheel) is independent of the outcomes of previous plays.

- a. Find the probability that a player who always bets on *joker* wins for the first time on the 15th play of the game.
- b. Find the probability that it takes a player who always bets on *joker* more than 70 plays to win for the first time.

**4.163** A Wired Equivalent Privacy (WEP) key is a security code that one must enter in order to access a secure WiFi network. The characters in the key are used from the numbers 0 to 9 and letters from A to F, which gives 16 possibilities for each character of the key. Note that repeats are allowed, that is, the same letter or number can be used more than once in a key. A WEP key for a WiFi network with 64-bit security is 10 characters long.

- a. How many different 64-bit WEP keys can be made by using the given numbers and letters?
- b. A specific 64-bit network has a WEP key in which the 2nd, 5th, 8th, and 9th characters are numbers and the other 6 characters are letters. How many different WEP keys are possible for this network?
- c. A hacker has determined that the WiFi network mentioned in part b will lock him out if he makes 20,000 unsuccessful attempts to break into the network. What is the probability that the hacker will be locked out of the network?

**4.164** A large university has 12,600 male students. Of these students, 5312 are members of so-called “Greek” social organizations (fraternities or sororities), 2844 are members of Greek service organizations, and the others are not members of either of these two types of Greek organizations. Similarly, the female students are members of Greek social organizations, Greek service organizations, or neither. Assuming that gender and membership are independent events, find the probabilities of the events in parts a to c:

- a. A student is a member of a Greek social organization given that the student is a female.
- b. A student is a member of a Greek service organization given that the student is a female.
- c. A student is not a member of either of these two types of Greek organizations given that the student is a female.
- d. If the university has 14,325 female students, is it possible that

$$P(\text{Greek social organization} \mid \text{male}) = P(\text{Greek social organization} \mid \text{female})?$$

Explain why/why not.

## Self-Review Test

---

1. The collection of all outcomes for an experiment is called
  - a. a sample space
  - b. the intersection of events
  - c. joint probability
2. A final outcome of an experiment is called
  - a. a compound event
  - b. a simple event
  - c. a complementary event
3. A compound event includes
  - a. all final outcomes
  - b. exactly two outcomes
  - c. more than one outcome for an experiment
4. Two equally likely events
  - a. have the same probability of occurrence
  - b. cannot occur together
  - c. have no effect on the occurrence of each other
5. Which of the following probability approaches can be applied only to experiments with equally likely outcomes?
  - a. Classical probability
  - b. Empirical probability
  - c. Subjective probability
6. Two mutually exclusive events
  - a. have the same probability
  - b. cannot occur together
  - c. have no effect on the occurrence of each other
7. Two independent events
  - a. have the same probability
  - b. cannot occur together
  - c. have no effect on the occurrence of each other
8. The probability of an event is always
  - a. less than 0
  - b. in the range 0 to 1.0
  - c. greater than 1.0
9. The sum of the probabilities of all final outcomes of an experiment is always
  - a. 100
  - b. 1.0
  - c. 0
10. The joint probability of two mutually exclusive events is always
  - a. 1.0
  - b. between 0 and 1
  - c. 0
11. Two independent events are
  - a. always mutually exclusive
  - b. never mutually exclusive
  - c. always complementary
12. A couple is planning their wedding reception. The bride's parents have given them a choice of four reception facilities, three caterers, five DJs, and two limo services. If the couple randomly selects one reception facility, one caterer, one DJ, and one limo service, how many different outcomes are possible?
13. Lucia graduated this year with an accounting degree from Eastern Connecticut State University. She has received job offers from an accounting firm, an insurance company, and an airline. She cannot decide which of the three job offers she should accept. Suppose she decides to randomly select one of these three job offers. Find the probability that the job offer selected is
  - a. from the insurance company
  - b. not from the accounting firm
14. There are 200 students in a particular graduate program at a state university. Of them, 110 are female and 125 are out-of-state students. Of the 110 females, 70 are out-of-state students.
  - a. Are the events "female" and "out-of-state student" independent? Are they mutually exclusive? Explain why or why not.
  - b. If one of these 200 students is selected at random, what is the probability that the student selected is
    - i. a male?
    - ii. an out-of-state student given that this student is a female?
15. Reconsider Problem 14. If one of these 200 students is selected at random, what is the probability that the selected student is a female *or* an out-of-state student?

- 16.** Reconsider Problem 14. If two of these 200 students are selected at random, what is the probability that both of them are out-of-state students?
- 17.** The probability that an adult has ever experienced a migraine headache is .35. If two adults are randomly selected, what is the probability that neither of them has ever experienced a migraine headache?
- 18.** A hat contains five green, eight red, and seven blue marbles. Let  $A$  be the event that a red marble is drawn if we randomly select one marble out of this hat. What is the probability of  $A$ ? What is the complementary event of  $A$ , and what is its probability?
- 19.** The probability that a randomly selected student from a college is a female is .55 and the probability that a student works for more than 10 hours per week is .62. If these two events are independent, find the probability that a randomly selected student is a
- male *and* works for more than 10 hours per week
  - female *or* works for more than 10 hours per week
- 20.** A sample was selected of 506 workers who currently receive two weeks of paid vacation per year. These workers were asked if they were willing to accept a small pay cut to get an additional week of paid vacation a year. The following table shows the responses of these workers.

	Yes	No	No Response
Man	77	140	32
Woman	104	119	34

- If one person is selected at random from these 506 workers, find the following probabilities.
  - $P(\text{yes})$
  - $P(\text{yes} \mid \text{woman})$
  - $P(\text{woman} \text{ and } \text{no})$
  - $P(\text{no response or man})$
- Are the events “woman” and “yes” independent? Are they mutually exclusive? Explain why or why not.

## Mini-Projects

### ■ MINI-PROJECT 4-1

Suppose that a small chest contains three drawers. The first drawer contains two \$1 bills, the second drawer contains two \$100 bills, and the third drawer contains one \$1 bill and one \$100 bill. Suppose that first a drawer is selected at random and then one of the two bills inside that drawer is selected at random. We can define these events:

$$\begin{array}{ll} A = \text{the first drawer is selected} & B = \text{the second drawer is selected} \\ C = \text{the third drawer is selected} & D = \text{a \$1 bill is selected} \end{array}$$

- Suppose when you randomly select one drawer and then one bill from that drawer, the bill you obtain is a \$1 bill. What is the probability that the second bill in this drawer is a \$100 bill? In other words, find the probability  $P(C \mid D)$  because for the second bill to be \$100, it has to be the third drawer. Answer this question intuitively without making any calculations.
- Use the relative frequency concept of probability to estimate  $P(C \mid D)$  as follows. First select a drawer by rolling a die once. If either 1 or 2 occurs, the first drawer is selected; if either 3 or 4 occurs, the second drawer is selected; and if either 5 or 6 occurs, the third drawer is selected. Whenever  $C$  occurs, then select a bill by tossing a coin once. (Note that if either  $A$  or  $B$  occurs, then you do not need to toss the coin because each of these drawers contains both bills of the same denomination.) If you obtain a head, assume that you select a \$1 bill; if you obtain a tail, assume that you select a \$100 bill. Repeat this process 100 times. How many

times in these 100 repetitions did the event  $D$  occur? What proportion of the time did  $C$  occur when  $D$  occurred? Use this proportion to estimate  $P(C | D)$ . Does this estimate support your guess of  $P(C | D)$  in part a?

- c. Calculate  $P(C | D)$  using the procedures developed in this chapter (a tree diagram may be helpful). Was your estimate in part b close to this value? Explain.

### ■ MINI-PROJECT 4-2

There are two families playing in a park, and each of these two families has two children. The Smith family has two daughters, and the Jones family has a daughter and a son. One family is selected at random, and one of the children from this family is chosen at random.

- a. Suppose that the selected child is a girl. What is the probability that the second child is also a girl? (Note: you need to determine this using conditional probability.)
- b. Use the relative frequency concept to estimate the probability that the second child in the family is a girl given that the selected child is a girl. Use the following process to do so. First toss a coin to determine whether the Smith family or the Jones family is chosen. If the Smith family is selected, then record that the second child in this family is a girl given that the selected child is a girl. This is so because both children in this family are girls. If the Jones family is selected, then toss the coin again to select a child and record the gender of the child selected and that of the second child in this family. Repeat this process 50 times, and then use the results to estimate the required probability. How close is your estimate to the probability calculated in part a?

### ■ MINI-PROJECT 4-3

The dice game Yahtzee® involves five standard dice. On your turn, you can roll all five or fewer dice up to three times to obtain different sets of numbers on the dice. For example, you will roll all five dice the first time; if you like two of the five numbers obtained, you can roll the other three dice a second time; now if you like three of the five numbers obtained, you can roll the other two dice the third time. Some of the sets of numbers obtained are similar to poker hands (three of a kind, four of a kind, full house, and so forth). However, a few other hands, like five of a kind (called a *yahtzee*), are not poker hands (or at least they are not the hands you would dare to show anyone).

For the purpose of this project, we will examine the outcomes on the first roll of the five dice. The five scenarios that we will consider are:

- i. Three of a kind—three numbers are the same and the remaining two numbers are both different, for example, 22254
- ii. Four of a kind—four numbers are the same and the fifth number is different, for example, 44442
- iii. Full house—three numbers are the same and the other two numbers are the same, for example, 33366
- iv. Large straight—five numbers in a row, for example, 12345
- v. Yahtzee—all five numbers are the same, for example, 33333

In the first two cases, the dice that are not part of the three or four of a kind must have different values than those in the three or four of a kind. For example, 22252 cannot be considered three of a kind, but 22254 is three of a kind. (Yahtzee players know that this situation differs from the rules of the actual game, but for the purpose of this project, we will change the rules.)

- a. Find the probability for each of these five cases for one roll of the five dice.
- b. In a regular game, you do not have to roll all five dice on each of the three rolls. You can leave some dice on the table and roll the others in an attempt to improve your score. For example, if you roll 13555 on the first roll of five dice, you can keep the three fives and roll the dice with 1 and 3 outcomes the second time in an attempt to get more fives, or possibly a pair of another number in order to get a full house. After your second roll, you are allowed to pick up any of the dice for your third roll. For example, suppose your first roll is 13644 and you keep the two fours. Then you roll the dice with 1, 3, and 6 outcomes the second time and obtain three fives. Thus, now you have 55544. Although you met your full house requirement before rolling the dice three times, you still need a yahtzee. So, you keep the fives and roll the two dice with fours the third time. Write a paragraph outlining all of the scenarios you will have to consider to calculate the probability of obtaining a yahtzee within your three rolls.

**DECIDE FOR YOURSELF****DECIDING ABOUT PRODUCTION PROCESSES**

Henry Ford was one of the major developers of mass production. Imagine if his factory had only one production line! If any component in that production line would have broken down, all production would have come to a halt. In order for mass production to be successful, the factory must be able to continue production when one or more machines in the production process break down. Automobile factories, like many other forms of production, have multiple production lines running side by side. So if one production line is shut down due to a breakdown, the other production lines can still operate. Probability theory can be used to study the reliability of production systems by determining the likelihood that a system will continue to operate even when some parts of the system fail.

In order to study such systems, we have to consider how they are set up. These systems comprise two types of arrangements: series and parallel. In a series system, a process is sequential. One part of the process must be completed before the item can move to the next part of the process. If any part of the system breaks down, none of the tasks that follow can be completed. In the auto example, if something in a series system breaks down while the chassis is being constructed, it will be impossible to install the seats, the windshield, the engine, and so forth.

In a parallel system, various processes work side by side. In some cases the processes are like toll collectors at a bridge or on a highway. As long as there is at least one toll collector working, traffic will continue moving, although more toll collectors would certainly speed up the process. In a computer network, different

servers are set up in parallel systems. If one server (such as the e-mail server) goes down, people on the network can still access the Web and file servers. However, if the servers were set up in a series system and the e-mail server failed, nobody would be able to do anything.

Let us consider a simplified example. Suppose a production line involves five tasks. Each of the machines that perform these tasks works successfully 97% of the time. In other words, the probability that a specific task can be completed (without interruption) is .97. For the sake of simplicity, let us assume that the machines work and fail independently of each other. Furthermore, suppose that the factory has three of these lines running in a parallel system. Following are some of the questions that arise. Act as if you are in charge of such a production process and try to answer these questions.

- 1.** What is the probability that all five tasks in a single line are completed without interruption?
- 2.** What is the probability that at least one of the three production lines is working properly?
- 3.** Why is the probability that a specific line works properly lower than the probability that at least one of the lines in the factory works properly?
- 4.** What happens to the reliability of the system if an additional task is added to each line?
- 5.** What happens when the number of tasks remains constant, but another line is added?

# T ECHNOLOGY INSTRUCTION

**Generating Random Numbers****TI-84**

- 1.** To generate a random number (not necessarily an integer) uniformly distributed between  $m$  and  $n$ , select **MATH >PRB** and type **rand\*(n-m)+m**.
- 2.** To generate a random number that is an integer uniformly distributed between  $m$  and  $n$ , select **MATH >PRB** and type **randInt(m,n)**.
- 3.** To create a sequence of random numbers (integer or noninteger) and store them in a list, you will need to use the **seq(** function in conjunction with the appropriate random number function from step 1 or step 2. Specifically, select **2nd >STAT >OPS seq(** to enter the sequence menu. In this menu, enter the function from step 1 or step 2 at the **Expr:** prompt, **X** at the **Variable:** prompt, **1** at the **start:** prompt, the quantity of random numbers you want at the **end:** prompt, and **1** at the **step:** prompt. Highlight **Paste** and press **ENTER**, which will paste the command on the home screen. Now, type **>STO >L1**

**>ENTER.** These instructions will store the data in list **L1** (see **Screen 4.1** and **Screen 4.2**). However, you can replace L1 by any other list you want in the above instructions.

- 4.** To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is *not* important, type the value of  $n$ , select **MATH > PRB > nCr**, type the value of  $x$ , and press **ENTER**. (See **Screen 4.3**.)

```
589
Expr:randInt(6,
Variable:X
start:1
end:50
step:1
Paste
```

**Screen 4.1**

```
40,X,1,50,1>L1■
```

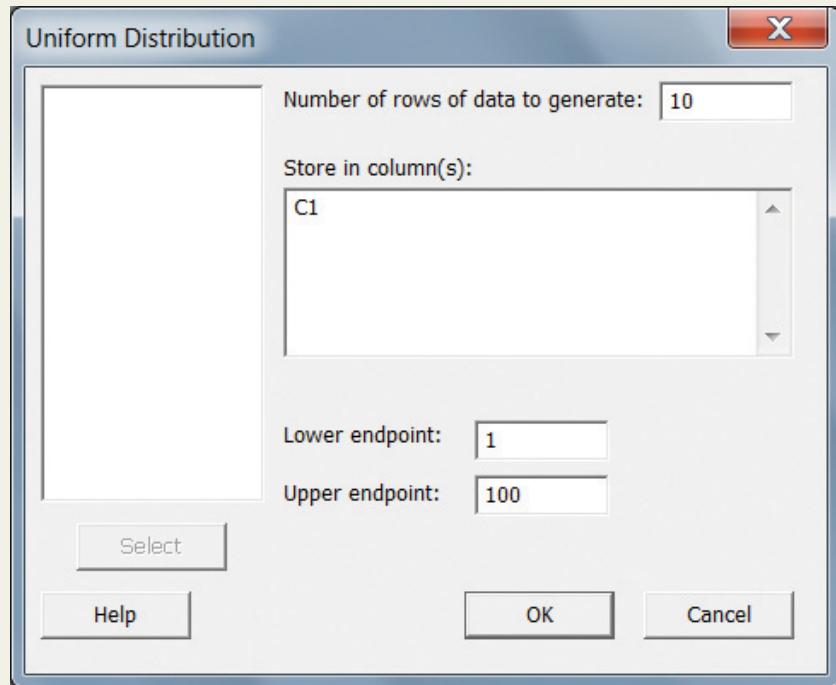
**Screen 4.2**

10 nCr 3	120
----------	-----

**Screen 4.3**

5. To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is important, type the value of  $n$ , select **MATH > PRB > nPr**, type the value of  $x$ , and press **ENTER**.

### Minitab

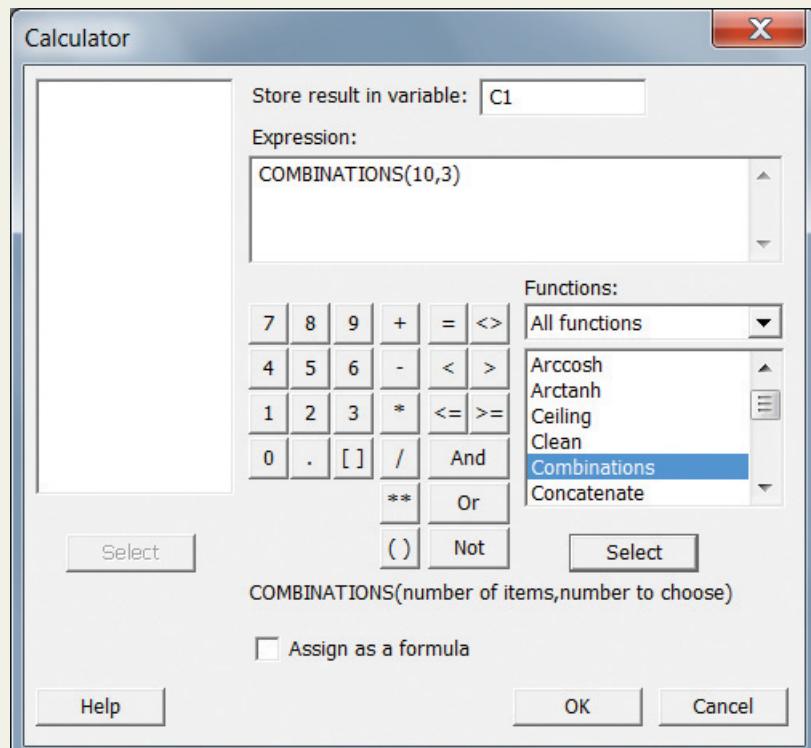


Screen 4.4

- To generate random numbers (not necessarily integers) uniformly distributed between  $m$  and  $n$ , select **Calc > Random Data > Uniform**. Enter the number of rows of data, the column where you wish to store the data, and the minimum  $m$  and maximum  $n$  values for the numbers (see Screens 4.4 and 4.5).

	C1
1	27.3425
2	8.4796
3	45.4208
4	27.3505
5	86.9136
6	39.6677
7	8.0859
8	2.6557
9	68.5980
10	14.6074

Screen 4.5



Screen 4.6

- To generate random integers uniformly distributed between  $m$  and  $n$ , select **Calc > Random Data > Integer**. Enter the number of rows of data, the column where you wish to store the data, and the minimum  $m$  and maximum  $n$  values for the integers.
- To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is not important, select **Calc > Calculator**. In the **Expression:** box, type **COMBINATIONS(value of  $n$ , value of  $x$ )**. Enter the column number where you want the result to appear in the **Store result in variable:** box. Click **OK**. (See Screen 4.6.)
- To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is important, select **Calc > Calculator**. In the **Expression:** box, type **PERMUTATIONS(value of  $n$ , value of  $x$ )**. Enter the column number where you want the result to appear in the **Store result in variable:** box. Click **OK**.

## Excel

A	B
1 =RAND()*(100-1)+1	

Screen 4.7

- To generate a random number (not necessarily an integer) uniformly distributed between  $m$  and  $n$ , enter the formula  $=\text{rand}()*(n-m)+m$ . If you need more than one random number, copy and paste the formula into as many cells as you need. The numbers will be recalculated every time any cell in the spreadsheet is calculated or recalculated (see Screen 4.7).
- To generate a random integer uniformly distributed between  $m$  and  $n$ , enter the formula  $=\text{floor}(\text{rand}()*(n-m+1)+m,1)$ . If you need more than one random number, copy and paste the formula into as many cells as you need. The numbers will be recalculated every time any cell in the spreadsheet is calculated or recalculated.
- To enter a random number (either type) that stays fixed after it is calculated, select the cell containing the formula, select the formula bar, and press **F9**. (Note that this procedure works only one cell at a time.) If you have a bunch of random numbers that you wish to keep fixed after being calculated, highlight all of the numbers, select **Edit >Copy**, go to an empty column, then select **Edit >Paste Special** and check the **Values** box.
- To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is not important, type  $=\text{COMBIN}(n, x)$ . (See Screens 4.8 and 4.9.)

SUM				
	A	B	C	D
1	=COMBIN(10,3)			
2	COMBIN(number, number_chosen)			

Screen 4.8

A1				
	A	B	C	D
1	120			

Screen 4.9

- To find the number of ways of choosing  $x$  objects out of  $n$  in which the order of selection is important, type  $=\text{PERMUT}(n, x)$ .

## TECHNOLOGY ASSIGNMENTS

**TA4.1** You want to simulate the tossing of a coin. Assign a value of 0 (zero) to Head and a value of 1 to Tail.

- Simulate 50 tosses of the coin by generating 50 random (integer) numbers between 0 and 1. Then calculate the mean of these 50 numbers. This mean gives you the proportion of 50 tosses that resulted in tails. Using this proportion, calculate the number of heads and tails you obtained in 50 simulated tosses. Prepare the frequency tables for expected (theoretical) frequencies and for actual frequencies you obtained.
- Repeat part a by simulating 600 tosses.
- Repeat part a by simulating 4000 tosses.

Comment on the percentage of Tails obtained as the number of tosses is increased.

**TA4.2** You want to simulate the rolling of a die. Assign the values 1 through 6 to the outcomes from 1-spot through 6-spots on the die, respectively.

- a. Simulate 200 rolls of the die by generating 200 random (integer) numbers between 1 and 6. Then make a histogram for these 200 numbers. Prepare the frequency tables for expected (theoretical) frequencies and for actual frequencies you obtained.
- b. Repeat part a by simulating 1000 rolls of the die.
- c. Repeat part a by simulating 6000 rolls of the die.

Comment on the histograms obtained in parts a through c.

**TA4.3** Random number generators can be used to simulate the behavior of many different types of events, including those that have an infinite set of possibilities.

- a. Generate a set of 200 random numbers on the interval 0 to 1 and save them to a column or list in the technology you are using.
- b. Generate a second set of 200 random numbers, but on the interval 12.3 to 13.3 and save them to a different column or list in the technology you are using.
- c. Create histograms of the simulated data in each of the two columns for parts a and b. Compare the shapes of the histograms.

**TA4.4** The data set *Simulated* (that is on the Web site of this text) contains four data sets named data1, data2, data3, and data4. Each of these four data sets consists of 1000 simulated values.

- a. Create a histogram and calculate the mean, median, and standard deviation for each of these four data sets.
- b. For data1, calculate the endpoints of the interval  $\mu \pm 1\sigma$ , that is, the interval from  $\mu - 1\sigma$  to  $\mu + 1\sigma$ . Calculate the probability that a randomly selected value from data1 falls in this interval. (Note: To find this probability, you can sort data1 and then count the number of values that fall within this interval.)
- c. Now repeat part b for data1 to find intervals  $\mu \pm 2\sigma$  and  $\mu \pm 3\sigma$ , respectively. Calculate the probability mentioned in part b for each of these intervals.
- d. Now repeat parts b and c for each of the other three data sets—data2, data3, and data4.
- e. How do the three probabilities (converted to percentages) for each data set compare with the probabilities (percentages) predicted by Chebyshev's Theorem and the Empirical Rule of Section 3.4? Do any of the four data sets appear to fit the percentages given by the Empirical Rule? If so, which data sets? Use the relevant histogram(s) that you created in part a to explain why this makes sense.
- f. All four of the data sets were simulated from four different populations that have the same mean. Based on the summary statistics, what does the value of the common mean of the four data sets appear to be?



## Discrete Random Variables and Their Probability Distributions

Now that you know a little about probability, do you feel lucky enough to play the lottery? If you have \$20 to spend on lunch today, are you willing to spend it all on four \$5 lottery tickets to increase your chance of winning? Do you think you will profit, on average, if you continue buying lottery tickets over time? Can lottery players beat the state, on average? Not a chance. (See Case Study 5–1 for answers.)

Chapter 4 discussed the concepts and rules of probability. This chapter extends the concept of probability to explain probability distributions. As we saw in Chapter 4, any given statistical experiment has more than one outcome. It is impossible to predict which of the many possible outcomes will occur if an experiment is performed. Consequently, decisions are made under uncertain conditions. For example, a lottery player does not know in advance whether or not he is going to win that lottery. If he knows that he is not going to win, he will definitely not play. It is the uncertainty about winning (some positive probability of winning) that makes him play. This chapter shows that if the outcomes and their probabilities for a statistical experiment are known, we can find out what will happen, on average, if that experiment is performed many times. For the lottery example, we can find out what a lottery player can expect to win (or lose), on average, if he continues playing this lottery again and again.

In this chapter, random variables and types of random variables are explained. Then, the concept of a probability distribution and its mean and standard deviation for a discrete random variable are discussed. Finally, three special probability distributions for a discrete random variable—the binomial probability distribution, the hypergeometric probability distribution, and the Poisson probability distribution—are developed.

### 5.1 Random Variables

### 5.2 Probability Distribution of a Discrete Random Variable

### 5.3 Mean and Standard Deviation of a Discrete Random Variable

#### Case Study 5–1 \$1,000 Downpour

### 5.4 The Binomial Probability Distribution

### 5.5 The Hypergeometric Probability Distribution

### 5.6 The Poisson Probability Distribution

#### Case Study 5–2 Global Birth and Death Rates

## 5.1 Random Variables

Suppose Table 5.1 gives the frequency and relative frequency distributions of the number of vehicles owned by all 2000 families living in a small town.

**Table 5.1** Frequency and Relative Frequency Distributions of the Number of Vehicles Owned by Families

Number of Vehicles Owned	Frequency	Relative Frequency
0	30	$30/2000 = .015$
1	470	$470/2000 = .235$
2	850	$850/2000 = .425$
3	490	$490/2000 = .245$
4	160	$160/2000 = .080$
	$N = 2000$	Sum = 1.000

Suppose one family is randomly selected from this population. The process of randomly selecting a family is called a *random* or *chance experiment*. Let  $x$  denote the number of vehicles owned by the selected family. Then  $x$  can assume any of the five possible values (0, 1, 2, 3, and 4) listed in the first column of Table 5.1. The value assumed by  $x$  depends on which family is selected. Thus, this value depends on the outcome of a random experiment. Consequently,  $x$  is called a **random variable** or a **chance variable**. In general, a random variable is denoted by  $x$  or  $y$ .

### Definition

**Random Variable** A *random variable* is a variable whose value is determined by the outcome of a random experiment.

As will be explained next, a random variable can be discrete or continuous.

### 5.1.1 Discrete Random Variable

A **discrete random variable** assumes values that can be counted. In other words, the consecutive values of a discrete random variable are separated by a certain gap.

### Definition

**Discrete Random Variable** A *random variable* that assumes countable values is called a *discrete random variable*.

In Table 5.1, the *number of vehicles owned by a family* is an example of a discrete random variable because the values of the random variable  $x$  are countable: 0, 1, 2, 3, and 4. Here are some other examples of discrete random variables:

1. The number of cars sold at a dealership during a given month
2. The number of houses in a certain block
3. The number of fish caught on a fishing trip
4. The number of complaints received at the office of an airline on a given day

5. The number of customers who visit a bank during any given hour
6. The number of heads obtained in three tosses of a coin

### 5.1.2 Continuous Random Variable

A random variable whose values are not countable is called a **continuous random variable**. A continuous random variable can assume any value over an interval or intervals.

#### Definition

**Continuous Random Variable** A random variable that can assume any value contained in one or more intervals is called a *continuous random variable*.

Because the number of values contained in any interval is infinite, the possible number of values that a continuous random variable can assume is also infinite. Moreover, we cannot count these values. Consider the life of a battery. We can measure it as precisely as we want. For instance, the life of this battery may be 40 hours, or 40.25 hours, or 40.247 hours. Assume that the maximum life of a battery is 200 hours. Let  $x$  denote the life of a randomly selected battery of this kind. Then,  $x$  can assume any value in the interval 0 to 200. Consequently,  $x$  is a continuous random variable. As shown in the diagram below, every point on the line representing the interval 0 to 200 gives a possible value of  $x$ .



Every point on this line represents a possible value of  $x$  that denotes the life of a battery. There is an infinite number of points on this line. The values represented by points on this line are uncountable.

The following are some examples of continuous random variables:

1. The length of a room
2. The time taken to commute from home to work
3. The amount of milk in a gallon (note that we do not expect “a gallon” to contain exactly one gallon of milk but either slightly more or slightly less than one gallon).
4. The weight of a letter
5. The price of a house

Note that amount of money is often treated as a continuous random variable, specifically when there are a large number of unique values.

This chapter is limited to a discussion of discrete random variables and their probability distributions. Continuous random variables will be discussed in Chapter 6.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 5.1** Explain the meaning of a random variable, a discrete random variable, and a continuous random variable. Give one example each of a discrete random variable and a continuous random variable.
- 5.2** Classify each of the following random variables as discrete or continuous.
- a. The time left on a parking meter
  - b. The number of bats broken by a major league baseball team in a season
  - c. The number of cars in a parking lot at a given time
  - d. The price of a car
  - e. The number of cars crossing a bridge on a given day
  - f. The time spent by a physician examining a patient

- 5.3** Indicate which of the following random variables are discrete and which are continuous.
- The amount of rainfall in a city during a specific month
  - The number of students on a waitlist to register for a class
  - The price of one ounce of gold at the close of trading on a given day
  - The number of vacation trips taken by a family during a given year
  - The amount of gasoline in your car's gas tank at a given time
  - The distance you walked to class this morning

## ■ APPLICATIONS

**5.4** A household can watch National news on any of the three networks—ABC, CBS, or NBC. On a certain day, five households randomly and independently decide which channel to watch. Let  $x$  be the number of households among these five that decide to watch news on ABC. Is  $x$  a discrete or a continuous random variable? Explain. What are the possible values that  $x$  can assume?

**5.5** One of the four gas stations located at an intersection of two major roads is a Texaco station. Suppose the next six cars that stop at any of these four gas stations make their selections randomly and independently. Let  $x$  be the number of cars in these six that stop at the Texaco station. Is  $x$  a discrete or a continuous random variable? Explain. What are the possible values that  $x$  can assume?

## 5.2 Probability Distribution of a Discrete Random Variable

Let  $x$  be a discrete random variable. The **probability distribution** of  $x$  describes how the probabilities are distributed over all the possible values of  $x$ .

### Definition

**Probability Distribution of a Discrete Random Variable** The *probability distribution of a discrete random variable* lists all the possible values that the random variable can assume and their corresponding probabilities.

Example 5–1 illustrates the concept of the probability distribution of a discrete random variable.

## ■ EXAMPLE 5–1

Writing the probability distribution of a discrete random variable.

Recall the frequency and relative frequency distributions of the number of vehicles owned by families given in Table 5.1. That table is reproduced as Table 5.2. Let  $x$  be the number of vehicles owned by a randomly selected family. Write the probability distribution of  $x$ .

**Table 5.2 Frequency and Relative Frequency Distributions of the Number of Vehicles Owned by Families**

Number of Vehicles Owned	Frequency	Relative Frequency
0	30	.015
1	470	.235
2	850	.425
3	490	.245
4	160	.080
	$N = 2000$	Sum = 1.000

**Solution** In Chapter 4, we learned that the relative frequencies obtained from an experiment or a sample can be used as approximate probabilities. However, when the relative frequencies represent the population, as in Table 5.2, they give the actual (theoretical) probabilities of outcomes. Using the relative frequencies of Table 5.2, we can write the *probability distribution* of the discrete random variable  $x$  in Table 5.3. Note that the values of  $x$  listed in Table 5.3 are pairwise mutually exclusive events.

**Table 5.3** Probability Distribution of the Number of Vehicles Owned by Families

Number of Vehicles Owned $x$	Probability $P(x)$
0	.015
1	.235
2	.425
3	.245
4	.080
$\Sigma P(x) = 1.000$	

The probability distribution of a discrete random variable possesses the following *two characteristics*:

1. The probability assigned to each value of a random variable  $x$  lies in the range 0 to 1; that is,  $0 \leq P(x) \leq 1$  for each  $x$ .
2. The sum of the probabilities assigned to all possible values of  $x$  is equal to 1.0; that is,  $\Sigma P(x) = 1$ . (Remember, if the probabilities are rounded, the sum may not be exactly 1.0.)

**Two Characteristics of a Probability Distribution** The probability distribution of a discrete random variable possesses the following two characteristics.

1.  $0 \leq P(x) \leq 1$  for each value of  $x$
2.  $\Sigma P(x) = 1$

These two characteristics are also called the *two conditions* that a probability distribution must satisfy. Notice that in Table 5.3 each probability listed in the column labeled  $P(x)$  is between 0 and 1. Also,  $\Sigma P(x) = 1.0$ . Because both conditions are satisfied, Table 5.3 represents the probability distribution of  $x$ .

From Table 5.3, we can read the probability for any value of  $x$ . For example, the probability that a randomly selected family from this town owns two vehicles is .425. This probability is written as

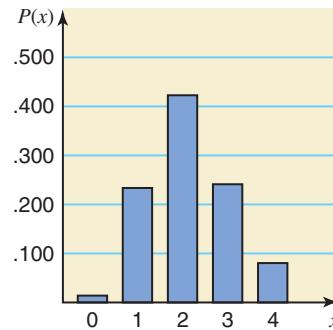
$$P(x = 2) = .425 \quad \text{or} \quad P(2) = .425$$

The probability that the selected family owns more than two vehicles is given by the sum of the probabilities of owning three and four vehicles. This probability is  $.245 + .080 = .325$ , which can be written as

$$P(x > 2) = P(x = 3) + P(x = 4) = P(3) + P(4) = .245 + .080 = .325$$

The probability distribution of a discrete random variable can be presented in the form of a *mathematical formula*, a *table*, or a *graph*. Table 5.3 presented the probability distribution in tabular form. Figure 5.1 shows the graphical presentation of the probability distribution of Table 5.3. In this figure, each value of  $x$  is marked on the horizontal axis. The probability for each

**Figure 5.1** Graphical presentation of the probability distribution of Table 5.3.



value of  $x$  is exhibited by the height of the corresponding bar. Such a graph is called a **bar graph**. This section does not discuss the presentation of a probability distribution using a mathematical formula.

### ■ EXAMPLE 5–2

Verifying the conditions of a probability distribution.

Each of the following tables lists certain values of  $x$  and their probabilities. Determine whether or not each table represents a valid probability distribution.

(a) $x$	$P(x)$	(b) $x$	$P(x)$	(c) $x$	$P(x)$
0	.08	2	.25	7	.70
1	.11	3	.34	8	.50
2	.39	4	.28	9	-.20
3	.27	5	.13		

### Solution

- (a) Because each probability listed in this table is in the range 0 to 1, it satisfies the first condition of a probability distribution. However, the sum of all probabilities is not equal to 1.0 because  $\Sigma P(x) = .08 + .11 + .39 + .27 = .85$ . Therefore, the second condition is not satisfied. Consequently, this table does not represent a valid probability distribution.
- (b) Each probability listed in this table is in the range 0 to 1. Also,  $\Sigma P(x) = .25 + .34 + .28 + .13 = 1.0$ . Consequently, this table represents a valid probability distribution.
- (c) Although the sum of all probabilities listed in this table is equal to 1.0, one of the probabilities is negative. This violates the first condition of a probability distribution. Therefore, this table does not represent a valid probability distribution. ■

### ■ EXAMPLE 5–3

The following table lists the probability distribution of the number of breakdowns per week for a machine based on past data.

Breakdowns per week	0	1	2	3
Probability	.15	.20	.35	.30

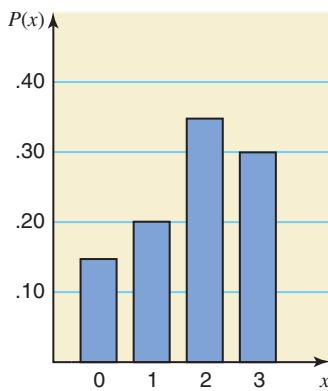
- (a) Present this probability distribution graphically.
- (b) Find the probability that the number of breakdowns for this machine during a given week is
  - i. exactly 2
  - ii. 0 to 2
  - iii. more than 1
  - iv. at most 1

**Solution** Let  $x$  denote the number of breakdowns for this machine during a given week. Table 5.4 lists the probability distribution of  $x$ .

**Table 5.4** Probability Distribution of the Number of Breakdowns

$x$	$P(x)$
0	.15
1	.20
2	.35
3	.30
$\Sigma P(x) = 1.00$	

- (a) Figure 5.2 shows the bar graph of the probability distribution of Table 5.4.



**Figure 5.2** Graphical presentation of the probability distribution of Table 5.4.

*Graphing a probability distribution.*

- (b) Using Table 5.4, we can calculate the required probabilities as follows.

- i. The probability of exactly two breakdowns is

$$P(\text{exactly 2 breakdowns}) = P(x = 2) = .35$$

- ii. The probability of 0 to 2 breakdowns is given by the sum of the probabilities of 0, 1, and 2 breakdowns:

$$\begin{aligned} P(0 \text{ to } 2 \text{ breakdowns}) &= P(0 \leq x \leq 2) \\ &= P(x = 0) + P(x = 1) + P(x = 2) \\ &= .15 + .20 + .35 = .70 \end{aligned}$$

- iii. The probability of more than 1 breakdown is obtained by adding the probabilities of 2 and 3 breakdowns:

$$\begin{aligned} P(\text{more than 1 breakdown}) &= P(x > 1) \\ &= P(x = 2) + P(x = 3) \\ &= .35 + .30 = .65 \end{aligned}$$

*Finding the probabilities of events for a discrete random variable.*

- iv. The probability of at most 1 breakdown is given by the sum of the probabilities of 0 and 1 breakdown:

$$\begin{aligned} P(\text{at most 1 breakdown}) &= P(x \leq 1) \\ &= P(x = 0) + P(x = 1) \\ &= .15 + .20 = .35 \end{aligned}$$



*Constructing a probability distribution.*

### ■ EXAMPLE 5–4

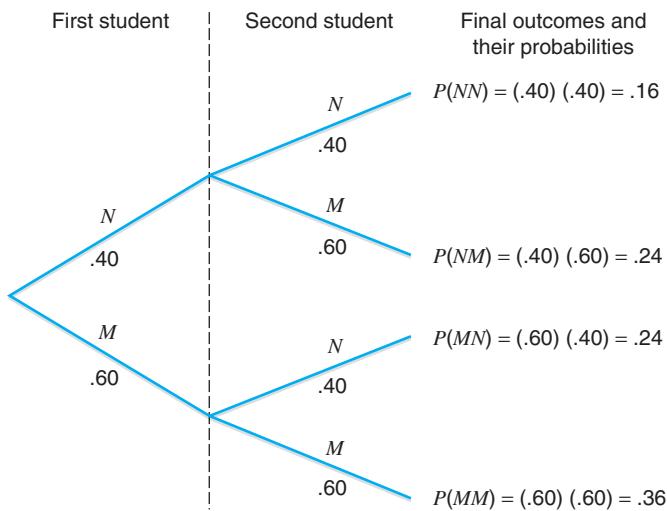
According to a survey, 60% of all students at a large university suffer from math anxiety. Two students are randomly selected from this university. Let  $x$  denote the number of students in this sample who suffer from math anxiety. Develop the probability distribution of  $x$ .

**Solution** Let us define the following two events:

$N$  = the student selected does not suffer from math anxiety

$M$  = the student selected suffers from math anxiety

As we can observe from the tree diagram of Figure 5.3, there are four possible outcomes for this experiment:  $NN$  (neither of the students suffers from math anxiety),  $NM$  (the first student does not suffer from math anxiety and the second does),  $MN$  (the first student suffers from math anxiety and the second does not), and  $MM$  (both students suffer from math anxiety). The probabilities of these four outcomes are listed in the tree diagram. Because 60% of the students suffer from math anxiety and 40% do not, the probability is .60 that any student selected suffers from math anxiety and .40 that he or she does not.



**Figure 5.3** Tree diagram.

In a sample of two students, the number who suffer from math anxiety can be 0 (given by  $NN$ ), 1 (given by  $NM$  or  $MN$ ), or 2 (given by  $MM$ ). Thus,  $x$  can assume any of three possible values: 0, 1, or 2. The probabilities of these three outcomes are calculated as follows:

$$P(x = 0) = P(NN) = .16$$

$$P(x = 1) = P(NM \text{ or } MN) = P(NM) + P(MN) = .24 + .24 = .48$$

$$P(x = 2) = P(MM) = .36$$

Using these probabilities, we can write the probability distribution of  $x$  as in Table 5.5.

**Table 5.5** Probability Distribution of the Number of Students with Math Anxiety

$x$	$P(x)$
0	.16
1	.48
2	.36
	$\Sigma P(x) = 1.00$

## EXERCISES

### CONCEPTS AND PROCEDURES

**5.6** Explain the meaning of the probability distribution of a discrete random variable. Give one example of such a probability distribution. What are the three ways to present the probability distribution of a discrete random variable?

**5.7** Briefly explain the two characteristics (conditions) of the probability distribution of a discrete random variable.

**5.8** Each of the following tables lists certain values of  $x$  and their probabilities. Verify whether or not each represents a valid probability distribution and explain why.

a. $x$	$P(x)$
0	.10
1	.05
2	.45
3	.40

b. $x$	$P(x)$
2	.35
3	.28
4	.20
5	.14

c. $x$	$P(x)$
7	-.25
8	.85
9	.40

**5.9** Each of the following tables lists certain values of  $x$  and their probabilities. Determine whether or not each one satisfies the two conditions required for a valid probability distribution and explain why.

a. $x$	$P(x)$
5	-.36
6	.48
7	.62
8	.26

b. $x$	$P(x)$
1	.27
2	.24
3	.49

c. $x$	$P(x)$
0	.15
1	.08
2	.20
3	.50

**5.10** The following table gives the probability distribution of a discrete random variable  $x$ .

$x$	0	1	2	3	4	5	6
$P(x)$	.11	.19	.28	.15	.12	.09	.06

Find the following probabilities.

- a.  $P(x = 3)$
- b.  $P(x \leq 2)$
- c.  $P(x \geq 4)$
- d.  $P(1 \leq x \leq 4)$
- e. Probability that  $x$  assumes a value less than 4
- f. Probability that  $x$  assumes a value greater than 2
- g. Probability that  $x$  assumes a value in the interval 2 to 5

**5.11** The following table gives the probability distribution of a discrete random variable  $x$ .

$x$	0	1	2	3	4	5
$P(x)$	.03	.17	.22	.31	.15	.12

Find the following probabilities.

- a.  $P(x = 1)$
- b.  $P(x \leq 1)$
- c.  $P(x \geq 3)$
- d.  $P(0 \leq x \leq 2)$
- e. Probability that  $x$  assumes a value less than 3
- f. Probability that  $x$  assumes a value greater than 3
- g. Probability that  $x$  assumes a value in the interval 2 to 4

### APPLICATIONS

**5.12** A review of emergency room records at rural Millard Fellmore Memorial Hospital was performed to determine the probability distribution of the number of patients entering the emergency room during a 1-hour period. The following table lists this probability distribution.

Patients per hour	0	1	2	3	4	5	6
Probability	.2725	.3543	.2303	.0998	.0324	.0084	.0023

- a. Graph the probability distribution.  
 b. Determine the probability that the number of patients entering the emergency room during a randomly selected 1-hour period is  
   i. 2 or more     ii. exactly 5     iii. fewer than 3     iv. at most 1

**5.13** Nathan Cheboygan, a singing gambler from northern Michigan, is famous for his loaded dice. The following table shows the probability distribution for the sum, denoted by  $x$ , of the faces on a pair of Nathan's dice.

$x$	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	.065	.065	.080	.095	.110	.170	.110	.095	.080	.065	.065

- a. Draw a bar graph for this probability distribution.  
 b. Determine the probability that the sum of the faces on a single roll of Nathan's dice is  
   i. an even number     ii. 7 or 11     iii. 4 to 6     iv. no less than 9

**5.14** The H2 Hummer limousine has eight tires on it. A fleet of 1300 H2 limos was fit with a batch of tires that mistakenly passed quality testing. The following table lists the frequency distribution of the number of defective tires on the 1300 H2 limos.

Number of defective tires	0	1	2	3	4	5	6	7	8
Number of H2 limos	59	224	369	347	204	76	18	2	1

- a. Construct a probability distribution table for the numbers of defective tires on these limos. Draw a bar graph for this probability distribution.  
 b. Are the probabilities listed in the table exact or approximate probabilities of the various outcomes? Explain.  
 c. Let  $x$  denote the number of defective tires on a randomly selected H2 limo. Find the following probabilities.  
   i.  $P(x = 0)$      ii.  $P(x < 4)$      iii.  $P(3 \leq x < 7)$      iv.  $P(x \geq 2)$

**5.15** One of the most profitable items at A1's Auto Security Shop is the remote starting system. Let  $x$  be the number of such systems installed on a given day at this shop. The following table lists the frequency distribution of  $x$  for the past 80 days.

$x$	1	2	3	4	5
$f$	8	20	24	16	12

- a. Construct a probability distribution table for the number of remote starting systems installed on a given day. Draw a graph of the probability distribution.  
 b. Are the probabilities listed in the table exact or approximate probabilities of various outcomes? Explain.  
 c. Find the following probabilities.  
   i.  $P(x = 3)$      ii.  $P(x \geq 3)$      iii.  $P(2 \leq x \leq 4)$      iv.  $P(x < 4)$

**5.16** Five percent of all cars manufactured at a large auto company are lemons. Suppose two cars are selected at random from the production line of this company. Let  $x$  denote the number of lemons in this sample. Write the probability distribution of  $x$ . Draw a tree diagram for this problem.

**5.17** According to the most recent data from the Insurance Research Council, 16.1% of motorists in the United States were uninsured in 2010 ([virginia.beach.injuryboard.com](http://virginia.beach.injuryboard.com)). Suppose that currently 16.1% of motorists in the United States are uninsured. Suppose that two motorists are selected at random. Let  $x$  denote the number of motorists in this sample of two who are uninsured. Construct the probability distribution table of  $x$ . Draw a tree diagram for this problem.

**5.18** According to a survey, 30% of adults are against using animals for research. Assume that this result holds true for the current population of all adults. Let  $x$  be the number of adults who are against using animals for research in a random sample of two adults. Obtain the probability distribution of  $x$ . Draw a tree diagram for this problem.

**5.19** According to the Alzheimer's Association ([www.alz.org/documents\\_custom/2011\\_Facts\\_Figures\\_Fact\\_Sheet.pdf](http://www.alz.org/documents_custom/2011_Facts_Figures_Fact_Sheet.pdf)), 3.7% of Americans with Alzheimer's disease were younger than the age of 65 years in 2011 (which means that they were diagnosed with early onset of Alzheimer's). Suppose that currently 3.7% of Americans with Alzheimer's disease are younger than the age of 65 years. Suppose that two Americans with Alzheimer's disease are selected at random. Let  $x$  denote the number in this sample of two Americans with Alzheimer's disease who are younger than the age of 65 years. Construct the probability distribution table of  $x$ . Draw a tree diagram for this problem.

**\*5.20** In a group of 12 persons, 3 are left-handed. Suppose that 2 persons are randomly selected from this group. Let  $x$  denote the number of left-handed persons in this sample. Write the probability distribution of  $x$ . You may draw a tree diagram and use it to write the probability distribution. (*Hint:* Note that the selections are made without replacement from a small population. Hence, the probabilities of outcomes do not remain constant for each selection.)

**\*5.21** In a group of 20 athletes, 6 have used performance-enhancing drugs that are illegal. Suppose that 2 athletes are randomly selected from this group. Let  $x$  denote the number of athletes in this sample who have used such illegal drugs. Write the probability distribution of  $x$ . You may draw a tree diagram and use that to write the probability distribution. (*Hint:* Note that the selections are made without replacement from a small population. Hence, the probabilities of outcomes do not remain constant for each selection.)

## 5.3 Mean and Standard Deviation of a Discrete Random Variable

In this section, we will learn how to calculate the mean and standard deviation of a discrete random variable and how to interpret them.

### 5.3.1 Mean of a Discrete Random Variable

The **mean of a discrete random variable**, denoted by  $\mu$ , is actually the mean of its probability distribution. The mean of a discrete random variable  $x$  is also called its *expected value* and is denoted by  $E(x)$ . The mean (or expected value) of a discrete random variable is the value that we expect to observe per repetition, on average, if we perform an experiment a large number of times. For example, we may expect a car salesperson to sell, on average, 2.4 cars per week. This does not mean that every week this salesperson will sell exactly 2.4 cars. (Obviously one cannot sell exactly 2.4 cars.) This simply means that if we observe for many weeks, this salesperson will sell a different number of cars during different weeks; however, the average for all these weeks will be 2.4 cars per week.

To calculate the mean of a discrete random variable  $x$ , we multiply each value of  $x$  by the corresponding probability and sum the resulting products. This sum gives the mean (or expected value) of the discrete random variable  $x$ .

**Mean of a Discrete Random Variable** The *mean of a discrete random variable  $x$*  is the value that is expected to occur per repetition, on average, if an experiment is repeated a large number of times. It is denoted by  $\mu$  and calculated as

$$\mu = \sum xP(x)$$

The mean of a discrete random variable  $x$  is also called its expected value and is denoted by  $E(x)$ ; that is,

$$E(x) = \sum xP(x)$$

Example 5–5 illustrates the calculation of the mean of a discrete random variable.

### ■ EXAMPLE 5–5

Recall Example 5–3 of Section 5.2. The probability distribution Table 5.4 from that example is reproduced here. In this table,  $x$  represents the number of breakdowns for a machine during a given week, and  $P(x)$  is the probability of the corresponding value of  $x$ .

*Calculating and interpreting the mean of a discrete random variable.*

$x$	$P(x)$
0	.15
1	.20
2	.35
3	.30

Find the mean number of breakdowns per week for this machine.

**Solution** To find the mean number of breakdowns per week for this machine, we multiply each value of  $x$  by its probability and add these products. This sum gives the mean of the probability distribution of  $x$ . The products  $xP(x)$  are listed in the third column of Table 5.6. The sum of these products gives  $\sum xP(x)$ , which is the mean of  $x$ .

**Table 5.6 Calculating the Mean for the Probability Distribution of Breakdowns**

$x$	$P(x)$	$xP(x)$
0	.15	$0(.15) = .00$
1	.20	$1(.20) = .20$
2	.35	$2(.35) = .70$
3	.30	$3(.30) = .90$
		$\sum xP(x) = 1.80$

The mean is

$$\mu = \sum xP(x) = 1.80$$

Thus, on average, this machine is expected to break down 1.80 times per week over a period of time. In other words, if this machine is used for many weeks, then for certain weeks we will observe no breakdowns; for some other weeks we will observe one breakdown per week; and for still other weeks we will observe two or three breakdowns per week. The mean number of breakdowns is expected to be 1.80 per week for the entire period.

Note that  $\mu = 1.80$  is also the expected value of  $x$ . It can also be written as

$$E(x) = 1.80$$

Case Study 5–1 illustrates the calculation of the mean amount that an instant lottery player is expected to win.

### 5.3.2 Standard Deviation of a Discrete Random Variable

The **standard deviation of a discrete random variable**, denoted by  $\sigma$ , measures the spread of its probability distribution. A higher value for the standard deviation of a discrete random variable indicates that  $x$  can assume values over a larger range about the mean. In contrast, a smaller value for the standard deviation indicates that most of the values that  $x$  can assume are clustered closely about the mean. The basic formula to compute the standard deviation of a discrete random variable is

$$\sigma = \sqrt{\sum [(x - \mu)^2 \cdot P(x)]}$$

However, it is more convenient to use the following shortcut formula to compute the standard deviation of a discrete random variable.

**Standard Deviation of a Discrete Random Variable** The *standard deviation of a discrete random variable*  $x$  measures the spread of its probability distribution and is computed as

$$\sigma = \sqrt{\sum x^2 P(x) - \mu^2}$$



Ticket with covered play symbols.



Ticket with uncovered play symbols.

The state of New Jersey has in circulation (as of 2011) an instant lottery game called *\$1,000 Downpour*. The cost of each ticket for this lottery game is \$5. A player can instantly win \$75,000, \$1000, \$100, \$50, \$20, \$10, or \$5. Each ticket has 19 spots covered by latex coating, and the top four spots contain numbers that, if matched by the player's numbers, win money. The remaining 15 spots belong to the player. A player wins if any of the numbers in the player's 15 spots matches any of the four winning numbers. The potential prize amounts are shown beneath the player's 15 numbers.

Based on the information available on this lottery game, the following table lists the number of tickets with different prizes in a total of 3,900,000 tickets printed. As is obvious from this table, of a total of 3,900,000 tickets, 2,853,533 are nonwinning tickets (the ones with a prize of \$0 in this table). Of the remaining 1,046,467 tickets with prizes, 621,075 tickets have a prize of \$5 each, 327,600 tickets contain a prize of \$10 each, and so forth.

Prize (dollars)	Number of Tickets
0	2,853,533
5	621,075
10	327,600
20	58,500
50	31,200
100	5525
1000	2561
75,000	6
<b>Total = 3,900,000</b>	

The net gain to a player for each of the winning tickets is equal to the amount of the prize minus \$5, which is the cost of the ticket. Thus, the net gain for each of the nonwinning tickets is  $-\$5$ , which is the cost of the ticket. Let

$x$  = the net amount a player wins by playing this lottery game

The following table shows the probability distribution of  $x$  and all the calculations required to compute the mean of  $x$  for this probability distribution. The probability of an outcome (net winnings) is calculated by dividing the number of tickets with that outcome by the total number of tickets.

$x$ (dollars)	$P(x)$	$xP(x)$
-5	$2,853,533 / 3,900,000 = .73167513$	$-.365837564$
0	$621,075 / 3,900,000 = .15925000$	$.00000000$
5	$327,600 / 3,900,000 = .08400000$	$.42000000$
10	$58,500 / 3,900,000 = .01500000$	$.22500000$
45	$31,200 / 3,900,000 = .00800000$	$.36000000$
95	$5525 / 3,900,000 = .00141667$	$.13458333$
995	$2561 / 3,900,000 = .00065667$	$.65338333$
74,995	$6 / 3,900,000 = .00000154$	$.11537692$
$\Sigma xP(x) = -1.75003206$		

Hence, the mean or expected value of  $x$  is

$$\mu = \sum xP(x) = -\$1.75$$

This mean gives the expected value of the random variable  $x$ , that is,

$$E(x) = \sum xP(x) = -\$1.75$$

Thus, the mean of net winnings for this lottery is  $-\$1.75$ . In other words, all players taken together lose an average of  $\$1.75$  per ticket. This means that of every  $\$5$  (the price of a ticket),  $\$3.25$  is returned to players in the form of prizes and  $\$1.75$  goes to the state of New Jersey, which covers the costs of operating the lottery, the commission paid to agents, and profit to the state. Because,  $\$1.75$  is approximately 35% of  $\$5$ , we can also say that 35% of the total money spent by players on this lottery goes to the state and  $100 - 35 = 65\%$  is returned to players in the form of prizes.

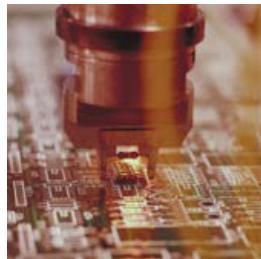
**Source:** Tickets are reproduced with the permission of The New Jersey Lottery.

Note that the variance  $\sigma^2$  of a discrete random variable is obtained by squaring its standard deviation.

Example 5–6 illustrates how to use the shortcut formula to compute the standard deviation of a discrete random variable.

## ■ EXAMPLE 5–6

*Calculating the standard deviation of a discrete random variable.*



Corbis Digital Stock

Baier's Electronics manufactures computer parts that are supplied to many computer companies. Despite the fact that two quality control inspectors at Baier's Electronics check every part for defects before it is shipped to another company, a few defective parts do pass through these inspections undetected. Let  $x$  denote the number of defective computer parts in a shipment of 400. The following table gives the probability distribution of  $x$ .

$x$	0	1	2	3	4	5
$P(x)$	.02	.20	.30	.30	.10	.08

Compute the standard deviation of  $x$ .

**Solution** Table 5.7 shows all the calculations required for the computation of the standard deviation of  $x$ .

**Table 5.7** Computations to Find the Standard Deviation

$x$	$P(x)$	$xP(x)$	$x^2$	$x^2P(x)$
0	.02	.00	0	.00
1	.20	.20	1	.20
2	.30	.60	4	1.20
3	.30	.90	9	2.70
4	.10	.40	16	1.60
5	.08	.40	25	2.00
$\Sigma xP(x) = 2.50$			$\Sigma x^2P(x) = 7.70$	

We perform the following steps to compute the standard deviation of  $x$ .

**Step 1.** Compute the mean of the discrete random variable.

The sum of the products  $xP(x)$ , recorded in the third column of Table 5.7, gives the mean of  $x$ .

$$\mu = \sum xP(x) = 2.50 \text{ defective computer parts in 400}$$

**Step 2.** Compute the value of  $\Sigma x^2 P(x)$ .

First we square each value of  $x$  and record it in the fourth column of Table 5.7. Then we multiply these values of  $x^2$  by the corresponding values of  $P(x)$ . The resulting values of  $x^2 P(x)$  are recorded in the fifth column of Table 5.7. The sum of this column is

$$\Sigma x^2 P(x) = 7.70$$

**Step 3.** Substitute the values of  $\mu$  and  $\Sigma x^2 P(x)$  in the formula for the standard deviation of  $x$  and simplify.

By performing this step, we obtain

$$\begin{aligned}\sigma &= \sqrt{\Sigma x^2 P(x) - \mu^2} = \sqrt{7.70 - (2.50)^2} = \sqrt{1.45} \\ &= 1.204 \text{ defective computer parts}\end{aligned}$$

Thus, a given shipment of 400 computer parts is expected to contain an average of 2.50 defective parts with a standard deviation of 1.204. ■

Because the standard deviation of a discrete random variable is obtained by taking the positive square root, its value is never negative.

◀ **Remember**

### ■ EXAMPLE 5-7

Lorraine Corporation is planning to market a new makeup product. According to the analysis made by the financial department of the company, it will earn an annual profit of \$4.5 million if this product has high sales, it will earn an annual profit of \$1.2 million if the sales are mediocre, and it will lose \$2.3 million a year if the sales are low. The probabilities of these three scenarios are .32, .51, and .17, respectively.

- (a) Let  $x$  be the profits (in millions of dollars) earned per annum from this product by the company. Write the probability distribution of  $x$ .
- (b) Calculate the mean and standard deviation of  $x$ .

#### Solution

- (a) The table below lists the probability distribution of  $x$ . Note that because  $x$  denotes profits earned by the company, the loss is written as a *negative profit* in the table.

Writing the probability distribution of a discrete random variable.

$x$	$P(x)$
4.5	.32
1.2	.51
-2.3	.17

- (b) Table 5.8 shows all the calculations needed for the computation of the mean and standard deviation of  $x$ .

Calculating the mean and standard deviation of a discrete random variable.

**Table 5.8** Computations to Find the Mean and Standard Deviation

$x$	$P(x)$	$xP(x)$	$x^2$	$x^2 P(x)$
4.5	.32	1.440	20.25	6.4800
1.2	.51	.612	1.44	.7344
-2.3	.17	-.391	5.29	.8993
$\Sigma xP(x) = 1.661$		$\Sigma x^2 P(x) = 8.1137$		

The mean of  $x$  is

$$\mu = \sum xP(x) = \$1.661 \text{ million}$$

The standard deviation of  $x$  is

$$\sigma = \sqrt{\sum x^2 P(x) - \mu^2} = \sqrt{8.1137 - (1.661)^2} = \$2.314 \text{ million}$$

Thus, it is expected that Loraine Corporation will earn an average of \$1.661 million in profits per year from the new product, with a standard deviation of \$2.314 million. ■

### Interpretation of the Standard Deviation

The standard deviation of a discrete random variable can be interpreted or used the same way as the standard deviation of a data set in Section 3.4 of Chapter 3. In that section, we learned that according to Chebyshev's theorem, at least  $[1 - (1/k^2)] \times 100\%$  of the total area under a curve lies within  $k$  standard deviations of the mean, where  $k$  is any number greater than 1. Thus, if  $k = 2$ , then at least 75% of the area under a curve lies between  $\mu - 2\sigma$  and  $\mu + 2\sigma$ . In Example 5–6,

$$\mu = 2.50 \quad \text{and} \quad \sigma = 1.204$$

Hence,

$$\mu - 2\sigma = 2.50 - 2(1.204) = .092$$

$$\mu + 2\sigma = 2.50 + 2(1.204) = 4.908$$

Using Chebyshev's theorem, we can state that at least 75% of the shipments (each containing 400 computer parts) are expected to contain .092 to 4.908 defective computer parts each.

## EXERCISES

### CONCEPTS AND PROCEDURES

**5.22** Briefly explain the concept of the mean and standard deviation of a discrete random variable.

**5.23** Find the mean and standard deviation for each of the following probability distributions.

a. $x$	$P(x)$
0	.16
1	.27
2	.39
3	.18

b. $x$	$P(x)$
6	.40
7	.26
8	.21
9	.13

**5.24** Find the mean and standard deviation for each of the following probability distributions.

a. $x$	$P(x)$
3	.09
4	.21
5	.34
6	.23
7	.13

b. $x$	$P(x)$
0	.43
1	.31
2	.17
3	.09

### APPLICATIONS

**5.25** Let  $x$  be the number of errors that appear on a randomly selected page of a book. The following table lists the probability distribution of  $x$ .

$x$	0	1	2	3	4
$P(x)$	.73	.16	.06	.04	.01

Find the mean and standard deviation of  $x$ .

- 5.26** Let  $x$  be the number of magazines a person reads every week. Based on a sample survey of adults, the following probability distribution table was prepared.

$x$	0	1	2	3	4	5
$P(x)$	.36	.24	.18	.10	.07	.05

Find the mean and standard deviation of  $x$ .

- 5.27** The following table gives the probability distribution of the number of camcorders sold on a given day at an electronics store.

Camcorders sold	0	1	2	3	4	5	6
Probability	.05	.12	.19	.30	.20	.10	.04

Calculate the mean and standard deviation for this probability distribution. Give a brief interpretation of the value of the mean.

- 5.28** The following table, reproduced from Exercise 5.12, lists the probability distribution of the number of patients entering the emergency room during a 1-hour period at Millard Fellmore Memorial Hospital.

Patients per hour	0	1	2	3	4	5	6
Probability	.2725	.3543	.2303	.0998	.0324	.0084	.0023

Calculate the mean and standard deviation for this probability distribution.

- 5.29** Let  $x$  be the number of heads obtained in two tosses of a coin. The following table lists the probability distribution of  $x$ .

$x$	0	1	2
$P(x)$	.25	.50	.25

Calculate the mean and standard deviation of  $x$ . Give a brief interpretation of the value of the mean.

- 5.30** Let  $x$  be the number of potential weapons detected by a metal detector at an airport on a given day. The following table lists the probability distribution of  $x$ .

$x$	0	1	2	3	4	5
$P(x)$	.14	.28	.22	.18	.12	.06

Calculate the mean and standard deviation for this probability distribution and give a brief interpretation of the value of the mean.

- 5.31** Refer to Exercise 5.14. Calculate the mean and standard deviation for the probability distribution you developed for the number of defective tires on all 1300 H2 Hummer limousines. Give a brief interpretation of the values of the mean and standard deviation.

- 5.32** Refer to Exercise 5.15. Find the mean and standard deviation of the probability distribution you developed for the number of remote starting systems installed per day by Al's Auto Security Shop over the past 80 days. Give a brief interpretation of the values of the mean and standard deviation.

- 5.33** Refer to the probability distribution you developed in Exercise 5.16 for the number of lemons in two selected cars. Calculate the mean and standard deviation of  $x$  for that probability distribution.

- 5.34** Refer to the probability distribution you developed in Exercise 5.17 for the number of uninsured motorists in a sample of two motorists. Calculate the mean and standard deviation of  $x$  for that probability distribution.

- 5.35** A contractor has submitted bids on three state jobs: an office building, a theater, and a parking garage. State rules do not allow a contractor to be offered more than one of these jobs. If this contractor is awarded any of these jobs, the profits earned from these contracts are \$10 million from the office building, \$5 million from the theater, and \$2 million from the parking garage. His profit is zero if he gets no contract. The contractor estimates that the probabilities of getting the office building contract, the theater contract, the parking garage contract, or nothing are .15, .30, .45, and .10, respectively. Let  $x$  be the random variable that represents the contractor's profits in millions of dollars. Write

the probability distribution of  $x$ . Find the mean and standard deviation of  $x$ . Give a brief interpretation of the values of the mean and standard deviation.

**5.36** An instant lottery ticket costs \$2. Out of a total of 10,000 tickets printed for this lottery, 1000 tickets contain a prize of \$5 each, 100 tickets have a prize of \$10 each, 5 tickets have a prize of \$1000 each, and 1 ticket has a prize of \$5000. Let  $x$  be the random variable that denotes the net amount a player wins by playing this lottery. Write the probability distribution of  $x$ . Determine the mean and standard deviation of  $x$ . How will you interpret the values of the mean and standard deviation of  $x$ ?

**\*5.37** Refer to the probability distribution you developed in Exercise 5.20 for the number of left-handed persons in a sample of two persons. Calculate the mean and standard deviation of  $x$  for that distribution.

**\*5.38** Refer to the probability distribution you developed in Exercise 5.21 for the number of athletes in a random sample of two who have used illegal performance-enhancing drugs. Calculate the mean and standard deviation of  $x$  for that distribution.

## 5.4 The Binomial Probability Distribution

The **binomial probability distribution** is one of the most widely used discrete probability distributions. It is applied to find the probability that an outcome will occur  $x$  times in  $n$  performances of an experiment. For example, given that the probability is .05 that a DVD player manufactured at a firm is defective, we may be interested in finding the probability that in a random sample of three DVD players manufactured at this firm, exactly one will be defective. As a second example, we may be interested in finding the probability that a baseball player with a batting average of .250 will have no hits in 10 trips to the plate.

To apply the binomial probability distribution, the random variable  $x$  must be a discrete dichotomous random variable. In other words, the variable must be a discrete random variable, and each repetition of the experiment must result in one of two possible outcomes. The binomial distribution is applied to experiments that satisfy the four conditions of a *binomial experiment*. (These conditions are described in Section 5.4.1.) Each repetition of a binomial experiment is called a **trial** or a **Bernoulli trial** (after Jacob Bernoulli). For example, if an experiment is defined as one toss of a coin and this experiment is repeated 10 times, then each repetition (toss) is called a trial. Consequently, there are 10 total trials for this experiment.

### 5.4.1 The Binomial Experiment

An experiment that satisfies the following four conditions is called a **binomial experiment**.

1. There are  $n$  identical trials. In other words, the given experiment is repeated  $n$  times, where  $n$  is a positive integer. All of these repetitions are performed under identical conditions.
2. Each trial has two and only two outcomes. These outcomes are usually called a *success* and a *failure*, respectively. In case there are more than two outcomes for an experiment, we can combine outcomes into two events and then apply binomial probability distribution.
3. The probability of success is denoted by  $p$  and that of failure by  $q$ , and  $p + q = 1$ . The probabilities  $p$  and  $q$  remain constant for each trial.
4. The trials are independent. In other words, the outcome of one trial does not affect the outcome of another trial.

**Conditions of a Binomial Experiment** A binomial experiment must satisfy the following four conditions.

1. There are  $n$  identical trials.
2. Each trial has only two possible outcomes (or events). In other words, the outcomes of a trial are divided into two mutually exclusive events.
3. The probabilities of the two outcomes (or events) remain constant.
4. The trials are independent.

Note that one of the two outcomes (or events) of a trial is called a *success* and the other a *failure*. Notice that a success does not mean that the corresponding outcome is considered favorable or desirable. Similarly, a failure does not necessarily refer to an unfavorable or undesirable outcome. Success and failure are simply the names used to denote the two possible outcomes of a trial. The outcome to which the question refers is usually called a success; the outcome to which it does not refer is called a failure.

### ■ EXAMPLE 5–8

Consider the experiment consisting of 10 tosses of a coin. Determine whether or not it is a binomial experiment.

*Verifying the conditions of a binomial experiment.*

**Solution** The experiment consisting of 10 tosses of a coin satisfies all four conditions of a binomial experiment as explained below.

1. There are a total of 10 trials (tosses), and they are all identical. All 10 tosses are performed under identical conditions. Here,  $n = 10$ .
2. Each trial (toss) has only two possible outcomes: a head and a tail. Let a head be called a success and a tail be called a failure.
3. The probability of obtaining a head (a success) is  $1/2$  and that of a tail (a failure) is  $1/2$  for any toss. That is,

$$p = P(H) = 1/2 \quad \text{and} \quad q = P(T) = 1/2$$

The sum of these two probabilities is 1.0. Also, these probabilities remain the same for each toss.

4. The trials (tosses) are independent. The result of any preceding toss has no bearing on the result of any succeeding toss.

Consequently, the experiment consisting of 10 tosses is a binomial experiment. ■

### ■ EXAMPLE 5–9

- (a) Five percent of all DVD players manufactured by a large electronics company are defective. Three DVD players are randomly selected from the production line of this company. The selected DVD players are inspected to determine whether each of them is defective or good. Is this experiment a binomial experiment?
- (b) A box contains 20 cellphones, and two of them are defective. Three cellphones are randomly selected from this box and inspected to determine whether each of them is good or defective. Is this experiment a binomial experiment?

*Verifying the conditions of a binomial experiment.*

**Solution**

- (a) Below we check whether all four conditions of a binomial experiment are satisfied.
  1. This example consists of three identical trials. A trial represents the selection of a DVD player.
  2. Each trial has two outcomes: a DVD player is defective or a DVD player is good. Let a defective DVD player be called a success and a good DVD player be called a failure.
  3. Five percent of all DVD players are defective. So, the probability  $p$  that a DVD player is defective is .05. As a result, the probability  $q$  that a DVD player is good is .95. These two probabilities add up to 1.
  4. Each trial (DVD player) is independent. In other words, if one DVD player is defective, it does not affect the outcome of another DVD player being defective or good. This is so because the size of the population is very large compared to the sample size.

Because all four conditions of a binomial experiment are satisfied, this is an example of a binomial experiment.

- (b) Below we check whether all four conditions of a binomial experiment are satisfied.
1. This example consists of three identical trials, where a trial represents the selection of a cellphone.
  2. Each trial has two outcomes: a cellphone is good, or a cellphone is defective. Let a good cellphone be called a success and a defective cellphone be called a failure.
  3. There are a total of 20 cellphones in the box, and two of them are defective. Let  $p$  be the probability that a cellphone is good and  $q$  the probability that a cellphone is defective. These two probabilities  $p$  and  $q$  do *not* remain constant for each selection. Due to the limited number (20) of cellphones, the probability of each outcome changes with each selection, depending on what happened in the previous selection.
  4. Because  $p$  and  $q$  do not remain constant for each selection, trials are not independent. The outcome of the first selection affects the outcome of the second selection, and so on.

Given that the third and fourth conditions of a binomial experiment are not satisfied, this example is not an example of a binomial experiment. ■

### 5.4.2 The Binomial Probability Distribution and Binomial Formula

The random variable  $x$  that represents the number of successes in  $n$  trials for a binomial experiment is called a *binomial random variable*. The probability distribution of  $x$  in such experiments is called the **binomial probability distribution** or simply the **binomial distribution**. Thus, the binomial probability distribution is applied to find the probability of  $x$  successes in  $n$  trials for a binomial experiment. The number of successes  $x$  in such an experiment is a discrete random variable. Consider Example 5–9(a). Let  $x$  be the number of defective DVD players in a sample of three. Because we can obtain any number of defective DVD players from zero to three in a sample of three,  $x$  can assume any of the values 0, 1, 2, and 3. Since the values of  $x$  are countable, it is a discrete random variable.

**Binomial Formula** For a binomial experiment, the probability of exactly  $x$  successes in  $n$  trials is given by the binomial formula

$$P(x) = {}_nC_x p^x q^{n-x}$$

where

$n$  = total number of trials

$p$  = probability of success

$q = 1 - p$  = probability of failure

$x$  = number of successes in  $n$  trials

$n - x$  = number of failures in  $n$  trials

In the binomial formula,  $n$  is the total number of trials and  $x$  is the total number of successes. The difference between the total number of trials and the total number of successes,  $n - x$ , gives the total number of failures in  $n$  trials. The value of  ${}_nC_x$  gives the number of ways to obtain  $x$  successes in  $n$  trials. As mentioned earlier,  $p$  and  $q$  are the probabilities of success and failure, respectively. Again, although it does not matter which of the two outcomes is called a success and which a failure, usually the outcome to which the question refers is called a success.

To solve a binomial problem, we determine the values of  $n$ ,  $x$ ,  $n - x$ ,  $p$ , and  $q$  and then substitute these values in the binomial formula. To find the value of  ${}_n C_x$ , we can use either the combinations formula from Section 4.6.3 or a calculator.

To find the probability of  $x$  successes in  $n$  trials for a binomial experiment, the only values needed are those of  $n$  and  $p$ . These are called the *parameters of the binomial probability distribution* or simply the **binomial parameters**. The value of  $q$  is obtained by subtracting the value of  $p$  from 1.0. Thus,  $q = 1 - p$ .

Next we solve a binomial problem, first without using the binomial formula and then by using the binomial formula.

### ■ EXAMPLE 5–10

Five percent of all DVD players manufactured by a large electronics company are defective. A quality control inspector randomly selects three DVD players from the production line. What is the probability that exactly one of these three DVD players is defective?

*Calculating the probability using a tree diagram and the binomial formula.*

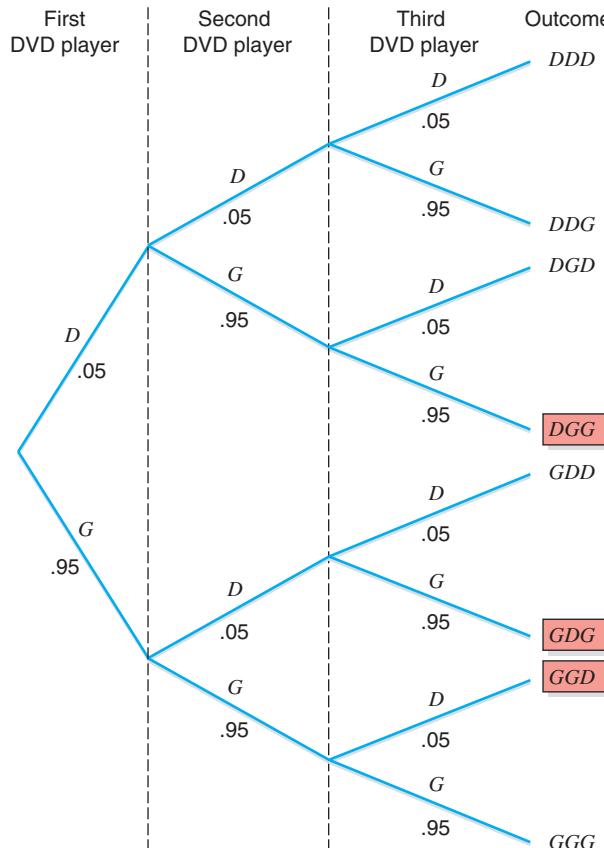
**Solution** Let

$D$  = a selected DVD player is defective

$G$  = a selected DVD player is good

As the tree diagram in Figure 5.4 shows, there are a total of eight outcomes, and three of them contain exactly one defective DVD player. These three outcomes are

$DGG$ ,  $GDG$ , and  $GGD$



**Figure 5.4** Tree diagram for selecting three DVD players.

We know that 5% of all DVD players manufactured at this company are defective. As a result, 95% of all DVD players are good. So the probability that a randomly selected DVD player is defective is .05 and the probability that it is good is .95.

$$P(D) = .05 \quad \text{and} \quad P(G) = .95$$

Because the size of the population is large (note that it is a large company), the selections can be considered to be independent. The probability of each of the three outcomes that give exactly one defective DVD player is calculated as follows:

$$P(DGG) = P(D) \cdot P(G) \cdot P(G) = (.05)(.95)(.95) = .0451$$

$$P(GDG) = P(G) \cdot P(D) \cdot P(G) = (.95)(.05)(.95) = .0451$$

$$P(GGD) = P(G) \cdot P(G) \cdot P(D) = (.95)(.95)(.05) = .0451$$

Note that  $DGG$  is simply the intersection of the three events  $D$ ,  $G$ , and  $G$ . In other words,  $P(DGG)$  is the joint probability of three events: the first DVD player selected is defective, the second is good, and the third is good. To calculate this probability, we use the multiplication rule for independent events we learned in Chapter 4. The same is true about the probabilities of the other two outcomes:  $GDG$  and  $GGD$ .

Exactly one defective DVD player will be selected if  $DGG$  or  $GDG$  or  $GGD$  occurs. These are three mutually exclusive outcomes. Therefore, from the addition rule of Chapter 4, the probability of the union of these three outcomes is simply the sum of their individual probabilities.

$$\begin{aligned} P(1 \text{ DVD player in 3 is defective}) &= P(DGG \text{ or } GDG \text{ or } GGD) \\ &= P(DGG) + P(GDG) + P(GGD) \\ &= .0451 + .0451 + .0451 = .1353 \end{aligned}$$

Now let us use the binomial formula to compute this probability. Let us call the selection of a defective DVD player a *success* and the selection of a good DVD player a *failure*. The reason we have called a defective DVD player a *success* is that the question refers to selecting exactly one defective DVD player. Then,

$$n = \text{total number of trials} = 3 \text{ DVD players}$$

$$x = \text{number of successes} = \text{number of defective DVD players} = 1$$

$$n - x = \text{number of failures} = \text{number of good DVD players} = 3 - 1 = 2$$

$$p = P(\text{success}) = .05$$

$$q = P(\text{failure}) = 1 - p = .95$$

The probability of one success is denoted by  $P(x = 1)$  or simply by  $P(1)$ . By substituting all of the values in the binomial formula, we obtain

$$P(x = 1) = {}_3C_1(.05)^1(.95)^2 = (3)(.05)(.9025) = .1354$$

Number of ways to obtain 1 success in 3 trials      Number of successes      Number of failures  
 ↓                          ↓                          ↓  
 Probability of success      Probability of failure

Note that the value of  ${}_3C_1$  in the formula either can be obtained from a calculator or can be computed as follows:

$${}_3C_1 = \frac{3!}{1!(3-1)!} = \frac{3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 1} = 3$$

In the above computation,  ${}_3C_1$  gives the three ways to select one defective DVD player in three selections. As listed previously, these three ways to select one defective DVD player are  $DGG$ ,  $GDG$ , and  $GGD$ . The probability .1354 is slightly different from the earlier calculation (.1353) because of rounding. ■

## ■ EXAMPLE 5-11

At the Express House Delivery Service, providing high-quality service to customers is the top priority of the management. The company guarantees a refund of all charges if a package it is delivering does not arrive at its destination by the specified time. It is known from past data that despite all efforts, 2% of the packages mailed through this company do not arrive at their destinations within the specified time. Suppose a corporation mails 10 packages through Express House Delivery Service on a certain day.

- (a) Find the probability that exactly one of these 10 packages will not arrive at its destination within the specified time.
- (b) Find the probability that at most one of these 10 packages will not arrive at its destination within the specified time.

**Solution** Let us call it a success if a package does not arrive at its destination within the specified time and a failure if it does arrive within the specified time. Then,

$$n = \text{total number of packages mailed} = 10$$

$$p = P(\text{success}) = .02$$

$$q = P(\text{failure}) = 1 - .02 = .98$$

- (a) For this part,

$$x = \text{number of successes} = 1$$

$$n - x = \text{number of failures} = 10 - 1 = 9$$

Substituting all values in the binomial formula, we obtain

$$\begin{aligned} P(x = 1) &= {}_{10}C_1(.02)^1(.98)^9 = \frac{10!}{1!(10-1)!}(.02)^1(.98)^9 \\ &= (10)(.02)(.83374776) = \mathbf{.1667} \end{aligned}$$

Thus, there is a .1667 probability that exactly one of the 10 packages mailed will not arrive at its destination within the specified time.

- (b) The probability that at most one of the 10 packages will not arrive at its destination within the specified time is given by the sum of the probabilities of  $x = 0$  and  $x = 1$ . Thus,

$$\begin{aligned} P(x \leq 1) &= P(x = 0) + P(x = 1) \\ &= {}_{10}C_0(.02)^0(.98)^{10} + {}_{10}C_1(.02)^1(.98)^9 \\ &= (1)(1)(.81707281) + (10)(.02)(.83374776) \\ &= .8171 + .1667 = \mathbf{.9838} \end{aligned}$$

Thus, the probability that at most one of the 10 packages will not arrive at its destination within the specified time is .9838. ■

*Calculating the probability using the binomial formula.*



PhotoDisc, Inc./Getty Images

## ■ EXAMPLE 5-12

In a Pew Research Center nationwide telephone survey conducted in March through April 2011, 74% of college graduates said that college provided them intellectual growth (*Time*, May 30, 2011). Assume that this result holds true for the current population of college graduates. Let  $x$  denote the number in a random sample of three college graduates who hold this opinion. Write the probability distribution of  $x$ , and draw a bar graph for this probability distribution.

*Constructing a binomial probability distribution and its graph.*

**Solution** Let  $x$  be the number of college graduates in a sample of three who hold the said opinion. Then,  $n - x$  is the number of college graduates who do not hold this opinion. From the given information,

$$n = \text{total college graduates in the sample} = 3$$

$$p = P(\text{a college graduate holds the said opinion}) = .74$$

$$q = P(\text{a college graduate does not hold the said opinion}) = 1 - .74 = .26$$

The possible values that  $x$  can assume are 0, 1, 2, and 3. In other words, the number of college graduates in a sample of three who hold the said opinion can be 0, 1, 2, or 3. The probability of each of these four outcomes is calculated as follows.

If  $x = 0$ , then  $n - x = 3$ . Using the binomial formula, we obtain the probability of  $x = 0$  as

$$P(x = 0) = {}_3C_0(.74)^0(.26)^3 = (1)(1)(.017576) = .0176$$

Note that  ${}_3C_0$  is equal to 1 by definition, and  $(.74)^0$  is equal to 1 because any number raised to the power zero is always 1.

If  $x = 1$ , then  $n - x = 2$ . Using the binomial formula, we obtain the probability of  $x = 1$  as

$$P(x = 1) = {}_3C_1(.74)^1(.26)^2 = (3)(.74)(.0676) = .1501$$

Similarly, if  $x = 2$ , then  $n - x = 1$ , and if  $x = 3$ , then  $n - x = 0$ . The probabilities of  $x = 2$  and  $x = 3$  are, respectively,

$$P(x = 2) = {}_3C_2(.74)^2(.26)^1 = (3)(.5476)(.26) = .4271$$

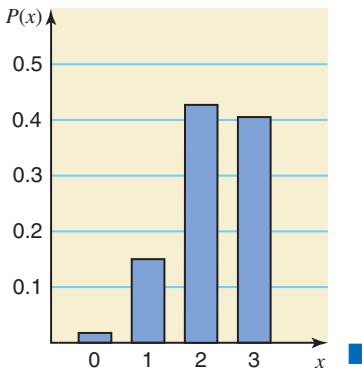
$$P(x = 3) = {}_3C_3(.74)^3(.26)^0 = (1)(.405224)(1) = .4052$$

These probabilities are written in Table 5.9. Figure 5.5 shows the bar graph for the probability distribution of Table 5.9.

**Table 5.9** Probability Distribution of  $x$

$x$	$P(x)$
0	.0176
1	.1501
2	.4271
3	.4052

**Figure 5.5** Bar graph of the probability distribution of  $x$ .



### 5.4.3 Using the Table of Binomial Probabilities

The probabilities for a binomial experiment can also be read from Table I, the table of binomial probabilities, in Appendix C. That table lists the probabilities of  $x$  for  $n = 1$  to  $n = 25$  and for selected values of  $p$ . Example 5–13 illustrates how to read Table I.

#### ■ EXAMPLE 5–13

In an NPD Group survey of adults, 30% of 50-year-old or older (let us call them 50-plus) adult Americans said that they would be willing to pay more for healthier options at restaurants (*USA TODAY*, July 20, 2011). Suppose this result holds true for the current population of 50-plus adult Americans. A random sample of six 50-plus adult Americans is selected. Using Table I of Appendix C, answer the following.

Using the binomial table to find probabilities and to construct the probability distribution and graph.

- Find the probability that exactly three persons in this sample hold the said opinion.
- Find the probability that at most two persons in this sample hold the said opinion.
- Find the probability that at least three persons in this sample hold the said opinion.
- Find the probability that one to three persons in this sample hold the said opinion.
- Let  $x$  be the number of 50-plus adult Americans in this sample who hold the said opinion. Write the probability distribution of  $x$ , and draw a bar graph for this probability distribution.

### Solution

- (a) To read the required probability from Table I of Appendix C, we first determine the values of  $n$ ,  $x$ , and  $p$ . For this example,

$$n = \text{number of persons in the sample} = 6$$

$$x = \text{number of persons in this sample who hold the said opinion} = 3$$

$$p = P(\text{a person holds the said opinion}) = .30$$

Then we locate  $n = 6$  in the column labeled  $n$  in Table I of Appendix C. The relevant portion of Table I with  $n = 6$  is reproduced as Table 5.10. Next, we locate 3 in the column for  $x$  in the portion of the table for  $n = 6$  and locate  $p = .30$  in the row for  $p$  at the top of the table. The entry at the intersection of the row for  $x = 3$  and the column for  $p = .30$  gives the probability of three successes in six trials when the probability of success is .30. From Table I or Table 5.10,

$$P(x = 3) = .1852$$

**Table 5.10** Determining  $P(x = 3)$  for  $n = 6$  and  $p = .30$

$n$	$x$	$p$					
		.05	.10	.20	.30	...	.95
$n = 6 \longrightarrow [6]$	0	.7351	.5314	.2621	.1176	...	.0000
	1	.2321	.3543	.3932	.3025	...	.0000
	2	.0305	.0984	.2458	.3241	...	.0001
$x = 3 \longrightarrow [3]$		.0021	.0146	.0819	.1852	...	.0021
	4	.0001	.0012	.0154	.0595	...	.0305
	5	.0000	.0001	.0015	.0102	...	.2321
	6	.0000	.0000	.0001	.0007	...	.7351

$p = .30$

$P(x = 3) = .1852$

Using Table I or Table 5.10, we write Table 5.11, which can be used to answer the remaining parts of this example.

**Table 5.11** Portion of Table I for  $n = 6$  and  $p = .30$

$n$	$x$	$p$	
		.30	
6	0	.1176	
	1	.3025	
	2	.3241	
	3	.1852	
	4	.0595	
	5	.0102	
	6	.0007	

- (b) The event that at most two 50-plus adult Americans in this sample hold the said opinion will occur if  $x$  is equal to 0, 1, or 2. From Table I of Appendix C or Table 5.11, the required probability is

$$\begin{aligned} P(\text{at most } 2) &= P(0 \text{ or } 1 \text{ or } 2) = P(x = 0) + P(x = 1) + P(x = 2) \\ &= .1176 + .3025 + .3241 = .7442 \end{aligned}$$

- (c) The probability that at least three 50-plus adult Americans in this sample hold the said opinion is given by the sum of the probabilities of 3, 4, 5, or 6. Using Table I of Appendix C or Table 5.11, we obtain

$$\begin{aligned} P(\text{at least } 3) &= P(3 \text{ or } 4 \text{ or } 5 \text{ or } 6) \\ &= P(x = 3) + P(x = 4) + P(x = 5) + P(x = 6) \\ &= .1852 + .0595 + .0102 + .0007 = .2556 \end{aligned}$$

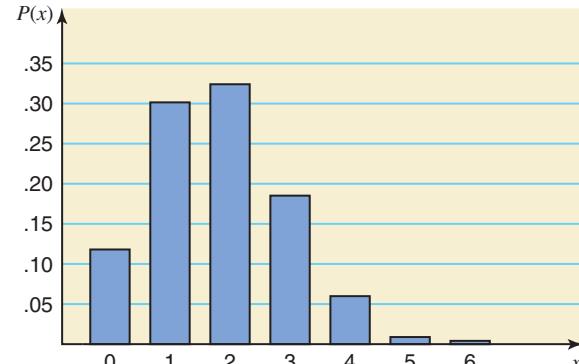
- (d) The probability that one to three 50-plus adult Americans in this sample hold the said opinion is given by the sum of the probabilities of  $x = 1, 2$ , and 3. Using Table I of Appendix C or Table 5.11, we obtain

$$\begin{aligned} P(1 \text{ to } 3) &= P(x = 1) + P(x = 2) + P(x = 3) \\ &= .3025 + .3241 + .1852 = .8118 \end{aligned}$$

- (e) Using Table I of Appendix C or Table 5.11, we list the probability distribution of  $x$  for  $n = 6$  and  $p = .30$  in Table 5.12. Figure 5.6 shows the bar graph of the probability distribution of  $x$ .

**Table 5.12** Probability Distribution of  $x$  for  $n = 6$  and  $p = .30$

$x$	$P(x)$
0	.1176
1	.3025
2	.3241
3	.1852
4	.0595
5	.0102
6	.0007



**Figure 5.6** Bar graph for the probability distribution of  $x$ .

#### 5.4.4 Probability of Success and the Shape of the Binomial Distribution

For any number of trials  $n$ :

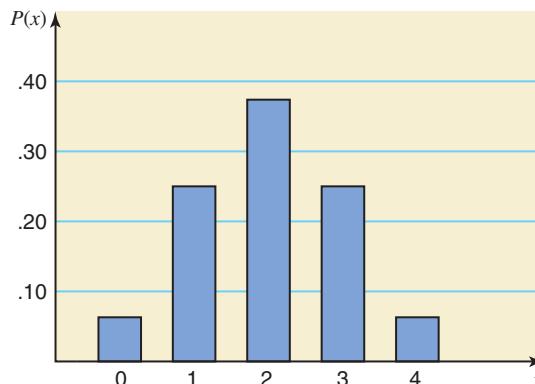
1. The binomial probability distribution is symmetric if  $p = .50$ .
2. The binomial probability distribution is skewed to the right if  $p$  is less than .50.
3. The binomial probability distribution is skewed to the left if  $p$  is greater than .50.

These three cases are illustrated next with examples and graphs.

1. Let  $n = 4$  and  $p = .50$ . Using Table I of Appendix C, we have written the probability distribution of  $x$  in Table 5.13 and plotted it in Figure 5.7. As we can observe from Table 5.13 and Figure 5.7, the probability distribution of  $x$  is symmetric.

**Table 5.13** Probability Distribution of  $x$  for  $n = 4$  and  $p = .50$

$x$	$P(x)$
0	.0625
1	.2500
2	.3750
3	.2500
4	.0625

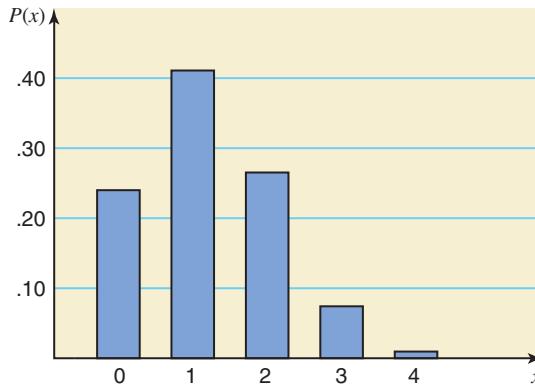


**Figure 5.7** Bar graph for the probability distribution of Table 5.13.

2. Let  $n = 4$  and  $p = .30$  (which is less than .50). Table 5.14, which is written by using Table I of Appendix C, and the graph of the probability distribution in Figure 5.8 show that the probability distribution of  $x$  for  $n = 4$  and  $p = .30$  is skewed to the right.

**Table 5.14** Probability Distribution of  $x$  for  $n = 4$  and  $p = .30$

$x$	$P(x)$
0	.2401
1	.4116
2	.2646
3	.0756
4	.0081

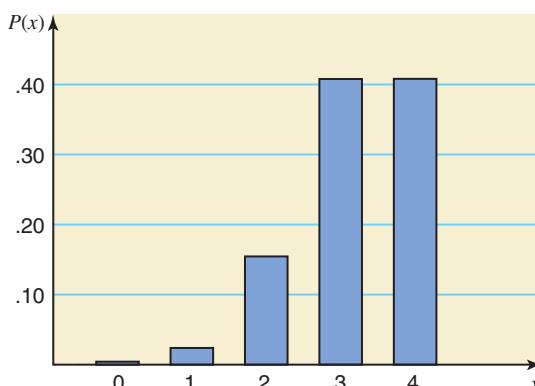


**Figure 5.8** Bar graph for the probability distribution of Table 5.14.

3. Let  $n = 4$  and  $p = .80$  (which is greater than .50). Table 5.15, which is written by using Table I of Appendix C, and the graph of the probability distribution in Figure 5.9 show that the probability distribution of  $x$  for  $n = 4$  and  $p = .80$  is skewed to the left.

**Table 5.15** Probability Distribution of  $x$  for  $n = 4$  and  $p = .80$

$x$	$P(x)$
0	.0016
1	.0256
2	.1536
3	.4096
4	.4096



**Figure 5.9** Bar graph for the probability distribution of Table 5.15.

### 5.4.5 Mean and Standard Deviation of the Binomial Distribution

Section 5.3 explained how to compute the mean and standard deviation, respectively, for a probability distribution of a discrete random variable. When a discrete random variable has a binomial distribution, the formulas learned in Section 5.3 could still be used to compute its mean and standard deviation. However, it is simpler and more convenient to use the following formulas to find the mean and standard deviation in such cases.

**Mean and Standard Deviation of a Binomial Distribution** The *mean and standard deviation of a binomial distribution* are, respectively,

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

where  $n$  is the total number of trials,  $p$  is the probability of success, and  $q$  is the probability of failure.

Example 5–14 describes the calculation of the mean and standard deviation of a binomial distribution.

#### ■ EXAMPLE 5–14

*Calculating the mean and standard deviation of a binomial random variable.*

In a 2011 *Time* magazine poll, American adults were asked, “When children today in the U.S. grow up, do you think they will be better off or worse off than people are now?” Of these adults, 52% said *worse*. Assume that this result is true for the current population of American adults. A sample of 50 adults is selected. Let  $x$  be the number of adults in this sample who hold the above-mentioned opinion. Find the mean and standard deviation of the probability distribution of  $x$ .

**Solution** This is a binomial experiment with a total of 50 trials (adults). Each trial has one of two outcomes: (1) The selected adult holds the said opinion, or (2) the selected adult does not hold the said opinion. The probabilities  $p$  and  $q$  for these two outcomes are .52 and .48, respectively. Thus,

$$n = 50, \quad p = .52, \quad \text{and} \quad q = .48$$

Using the formulas for the mean and standard deviation of the binomial distribution, we obtain

$$\begin{aligned}\mu &= np = 50(.52) = 26 \\ \sigma &= \sqrt{npq} = \sqrt{(50)(.52)(.48)} = 3.5327\end{aligned}$$

Thus, the mean of the probability distribution of  $x$  is 26, and the standard deviation is 3.5327. The value of the mean is what we expect to obtain, on average, per repetition of the experiment. In this example, if we select many samples of 50 adults each, we expect that each sample will contain an average of 26 adults, with a standard deviation of 3.5327, who will hold the said opinion. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**5.39** Briefly explain the following.

- a. A binomial experiment
- b. A trial
- c. A binomial random variable

**5.40** What are the parameters of the binomial probability distribution, and what do they mean?

**5.41** Which of the following are binomial experiments? Explain why.

- a. Rolling a die many times and observing the number of spots
- b. Rolling a die many times and observing whether the number obtained is even or odd



- c. Selecting a few voters from a very large population of voters and observing whether or not each of them favors a certain proposition in an election when 54% of all voters are known to be in favor of this proposition.

**5.42** Which of the following are binomial experiments? Explain why.

- Drawing 3 balls with replacement from a box that contains 10 balls, 6 of which are red and 4 are blue, and observing the colors of the drawn balls
- Drawing 3 balls without replacement from a box that contains 10 balls, 6 of which are red and 4 are blue, and observing the colors of the drawn balls
- Selecting a few households from New York City and observing whether or not they own stocks when it is known that 28% of all households in New York City own stocks

**5.43** Let  $x$  be a discrete random variable that possesses a binomial distribution. Using the binomial formula, find the following probabilities.

- $P(x = 5)$  for  $n = 8$  and  $p = .70$
- $P(x = 3)$  for  $n = 4$  and  $p = .40$
- $P(x = 2)$  for  $n = 6$  and  $p = .30$

Verify your answers by using Table I of Appendix C.

**5.44** Let  $x$  be a discrete random variable that possesses a binomial distribution. Using the binomial formula, find the following probabilities.

- $P(x = 0)$  for  $n = 5$  and  $p = .05$
- $P(x = 4)$  for  $n = 7$  and  $p = .90$
- $P(x = 7)$  for  $n = 10$  and  $p = .60$

Verify your answers by using Table I of Appendix C.

**5.45** Let  $x$  be a discrete random variable that possesses a binomial distribution.

- Using Table I of Appendix C, write the probability distribution of  $x$  for  $n = 7$  and  $p = .30$  and graph it.
- What are the mean and standard deviation of the probability distribution developed in part a?

**5.46** Let  $x$  be a discrete random variable that possesses a binomial distribution.

- Using Table I of Appendix C, write the probability distribution of  $x$  for  $n = 5$  and  $p = .80$  and graph it.
- What are the mean and standard deviation of the probability distribution developed in part a?

**5.47** The binomial probability distribution is symmetric for  $p = .50$ , skewed to the right for  $p < .50$ , and skewed to the left for  $p > .50$ . Illustrate each of these three cases by writing a probability distribution table and drawing a graph. Choose any values of  $n$  (equal to 4 or higher) and  $p$  and use the table of binomial probabilities (Table I of Appendix C) to write the probability distribution tables.

## ■ APPLICATIONS

**5.48** The most recent data from the Department of Education show that 34.8% of students who submitted otherwise valid applications for a Title IV Pell Grant in 2005–2006 were ineligible to receive such a grant ([www2.ed.gov/finaid/prof/resources/data/pell-2005-06/eoy-05-06.pdf](http://www2.ed.gov/finaid/prof/resources/data/pell-2005-06/eoy-05-06.pdf)). Suppose that this result is true for the current population of students who submitted otherwise valid applications for this grant.

- Let  $x$  be a binomial random variable that denotes the number of students in a random sample of 20 who submitted otherwise valid applications for a Title IV Pell Grant but were ineligible to receive one. What are the possible values that  $x$  can assume?
- Find the probability that exactly 6 students are ineligible to receive a Title IV Pell Grant in a random sample of 20 who submitted otherwise valid applications for this grant. Use the binomial probability distribution formula.

**5.49** According to a 2011 poll, 55% of Americans do not know that GOP stands for Grand Old Party (*Time*, October 17, 2011). Suppose that this result is true for the current population of Americans.

- Let  $x$  be a binomial random variable that denotes the number of people in a random sample of 17 Americans who do not know that GOP stands for Grand Old Party. What are the possible values that  $x$  can assume?
- Find the probability that exactly 8 people in a random sample of 17 Americans do not know that GOP stands for Grand Old Party. Use the binomial probability distribution formula.

**5.50** In a poll, men and women were asked, “When someone yelled or snapped at you at work, how did you want to respond?” Twenty percent of the women in the survey said that they felt like crying (*Time*, April 4, 2011). Suppose that this result is true for the current population of women employees. A random

sample of 24 women employees is selected. Use the binomial probabilities table (Table I of Appendix C) or technology to find the probability that the number of women employees in this sample of 24 who will hold the above opinion in response to the said question is

- a. at least 5
- b. 1 to 3
- c. at most 6

**5.51** According to a Wakefield Research survey of adult women, 50% of the women said that they had tried five or more diets in their lifetime (*USA TODAY*, June 21, 2011). Suppose that this result is true for the current population of adult women. A random sample of 13 adult women is selected. Use the binomial probabilities table (Table I of Appendix C) or technology to find the probability that the number of women in this sample of 13 who had tried five or more diets in their lifetime is

- a. at most 7
- b. 5 to 8
- c. at least 7

**5.52** Magnetic resonance imaging (MRI) is a process that produces internal body images using a strong magnetic field. Some patients become claustrophobic and require sedation because they are required to lie within a small, enclosed space during the MRI test. Suppose that 20% of all patients undergoing MRI testing require sedation due to claustrophobia. If five patients are selected at random, using the binomial probability distribution formula, find the probability that the number of patients in these five who require sedation is

- a. exactly 2
- b. none
- c. exactly 4

**5.53** In a 2011 *Time/Money Magazine* survey of adult Americans, 61% said that they were *less sure* that their children will achieve the American Dream (*Time*, October 10, 2011). Suppose that this result is true for the current population of adult Americans. A random sample of 16 adult Americans is selected. Using the binomial probability distribution formula, find the probability that the number of adult Americans in this sample of 16 who hold the above opinion is

- a. exactly 7
- b. none
- c. exactly 9

**5.54** During the 2011 NFL regular season, kickers converted 83.5% of the field goals attempted. Assume that this percentage is true for all kickers in the upcoming NFL season. What is the probability that a randomly selected kicker who will try 4 field goal attempts in a game will

- a. convert all 4 field goal attempts
- b. miss all 4 field goal attempts

**5.55** A professional basketball player makes 85% of the free throws he tries. Assuming this percentage holds true for future attempts, use the binomial formula to find the probability that in the next eight tries, the number of free throws he will make is

- a. exactly 8
- b. exactly 5

**5.56** Although Microsoft Windows is the primary operating system for desktop and laptop PC computers, Microsoft's Windows Phone operating system is installed in only 1.6% of smartphones ([www.latimes.com/business/la-fi-google-mobile-20110817,0,6230477.story](http://www.latimes.com/business/la-fi-google-mobile-20110817,0,6230477.story)).

- a. Assuming that 1.6% of all current smartphones have Microsoft's Windows Phone operating system, using the binomial formula, find the probability that the number of smartphones in a sample of 80 that have Microsoft's Windows Phone operating system is
  - i. exactly 2
  - ii. exactly 4
- b. Suppose that 5% of all current smartphones have Microsoft's Windows Phone operating system. Use the binomial probabilities table (Table I of Appendix C) or technology to find the probability that in a random sample of 20 smartphones, the number of smartphones that have Microsoft's Windows Phone operating system is
  - i. at most 2
  - ii. 2 to 3
  - iii. at least 2

**5.57** An office supply company conducted a survey before marketing a new paper shredder designed for home use. In the survey, 80% of the people who used the shredder were satisfied with it. Because of this high acceptance rate, the company decided to market the new shredder. Assume that 80% of all people who will use it will be satisfied. On a certain day, seven customers bought this shredder.

- a. Let  $x$  denote the number of customers in this sample of seven who will be satisfied with this shredder. Using the binomial probabilities table (Table I, Appendix C), obtain the probability distribution of  $x$  and draw a graph of the probability distribution. Find the mean and standard deviation of  $x$ .
- b. Using the probability distribution of part a, find the probability that exactly four of the seven customers will be satisfied.

**5.58** Johnson Electronics makes calculators. Consumer satisfaction is one of the top priorities of the company's management. The company guarantees a refund or a replacement for any calculator that malfunctions within 2 years from the date of purchase. It is known from past data that despite all efforts, 5% of the calculators manufactured by the company malfunction within a 2-year period. The company mailed a package of 10 randomly selected calculators to a store.

- a. Let  $x$  denote the number of calculators in this package of 10 that will be returned for refund or replacement within a 2-year period. Using the binomial probabilities table, obtain the probability distribution of  $x$  and draw a graph of the probability distribution. Determine the mean and standard deviation of  $x$ .
- b. Using the probability distribution of part a, find the probability that exactly 2 of the 10 calculators will be returned for refund or replacement within a 2-year period.

**5.59** A fast food chain store conducted a taste survey before marketing a new hamburger. The results of the survey showed that 70% of the people who tried this hamburger liked it. Encouraged by this result, the company decided to market the new hamburger. Assume that 70% of all people like this hamburger. On a certain day, eight customers bought it for the first time.

- a. Let  $x$  denote the number of customers in this sample of eight who will like this hamburger. Using the binomial probabilities table, obtain the probability distribution of  $x$  and draw a graph of the probability distribution. Determine the mean and standard deviation of  $x$ .
- b. Using the probability distribution of part a, find the probability that exactly three of the eight customers will like this hamburger.

## 5.5 The Hypergeometric Probability Distribution

In Section 5.4, we learned that one of the conditions required to apply the binomial probability distribution is that the trials are independent, so that the probabilities of the two outcomes or events (success and failure) remain constant. If the trials are not independent, we cannot apply the binomial probability distribution to find the probability of  $x$  successes in  $n$  trials. In such cases we replace the binomial by the **hypergeometric probability distribution**. Such a case occurs when a sample is drawn without replacement from a finite population.

As an example, suppose 20% of all auto parts manufactured at a company are defective. Four auto parts are selected at random. What is the probability that three of these four parts are good? Note that we are to find the probability that three of the four auto parts are good and one is defective. In this case, the population is very large and the probability of the first, second, third, and fourth auto parts being defective remains the same at .20. Similarly, the probability of any of the parts being good remains unchanged at .80. Consequently, we will apply the binomial probability distribution to find the probability of three good parts in four.

Now suppose this company shipped 25 auto parts to a dealer. Later, it finds out that 5 of those parts were defective. By the time the company manager contacts the dealer, 4 auto parts from that shipment have already been sold. What is the probability that 3 of those 4 parts were good parts and 1 was defective? Here, because the 4 parts were selected without replacement from a small population, the probability of a part being good changes from the first selection to the second selection, to the third selection, and to the fourth selection. In this case we cannot apply the binomial probability distribution. In such instances, we use the hypergeometric probability distribution to find the required probability.

### Hypergeometric Probability Distribution

Let

$$N = \text{total number of elements in the population}$$

$$r = \text{number of successes in the population}$$

$$N - r = \text{number of failures in the population}$$

$$n = \text{number of trials (sample size)}$$

$$x = \text{number of successes in } n \text{ trials}$$

$$n - x = \text{number of failures in } n \text{ trials}$$

The probability of  $x$  successes in  $n$  trials is given by

$$P(x) = \frac{rC_x N-rC_{n-x}}{NC_n}$$

Examples 5–15 and 5–16 provide applications of the hypergeometric probability distribution.

*Calculating probability by using hypergeometric distribution formula.*

### ■ EXAMPLE 5–15

Brown Manufacturing makes auto parts that are sold to auto dealers. Last week the company shipped 25 auto parts to a dealer. Later, it found out that 5 of those parts were defective. By the time the company manager contacted the dealer, 4 auto parts from that shipment had already been sold. What is the probability that 3 of those 4 parts were good parts and 1 was defective?

**Solution** Let a good part be called a success and a defective part be called a failure. From the given information,

$$N = \text{total number of elements (auto parts) in the population} = 25$$

$$r = \text{number of successes (good parts) in the population} = 20$$

$$N - r = \text{number of failures (defective parts) in the population} = 5$$

$$n = \text{number of trials (sample size)} = 4$$

$$x = \text{number of successes in four trials} = 3$$

$$n - x = \text{number of failures in four trials} = 1$$

Using the hypergeometric formula, we calculate the required probability as follows:

$$\begin{aligned} P(x = 3) &= \frac{rC_x N-rC_{n-x}}{N C_n} = \frac{20C_3 5C_1}{25C_4} = \frac{\frac{20!}{3!(20-3)!} \cdot \frac{5!}{1!(5-1)!}}{\frac{25!}{4!(25-4)!}} \\ &= \frac{(1140)(5)}{12,650} = .4506 \end{aligned}$$

Thus, the probability that 3 of the 4 parts sold are good and 1 is defective is .4506.

In the above calculations, the values of combinations can either be calculated using the formula learned in Section 4.6.3 (as done here) or by using a calculator. ■

### ■ EXAMPLE 5–16

*Calculating probability by using hypergeometric distribution formula.*

Dawn Corporation has 12 employees who hold managerial positions. Of them, 7 are female and 5 are male. The company is planning to send 3 of these 12 managers to a conference. If 3 managers are randomly selected out of 12,

- (a) find the probability that all 3 of them are female
- (b) find the probability that at most 1 of them is a female

**Solution** Let the selection of a female be called a success and the selection of a male be called a failure.

- (a) From the given information,

$$N = \text{total number of managers in the population} = 12$$

$$r = \text{number of successes (females) in the population} = 7$$

$$N - r = \text{number of failures (males) in the population} = 5$$

$$n = \text{number of selections (sample size)} = 3$$

$$x = \text{number of successes (females) in three selections} = 3$$

$$n - x = \text{number of failures (males) in three selections} = 0$$

Using the hypergeometric formula, we calculate the required probability as follows:

$$P(x = 3) = \frac{rC_x N-rC_{n-x}}{N C_n} = \frac{7C_3 5C_0}{12C_3} = \frac{(35)(1)}{220} = .1591$$

Thus, the probability that all 3 of the managers selected are female is .1591.

- (b) The probability that at most 1 of them is a female is given by the sum of the probabilities that either none or 1 of the selected managers is a female.

To find the probability that none of the selected managers is a female, we use

$$N = \text{total number of managers in the population} = 12$$

$$r = \text{number of successes (females) in the population} = 7$$

$$N - r = \text{number of failures (males) in the population} = 5$$

$$n = \text{number of selections (sample size)} = 3$$

$$x = \text{number of successes (females) in three selections} = 0$$

$$n - x = \text{number of failures (males) in three selections} = 3$$

Using the hypergeometric formula, we calculate the required probability as follows:

$$P(x = 0) = \frac{{}_rC_x {}_{N-r}C_{n-x}}{{}_NC_n} = \frac{{}_7C_0 {}_5C_3}{{}_{12}C_3} = \frac{(1)(10)}{220} = .0455$$

To find the probability that 1 of the selected managers is a female, we use

$$N = \text{total number of managers in the population} = 12$$

$$r = \text{number of successes (females) in the population} = 7$$

$$N - r = \text{number of failures (males) in the population} = 5$$

$$n = \text{number of selections (sample size)} = 3$$

$$x = \text{number of successes (females) in three selections} = 1$$

$$n - x = \text{number of failures (males) in three selections} = 2$$

Using the hypergeometric formula, we obtain the required probability as follows:

$$P(x = 1) = \frac{{}_rC_x {}_{N-r}C_{n-x}}{{}_NC_n} = \frac{{}_7C_1 {}_5C_2}{{}_{12}C_3} = \frac{(7)(10)}{220} = .3182$$

The probability that at most 1 of the 3 managers selected is a female is

$$P(x \leq 1) = P(x = 0) + P(x = 1) = .0455 + .3182 = \mathbf{.3637}$$

## EXERCISES

### CONCEPTS AND PROCEDURES

- 5.60** Explain the hypergeometric probability distribution. Under what conditions is this probability distribution applied to find the probability of a discrete random variable  $x$ ? Give one example of the application of the hypergeometric probability distribution.

- 5.61** Let  $N = 8$ ,  $r = 3$ , and  $n = 4$ . Using the hypergeometric probability distribution formula, find

- a.  $P(x = 2)$       b.  $P(x = 0)$       c.  $P(x \leq 1)$

- 5.62** Let  $N = 14$ ,  $r = 6$ , and  $n = 5$ . Using the hypergeometric probability distribution formula, find

- a.  $P(x = 4)$       b.  $P(x = 5)$       c.  $P(x \leq 1)$

- 5.63** Let  $N = 11$ ,  $r = 4$ , and  $n = 4$ . Using the hypergeometric probability distribution formula, find

- a.  $P(x = 2)$       b.  $P(x = 4)$       c.  $P(x \leq 1)$

- 5.64** Let  $N = 16$ ,  $r = 10$ , and  $n = 5$ . Using the hypergeometric probability distribution formula, find

- a.  $P(x = 5)$       b.  $P(x = 0)$       c.  $P(x \leq 1)$

### APPLICATIONS

- 5.65** An Internal Revenue Service inspector is to select 3 corporations from a list of 15 for tax audit purposes. Of the 15 corporations, 6 earned profits and 9 incurred losses during the year for which the tax returns are to be audited. If the IRS inspector decides to select 3 corporations randomly, find the probability that the number of corporations in these 3 that incurred losses during the year for which the tax returns are to be audited is

- a. exactly 2      b. none      c. at most 1

**5.66** Six jurors are to be selected from a pool of 20 potential candidates to hear a civil case involving a lawsuit between two families. Unknown to the judge or any of the attorneys, 4 of the 20 prospective jurors are potentially prejudiced by being acquainted with one or more of the litigants. They will not disclose this during the jury selection process. If 6 jurors are selected at random from this group of 20, find the probability that the number of potentially prejudiced jurors among the 6 selected jurors is

- a. exactly 1
- b. none
- c. at most 2

**5.67** A really bad carton of 18 eggs contains 7 spoiled eggs. An unsuspecting chef picks 4 eggs at random for his “Mega-Omelet Surprise.” Find the probability that the number of *unspoiled* eggs among the 4 selected is

- a. exactly 4
- b. 2 or fewer
- c. more than 1

**5.68** Bender Electronics buys keyboards for its computers from another company. The keyboards are received in shipments of 100 boxes, each box containing 20 keyboards. The quality control department at Bender Electronics first randomly selects one box from each shipment and then randomly selects 5 keyboards from that box. The shipment is accepted if not more than 1 of the 5 keyboards is defective. The quality control inspector at Bender Electronics selected a box from a recently received shipment of keyboards. Unknown to the inspector, this box contains 6 defective keyboards.

- a. What is the probability that this shipment will be accepted?
- b. What is the probability that this shipment will not be accepted?

## 5.6 The Poisson Probability Distribution

The **Poisson probability distribution**, named after the French mathematician Siméon-Denis Poisson, is another important probability distribution of a discrete random variable that has a large number of applications. Suppose a washing machine in a laundromat breaks down an average of three times a month. We may want to find the probability of exactly two breakdowns during the next month. This is an example of a Poisson probability distribution problem. Each breakdown is called an *occurrence* in Poisson probability distribution terminology. The Poisson probability distribution is applied to experiments with random and independent occurrences. The occurrences are random in the sense that they do not follow any pattern, and, hence, they are unpredictable. Independence of occurrences means that one occurrence (or nonoccurrence) of an event does not influence the successive occurrences or nonoccurrences of that event. The occurrences are always considered with respect to an interval. In the example of the washing machine, the interval is one month. The interval may be a time interval, a space interval, or a volume interval. The actual number of occurrences within an interval is random and independent. If the average number of occurrences for a given interval is known, then by using the Poisson probability distribution, we can compute the probability of a certain number of occurrences,  $x$ , in that interval. Note that the number of actual occurrences in an interval is denoted by  $x$ .

**Conditions to Apply the Poisson Probability Distribution** The following three conditions must be satisfied to apply the Poisson probability distribution.

1.  $x$  is a discrete random variable.
2. The occurrences are random.
3. The occurrences are independent.

The following are three examples of discrete random variables for which the occurrences are random and independent. Hence, these are examples to which the Poisson probability distribution can be applied.

1. Consider the number of telemarketing phone calls received by a household during a given day. In this example, the receiving of a telemarketing phone call by a household is called an occurrence, the interval is one day (an interval of time), and the occurrences are random (that is, there is no specified time for such a phone call to come in) and discrete. The total number of telemarketing phone calls received by a household during a given day may be 0, 1, 2, 3, 4, and so forth. The independence of occurrences in this example means that the telemarketing phone calls are received individually and none of two (or more) of these phone calls are related.

2. Consider the number of defective items in the next 100 items manufactured on a machine. In this case, the interval is a volume interval (100 items). The occurrences (number of defective items) are random and discrete because there may be 0, 1, 2, 3,..., 100 defective items in 100 items. We can assume the occurrence of defective items to be independent of one another.
3. Consider the number of defects in a 5-foot-long iron rod. The interval, in this example, is a space interval (5 feet). The occurrences (defects) are random because there may be any number of defects in a 5-foot iron rod. We can assume that these defects are independent of one another.

The following examples also qualify for the application of the Poisson probability distribution.

1. The number of accidents that occur on a given highway during a 1-week period
2. The number of customers entering a grocery store during a 1-hour interval
3. The number of television sets sold at a department store during a given week

In contrast, consider the arrival of patients at a physician's office. These arrivals are non-random if the patients have to make appointments to see the doctor. The arrival of commercial airplanes at an airport is nonrandom because all planes are scheduled to arrive at certain times, and airport authorities know the exact number of arrivals for any period (although this number may change slightly because of late or early arrivals and cancellations). The Poisson probability distribution cannot be applied to these examples.

In the Poisson probability distribution terminology, the average number of occurrences in an interval is denoted by  $\lambda$  (Greek letter *lambda*). The actual number of occurrences in that interval is denoted by  $x$ . Then, using the Poisson probability distribution, we find the probability of  $x$  occurrences during an interval given that the mean occurrences during that interval are  $\lambda$ .

**Poisson Probability Distribution Formula** According to the *Poisson probability distribution*, the probability of  $x$  occurrences in an interval is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where  $\lambda$  (pronounced *lambda*) is the mean number of occurrences in that interval and the value of  $e$  is approximately 2.71828.

The mean number of occurrences in an interval, denoted by  $\lambda$ , is called the *parameter of the Poisson probability distribution* or the **Poisson parameter**. As is obvious from the Poisson probability distribution formula, we need to know only the value of  $\lambda$  to compute the probability of any given value of  $x$ . We can read the value of  $e^{-\lambda}$  for a given  $\lambda$  from Table II of Appendix C. Examples 5–17 through 5–19 illustrate the use of the Poisson probability distribution formula.

## ■ EXAMPLE 5–17

On average, a household receives 9.5 telemarketing phone calls per week. Using the Poisson probability distribution formula, find the probability that a randomly selected household receives exactly 6 telemarketing phone calls during a given week.

Using the Poisson formula:  
 $x$  equals a specific value.

**Solution** Let  $\lambda$  be the mean number of telemarketing phone calls received by a household per week. Then,  $\lambda = 9.5$ . Let  $x$  be the number of telemarketing phone calls received by a household during a given week. We are to find the probability of  $x = 6$ . Substituting all of the values in the Poisson formula, we obtain

$$P(x = 6) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(9.5)^6 e^{-9.5}}{6!} = \frac{(735,091.8906)(.00007485)}{720} = .0764$$

To do these calculations, we can find the value of  $6!$  either by using the factorial key on a calculator or by multiplying all integers from 1 to 6, and we can find the value of  $e^{-9.5}$  by using the  $e^x$  key on a calculator or from Table II in Appendix C. ■

*Calculating probabilities using the Poisson formula.*

### ■ EXAMPLE 5–18

A washing machine in a laundromat breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have

- (a) exactly two breakdowns
- (b) at most one breakdown

**Solution** Let  $\lambda$  be the mean number of breakdowns per month, and let  $x$  be the actual number of breakdowns observed during the next month for this machine. Then,

$$\lambda = 3$$

- (a) The probability that exactly two breakdowns will be observed during the next month is

$$P(x = 2) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(3)^2 e^{-3}}{2!} = \frac{(9)(.04978707)}{2} = .2240$$

- (b) The probability that at most one breakdown will be observed during the next month is given by the sum of the probabilities of zero and one breakdown. Thus,

$$\begin{aligned} P(\text{at most 1 breakdown}) &= P(0 \text{ or } 1 \text{ breakdown}) = P(x = 0) + P(x = 1) \\ &= \frac{(3)^0 e^{-3}}{0!} + \frac{(3)^1 e^{-3}}{1!} \\ &= \frac{(1)(.04978707)}{1} + \frac{(3)(.04978707)}{1} \\ &= .0498 + .1494 = .1992 \end{aligned}$$

**Remember ▶**

One important point about the Poisson probability distribution is that *the intervals for  $\lambda$  and  $x$  must be equal*. If they are not, the mean  $\lambda$  should be redefined to make them equal. Example 5–19 illustrates this point.

### ■ EXAMPLE 5–19

*Calculating a probability using the Poisson formula.*

Cynthia's Mail Order Company provides free examination of its products for 7 days. If not completely satisfied, a customer can return the product within that period and get a full refund. According to past records of the company, an average of 2 of every 10 products sold by this company are returned for a refund. Using the Poisson probability distribution formula, find the probability that exactly 6 of the 40 products sold by this company on a given day will be returned for a refund.

**Solution** Let  $x$  denote the number of products in 40 that will be returned for a refund. We are to find  $P(x = 6)$ . The given mean is defined per 10 products, but  $x$  is defined for 40 products. As a result, we should first find the mean for 40 products. Because, on average, 2 out of 10 products are returned, the mean number of products returned out of 40 will be 8. Thus,  $\lambda = 8$ . Substituting  $x = 6$  and  $\lambda = 8$  in the Poisson probability distribution formula, we obtain

$$P(x = 6) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(8)^6 e^{-8}}{6!} = \frac{(262,144)(.00033546)}{720} = .1221$$

Thus, the probability is .1221 that exactly 6 products out of 40 sold on a given day will be returned.

Note that Example 5–19 is actually a binomial problem with  $p = 2/10 = .20$ ,  $n = 40$ , and  $x = 6$ . In other words, the probability of success (that is, the probability that a product is returned) is .20 and the number of trials (products sold) is 40. We are to find the probability of six successes (returns). However, we used the Poisson distribution to solve this problem. This

is referred to as *using the Poisson distribution as an approximation to the binomial distribution*. We can also use the binomial distribution to find this probability as follows:

$$\begin{aligned} P(x = 6) &= {}_{40}C_6 (.20)^6 (.80)^{34} = \frac{40!}{6!(40 - 6)!} (.20)^6 (.80)^{34} \\ &= (3,838,380)(.000064)(.00050706) = .1246 \end{aligned}$$

Thus the probability  $P(x = 6)$  is .1246 when we use the binomial distribution.

As we can observe, simplifying the above calculations for the binomial formula is quite complicated when  $n$  is large. It is much easier to solve this problem using the Poisson probability distribution. As a general rule, if it is a binomial problem with  $n > 25$  but  $\mu \leq 25$ , then we can use the Poisson probability distribution as an approximation to the binomial distribution. However, if  $n > 25$  and  $\mu > 25$ , we prefer to use the normal distribution as an approximation to the binomial. The latter case will be discussed in Chapter 6. However, if you are using technology, it does not matter how large  $n$  is. You can always use the binomial probability distribution if it is a binomial problem.

Case Study 5–2 presents an application of the Poisson probability distribution.

### 5.6.1 Using the Table of Poisson Probabilities

The probabilities for a Poisson distribution can also be read from Table III in Appendix C, the table of Poisson probabilities. The following example describes how to read that table.

#### ■ EXAMPLE 5–20

On average, two new accounts are opened per day at an Imperial Savings Bank branch. Using Table III of Appendix C, find the probability that on a given day the number of new accounts opened at this bank will be

- (a) exactly 6      (b) at most 3      (c) at least 7

**Solution** Let

$\lambda$  = mean number of new accounts opened per day at this bank

$x$  = number of new accounts opened at this bank on a given day

- (a) The values of  $\lambda$  and  $x$  are

$$\lambda = 2 \quad \text{and} \quad x = 6$$

In Table III of Appendix C, we first locate the column that corresponds to  $\lambda = 2$ . In this column, we then read the value for  $x = 6$ . The relevant portion of that table is shown here as Table 5.16. The probability that exactly 6 new accounts will be opened on a given day is .0120. Therefore,

$$P(x = 6) = .0120$$

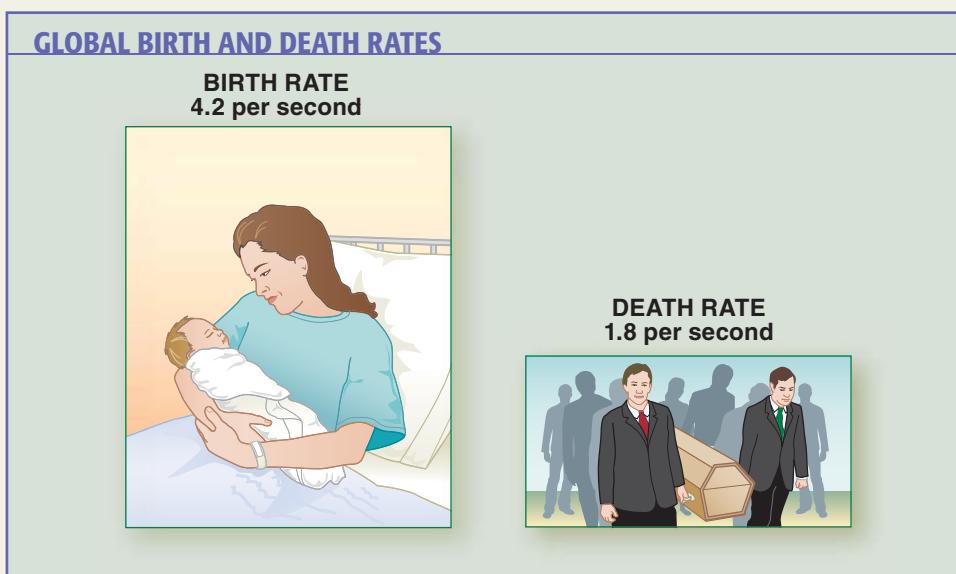
**Table 5.16** Portion of Table III for  $\lambda = 2.0$

$x$	1.1	1.2	...	<b>2.0</b>	$\leftarrow \lambda = 2.0$
0				.1353	
1				.2707	
2				.2707	
3				.1804	
4				.0902	
5				.0361	
$x = 6 \longrightarrow$	<b>6</b>			<b>.0120</b>	$\leftarrow P(x = 6)$
	7			.0034	
	8			.0009	
	9			.0002	

*Using the table of Poisson probabilities.*

Actually, Table 5.16 gives the probability distribution of  $x$  for  $\lambda = 2.0$ . Note that the sum of the 10 probabilities given in Table 5.16 is .9999 and not 1.0. This is so for two

## GLOBAL BIRTH AND DEATH RATES



Data source: The International Data Base and U.S. Census Bureau.

The accompanying graph shows the average global birth and death rates. According to this information, on average, 4.2 children are born per second and 1.8 persons die per second in the world. These rates are based on data collected by the International Data Base and the U.S. Census Bureau (<http://www.census.gov/population/international/data/idb/worldvitalevents.php>). If we assume that the global birth and death rates follow the Poisson probability distribution, we can find the probability of any given number of global births or deaths for a given time interval. For example, if  $x$  is the actual number of global births in a 1-second interval, then  $x$  can assume any (nonnegative integer) value, such as 0, 1, 2, 3,.... The same is true for the number of global deaths per 1-second interval. For example, if  $y$  is the actual number of global deaths in a 1-second interval, then  $y$  can assume any (nonnegative integer) value, such as 0, 1, 2, 3,.... Here  $x$  and  $y$  are both discrete random variables.

Using the Poisson formula or Table III of Appendix C, we can find the probability of any values of  $x$  and  $y$ . For example, if we want to find the probability of at most three global births during any given 1-second interval, then, using  $\lambda = 4.2$ , we find this probability from Table III as

$$P(x \leq 3) = P(0) + P(1) + P(2) + P(3) = .0150 + .0630 + .1323 + .1852 = .3955$$

Now suppose we want to find the probability of exactly six global births using the Poisson formula. This probability is

$$P(x = 6) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{4.2^6 e^{-4.2}}{6!} = .1143$$

As mentioned earlier, let  $y$  be the number of global deaths in a given 1-second interval. If we want to find the probability of at most two global deaths during any given 1-second interval, then, using  $\lambda = 1.8$ , we find this probability from Table III as

$$P(y \leq 2) = P(0) + P(1) + P(2) = .1653 + .2975 + .2678 = .7306$$

Now suppose we want to find the probability of exactly three global deaths in a given 1-second interval using the Poisson formula. This probability is

$$P(y = 3) = \frac{\lambda^y e^{-\lambda}}{y!} = \frac{1.8^3 e^{-1.8}}{3!} = .1607$$

Using Table III of Appendix C, we can prepare the probability distributions of  $x$  and  $y$ .

reasons. First, these probabilities are rounded to four decimal places. Second, on a given day more than 9 new accounts might be opened at this bank. However, the probabilities of 10, 11, 12, . . . new accounts are very small, and they are not listed in the table.

- (b) The probability that at most three new accounts are opened on a given day is obtained by adding the probabilities of 0, 1, 2, and 3 new accounts. Thus, using Table III of Appendix C or Table 5.16, we obtain

$$\begin{aligned} P(\text{at most } 3) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &= .1353 + .2707 + .2707 + .1804 = \mathbf{.8571} \end{aligned}$$

- (c) The probability that at least 7 new accounts are opened on a given day is obtained by adding the probabilities of 7, 8, and 9 new accounts. Note that 9 is the last value of  $x$  for  $\lambda = 2.0$  in Table III of Appendix C or Table 5.16. Hence, 9 is the last value of  $x$  whose probability is included in the sum. However, this does not mean that on a given day more than 9 new accounts cannot be opened. It simply means that the probability of 10 or more accounts is close to zero. Thus,

$$\begin{aligned} P(\text{at least } 7) &= P(x = 7) + P(x = 8) + P(x = 9) \\ &= .0034 + .0009 + .0002 = \mathbf{.0045} \end{aligned}$$

*Constructing a Poisson probability distribution and graphing it.*

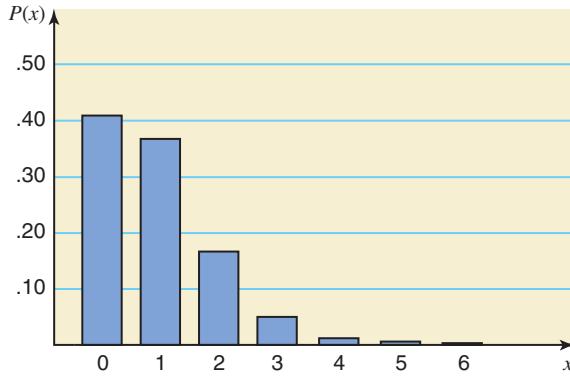
### ■ EXAMPLE 5–21

An auto salesperson sells an average of .9 car per day. Let  $x$  be the number of cars sold by this salesperson on any given day. Using the Poisson probability distribution table, write the probability distribution of  $x$ . Draw a graph of the probability distribution.

**Solution** Let  $\lambda$  be the mean number of cars sold per day by this salesperson. Hence,  $\lambda = .9$ . Using the portion of Table III of Appendix C that corresponds to  $\lambda = .9$ , we write the probability distribution of  $x$  in Table 5.17. Figure 5.10 shows the bar graph for the probability distribution of Table 5.17.

**Table 5.17** Probability Distribution of  $x$  for  $\lambda = .9$

$x$	$P(x)$
0	.4066
1	.3659
2	.1647
3	.0494
4	.0111
5	.0020
6	.0003



**Figure 5.10** Bar graph for the probability distribution of Table 5.17.

Note that 6 is the largest value of  $x$  for  $\lambda = .9$  listed in Table III for which the probability is greater than zero. However, this does not mean that this salesperson cannot sell more than six cars on a given day. What this means is that the probability of selling seven or more cars is very small. Actually, the probability of  $x = 7$  for  $\lambda = .9$  calculated by using the Poisson formula is .000039. When rounded to four decimal places, this probability is .0000, as listed in Table III. ■

## 5.6.2 Mean and Standard Deviation of the Poisson Probability Distribution

For the Poisson probability distribution, the mean and variance both are equal to  $\lambda$ , and the standard deviation is equal to  $\sqrt{\lambda}$ . That is, for the Poisson probability distribution,

$$\mu = \lambda, \quad \sigma^2 = \lambda, \quad \text{and} \quad \sigma = \sqrt{\lambda}$$

For Example 5–21,  $\lambda = .9$ . Therefore, for the probability distribution of  $x$  in Table 5.17, the mean, variance, and standard deviation are, respectively,

$$\mu = \lambda = .9 \text{ car}$$

$$\sigma^2 = \lambda = .9$$

$$\sigma = \sqrt{\lambda} = \sqrt{.9} = .949 \text{ car}$$



## EXERCISES

### CONCEPTS AND PROCEDURES

**5.69** What are the conditions that must be satisfied to apply the Poisson probability distribution?

**5.70** What is the parameter of the Poisson probability distribution, and what does it mean?

**5.71** Using the Poisson formula, find the following probabilities.

- a.  $P(x \leq 1)$  for  $\lambda = 5$
- b.  $P(x = 2)$  for  $\lambda = 2.5$

Verify these probabilities using Table III of Appendix C.

**5.72** Using the Poisson formula, find the following probabilities.

- a.  $P(x < 2)$  for  $\lambda = 3$
- b.  $P(x = 8)$  for  $\lambda = 5.5$

Verify these probabilities using Table III of Appendix C.

**5.73** Let  $x$  be a Poisson random variable. Using the Poisson probabilities table, write the probability distribution of  $x$  for each of the following. Find the mean, variance, and standard deviation for each of these probability distributions. Draw a graph for each of these probability distributions.

- a.  $\lambda = 1.3$
- b.  $\lambda = 2.1$

**5.74** Let  $x$  be a Poisson random variable. Using the Poisson probabilities table, write the probability distribution of  $x$  for each of the following. Find the mean, variance, and standard deviation for each of these probability distributions. Draw a graph for each of these probability distributions.

- a.  $\lambda = .6$
- b.  $\lambda = 1.8$

### APPLICATIONS

**5.75** A household receives an average of 1.7 pieces of junk mail per day. Find the probability that this household will receive exactly 3 pieces of junk mail on a certain day. Use the Poisson probability distribution formula.

**5.76** A commuter airline receives an average of 9.7 complaints per day from its passengers. Using the Poisson formula, find the probability that on a certain day this airline will receive exactly 6 complaints.

**5.77** On average, 5.4 shoplifting incidents occur per week at an electronics store. Find the probability that exactly 3 such incidents will occur during a given week at this store.

**5.78** On average, 12.5 rooms stay vacant per day at a large hotel in a city. Find the probability that on a given day exactly 3 rooms will be vacant. Use the Poisson formula.

**5.79** A university police department receives an average of 3.7 reports per week of lost student ID cards.

- a. Find the probability that at most 1 such report will be received during a given week by this police department. Use the Poisson probability distribution formula.
- b. Using the Poisson probabilities table, find the probability that during a given week the number of such reports received by this police department is
  - i. 1 to 4
  - ii. at least 6
  - iii. at most 3

**5.80** A large proportion of small businesses in the United States fail during the first few years of operation. On average, 1.6 businesses file for bankruptcy per day in a particular large city.

- a. Using the Poisson formula, find the probability that exactly 3 businesses will file for bankruptcy on a given day in this city.
- b. Using the Poisson probabilities table, find the probability that the number of businesses that will file for bankruptcy on a given day in this city is
  - i. 2 to 3
  - ii. more than 3
  - iii. less than 3

**5.81** Despite all efforts by the quality control department, the fabric made at Benton Corporation always contains a few defects. A certain type of fabric made at this corporation contains an average of .5 defect per 500 yards.

- a. Using the Poisson formula, find the probability that a given piece of 500 yards of this fabric will contain exactly 1 defect.
- b. Using the Poisson probabilities table, find the probability that the number of defects in a given 500-yard piece of this fabric will be
  - i. 2 to 4
  - ii. more than 3
  - iii. less than 2

**5.82** The number of students who log in to a randomly selected computer in a college computer lab follows a Poisson probability distribution with a mean of 19 students per day.

- a. Using the Poisson probability distribution formula, determine the probability that exactly 12 students will log in to a randomly selected computer at this lab on a given day.
- b. Using the Poisson probability distribution table, determine the probability that the number of students who will log in to a randomly selected computer at this lab on a given day is
  - i. from 13 to 16
  - ii. fewer than 8

**5.83** According to a study performed by the NCAA, the average rate of injuries occurring in collegiate women's soccer is 8.6 per 1000 participants ([www.fastsports.com/tips/tip12/](http://www.fastsports.com/tips/tip12/)).

- a. Using the Poisson formula, find the probability that the number of injuries in a sample of 1000 women's soccer participants is
  - i. exactly 12
  - ii. exactly 5
- b. Using the Poisson probabilities table, find the probability that the number of injuries in a sample of 1000 women's soccer participants is
  - i. more than 3
  - ii. less than 10
  - iii. 8 to 13

**5.84** Although Borok's Electronics Company has no openings, it still receives an average of 3.2 unsolicited applications per week from people seeking jobs.

- a. Using the Poisson formula, find the probability that this company will receive no applications next week.
- b. Let  $x$  denote the number of applications this company will receive during a given week. Using the Poisson probabilities table from Appendix C, write the probability distribution table of  $x$ .
- c. Find the mean, variance, and standard deviation of the probability distribution developed in part b.

**5.85** An insurance salesperson sells an average of 1.4 policies per day.

- a. Using the Poisson formula, find the probability that this salesperson will sell no insurance policy on a certain day.
- b. Let  $x$  denote the number of insurance policies that this salesperson will sell on a given day. Using the Poisson probabilities table, write the probability distribution of  $x$ .
- c. Find the mean, variance, and standard deviation of the probability distribution developed in part b.

**5.86** An average of .8 accident occur per day in a particular large city.

- a. Find the probability that no accident will occur in this city on a given day.
- b. Let  $x$  denote the number of accidents that will occur in this city on a given day. Write the probability distribution of  $x$ .
- c. Find the mean, variance, and standard deviation of the probability distribution developed in part b.

**\*5.87** On average, 20 households in 50 own answering machines.

- a. Using the Poisson formula, find the probability that in a random sample of 50 households, exactly 25 will own answering machines.
- b. Using the Poisson probabilities table, find the probability that the number of households in 50 who own answering machines is
  - i. at most 12
  - ii. 13 to 17
  - iii. at least 30

**\*5.88** Twenty percent of the cars passing through a school zone are exceeding the speed limit by more than 10 mph.

- a. Using the Poisson formula, find the probability that in a random sample of 100 cars passing through this school zone, exactly 25 will exceed the speed limit by more than 10 mph.
- b. Using the Poisson probabilities table, find the probability that the number of cars exceeding the speed limit by more than 10 mph in a random sample of 100 cars passing through this school zone is
  - i. at most 8
  - ii. 15 to 20
  - iii. at least 30

## USES AND MISUSES...

### 1. PUT ON YOUR GAME FACE

Gambling would be nothing without probability. A gambler always has a positive probability of winning. Unfortunately, the house always plays with better odds. A classic discrete probability distribution applies to the hands in straight poker. Using the tools you have learned in this chapter and a bit of creativity, you can derive the probability of being dealt a certain hand. However, this probability distribution is only going to be of limited use when you begin to play poker.

The hands in descending order of rank and increasing order of probability are straight flush, four-of-a-kind, full house, flush, straight, three-of-a-kind, two pair, pair, and high cards. To begin, let us determine how many hands there are. As we know, there are 52 cards in a deck, and any 5 cards can be a valid hand. Using the combinations notation, we see that there are  ${}_{52}C_5$  or 2,598,960 hands.

We can count the highest hands based on their composition. The straight flush is any 5 cards in rank order from the same suit. Because an ace can be high or low, there are 10 straight flushes per suit. Because there are 4 suits, this gives us 40 straight flushes. Once you have chosen your rank for four of a kind, for example, a jack, there are  $52 - 4 = 48$  remaining cards. Hence, there are  $13 \times 48 = 624$  possible four-of-a-kind hands.

The rest of the hands require us to use the combinations notation to determine their numbers. A full house is three of a kind and a pair (for example, three kings and a pair of 7s). There are 13 choices for three of a kind (for example, three aces, three kings, and so on); then there are  ${}_4C_3 = 4$  ways to choose each set of three of a kind from 4 cards (for example, 3 kings out of 4). Once three cards of a kind have been selected, there are 12 possibilities for a pair, and  ${}_4C_2 = 6$  ways to choose any 2 cards for a pair out of 4 cards (e.g., two 9s out of 4). Thus, there are  $13 \times 4 \times 12 \times 6 = 3744$  full houses. A flush is five cards drawn from the same suit. Hence, we have 4 suits multiplied by  ${}_{13}C_5$  ways to choose the members, which gives 5148 flushes. However, 40 of those are straight flushes, so 5108 flushes are not straight flushes. For brevity, we omit the calculation of the remainder of the hands and present the results and the probability of being dealt the hand in the table below.

	Number of Hands	Probability
Straight flush	40	.0000154
Four of a kind	624	.0002401
Full house	3744	.0014406
Flush	5108	.0019654
Straight	10,200	.0039246
Three of a kind	54,912	.0211285
Two pair	123,552	.0475390
Pair	1,098,240	.4225690
High card	1,302,540	.5011774
Total	2,598,960	1.000000

Memorizing this table is only the beginning of poker. Any table entry represents the probability that the five cards you have been dealt constitute one of the nine poker hands. Suppose that you are playing poker with four people and you are dealt a pair of sevens. The probability of being dealt a hand that is classified as a pair is .4225690, but that is not the probability in which you are interested. You want to know the probability that the pair of sevens that you hold will beat the hands your opponents were dealt. Despite your intimate knowledge of the probability of your hand, the above table gives information for only one player. Be very careful when working with probability distributions, and make sure you understand exactly what the probabilities represent.

### 2. ACTUARIAL SCIENCE

Although many people who take a course in statistics do so because they need to perform or understand data analyses related to their major, there are a number of people who go into careers that heavily involve probability and statistics on a daily basis. One field that has received a great deal of attention in the public press over the last quarter century is Actuarial Science. A career as an actuary is consistently highly rated among the most desirable jobs to have. In its 2011 jobs review, *Forbes* magazine rated an actuarial career as being the third-most-desirable job, with a median income of more than \$87,000 and a strong hiring outlook.

Before you go changing your major, it is important to know what an actuary does and what types of skills are required to become one. An actuary is effectively a risk analyst, which means that actuaries assess the likelihood of specific types of events and their financial ramifications. Often these events are considered to be undesirable events. Although actuaries can be employed in any field that has inherent risk, most people who have heard about actuaries associate them with insurance companies and pension management firms.

As mentioned on [www.BeAnActuary.com](http://www.BeAnActuary.com), here are some questions that would arise about societal behavior if actuaries did not exist:

1. Would as many people be willing to own a home if fire insurance did not exist?
2. Would a company build a factory that could be destroyed in an earthquake if it were not protected by insurance?
3. Would people spend money today and still be confident about their future if there were no retirement programs or Social Security?
4. Would automobiles be safe if their parts were not rigorously tested to last for many years using the mathematical techniques that actuaries routinely use?
5. Would parents enjoy risky and adventurous recreational activities such as rock climbing or skiing if their children faced financial disaster in the event of an accident?

Although there are a number of subjects in which an actuary must be well versed, such as finance and certain types of law, every actuary must have a strong background in mathematics, including calculus, linear

algebra, probability, data analysis, and mathematical statistics. To become an actuarial fellow, an applicant must pass a series of nine complex exams, often spending 300 to 400 hours preparing for each exam, each of which often has a passing rate between 25 and 40 percent. These days, students need to pass at least one exam while in college to have a good chance to be hired as an actuary and to

pass the remaining exams while they are on the job. Many larger firms will give employees time off to study and prepare for exams while at work, but actuaries need to spend a good deal of time preparing away from work.

If you are interested in learning more about becoming an actuary, a very good resource can be found at [www.BeAnActuary.com](http://www.BeAnActuary.com).

## Glossary

**Bernoulli trial** One repetition of a binomial experiment. Also called a *trial*.

**Binomial experiment** An experiment that contains  $n$  identical trials such that each of these  $n$  trials has only two possible outcomes (or events), the probabilities of these two outcomes (or events) remain constant for each trial, and the trials are independent.

**Binomial parameters** The total trials  $n$  and the probability of success  $p$  for the binomial probability distribution.

**Binomial probability distribution** The probability distribution that gives the probability of  $x$  successes in  $n$  trials when the probability of success is  $p$  for each trial of a binomial experiment.

**Continuous random variable** A random variable that can assume any value in one or more intervals.

**Discrete random variable** A random variable whose values are countable.

**Hypergeometric probability distribution** The probability distribution that is applied to determine the probability of  $x$  successes in  $n$  trials when the trials are not independent.

**Mean of a discrete random variable** The mean of a discrete random variable  $x$  is the value that is expected to occur per repetition, on average, if an experiment is performed a large number of times. The mean of a discrete random variable is also called its *expected value*.

**Poisson parameter** The average occurrences, denoted by  $\lambda$ , during an interval for a Poisson probability distribution.

**Poisson probability distribution** The probability distribution that gives the probability of  $x$  occurrences in an interval when the average occurrences in that interval are  $\lambda$ .

**Probability distribution of a discrete random variable** A list of all the possible values that a discrete random variable can assume and their corresponding probabilities.

**Random variable** A variable, denoted by  $x$ , whose value is determined by the outcome of a random experiment. Also called a *chance variable*.

**Standard deviation of a discrete random variable** A measure of spread for the probability distribution of a discrete random variable.

## Supplementary Exercises

**5.89** Let  $x$  be the number of cars that a randomly selected auto mechanic repairs on a given day. The following table lists the probability distribution of  $x$ .

$x$	2	3	4	5	6
$P(x)$	.05	.22	.40	.23	.10

Find the mean and standard deviation of  $x$ . Give a brief interpretation of the value of the mean.

**5.90** Let  $x$  be the number of emergency root canal surgeries performed by Dr. Sharp on a given Monday. The following table lists the probability distribution of  $x$ .

$x$	0	1	2	3	4	5
$P(x)$	.13	.28	.30	.17	.08	.04

Calculate the mean and standard deviation of  $x$ . Give a brief interpretation of the value of the mean.

**5.91** Based on its analysis of the future demand for its products, the financial department at Tipper Corporation has determined that there is a .17 probability that the company will lose \$1.2 million during the next year, a .21 probability that it will lose \$.7 million, a .37 probability that it will make a profit of \$.9 million, and a .25 probability that it will make a profit of \$2.3 million.

- a. Let  $x$  be a random variable that denotes the profit earned by this corporation during the next year. Write the probability distribution of  $x$ .

- b. Find the mean and standard deviation of the probability distribution of part a. Give a brief interpretation of the value of the mean.

**5.92** GESCO Insurance Company charges a \$350 premium per annum for a \$100,000 life insurance policy for a 40-year-old female. The probability that a 40-year-old female will die within 1 year is .002.

- a. Let  $x$  be a random variable that denotes the gain of the company for next year from a \$100,000 life insurance policy sold to a 40-year-old female. Write the probability distribution of  $x$ .  
 b. Find the mean and standard deviation of the probability distribution of part a. Give a brief interpretation of the value of the mean.

**5.93** Spoke Weaving Corporation has eight weaving machines of the same kind and of the same age. The probability is .04 that any weaving machine will break down at any time. Find the probability that at any given time

- a. all eight weaving machines will be broken down  
 b. exactly two weaving machines will be broken down  
 c. none of the weaving machines will be broken down

**5.94** At the Bank of California, past data show that 8% of all credit card holders default at some time in their lives. On one recent day, this bank issued 12 credit cards to new customers. Find the probability that of these 12 customers, eventually

- a. exactly 3 will default      b. exactly 1 will default      c. none will default

**5.95** Maine Corporation buys motors for electric fans from another company that guarantees that at most 5% of its motors are defective and that it will replace all defective motors at no cost to Maine Corporation. The motors are received in large shipments. The quality control department at Maine Corporation randomly selects 20 motors from each shipment and inspects them for being good or defective. If this sample contains more than 2 defective motors, the entire shipment is rejected.

- a. Using the appropriate probabilities table from Appendix C, find the probability that a given shipment of motors received by Maine Corporation will be accepted. Assume that 5% of all motors received are defective.  
 b. Using the appropriate probabilities table from Appendix C, find the probability that a given shipment of motors received by Maine Corporation will be rejected.

**5.96** One of the toys made by Dillon Corporation is called Speaking Joe, which is sold only by mail. Consumer satisfaction is one of the top priorities of the company's management. The company guarantees a refund or a replacement for any Speaking Joe toy if the chip that is installed inside becomes defective within 1 year from the date of purchase. It is known from past data that 10% of these chips become defective within a 1-year period. The company sold 15 Speaking Joes on a given day.

- a. Let  $x$  denote the number of Speaking Joes in these 15 that will be returned for a refund or a replacement within a 1-year period. Using the appropriate probabilities table from Appendix C, obtain the probability distribution of  $x$  and draw a graph of the probability distribution. Determine the mean and standard deviation of  $x$ .  
 b. Using the probability distribution constructed in part a, find the probability that exactly 5 of the 15 Speaking Joes will be returned for a refund or a replacement within a 1-year period.

**5.97** In a list of 15 households, 9 own homes and 6 do not own homes. Four households are randomly selected from these 15 households. Find the probability that the number of households in these 4 who own homes is

- a. exactly 3      b. at most 1      c. exactly 4

**5.98** Twenty corporations were asked whether or not they provide retirement benefits to their employees. Fourteen of the corporations said they do provide retirement benefits to their employees, and 6 said they do not. Five corporations are randomly selected from these 20. Find the probability that

- a. exactly 2 of them provide retirement benefits to their employees.  
 b. none of them provides retirement benefits to their employees.  
 c. at most one of them provides retirement benefits to employees.

**5.99** Uniroyal Electronics Company buys certain parts for its refrigerators from Bob's Corporation. The parts are received in shipments of 400 boxes, each box containing 16 parts. The quality control department at Uniroyal Electronics first randomly selects 1 box from each shipment and then randomly selects 4 parts from that box. The shipment is accepted if at most 1 of the 4 parts is defective. The quality control inspector at Uniroyal Electronics selected a box from a recently received shipment of such parts. Unknown to the inspector, this box contains 3 defective parts.

- a. What is the probability that this shipment will be accepted?  
 b. What is the probability that this shipment will not be accepted?

**5.100** Alison Bender works for an accounting firm. To make sure her work does not contain errors, her manager randomly checks on her work. Alison recently filled out 12 income tax returns for the company's clients. Unknown to anyone, 2 of these 12 returns have minor errors. Alison's manager randomly selects 3 returns from these 12 returns. Find the probability that

- a. exactly 1 of them contains errors.
- b. none of them contains errors.
- c. exactly 2 of them contain errors.

**5.101** The student health center at a university treats an average of seven cases of mononucleosis per day during the week of final examinations.

- a. Using the appropriate formula, find the probability that on a given day during the finals week exactly four cases of mononucleosis will be treated at this health center.
- b. Using the appropriate probabilities table from Appendix C, find the probability that on a given day during the finals week the number of cases of mononucleosis treated at this health center will be
  - i. at least 7
  - ii. at most 3
  - iii. 2 to 5

**5.102** An average of 6.3 robberies occur per day in a large city.

- a. Using the Poisson formula, find the probability that on a given day exactly 3 robberies will occur in this city.
- b. Using the appropriate probabilities table from Appendix C, find the probability that on a given day the number of robberies that will occur in this city is
  - i. at least 12
  - ii. at most 3
  - iii. 2 to 6

**5.103** An average of 1.4 private airplanes arrive per hour at an airport.

- a. Find the probability that during a given hour no private airplane will arrive at this airport.
- b. Let  $x$  denote the number of private airplanes that will arrive at this airport during a given hour. Write the probability distribution of  $x$ . Use the appropriate probabilities table from Appendix C.

**5.104** A high school boys' basketball team averages 1.2 technical fouls per game.

- a. Using the appropriate formula, find the probability that in a given basketball game this team will commit exactly 3 technical fouls.
- b. Let  $x$  denote the number of technical fouls that this team will commit during a given basketball game. Using the appropriate probabilities table from Appendix C, write the probability distribution of  $x$ .

## Advanced Exercises

**5.105** Scott offers you the following game: You will roll two fair dice. If the sum of the two numbers obtained is 2, 3, 4, 9, 10, 11, or 12, Scott will pay you \$20. However, if the sum of the two numbers is 5, 6, 7, or 8, you will pay Scott \$20. Scott points out that you have seven winning numbers and only four losing numbers. Is this game fair to you? Should you accept this offer? Support your conclusion with appropriate calculations.

**5.106** Suppose the owner of a salvage company is considering raising a sunken ship. If successful, the venture will yield a net profit of \$10 million. Otherwise, the owner will lose \$4 million. Let  $p$  denote the probability of success for this venture. Assume the owner is willing to take the risk to go ahead with this project provided the expected net profit is at least \$500,000.

- a. If  $p = .40$ , find the expected net profit. Will the owner be willing to take the risk with this probability of success?
- b. What is the smallest value of  $p$  for which the owner will take the risk to undertake this project?

**5.107** Two teams, A and B, will play a best-of-seven series, which will end as soon as one of the teams wins four games. Thus, the series may end in four, five, six, or seven games. Assume that each team has an equal chance of winning each game and that all games are independent of one another. Find the following probabilities.

- a. Team A wins the series in four games.
- b. Team A wins the series in five games.
- c. Seven games are required for a team to win the series.

**5.108** York Steel Corporation produces a special bearing that must meet rigid specifications. When the production process is running properly, 10% of the bearings fail to meet the required specifications. Sometimes problems develop with the production process that cause the rejection rate to exceed 10%. To guard against this higher rejection rate, samples of 15 bearings are taken periodically and carefully inspected. If

more than 2 bearings in a sample of 15 fail to meet the required specifications, production is suspended for necessary adjustments.

- If the true rate of rejection is 10% (that is, the production process is working properly), what is the probability that the production will be suspended based on a sample of 15 bearings?
- What assumptions did you make in part a?

**5.109** Residents in an inner-city area are concerned about drug dealers entering their neighborhood. Over the past 14 nights, they have taken turns watching the street from a darkened apartment. Drug deals seem to take place randomly at various times and locations on the street and average about three per night. The residents of this street contacted the local police, who informed them that they do not have sufficient resources to set up surveillance. The police suggested videotaping the activity on the street, and if the residents are able to capture five or more drug deals on tape, the police will take action. Unfortunately, none of the residents on this street owns a video camera and, hence, they would have to rent the equipment. Inquiries at the local dealers indicated that the best available rate for renting a video camera is \$75 for the first night and \$40 for each additional night. To obtain this rate, the residents must sign up in advance for a specified number of nights. The residents hold a neighborhood meeting and invite you to help them decide on the length of the rental period. Because it is difficult for them to pay the rental fees, they want to know the probability of taping at least five drug deals on a given number of nights of videotaping.

- Which of the probability distributions you have studied might be helpful here?
- What assumption(s) would you have to make?
- If the residents tape for two nights, what is the probability they will film at least five drug deals?
- For how many nights must the camera be rented so that there is at least .90 probability that five or more drug deals will be taped?

**5.110** A high school history teacher gives a 50-question multiple-choice examination in which each question has four choices. The scoring includes a penalty for guessing. Each correct answer is worth 1 point, and each wrong answer costs 1/2 point. For example, if a student answers 35 questions correctly, 8 questions incorrectly, and does not answer 7 questions, the total score for this student will be  $35 - (1/2)(8) = 31$ .

- What is the expected score of a student who answers 38 questions correctly and guesses on the other 12 questions? Assume that the student randomly chooses one of the four answers for each of the 12 guessed questions.
- Does a student increase his expected score by guessing on a question if he has no idea what the correct answer is? Explain.
- Does a student increase her expected score by guessing on a question for which she can eliminate one of the wrong answers? Explain.

**5.111** A baker who makes fresh cheesecakes daily sells an average of five such cakes per day. How many cheesecakes should he make each day so that the probability of running out and losing one or more sales is less than .10? Assume that the number of cheesecakes sold each day follows a Poisson probability distribution. You may use the Poisson probabilities table from Appendix C.

**5.112** Suppose that a certain casino has the “money wheel” game. The money wheel is divided into 50 sections, and the wheel has an equal probability of stopping on each of the 50 sections when it is spun. Twenty-two of the sections on this wheel show a \$1 bill, 14 show a \$2 bill, 7 show a \$5 bill, 3 show a \$10 bill, 2 show a \$20 bill, 1 shows a flag, and 1 shows a joker. A gambler may place a bet on any of the seven possible outcomes. If the wheel stops on the outcome that the gambler bet on, he or she wins. The net payoffs for these outcomes for \$1 bets are as follows.

Symbol bet on	\$1	\$2	\$5	\$10	\$20	Flag	Joker
Payoff (dollars)	1	2	5	10	20	40	40

- If the gambler bets on the \$1 outcome, what is the expected net payoff?
- Calculate the expected net payoffs for each of the other six outcomes.
- Which bet(s) is (are) best in terms of expected net payoff? Which is (are) worst?

**5.113** A history teacher has given her class a list of seven essay questions to study before the next test. The teacher announced that she will choose four of the seven questions to give on the test, and each student will have to answer three of those four questions.

- In how many ways can the teacher choose four questions from the set of seven?
- Suppose that a student has enough time to study only five questions. In how many ways can the teacher choose four questions from the set of seven so that the four selected questions include both questions that the student did not study?

- c. What is the probability that the student in part b will have to answer a question that he or she did not study? That is, what is the probability that the four questions on the test will include both questions that the student did not study?

**5.114** Consider the following three games. Which one would you be most likely to play? Which one would you be least likely to play? Explain your answer mathematically.

- Game I: You toss a fair coin once. If a head appears you receive \$3, but if a tail appears you have to pay \$1.
- Game II: You buy a single ticket for a raffle that has a total of 500 tickets. Two tickets are chosen without replacement from the 500. The holder of the first ticket selected receives \$300, and the holder of the second ticket selected receives \$150.
- Game III: You toss a fair coin once. If a head appears you receive \$1,000,002, but if a tail appears you have to pay \$1,000,000.

**5.115** Brad Henry is a stone products salesman. Let  $x$  be the number of contacts he visits on a particular day. The following table gives the probability distribution of  $x$ .

$x$	1	2	3	4
$P(x)$	.12	.25	.56	.07

Let  $y$  be the total number of contacts Brad visits on two randomly selected days. Write the probability distribution for  $y$ .

**5.116** The number of calls that come into a small mail-order company follows a Poisson distribution. Currently, these calls are serviced by a single operator. The manager knows from past experience that an additional operator will be needed if the rate of calls exceeds 20 per hour. The manager observes that 9 calls came into the mail-order company during a randomly selected 15-minute period.

- a. If the rate of calls is actually 20 per hour, what is the probability that 9 or more calls will come in during a given 15-minute period?
- b. If the rate of calls is really 30 per hour, what is the probability that 9 or more calls will come in during a given 15-minute period?
- c. Based on the calculations in parts a and b, do you think that the rate of incoming calls is more likely to be 20 or 30 per hour?
- d. Would you advise the manager to hire a second operator? Explain.

**5.117** Many of you probably played the game “Rock, Paper, Scissors” as a child. Consider the following variation of that game. Instead of two players, suppose three players play this game, and let us call these players A, B, and C. Each player selects one of these three items—Rock, Paper, or Scissors—-independent of each other. Player A will win the game if all three players select the same item, for example, rock. Player B will win the game if exactly two of the three players select the same item and the third player selects a different item. Player C will win the game if every player selects a different item. If Player B wins the game, he or she will be paid \$1. If Player C wins the game, he or she will be paid \$3. Assuming that the expected winnings should be the same for each player to make this a fair game, how much should Player A be paid if he or she wins the game?

**5.118** Customers arrive at the checkout counter of a supermarket at an average rate of 10 per hour, and these arrivals follow a Poisson distribution. Using each of the following two methods, find the probability that exactly 4 customers will arrive at this checkout counter during a 2-hour period.

- a. Use the arrivals in each of the two nonoverlapping 1-hour periods and then add these. (Note that the numbers of arrivals in two nonoverlapping periods are independent of each other.)
- b. Use the arrivals in a single 2-hour period.

**5.119** Consider the Uses and Misuses section in this chapter on poker, where we learned how to calculate the probabilities of specific poker hands. Find the probability of being dealt

- a. three of a kind    b. two pairs    c. one pair

## Self-Review Test

- Briefly explain the meaning of a random variable, a discrete random variable, and a continuous random variable. Give one example each of a discrete and a continuous random variable.
- What name is given to a table that lists all of the values that a discrete random variable  $x$  can assume and their corresponding probabilities?

3. For the probability distribution of a discrete random variable, the probability of any single value of  $x$  is always
  - a. in the range 0 to 1
  - b. 1.0
  - c. less than zero
4. For the probability distribution of a discrete random variable, the sum of the probabilities of all possible values of  $x$  is always
  - a. greater than 1
  - b. 1.0
  - c. less than 1.0
5. State the four conditions of a binomial experiment. Give one example of such an experiment.
6. The parameters of the binomial probability distribution are
  - a.  $n, p$ , and  $q$
  - b.  $n$  and  $p$
  - c.  $n, p$ , and  $x$
7. The mean and standard deviation of a binomial probability distribution with  $n = 25$  and  $p = .20$  are
  - a. 5 and 2
  - b. 8 and 4
  - c. 4 and 3
8. The binomial probability distribution is symmetric if
  - a.  $p < .5$
  - b.  $p = .5$
  - c.  $p > .5$
9. The binomial probability distribution is skewed to the right if
  - a.  $p < .5$
  - b.  $p = .5$
  - c.  $p > .5$
10. The binomial probability distribution is skewed to the left if
  - a.  $p < .5$
  - b.  $p = .5$
  - c.  $p > .5$
11. Briefly explain when a hypergeometric probability distribution is used. Give one example of a hypergeometric probability distribution.
12. The parameter/parameters of the Poisson probability distribution is/are
  - a.  $\lambda$
  - b.  $\lambda$  and  $x$
  - c.  $\lambda$  and  $e$
13. Describe the three conditions that must be satisfied to apply the Poisson probability distribution.
14. Let  $x$  be the number of homes sold per week by all four real estate agents who work at a realty office. The following table lists the probability distribution of  $x$ .

$x$	0	1	2	3	4	5
$P(x)$	.15	.24	.29	.14	.10	.08

Calculate the mean and standard deviation of  $x$ . Give a brief interpretation of the value of the mean.

15. According to a survey, 60% of adults believe that all college students should be required to perform a specified number of hours of community service to graduate. Assume that this percentage is true for the current population of all adults.
  - a. Find the probability that the number of adults in a random sample of 12 who hold this view is
    - i. exactly 8 (use the appropriate formula)
    - ii. at least 6 (use the appropriate table from Appendix C)
    - iii. less than 4 (use the appropriate table from Appendix C)
  - b. Let  $x$  be the number of adults in a random sample of 12 who believe that all college students should be required to perform a specified number of hours of community service to graduate. Using the appropriate table from Appendix C, write the probability distribution of  $x$ . Find the mean and standard deviation of  $x$ .
16. The Red Cross honors and recognizes its best volunteers from time to time. One of the Red Cross offices has received 12 nominations for the next group of 4 volunteers to be recognized. Eight of these 12 nominated volunteers are female. If the Red Cross office decides to randomly select 4 names out of these 12 nominated volunteers, find the probability that of these 4 volunteers
  - a. exactly 3 are female.
  - b. exactly 1 is female.
  - c. at most 1 is female.
17. The police department in a large city has installed a traffic camera at a busy intersection. Any car that runs a red light will be photographed with its license plate visible, and the driver will receive a citation. Suppose that during the morning rush hour of weekdays, an average of 10 drivers are caught running the red light per day by this system.
  - a. Find the probability that during the morning rush hour on a given weekday this system will catch
    - i. exactly 14 drivers (use the appropriate formula)
    - ii. at most 7 drivers (use the appropriate table from Appendix C)
    - iii. 13 to 18 drivers (use the appropriate table from Appendix C)

- b. Let  $x$  be the number of drivers caught by this system during the morning rush hour on a given weekday. Write the probability distribution of  $x$ . Use the appropriate table from Appendix C.
18. The binomial probability distribution is symmetric when  $p = .50$ , it is skewed to the right when  $p < .50$ , and it is skewed to the left when  $p > .50$ . Illustrate these three cases by writing three probability distributions and graphing them. Choose any values of  $n$  (4 or higher) and  $p$  and use the table of binomial probabilities (Table I of Appendix C).

## Mini-Projects

### MINI-PROJECT 5-1

Consider the NFL data (Data Set III) given on the Web site of this text.

- a. What is the proportion of these players who have 10 or more years of playing experience in the NFL?
- b. Suppose a random sample of 22 of these NFL players is taken, and  $x$  is the number of players in the sample who have 10 or more years of playing experience in the NFL. Find  $P(x = 0)$ ,  $P(x = 1)$ ,  $P(x = 2), \dots$ , through  $P(x = 8)$ .
- c. Note that  $x$  in part b has a binomial distribution with  $\mu = np$ . Use the Poisson probabilities table of Appendix C to approximate the probabilities that you calculated in part b.
- d. Are the probabilities of parts b and c consistent, or is the Poisson approximation poor? Explain why.

### MINI-PROJECT 5-2

Obtain information on the odds and payoffs of one of the instant lottery games in your state or a nearby state. Let the random variable  $x$  be the net amount won on one ticket (payoffs minus purchase price). Using the concepts presented in this chapter, find the probability distribution of  $x$ . Then calculate the mean and standard deviation of  $x$ . What is the player's average net gain (or loss) per ticket purchased?

### MINI-PROJECT 5-3

For this project, first collect data by doing the following. Select an intersection in your town that is controlled by traffic light. For a specific time period (e.g., 9–10 A.M. or 5–6 P.M.), count the number of cars that arrive at that intersection from any one direction during each light cycle. Make sure that you do not count a car twice if it has to sit through two red lights before getting through the intersection. Perform the following tasks with your data.

- a. Create a graphical display of your data. Describe the shape of the distribution. Also discuss which of the following graphs is more useful for displaying the data you collected: a dotplot, a bar graph, or a histogram.
- b. Calculate the mean and variance for your data for light cycles. Note that your sample size is the number of light cycles you observed. Do you notice a relationship between these two summary measures? If so, explain what it is.
- c. For each unique number of arrivals in your data, calculate the proportion of light cycles that had that number of arrivals. For example, suppose you collected these data for 100 light cycles, and you observed 8 cars arriving for each of 12 light cycles. Then,  $12/100 = .12$  of the light cycles had 8 arrivals. Also calculate the theoretical probabilities for each number of arrivals using the Poisson distribution with  $\lambda$  equal to the sample mean that you obtained in part b. How do the two sets of probabilities compare? Is the Poisson a satisfactory model for your data?

## DECIDE FOR YOURSELF

## DECIDING ABOUT INVESTING

If you are a traditional college student, it is quite likely that your financial portfolio includes a checking account and, possibly, a savings account. However, before you know it, you will graduate from college and take a job. On your first day of work, you will have a meeting with your personnel/human resource manager to discuss, among other things, your retirement plans. You may decide to invest a portion of your earnings in a variety of accounts (usually mutual funds) with the hope that you will have enough money to carry you

through your golden years. But wait—How does one decide which mutual funds to invest in? Moreover, how does this relate to the concepts of expected value and variance?

The following table lists the top 10 (as of May 30, 2009) mid-cap growth mutual funds based on the 5-year average return (*Source: <http://biz.yahoo.com/p/tops/mg.html>*). The table also lists the standard deviations of the annual returns for these funds.

By looking at and analyzing the annual returns and the standard deviations of the annual returns for the mutual funds listed in the table, some questions arise that you should try to answer.

- If you decide to invest in a mutual fund based solely on these average annual returns, which fund would you invest in and why? Is this a wise decision?
- The Integrity Williston Bsn/Md-N Amer Stk A fund has the highest average annual return over the 5-year period as shown in the table. Does this imply that the fund is still doing better than all of the other funds listed in the table? Why or why not? Do you think this fund will continue to do better than other funds in the future?
- By considering both the average annual return and the standard deviation of the annual returns, why might a person choose to invest

in the Brown Capital Mgmt Mid-Cap Instl fund over the Integrity Williston Bsn/Md-N Amer Stk A fund, even though the average annual return is more than 10 percent lower for the Brown Capital Mgmt Mid-Cap Instl fund?

- Which of these funds would you invest in and why?
- People who are in their 20s and 30s can afford to take more risks with their investment portfolios because they have plenty of time to offset short-term losses. However, people who are closer to retirement age are less likely to take such high risks. Assuming that the future behavior of the mutual funds is comparable to that of the past 5 years, which of the mutual funds listed in the table would be better to invest in if you are in your 20s or 30s and why? What if you are close to retiring?

Fund Name	Symbol	Annual Return (%)	Standard Deviation (%)
Integrity Williston Bsn/Md-N Amer Stk A	ICPAX	10.99	20.91
Needham Aggressive Growth	NEAGX	10.95	19.55
Delaware Pooled Focus Smid-Cap Gr Eq	DCGTX	10.95	22.37
Westcore Select	WTSIX	10.15	20.87
Eaton Vance Atlanta Capital SMID-Cap I	EISMX	9.83	19.72
Brown Capital Mgmt Mid-Cap Instl	BCMSX	9.82	18.49
American Century Heritage Inst	ATHIX	9.68	22.19
Eaton Vance Atlanta Capital SMID-Cap A	EAASX	9.58	24.19
American Century Heritage Inv	TWHIX	9.46	22.17
Eaton Vance Atlanta Capital SMID-Cap R	ERSMX	9.27	19.68

## TECHNOLOGY INSTRUCTION

### Binomial Distribution, Hypergeometric Distribution, and Poisson Distribution

#### TI-84

- To find the number of ways to choose  $x$  objects out of  $n$ , type  $n$ , select **MATH** > **PRB** > **nCr**, and then type  $x$  and press **ENTER**. To find the probability of  $x$  successes in  $n$  trials for a population with  $N$  elements and  $r$  successes using the hypergeometric probability distribution, you will need to use the **nCr** function three times. For example, if a population has  $N = 25$  elements and  $r = 20$  successes, then to calculate the probability of  $x = 3$  successes in  $n = 4$  trials, the calculator entry would appear as  $(20 \text{ nCr } 3)*(5 \text{ nCr } 1)/(25 \text{ nCr } 4)$ . (See Screen 5.1.)
- To find the probability of  $x$  successes in  $n$  trials using the binomial probability distribution with  $p$  as the probability of success, select **DISTR>binompdf**. In the **binompdf** menu, enter  $n$  at the **trials:** prompt,  $p$  at the **p:** prompt,  $x$  at the **x value:** prompt, and then highlight **Paste** and press **ENTER** twice. To find the cumulative probability of  $x$  or fewer successes in  $n$  trials using the binomial probability distribution with  $p$  as the probability of success, select **DISTR>binomcdf**. (See Screens 5.2 and 5.3.)

```
(20 nCr 3)*(5 nCr 1)/(25 nCr 4)
.4505928854
```

Screen 5.1

```
binompdf
trials:10
P:.3
x value:3
Paste
```

Screen 5.2

```
binompdf(10,.3,3)
.266827932
```

Screen 5.3

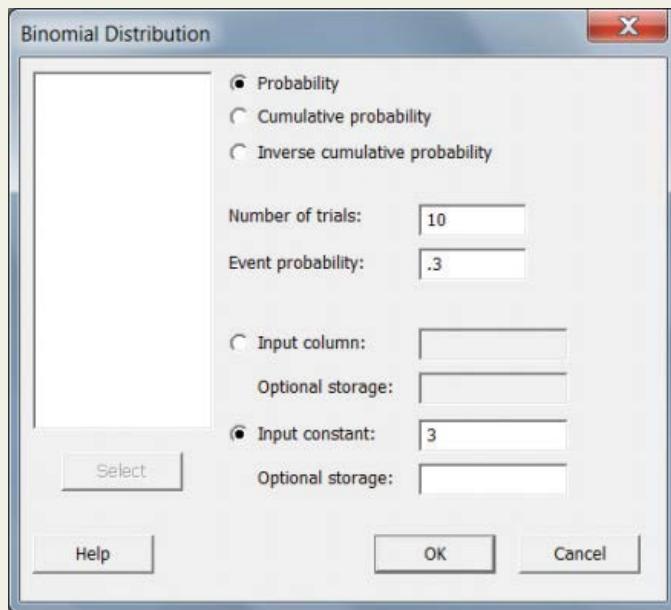
```
poissonpdf
lambda:3.2
x value:5
Paste
```

Screen 5.4

3. To find the probability of  $x$  occurrences in a Poisson probability distribution with a mean of  $\lambda$ , select **DISTR>poissonpdf( $\lambda$ , $x$ )** and press **ENTER**. In the **poissonpdf(** menu, enter  $\lambda$  at the  **$\lambda:$**  prompt,  $x$  at the  **$x$  value:** prompt, and then highlight **Paste** and press **ENTER** twice. To find the cumulative probability of  $x$  or fewer occurrences in a Poisson probability distribution with a mean of  $\lambda$ , select **DISTR>poissoncdf( $\lambda$ , $x$ )** and press **ENTER**. (See Screen 5.4.)

## Minitab

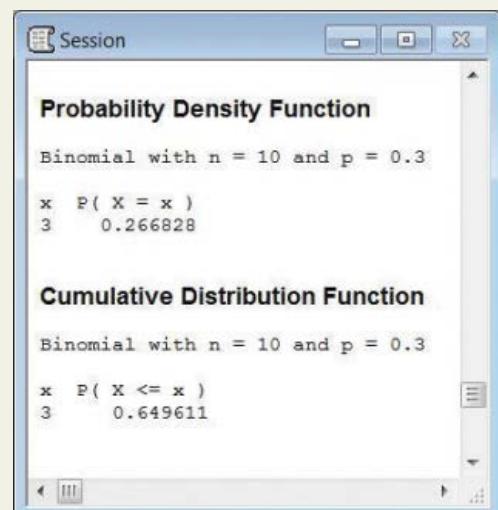
1. To find the probability of  $x$  successes in  $n$  trials for a binomial random variable with probability of success  $p$ , select **Calc >Probability Distributions >Binomial**. In the dialog box, make sure that **Probability** is selected, then enter the number of trials  $n$ , as well as the probability  $p$  of success. Select **Input constant**, and enter the value of  $x$ .



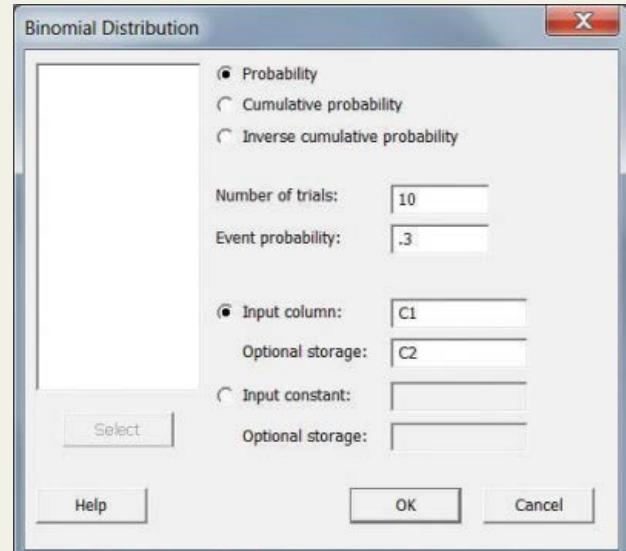
Screen 5.5

To find the probability of  $x$  or fewer successes in  $n$  trials, use the aforementioned process, but click next to **Cumulative probability** instead of **Probability** in the dialog box. (See Screens 5.5 and 5.6.)

If you need to create a table of probabilities or cumulative probabilities for various values of  $x$ , first enter the values of  $x$  into a column in worksheet. Then select **Calc > Probability Distributions > Binomial**, enter the values of  $n$  and  $p$  in the dialog box, and click next to **Probability** or **Cumulative probability** (whichever is relevant). Now select **Input column**, and enter the name of the column where you entered the desired values of  $x$ . If you wish to store the resulting probabilities, enter the name of a column under **Optional storage**. (See Screen 5.7 and Columns C1 and C2 of Screen 5.8.)



Screen 5.6

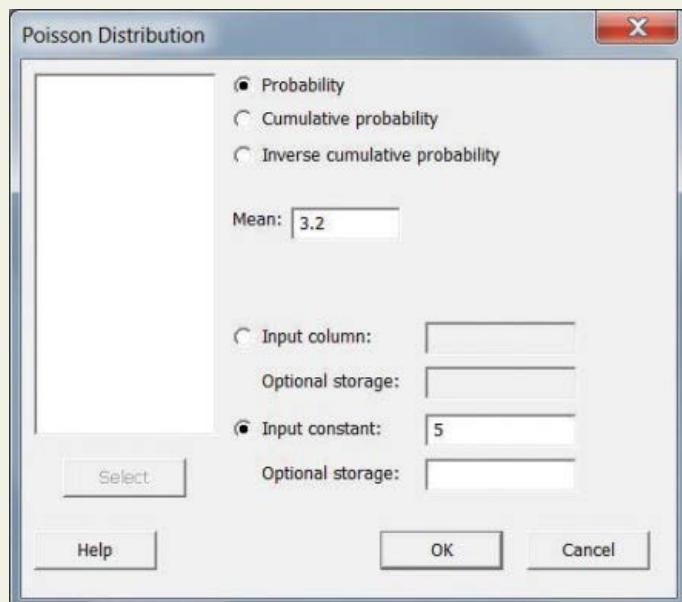


Screen 5.7

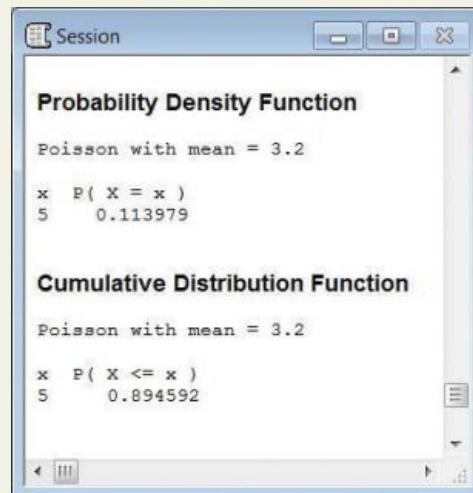
	C1	C2	C3	C4	C5	C6
	binom. x	binom. prob	poisson x	poisson prob	hypgeom x	hypgeom prob
1	0	0.028248	0	0.040762	0	0.000023
2	1	0.121061	1	0.130439	1	0.001828
3	2	0.233474	2	0.208702	2	0.028795
4	3	0.266828	3	0.222616	3	0.153572
5	4	0.200121	4	0.178093	4	0.335939
6	5	0.102919	5	0.113979	5	0.322501
7	6	0.036757	6	0.060789	6	0.134375
8	7	0.009002	7	0.027789	7	0.021939
9	8	0.001447	8	0.011116	8	0.001028
10	9	0.000138	9	0.003952		
11	10	0.000006	10	0.001265		
12			11	0.000368		
13			12	0.000098		

Screen 5.8

2. To find the probability of  $x$  for a Poisson random variable, select **Calc >Probability Distributions > Poisson**. In the dialog box, make sure that **Probability** is selected, then enter the value of mean  $\lambda$ . Select **Input constant**, and enter  $x$ . To find the probability of  $x$  or fewer occurrences for a Poisson random variable, use the aforementioned process, but click next to **Cumulative probability** instead of **Probability** in the dialog box. (See Screens 5.9 and 5.10.)



Screen 5.9

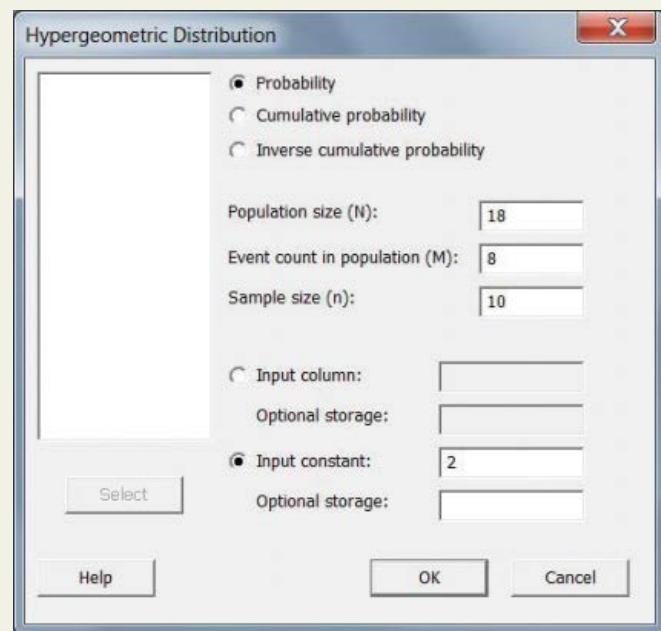


Screen 5.10

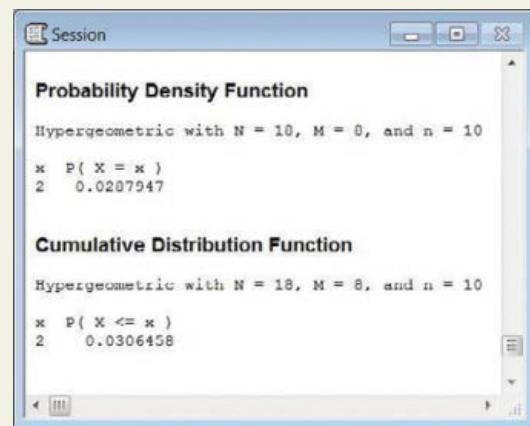
If you need to create a table of probabilities or cumulative probabilities for various values of  $x$ , first enter these values of  $x$  into a column in the worksheet. Then select **Calc > Probability Distributions > Poisson**, click next to **Probability** or **Cumulative probability** (whichever is relevant), and enter the value of  $\lambda$  next to **Mean**. Now select **Input column**, and enter the name of the column where you entered the values of  $x$ . If you wish to store the resulting probabilities, enter the name of a column under **Optional storage**. (See **Columns C3 and C4 of Screen 5.8.**)

- To find the probability of  $x$  successes in  $n$  trials in a population with  $N$  elements and  $r$  successes (denoted  $M$  in Minitab) for a hypergeometric random variable, select **Calc > Probability Distributions > Hypergeometric**. In the dialog box, make sure to click next to **Probability**, then enter the values of  $N$ ,  $r$ , and  $n$  in their respective boxes. Select **Input constant**, and enter the value of  $x$ . To find the probability of  $x$  or fewer occurrences for a Hypergeometric random variable, use the aforementioned process, but click next to **Cumulative probability** instead of **Probability**. (See **Screens 5.11 and 5.12.**.)

If you need to create a table of probabilities or cumulative probabilities for various values of  $x$ , first enter the values of  $x$  into a column in the worksheet. Then select **Calc > Probability Distributions > Hypergeometric**, click next to **Probability** or **Cumulative probability** (whichever is relevant), and enter the values of  $N$ ,  $r$ , and  $n$ . Now select **Input column**, and enter the name of the column where you entered the values of  $x$ . If you wish to store the resulting probabilities, enter the name of a column under **Optional storage**. (See **Columns C5 and C6 of Screen 5.8.**)



Screen 5.11



Screen 5.12

## Excel

- To find the binomial probability of  $x$  successes in  $n$  trials with probability of success  $p$ , type **=BINOM.DIST(x,n,p,0)**. To find the binomial probability of  $x$  or fewer successes in  $n$  trials with probability of success  $p$ , type **=BINOM.DIST(x,n,p,1)**. (See **Screens 5.13 and 5.14.**) (Note: For Excel 2007 and earlier versions, type **BINOMDIST** instead of **BINOM.DIST**.)

	A	B	C	D	E	F
1	Prob. Of 3 successes in a binomial experiment with n=10 and p=.3					
2						
3	=BINOM.DIST(3,10,0.3,0)					
4		BINOM.DIST(number_s, trials, probability_s, cumulative)				
5						

Screen 5.13

	A	B	C	D	E	F
1	Prob. Of 3 successes in a binomial experiment with n=10 and p=.3					
2						
3	0.266828					

Screen 5.14

2. To find the Poisson probability of  $x$  occurrences with a mean of  $\lambda$ , type **=POISSON**.  
 $\text{DIST}(x, \lambda, p, 0)$ . To find the Poisson probability of  $x$  or fewer occurrences with a mean of  $\lambda$ , type **=POISSON.DIST(x, λ, p, 1)**. (Note: For Excel 2007 and earlier versions, type **POISSONDIST** instead of **POISSON.DIST**.)
3. To find the probability of  $x$  successes in  $n$  trials from a population with  $N$  elements and  $r$  successes for a hypergeometric random variable, type **=HYPGEOM.DIST(x, n, r, N, 0)**. To find the probability of  $x$  or fewer successes in  $n$  trials from a population with  $N$  elements and  $r$  successes for a hypergeometric random variable, type **=HYPGEOM.DIST(x, n, r, N, 1)**. (Note: For Excel 2007 and earlier versions, type **HYPGEOMDIST** instead of **HYPGEOM.DIST**.)

## TECHNOLOGY ASSIGNMENTS

**TA5.1** Forty-five percent of the adult population in a particular large city are women. A court is to randomly select a jury of 12 adults from the population of all adults of this city.

- a. Find the probability that none of the 12 jurors is a woman.
  - b. Find the probability that at most 4 of the 12 jurors are women.
  - c. Let  $x$  denote the number of women in 12 adults selected for this jury. Obtain the probability distribution of  $x$ .
  - d. Using the probability distribution obtained in part c, find the following probabilities.
- i.  $P(x > 6)$       ii.  $P(x \leq 3)$       iii.  $P(2 \leq x \leq 7)$

**TA5.2** According to an October 2010 *Consumer Reports* survey, 39% of car owners in the United States are considering a hybrid or a plug-in for their next car purchase ([news.consumerreports.org/cars/2010/10/consumer-reports-shares-preliminary-green-car-survey-findings-at-gridweek-conference.html](http://news.consumerreports.org/cars/2010/10/consumer-reports-shares-preliminary-green-car-survey-findings-at-gridweek-conference.html)).

- a. Find the probability that in a random sample of 70 car owners in the United States, exactly 32 would be considering a hybrid or a plug-in for their next car purchase.
- b. Find the probability that in a random sample of 70 car owners in the United States, 31 or more would be considering a hybrid or a plug-in for their next car purchase.
- c. Find the probability that in a random sample of 700 car owners in the United States, 301 or more would be considering a hybrid or a plug-in for their next car purchase.
- d. A reporter states that he believes that the sample results in parts b and c imply that the percentage of car owners in the United States who are considering a hybrid or a plug-in for their next car purchase is higher than 39%. Using the probabilities from parts b and c, state whether you believe that the reporter's inference is reasonable and explain why.

**TA5.3** A mail-order company receives an average of 40 orders per day.

- a. Find the probability that it will receive exactly 55 orders on a certain day.
- b. Find the probability that it will receive at most 29 orders on a certain day.

- c. Let  $x$  denote the number of orders received by this company on a given day. Obtain the probability distribution of  $x$ .
- d. Using the probability distribution obtained in part c, find the following probabilities.
  - i.  $P(x \geq 45)$
  - ii.  $P(x < 33)$
  - iii.  $P(36 < x < 52)$

**TA5.4** A commuter airline receives an average of 13 complaints per week from its passengers. Let  $x$  denote the number of complaints received by this airline during a given week.

- a. Find  $P(x = 0)$ . If your answer is zero, does it mean that this cannot happen? Explain.
- b. Find  $P(x \leq 10)$ .
- c. Obtain the probability distribution of  $x$ .
- d. Using the probability distribution obtained in part c, find the following probabilities.
  - i.  $P(x > 18)$
  - ii.  $P(x \leq 9)$
  - iii.  $P(10 \leq x \leq 17)$

# CHAPTER 6



© Mauritus/SuperStock

## Continuous Random Variables and the Normal Distribution

### 6.1 Continuous Probability Distribution and the Normal Probability Distribution

#### Case Study 6-1 Distribution of Time Taken to Run a Road Race

### 6.2 Standardizing a Normal Distribution

### 6.3 Applications of the Normal Distribution

### 6.4 Determining the $z$ and $x$ Values When an Area Under the Normal Distribution Curve Is Known

### 6.5 The Normal Approximation to the Binomial Distribution

### Appendix 6-1 Normal Quantile Plots

Have you ever participated in a road race? If you have, where did you stand in comparison to the other runners? Do you think the time taken to finish a road race varies as much among runners as the runners themselves? See Case Study 6-1 for the distribution of times for runners who completed the Beach to Beacon 10K Road Race in 2011.

Discrete random variables and their probability distributions were presented in Chapter 5. Section 5.1 defined a continuous random variable as a variable that can assume any value in one or more intervals.

The possible values that a continuous random variable can assume are infinite and uncountable. For example, the variable that represents the time taken by a worker to commute from home to work is a continuous random variable. Suppose 5 minutes is the minimum time and 130 minutes is the maximum time taken by all workers to commute from home to work. Let  $x$  be a continuous random variable that denotes the time taken to commute from home to work by a randomly selected worker. Then  $x$  can assume any value in the interval 5 to 130 minutes. This interval contains an infinite number of values that are uncountable.

A continuous random variable can possess one of many probability distributions. In this chapter, we discuss the normal probability distribution and the normal distribution as an approximation to the binomial distribution.

## 6.1 Continuous Probability Distribution and the Normal Probability Distribution

In this section we will learn about the continuous probability distribution and its properties and then discuss the normal probability distribution.

### 6.1.1 Continuous Probability Distribution

In Chapter 5, we defined a **continuous random variable** as a random variable whose values are not countable. A continuous random variable can assume any value over an interval or intervals. Because the number of values contained in any interval is infinite, the possible number of values that a continuous random variable can assume is also infinite. Moreover, we cannot count these values. In Chapter 5, it was stated that the life of a battery, heights of people, time taken to complete an examination, amount of milk in a gallon, weights of babies, and prices of houses are all examples of continuous random variables. Note that although money can be counted, variables involving money are often represented by continuous random variables. This is so because a variable involving money often has a very large number of outcomes.

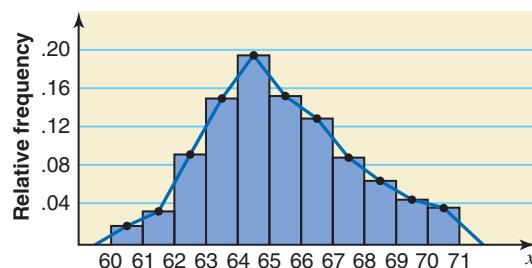
Suppose 5000 female students are enrolled at a university, and  $x$  is the continuous random variable that represents the heights of these female students. Table 6.1 lists the frequency and relative frequency distributions of  $x$ .

**Table 6.1** Frequency and Relative Frequency Distributions of Heights of Female Students

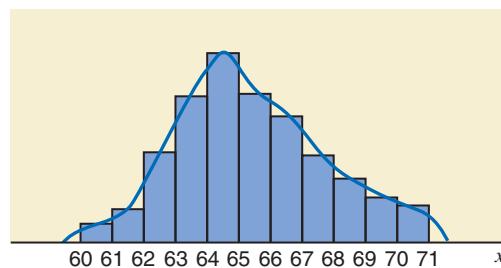
Height of a Female Student (inches)	$f$	Relative Frequency
$x$		
60 to less than 61	90	.018
61 to less than 62	170	.034
62 to less than 63	460	.092
63 to less than 64	750	.150
64 to less than 65	970	.194
65 to less than 66	760	.152
66 to less than 67	640	.128
67 to less than 68	440	.088
68 to less than 69	320	.064
69 to less than 70	220	.044
70 to less than 71	180	.036
$N = 5000$		Sum = 1.0

The relative frequencies given in Table 6.1 can be used as the probabilities of the respective classes. Note that these are exact probabilities because we are considering the population of all female students.

Figure 6.1 displays the histogram and polygon for the relative frequency distribution of Table 6.1. Figure 6.2 shows the smoothed polygon for the data of Table 6.1. The smoothed



**Figure 6.1** Histogram and polygon for Table 6.1.

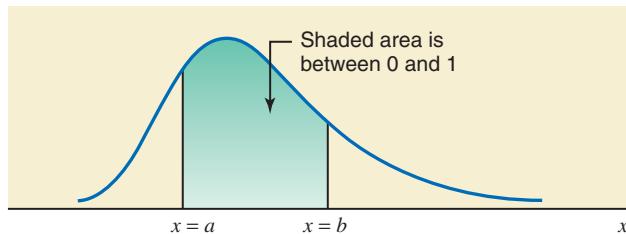
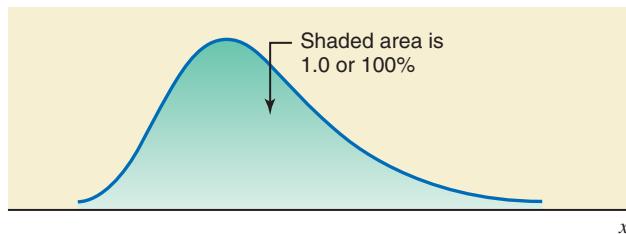
**Figure 6.2** Probability distribution curve for heights.

polygon is an approximation of the *probability distribution curve* of the continuous random variable  $x$ . Note that each class in Table 6.1 has a width equal to 1 inch. If the width of classes is more (or less) than 1 unit, we first obtain the *relative frequency densities* and then graph these relative frequency densities to obtain the distribution curve. The relative frequency density of a class is obtained by dividing the relative frequency of that class by the class width. The relative frequency densities are calculated to make the sum of the areas of all rectangles in the histogram equal to 1.0. Case Study 6–1, which appears later in this section, illustrates this procedure. The probability distribution curve of a continuous random variable is also called its *probability density function*.

The probability distribution of a continuous random variable possesses the following *two characteristics*.

1. The probability that  $x$  assumes a value in any interval lies in the range 0 to 1.
2. The total probability of all the (mutually exclusive) intervals within which  $x$  can assume a value is 1.0.

The first characteristic states that the area under the probability distribution curve of a continuous random variable between any two points is between 0 and 1, as shown in Figure 6.3. The second characteristic indicates that the total area under the probability distribution curve of a continuous random variable is always 1.0, or 100%, as shown in Figure 6.4.

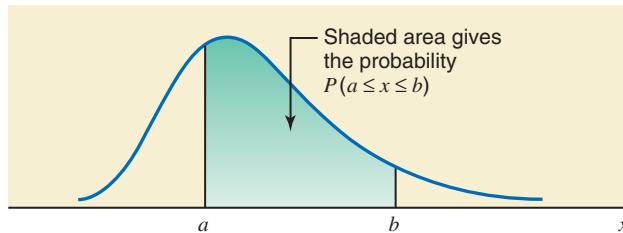
**Figure 6.3** Area under a curve between two points.**Figure 6.4** Total area under a probability distribution curve.

The probability that a continuous random variable  $x$  assumes a value within a certain interval is given by the area under the curve between the two limits of the interval, as shown in

Figure 6.5. The shaded area under the curve from  $a$  to  $b$  in this figure gives the probability that  $x$  falls in the interval  $a$  to  $b$ ; that is,

$$P(a \leq x \leq b) = \text{Area under the curve from } a \text{ to } b$$

Note that the interval  $a \leq x \leq b$  states that  $x$  is greater than or equal to  $a$  but less than or equal to  $b$ .

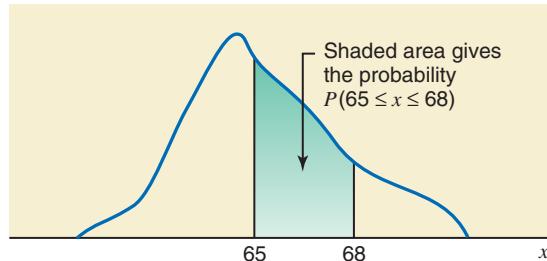


**Figure 6.5** Area under the curve as probability.

Reconsider the example on the heights of all female students at a university. The probability that the height of a randomly selected female student from this university lies in the interval 65 to 68 inches is given by the area under the distribution curve of the heights of all female students from  $x = 65$  to  $x = 68$ , as shown in Figure 6.6. This probability is written as

$$P(65 \leq x \leq 68)$$

which states that  $x$  is greater than or equal to 65 but less than or equal to 68.



**Figure 6.6** Probability that  $x$  lies in the interval 65 to 68.

For a continuous probability distribution, the probability is always calculated for an interval. For example, in Figure 6.6, the interval representing the shaded area is from 65 to 68. Consequently, the shaded area in that figure gives the probability for the interval  $65 \leq x \leq 68$ .

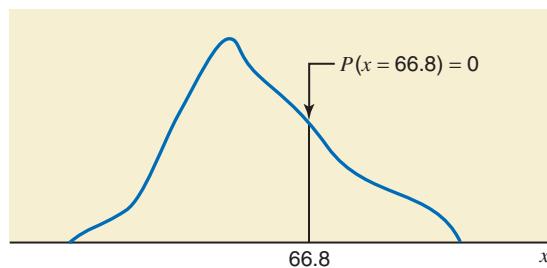
**The probability that a continuous random variable  $x$  assumes a single value is always zero.** This is so because the area of a line, which represents a single point, is zero. For example, if  $x$  is the height of a randomly selected female student from that university, then the probability that this student is exactly 66.8 inches tall is zero; that is,

$$P(x = 66.8) = 0$$

This probability is shown in Figure 6.7. Similarly, the probability for  $x$  to assume any other single value is zero.

In general, if  $a$  and  $b$  are two of the values that  $x$  can assume, then

$$P(a) = 0 \quad \text{and} \quad P(b) = 0$$

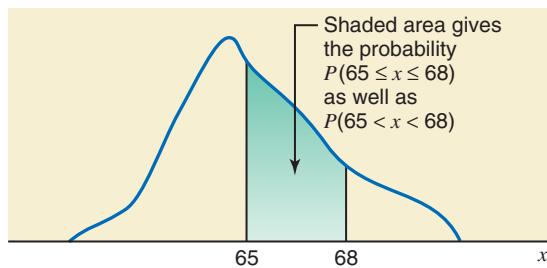


**Figure 6.7** The probability of a single value of  $x$  is zero.

From this we can deduce that for a continuous random variable,

$$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$

In other words, the probability that  $x$  assumes a value in the interval  $a$  to  $b$  is the same whether or not the values  $a$  and  $b$  are included in the interval. For the example on the heights of female students, the probability that a randomly selected female student is between 65 and 68 inches tall is the same as the probability that this female is 65 to 68 inches tall. This is shown in Figure 6.8.



**Figure 6.8** Probability “from 65 to 68” and “between 65 and 68.”

Note that the interval “between 65 and 68” represents “ $65 < x < 68$ ” and it does not include 65 and 68. On the other hand, the interval “from 65 to 68” represents “ $65 \leq x \leq 68$ ” and it does include 65 and 68. However, as mentioned previously, in the case of a continuous random variable, both of these intervals contain the same probability or area under the curve.

Case Study 6–1 on the next page describes how we obtain the probability distribution curve of a continuous random variable.

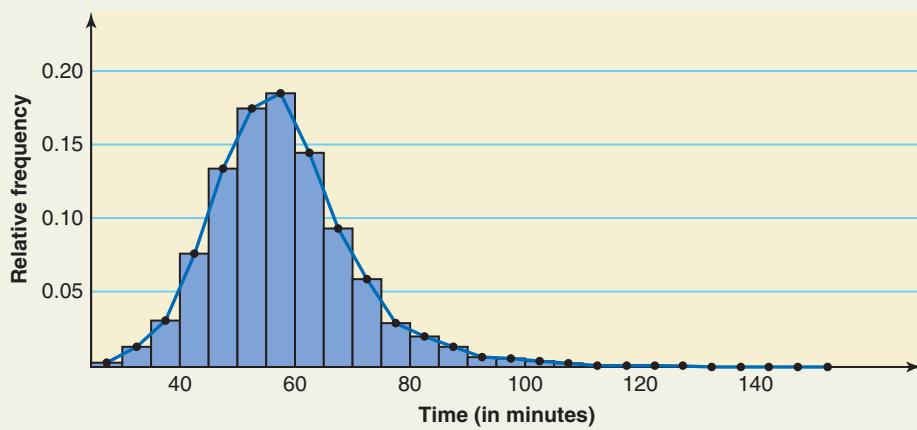
### 6.1.2 The Normal Distribution

The normal distribution is one of the many probability distributions that a continuous random variable can possess. The normal distribution is the most important and most widely used of all probability distributions. A large number of phenomena in the real world are approximately normally distributed. The continuous random variables representing heights and weights of people, scores on an examination, weights of packages (e.g., cereal boxes, boxes of cookies), amount of milk in a gallon, life of an item (such as a light-bulb or a television set), and time taken to complete a certain job have all been observed to have an approximate normal distribution.

## DISTRIBUTION OF TIME TAKEN TO RUN A ROAD RACE

The accompanying table gives the frequency and relative frequency distributions for the time (in minutes) taken to complete the 14th Beach to Beacon 10K Road Race (held on August 6, 2011) by 5875 participants who finished that race. This event is held every year on the first Saturday in August in Cape Elizabeth, Maine. The total distance of the course is 10 kilometers (which is approximately 6.214 miles). The relative frequencies in the table are used to construct the histogram and polygon in Figure 6.9.

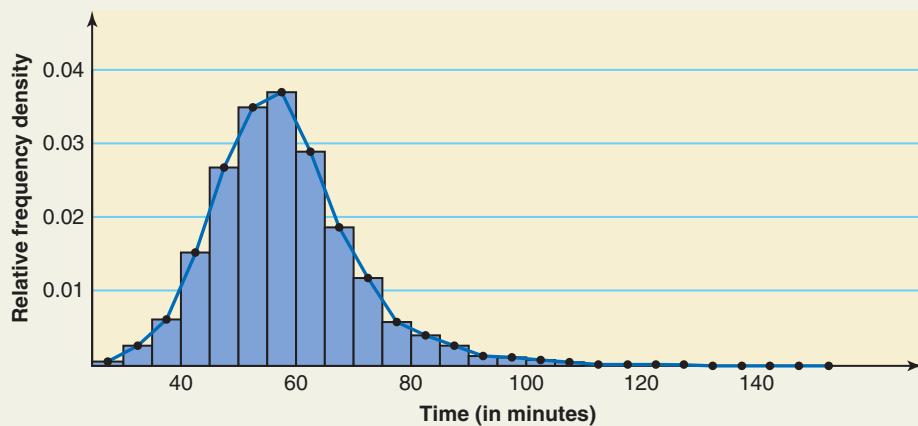
Class	Frequency	Relative Frequency
25 to less than 30	11	.0019
30 to less than 35	76	.0129
35 to less than 40	183	.0311
40 to less than 45	449	.0764
45 to less than 50	787	.1340
50 to less than 55	1030	.1753
55 to less than 60	1088	.1852
60 to less than 65	855	.1455
65 to less than 70	551	.0938
70 to less than 75	347	.0591
75 to less than 80	175	.0298
80 to less than 85	123	.0209
85 to less than 90	76	.0129
90 to less than 95	38	.0065
95 to less than 100	35	.0060
100 to less than 105	22	.0037
105 to less than 110	15	.0026
110 to less than 115	6	.0010
115 to less than 120	3	.0005
120 to less than 125	3	.0005
125 to less than 130	0	.0000
130 to less than 135	1	.0002
135 to less than 140	0	.0000
140 to less than 145	0	.0000
145 to less than 150	0	.0000
150 to less than 155	1	.0002
$\Sigma f = 5875$		Sum = 1.000



**Figure 6.9** Histogram and polygon for the Beach to Beacon 10K Road Race data.

To derive the probability distribution curve for these data, we calculate the relative frequency densities by dividing the relative frequencies by the class widths. The width of each class in the table is 5. By dividing the relative frequencies by 5, we obtain the relative frequency densities, which are recorded in the table on the next page. Using the relative frequency densities, we draw a histogram and smoothed polygon, as shown in Figure 6.10. The smoothed polygon in this figure gives the probability distribution curve for the Beach to Beacon Road Race data.

Class		Relative Frequency Density
25 to less than 30	30	.00038
30 to less than 35	35	.00258
35 to less than 40	40	.00622
40 to less than 45	45	.01528
45 to less than 50	50	.02680
50 to less than 55	55	.03506
55 to less than 60	60	.03704
60 to less than 65	65	.02910
65 to less than 70	70	.01876
70 to less than 75	75	.01182
75 to less than 80	80	.00596
80 to less than 85	85	.00418
85 to less than 90	90	.00258
90 to less than 95	95	.00130
95 to less than 100	100	.00120
100 to less than 105	105	.00074
105 to less than 110	110	.00052
110 to less than 115	115	.00020
115 to less than 120	120	.00010
120 to less than 125	125	.00010
125 to less than 130	130	.00000
130 to less than 135	135	.00004
135 to less than 140	140	.00000
140 to less than 145	145	.00000
145 to less than 150	150	.00000
150 to less than 155	155	.00004

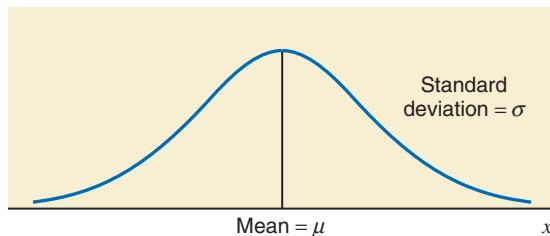


**Figure 6.10** Probability distribution for the Beach to Beacon 10K Road Race Data.

Note that the areas of the rectangles in Figure 6.9 do not give probabilities (which are approximated by relative frequencies). Rather, it is the heights of these rectangles that give the probabilities. This is so because the base of each rectangle is 5 in the histogram. Consequently, the area of each rectangle is given by its height multiplied by 5. Thus, the total area of all the rectangles in Figure 6.9 is 5.0, not 1.0. However, in Figure 6.10, it is the areas, not the heights, of the rectangles that give the probabilities of the respective classes. Thus, if we add the areas of all the rectangles in Figure 6.10, we obtain the sum of all the probabilities equal to 1.0. Consequently, the total area under the curve is equal to 1.0.

The probability distribution of a continuous random variable has a mean and a standard deviation, denoted by  $\mu$  and  $\sigma$ , respectively. The mean and standard deviation of the probability distribution curve of Figure 6.10 are 58.105 and 12.603 minutes, respectively. These values of  $\mu$  and  $\sigma$  are calculated by using the raw data on 5875 participants.

**A normal probability distribution** or a *normal curve* is a bell-shaped (symmetric) curve. Such a curve is shown in Figure 6.11. Its mean is denoted by  $\mu$  and its standard deviation by  $\sigma$ . A continuous random variable  $x$  that has a normal distribution is called a *normal random variable*. Note that not all bell-shaped curves represent a normal distribution curve. Only a specific kind of bell-shaped curve represents a normal curve.



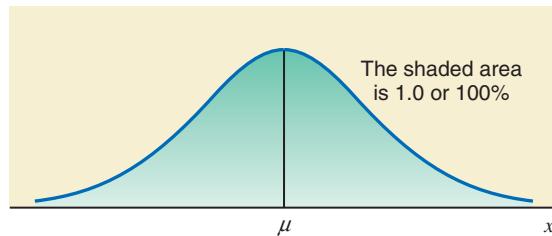
**Figure 6.11** Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

**Normal Probability Distribution** A *normal probability distribution*, when plotted, gives a bell-shaped curve such that:

1. The total area under the curve is 1.0.
2. The curve is symmetric about the mean.
3. The two tails of the curve extend indefinitely.

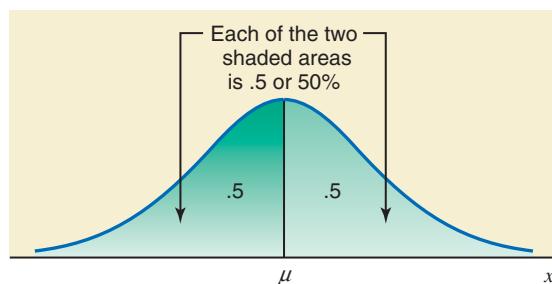
A normal distribution possesses the following three characteristics:

1. The total area under a normal distribution curve is 1.0, or 100%, as shown in Figure 6.12.



**Figure 6.12** Total area under a normal curve.

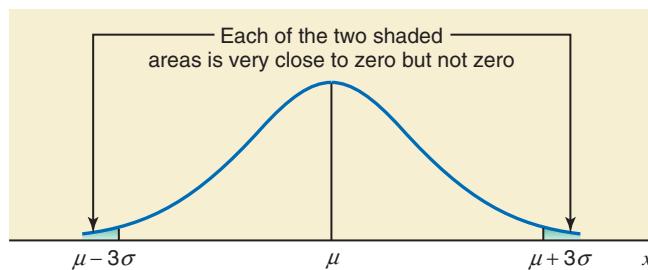
2. A normal distribution curve is symmetric about the mean, as shown in Figure 6.13. Consequently, 50% of the total area under a normal distribution curve lies on the left side of the mean, and 50% lies on the right side of the mean.



**Figure 6.13** A normal curve is symmetric about the mean.

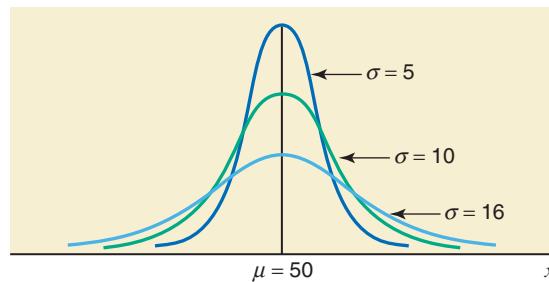
3. The tails of a normal distribution curve extend indefinitely in both directions without touching or crossing the horizontal axis. Although a normal distribution curve never meets the horizontal axis, beyond the points represented by  $\mu - 3\sigma$  and  $\mu + 3\sigma$  it becomes so close to this axis that the area under the curve beyond these points in both directions is very small and can be taken as very close to zero (but not zero). These areas are shown in Figure 6.14.

**Figure 6.14** Areas of the normal curve beyond  $\mu \pm 3\sigma$ .

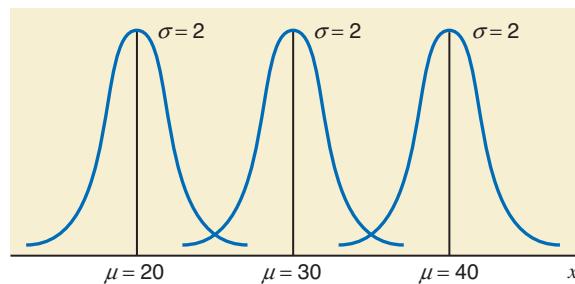


The mean,  $\mu$ , and the standard deviation,  $\sigma$ , are the *parameters* of the normal distribution. Given the values of these two parameters, we can find the area under a normal distribution curve for any interval. Remember, there is not just one normal distribution curve but a *family* of normal distribution curves. Each different set of values of  $\mu$  and  $\sigma$  gives a different normal distribution. The value of  $\mu$  determines the center of a normal distribution curve on the horizontal axis, and the value of  $\sigma$  gives the spread of the normal distribution curve. The three normal distribution curves shown in Figure 6.15 have the same mean but different standard deviations. By contrast, the three normal distribution curves in Figure 6.16 have different means but the same standard deviation.

**Figure 6.15** Three normal distribution curves with the same mean but different standard deviations.



**Figure 6.16** Three normal distribution curves with different means but the same standard deviation.



Like the binomial and Poisson probability distributions discussed in Chapter 5, the normal probability distribution can also be expressed by a mathematical equation.<sup>1</sup> However, we will not use this equation to find the area under a normal distribution curve. Instead, we will use Table IV of Appendix C.

<sup>1</sup>The equation of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2}$$

where  $e = 2.71828$  and  $\pi = 3.14159$  approximately;  $f(x)$ , called the probability density function, gives the vertical distance between the horizontal axis and the curve at point  $x$ . For the information of those who are familiar with integral calculus, the definite integral of this equation from  $a$  to  $b$  gives the probability that  $x$  assumes a value between  $a$  and  $b$ .

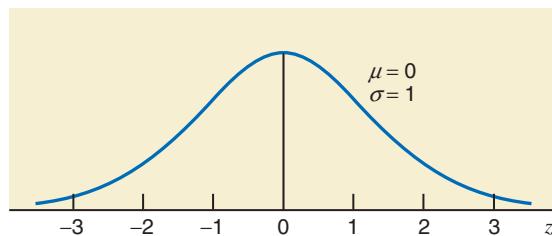
### 6.1.3 The Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution. For the standard normal distribution, the value of the mean is equal to zero, and the value of the standard deviation is equal to 1.

#### Definition

**Standard Normal Distribution** The normal distribution with  $\mu = 0$  and  $\sigma = 1$  is called the *standard normal distribution*.

Figure 6.17 displays the standard normal distribution curve. The random variable that possesses the standard normal distribution is denoted by  $z$ . In other words, the units for the standard normal distribution curve are denoted by  $z$  and are called the ***z values*** or ***z scores***. They are also called *standard units* or *standard scores*.



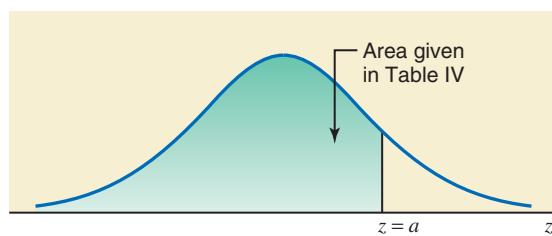
**Figure 6.17** The standard normal distribution curve.

#### Definition

***z Values or z Scores*** The units marked on the horizontal axis of the standard normal curve are denoted by  $z$  and are called the *z values* or *z scores*. A specific value of  $z$  gives the distance between the mean and the point represented by  $z$  in terms of the standard deviation.

In Figure 6.17, the horizontal axis is labeled  $z$ . The  $z$  values on the right side of the mean are positive and those on the left side are negative. *The z value for a point on the horizontal axis gives the distance between the mean and that point in terms of the standard deviation.* For example, a point with a value of  $z = 2$  is two standard deviations to the right of the mean. Similarly, a point with a value of  $z = -2$  is two standard deviations to the left of the mean.

The standard normal distribution table, Table IV of Appendix C, lists the areas under the standard normal curve to the left of  $z$  values from  $-3.49$  to  $3.49$ . To read the standard normal distribution table, we look for the given  $z$  value in the table and record the value corresponding to that  $z$  value. As shown in Figure 6.18, Table IV gives what is called the cumulative probability to the left of any  $z$  value.



**Figure 6.18** Area under the standard normal curve.

**Remember ►** Although the values of  $z$  on the left side of the mean are negative, the area under the curve is always positive.

The area under the standard normal curve between any two points can be interpreted as the probability that  $z$  assumes a value within that interval. Examples 6–1 through 6–4 describe how to read Table IV of Appendix C to find areas under the standard normal curve.

### ■ EXAMPLE 6–1

Finding the area to the left of a positive  $z$ .

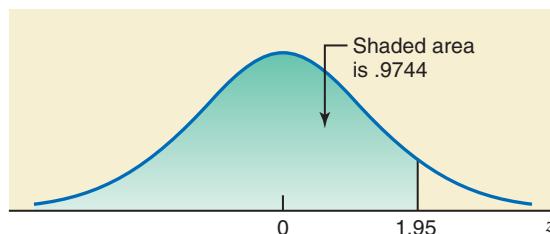
Find the area under the standard normal curve to the left of  $z = 1.95$ .

**Solution** We divide the given number 1.95 into two portions: 1.9 (the digit before the decimal and one digit after the decimal) and .05 (the second digit after the decimal). (Note that  $1.95 = 1.9 + .05$ .) To find the required area under the standard normal curve, we locate 1.9 in the column for  $z$  on the left side of Table IV and .05 in the row for  $z$  at the top of Table IV. The entry where the row for 1.9 and the column for .05 intersect gives the area under the standard normal curve to the left of  $z = 1.95$ . The relevant portion of Table IV is reproduced as Table 6.2. From Table IV or Table 6.2, the entry where the row for 1.9 and the column for .05 cross is .9744. Consequently, the area under the standard normal curve to the left of  $z = 1.95$  is .9744. This area is shown in Figure 6.19. (It is always helpful to sketch the curve and mark the area you are determining.)

**Table 6.2** Area Under the Standard Normal Curve to the Left of  $z = 1.95$

$z$	.00	.01	...	.05	...	.09
−3.4	.0003	.0003	...	.0003	...	.0002
−3.3	.0005	.0005	...	.0004	...	.0003
−3.2	.0007	.0007	...	.0006	...	.0005
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
1.9	.9713	.9719	...	.9744	...	.9767
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
3.4	.9997	.9997	...	.9997	...	.9998

Required area



**Figure 6.19** Area to the left of  $z = 1.95$ .

The area to the left of  $z = 1.95$  can be interpreted as the probability that  $z$  assumes a value less than 1.95; that is,

$$\text{Area to the left of } 1.95 = P(z < 1.95) = .9744$$

As mentioned in Section 6.1, the probability that a continuous random variable assumes a single value is zero. Therefore,

$$P(z = 1.95) = 0$$

Hence,

$$P(z < 1.95) = P(z \leq 1.95) = .9744$$



## ■ EXAMPLE 6–2

Find the area under the standard normal curve from  $z = -2.17$  to  $z = 0$ .

**Solution** To find the area from  $z = -2.17$  to  $z = 0$ , first we find the areas to the left of  $z = 0$  and to the left of  $z = -2.17$  in the standard normal distribution table (Table IV). As shown in Table 6.3, these two areas are .5 and .0150, respectively. Next we subtract .0150 from .5 to find the required area.

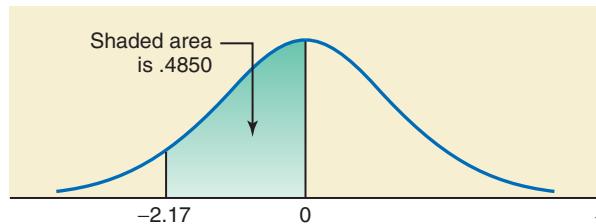
Finding the area  
between a negative  $z$  and  $z = 0$

**Table 6.3** Area Under the Standard Normal Curve

$z$	.00	.01	...	.07	...	.09
-3.4	.0003	.0003	...	.0003	...	.0002
-3.3	.0005	.0005	...	.0004	...	.0003
-3.2	.0007	.0007	...	.0005	...	.0005
...	...	...	...	...	...	...
-2.1	.0179	.0174	...	.0150	...	.0143
...	...	...	...	...	...	...
0.0	.5000	...	.5040	...	.5279	...
...	...	...	...	...	...	...
3.4	.9997	...	.9997	...	.9997	...
	Area to the left of $z = 0$			Area to the left of $z = -2.17$		

The area from  $z = -2.17$  to  $z = 0$  gives the probability that  $z$  lies in the interval  $-2.17$  to 0 (see Figure 6.20); that is,

$$\begin{aligned} \text{Area from } -2.17 \text{ to } 0 &= P(-2.17 \leq z \leq 0) \\ &= P(z \leq 0) - P(z \leq -2.17) = .5000 - .0150 = .4850 \end{aligned}$$



**Figure 6.20** Area from  $z = -2.17$  to  $z = 0$ .



Finding the areas in the right and left tails.

### ■ EXAMPLE 6–3

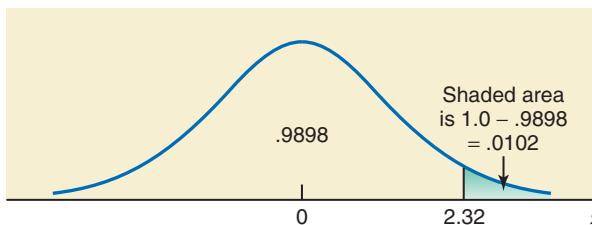
Find the following areas under the standard normal curve.

- Area to the right of  $z = 2.32$
- Area to the left of  $z = -1.54$

#### Solution

- (a) As mentioned earlier, the normal distribution table gives the area to the left of a  $z$  value. To find the area to the right of  $z = 2.32$ , first we find the area to the left of  $z = 2.32$ . Then we subtract this area from 1.0, which is the total area under the curve. From Table IV, the area to the left of  $z = 2.32$  is .9898. Consequently, the required area is  $1.0 - .9898 = .0102$ , as shown in Figure 6.21.

**Figure 6.21** Area to the right of  $z = 2.32$ .

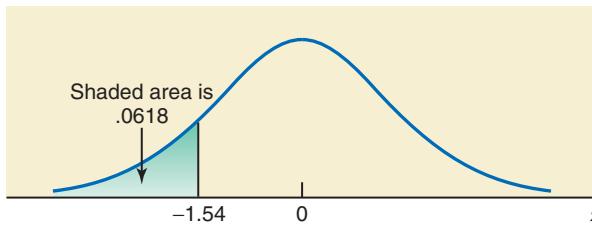


The area to the right of  $z = 2.32$  gives the probability that  $z$  is greater than 2.32. Thus,

$$\text{Area to the right of } 2.32 = P(z > 2.32) = 1.0 - .9898 = \mathbf{.0102}$$

- (b) To find the area under the standard normal curve to the left of  $z = -1.54$ , we find the area in Table IV that corresponds to  $-1.5$  in the  $z$  column and  $.04$  in the top row. This area is .0618. This area is shown in Figure 6.22.

**Figure 6.22** Area to the left of  $z = -1.54$ .



The area to the left of  $z = -1.54$  gives the probability that  $z$  is less than  $-1.54$ . Thus,

$$\text{Area to the left of } -1.54 = P(z < -1.54) = \mathbf{.0618}$$

### ■ EXAMPLE 6–4

Find the following probabilities for the standard normal curve.

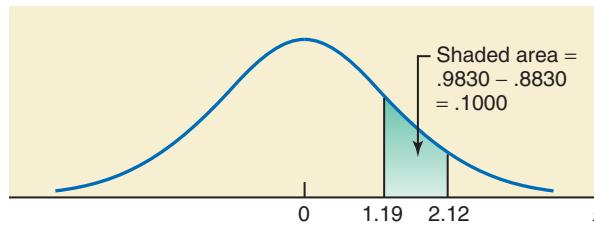
- $P(1.19 < z < 2.12)$
- $P(-1.56 < z < 2.31)$
- $P(z > -0.75)$

#### Solution

Finding the area between two positive values of  $z$ .

- (a) The probability  $P(1.19 < z < 2.12)$  is given by the area under the standard normal curve between  $z = 1.19$  and  $z = 2.12$ , which is the shaded area in Figure 6.23.

To find the area between  $z = 1.19$  and  $z = 2.12$ , first we find the areas to the left of  $z = 1.19$  and  $z = 2.12$ . Then we subtract the smaller area (to the left of  $z = 1.19$ ) from the larger area (to the left of  $z = 2.12$ ).



**Figure 6.23** Finding  $P(1.19 < z < 2.12)$ .

From Table IV for the standard normal distribution, we find

$$\text{Area to the left of } 1.19 = .8830$$

$$\text{Area to the left of } 2.12 = .9830$$

Then, the required probability is

$$\begin{aligned} P(1.19 < z < 2.12) &= \text{Area between } 1.19 \text{ and } 2.12 \\ &= .9830 - .8830 = \mathbf{.1000} \end{aligned}$$

- (b) The probability  $P(-1.56 < z < 2.31)$  is given by the area under the standard normal curve between  $z = -1.56$  and  $z = 2.31$ , which is the shaded area in Figure 6.24.  
From Table IV for the standard normal distribution, we have

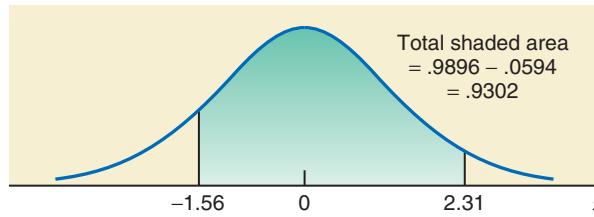
*Finding the area between a positive and a negative value of  $z$ .*

$$\text{Area to the left of } -1.56 = .0594$$

$$\text{Area to the left of } 2.31 = .9896$$

The required probability is

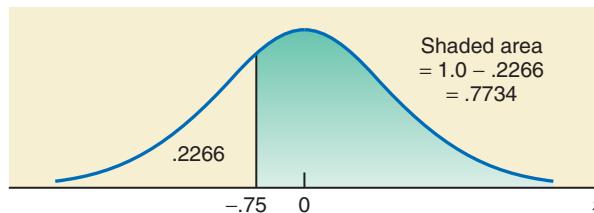
$$\begin{aligned} P(-1.56 < z < 2.31) &= \text{Area between } -1.56 \text{ and } 2.31 \\ &= .9896 - .0594 = \mathbf{.9302} \end{aligned}$$



**Figure 6.24** Finding  $P(-1.56 < z < 2.31)$ .

- (c) The probability  $P(z > -.75)$  is given by the area under the standard normal curve to the right of  $z = -.75$ , which is the shaded area in Figure 6.25.

*Finding the area to the right of a negative value of  $z$ .*



**Figure 6.25** Finding  $P(z > -.75)$ .

From Table IV for the standard normal distribution.

Area to the left of  $-0.75 = .2266$

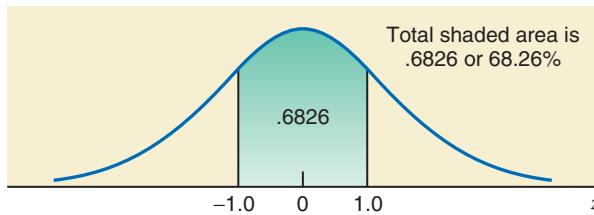
The required probability is

$$P(z > -0.75) = \text{Area to the right of } -0.75 = 1.0 - .2266 = .7734 \blacksquare$$

In the discussion in Section 3.4 of Chapter 3 on the use of the standard deviation, we discussed the empirical rule for a bell-shaped curve. That empirical rule is based on the standard normal distribution. By using the normal distribution table, we can now verify the empirical rule as follows.

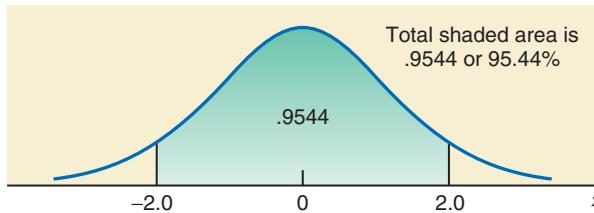
1. The total area within one standard deviation of the mean is 68.26%. This area is given by the difference between the area to the left of  $z = 1.0$  and the area to the left of  $z = -1.0$ . As shown in Figure 6.26, this area is  $.8413 - .1587 = .6826$ , or 68.26%.

**Figure 6.26** Area within one standard deviation of the mean.



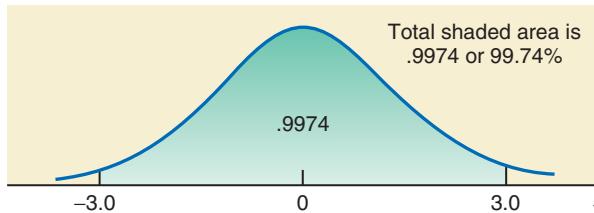
2. The total area within two standard deviations of the mean is 95.44%. This area is given by the difference between the area to the left of  $z = 2.0$  and the area to the left of  $z = -2.0$ . As shown in Figure 6.27, this area is  $.9772 - .0228 = .9544$ , or 95.44%.

**Figure 6.27** Area within two standard deviations of the mean.



3. The total area within three standard deviations of the mean is 99.74%. This area is given by the difference between the area to the left of  $z = 3.0$  and the area to the left of  $z = -3.0$ . As shown in Figure 6.28, this area is  $.9987 - .0013 = .9974$ , or 99.74%.

**Figure 6.28** Area within three standard deviations of the mean.



Again, as mentioned earlier, only a specific bell-shaped curve represents the normal distribution. Now we can state that a bell-shaped curve that contains (about) 68.26% of the total area within one standard deviation of the mean, (about) 95.44% of the total area within two standard deviations of the mean, and (about) 99.74% of the total area within three standard deviations of the mean represents a normal distribution curve.

The standard normal distribution table, Table IV of Appendix C, goes from  $z = -3.49$  to  $z = 3.49$ . Consequently, if we need to find the area to the left of  $z = -3.50$  or a smaller value of  $z$ , we can assume it to be approximately 0.0. If we need to find the area to the left of  $z = 3.50$  or a larger number, we can assume it to be approximately 1.0. Example 6–5 illustrates this procedure.

## ■ EXAMPLE 6–5

Find the following probabilities for the standard normal curve.

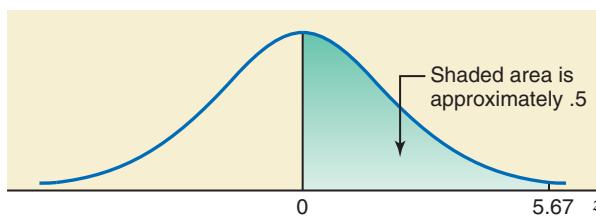
$$(a) P(0 < z < 5.67) \quad (b) P(z < -5.35)$$

### Solution

- (a) The probability  $P(0 < z < 5.67)$  is given by the area under the standard normal curve between  $z = 0$  and  $z = 5.67$ . Because  $z = 5.67$  is greater than 3.49 and is not in Table IV, the area under the standard normal curve to the left of  $z = 5.67$  can be approximated by 1.0. Also, the area to the left of  $z = 0$  is .5. Hence, the required probability is

$$P(0 < z < 5.67) = \text{Area between } 0 \text{ and } 5.67 = 1.0 - .5 = .5 \text{ approximately}$$

Note that the area between  $z = 0$  and  $z = 5.67$  is not exactly .5 but very close to .5. This area is shown in Figure 6.29.

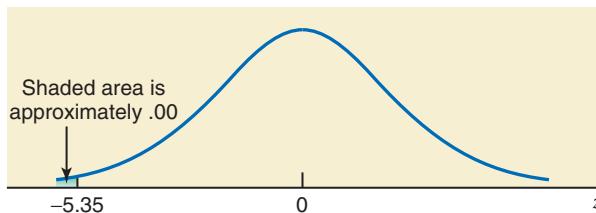


**Figure 6.29** Area between  $z = 0$  and  $z = 5.67$ .

*Finding the area between  $z = 0$  and a value of  $z$  greater than 3.49.*

- (b) The probability  $P(z < -5.35)$  represents the area under the standard normal curve to the left of  $z = -5.35$ . Since  $z = -5.35$  is not in the table, we can assume that this area is approximately .00. This is shown in Figure 6.30.

*Finding the area to the left of a  $z$  that is less than -3.49.*



**Figure 6.30** Area to the left of  $z = -5.35$ .

The required probability is

$$P(z < -5.35) = \text{Area to the left of } -5.35 = .00 \text{ approximately}$$

Again, note that the area to the left of  $z = -5.35$  is not exactly .00 but very close to .00.

We can find the exact areas for parts (a) and (b) of this example by using technology. The reader should do that. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 6.1** What is the difference between the probability distribution of a discrete random variable and that of a continuous random variable? Explain.

- 6.2** Let  $x$  be a continuous random variable. What is the probability that  $x$  assumes a single value, such as  $a$ ?

- 6.3** For a continuous probability distribution, explain why the following holds true.

$$P(a < x < b) = P(a < x \leq b) = P(a \leq x < b) = P(a \leq x \leq b)$$

- 6.4** Briefly explain the main characteristics of a normal distribution. Illustrate with the help of graphs.

- 6.5** Briefly describe the standard normal distribution curve.

**6.6** What are the parameters of the normal distribution?

**6.7** How do the width and height of a normal distribution change when its mean remains the same but its standard deviation decreases?

**6.8** Do the width and/or height of a normal distribution change when its standard deviation remains the same but its mean increases?

**6.9** For the standard normal distribution, what does  $z$  represent?

**6.10** For the standard normal distribution, find the area within one standard deviation of the mean—that is, the area between  $\mu - \sigma$  and  $\mu + \sigma$ .

**6.11** For the standard normal distribution, find the area within 1.5 standard deviations of the mean—that is, the area between  $\mu - 1.5\sigma$  and  $\mu + 1.5\sigma$ .

**6.12** For the standard normal distribution, what is the area within two standard deviations of the mean?

**6.13** For the standard normal distribution, what is the area within 2.5 standard deviations of the mean?

**6.14** For the standard normal distribution, what is the area within three standard deviations of the mean?

**6.15** Find the area under the standard normal curve

- |  |  |
|--|--|
| a. between $z = 0$ and $z = 1.95$<br>c. between $z = 1.15$ and $z = 2.37$<br>e. from $z = -1.67$ to $z = 2.24$ | b. between $z = 0$ and $z = -2.05$<br>d. from $z = -1.53$ to $z = -2.88$ |
|--|--|

**6.16** Find the area under the standard normal curve

- |   |   |
|---|---|
| a. from $z = 0$ to $z = 2.34$<br>c. from $z = .84$ to $z = 1.95$<br>e. between $z = -2.15$ and $z = 1.87$ | b. between $z = 0$ and $z = -2.58$<br>d. between $z = -.57$ and $z = -2.49$ |
|---|---|

**6.17** Find the area under the standard normal curve

- |   |   |
|---|---|
| a. to the right of $z = 1.36$<br>c. to the right of $z = -2.05$ | b. to the left of $z = -1.97$<br>d. to the left of $z = 1.76$ |
|---|---|

**6.18** Obtain the area under the standard normal curve

- |  |  |
|--|--|
| a. to the right of $z = 1.43$<br>c. to the right of $z = -.65$ | b. to the left of $z = -1.65$<br>d. to the left of $z = .89$ |
|--|--|

**6.19** Find the area under the standard normal curve

- |  |   |
|--|---|
| a. between $z = 0$ and $z = 4.28$<br>c. to the right of $z = 7.43$ | b. from $z = 0$ to $z = -3.75$<br>d. to the left of $z = -4.69$ |
|--|---|

**6.20** Find the area under the standard normal curve

- |  |   |
|--|---|
| a. from $z = 0$ to $z = 3.94$<br>c. to the right of $z = 5.42$ | b. between $z = 0$ and $z = -5.16$<br>d. to the left of $z = -3.68$ |
|--|---|

**6.21** Determine the following probabilities for the standard normal distribution.

- |   |   |
|---|---|
| a. $P(-1.83 \leq z \leq 2.57)$<br>c. $P(-1.99 \leq z \leq 0)$ | b. $P(0 \leq z \leq 2.02)$<br>d. $P(z \geq 1.48)$ |
|---|---|

**6.22** Determine the following probabilities for the standard normal distribution.

- |   |  |
|---|--|
| a. $P(-2.46 \leq z \leq 1.88)$<br>c. $P(-2.58 \leq z \leq 0)$ | b. $P(0 \leq z \leq 1.96)$<br>d. $P(z \geq .73)$ |
|---|--|

**6.23** Find the following probabilities for the standard normal distribution.

- |   |  |
|---|--|
| a. $P(z < -2.34)$<br>c. $P(-2.07 \leq z \leq -.93)$ | b. $P(.67 \leq z \leq 2.59)$<br>d. $P(z < 1.78)$ |
|---|--|

**6.24** Find the following probabilities for the standard normal distribution.

- |  |   |
|--|---|
| a. $P(z < -1.31)$<br>c. $P(-2.24 \leq z \leq -1.19)$ | b. $P(1.23 \leq z \leq 2.89)$<br>d. $P(z < 2.02)$ |
|--|---|

**6.25** Obtain the following probabilities for the standard normal distribution.

- |  |  |
|--|--|
| a. $P(z > -.98)$<br>c. $P(0 \leq z \leq 4.25)$<br>e. $P(z > 6.07)$ | b. $P(-2.47 \leq z \leq 1.29)$<br>d. $P(-5.36 \leq z \leq 0)$<br>f. $P(z < -5.27)$ |
|--|--|

**6.26** Obtain the following probabilities for the standard normal distribution.

- |   |   |
|---|---|
| a. $P(z > -1.86)$<br>c. $P(0 \leq z \leq 3.85)$<br>e. $P(z > 4.82)$ | b. $P(-.68 \leq z \leq 1.94)$<br>d. $P(-4.34 \leq z \leq 0)$<br>f. $P(z < -6.12)$ |
|---|---|

## 6.2 Standardizing a Normal Distribution

As was shown in the previous section, Table IV of Appendix C can be used to find areas under the standard normal curve. However, in real-world applications, a (continuous) random variable may have a normal distribution with values of the mean and standard deviation that are different from 0 and 1, respectively. The first step in such a case is to convert the given normal distribution to the standard normal distribution. This procedure is called *standardizing a normal distribution*. The units of a normal distribution (which is not the standard normal distribution) are denoted by  $x$ . We know from Section 6.1.3 that units of the standard normal distribution are denoted by  $z$ .

**Converting an  $x$  Value to a  $z$  Value** For a normal random variable  $x$ , a particular value of  $x$  can be converted to its corresponding  $z$  value by using the formula

$$z = \frac{x - \mu}{\sigma}$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the normal distribution of  $x$ , respectively. When  $x$  follows a normal distribution,  $z$  follows the standard normal distribution.

Thus, to find the  $z$  value for an  $x$  value, we calculate the difference between the given  $x$  value and the mean,  $\mu$ , and divide this difference by the standard deviation,  $\sigma$ . If the value of  $x$  is equal to  $\mu$ , then its  $z$  value is equal to zero. Note that we will always round  $z$  values to two decimal places.

The  $z$  value for the mean of a normal distribution is always zero. The  $z$  value for an  $x$  greater than the mean is positive and the  $z$  value for an  $x$  smaller than the mean is negative.

◀ **Remember**

Examples 6–6 through 6–10 describe how to convert  $x$  values to the corresponding  $z$  values and how to find areas under a normal distribution curve.

### ■ EXAMPLE 6–6

Let  $x$  be a continuous random variable that has a normal distribution with a mean of 50 and a standard deviation of 10. Convert the following  $x$  values to  $z$  values and find the probability to the left of these points.

- (a)  $x = 55$       (b)  $x = 35$

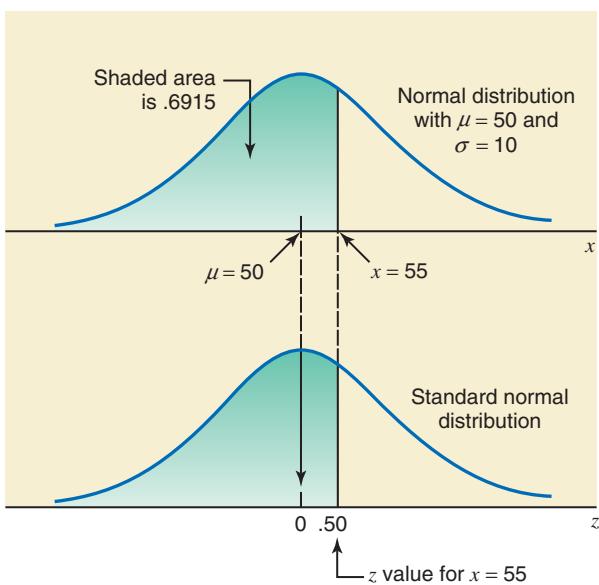
*Converting  $x$  values  
to the corresponding  $z$  values.*

**Solution** For the given normal distribution,  $\mu = 50$  and  $\sigma = 10$ .

- (a) The  $z$  value for  $x = 55$  is computed as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{55 - 50}{10} = .50$$

Thus, the  $z$  value for  $x = 55$  is .50. The  $z$  values for  $\mu = 50$  and  $x = 55$  are shown in Figure 6.31. Note that the  $z$  value for  $\mu = 50$  is zero. The value  $z = .50$  for  $x = 55$  indicates that the distance between  $\mu = 50$  and  $x = 55$  is  $1/2$  of the standard deviation, which is 10. Consequently, we can state that the  $z$  value represents the distance between  $\mu$  and  $x$  in terms of the standard deviation. Because  $x = 55$  is greater than  $\mu = 50$ , its  $z$  value is positive.

**Figure 6.31**  $z$  value for  $x = 55$ .

From this point on, we will usually show only the  $z$  axis below the  $x$  axis and not the standard normal curve itself.

To find the probability to the left of  $x = 55$ , we find the probability to the left of  $z = .50$  from Table IV. This probability is .6915. Therefore,

$$P(x < 55) = P(z < .50) = .6915$$

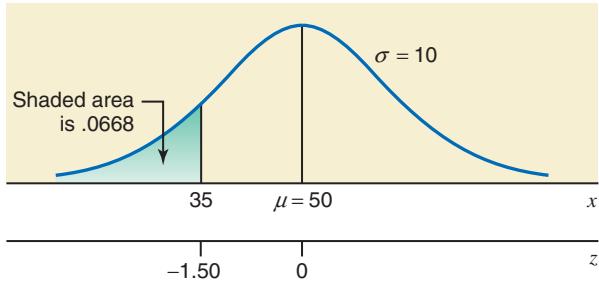
(b) The  $z$  value for  $x = 35$  is computed as follows and is shown in Figure 6.32:

$$z = \frac{x - \mu}{\sigma} = \frac{35 - 50}{10} = -1.50$$

Because  $x = 35$  is on the left side of the mean (i.e., 35 is less than  $\mu = 50$ ), its  $z$  value is negative. As a general rule, whenever an  $x$  value is less than the value of  $\mu$ , its  $z$  value is negative.

To find the probability to the left of  $x = 35$ , we find the area under the normal curve to the left of  $z = -1.50$ . This area from Table IV is .0668. Hence,

$$P(x < 35) = P(z < -1.50) = .0668$$

**Figure 6.32**  $z$  value for  $x = 35$ .

**Remember ▶** The  $z$  value for an  $x$  value that is greater than  $\mu$  is positive, the  $z$  value for an  $x$  value that is equal to  $\mu$  is zero, and the  $z$  value for an  $x$  value that is less than  $\mu$  is negative.

To find the area between two values of  $x$  for a normal distribution, we first convert both values of  $x$  to their respective  $z$  values. Then we find the area under the standard normal curve

between those two  $z$  values. The area between the two  $z$  values gives the area between the corresponding  $x$  values. Example 6–7 illustrates this case.

### ■ EXAMPLE 6–7

Let  $x$  be a continuous random variable that is normally distributed with a mean of 25 and a standard deviation of 4. Find the area

- (a) between  $x = 25$  and  $x = 32$     (b) between  $x = 18$  and  $x = 34$

**Solution** For the given normal distribution,  $\mu = 25$  and  $\sigma = 4$ .

- (a) The first step in finding the required area is to standardize the given normal distribution by converting  $x = 25$  and  $x = 32$  to their respective  $z$  values using the formula

$$z = \frac{x - \mu}{\sigma}$$

Finding the area  
between the mean and a  
point to its right.

The  $z$  value for  $x = 25$  is zero because it is the mean of the normal distribution. The  $z$  value for  $x = 32$  is

$$z = \frac{32 - 25}{4} = 1.75$$

The area between  $x = 25$  and  $x = 32$  under the given normal distribution curve is equivalent to the area between  $z = 0$  and  $z = 1.75$  under the standard normal curve. From Table IV, the area to the left of  $z = 1.75$  is .9599, and the area to the left of  $z = 0$  is .50. Hence, the required area is  $.9599 - .50 = .4599$ , which is shown in Figure 6.33.

The area between  $x = 25$  and  $x = 32$  under the normal curve gives the probability that  $x$  assumes a value between 25 and 32. This probability can be written as

$$P(25 < x < 32) = P(0 < z < 1.75) = .4599$$

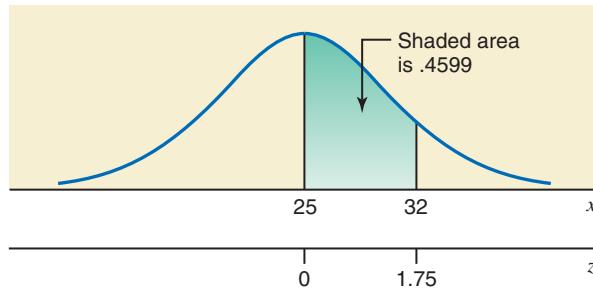


Figure 6.33 Area between  $x = 25$  and  $x = 32$ .

- (b) First, we calculate the  $z$  values for  $x = 18$  and  $x = 34$  as follows:

$$\text{For } x = 18: z = \frac{18 - 25}{4} = -1.75$$

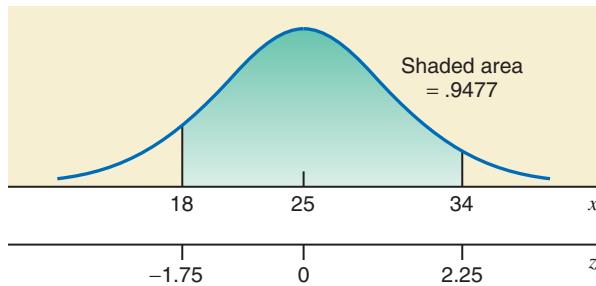
$$\text{For } x = 34: z = \frac{34 - 25}{4} = 2.25$$

Finding the area  
between two points on different  
sides of the mean

The area under the given normal distribution curve between  $x = 18$  and  $x = 34$  is given by the area under the standard normal curve between  $z = -1.75$  and  $z = 2.25$ . From Table IV, the area to the left of  $z = 2.25$  is .9878, and the area to the left of  $z = -1.75$  is .0401. Hence, the required area is

$$P(18 < x < 34) = P(-1.75 < z < 2.25) = .9878 - .0401 = .9477$$

This area is shown in Figure 6.34.

**Figure 6.34** Area between  $x = 18$  and  $x = 34$ .**■ EXAMPLE 6-8**

Let  $x$  be a normal random variable with its mean equal to 40 and standard deviation equal to 5. Find the following probabilities for this normal distribution.

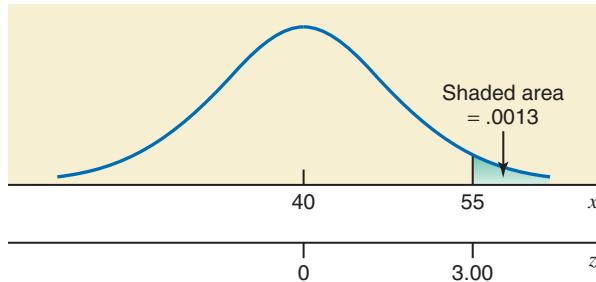
- (a)  $P(x > 55)$     (b)  $P(x < 49)$

**Solution** For the given normal distribution,  $\mu = 40$  and  $\sigma = 5$ .

*Calculating the probability to the right of a value of  $x$ .*

- (a) The probability that  $x$  assumes a value greater than 55 is given by the area under the normal distribution curve to the right of  $x = 55$ , as shown in Figure 6.35. This area is calculated by subtracting the area to the left of  $x = 55$  from 1.0, which is the total area under the curve.

$$\text{For } x = 55: z = \frac{55 - 40}{5} = 3.00$$

**Figure 6.35** Finding  $P(x > 55)$ .

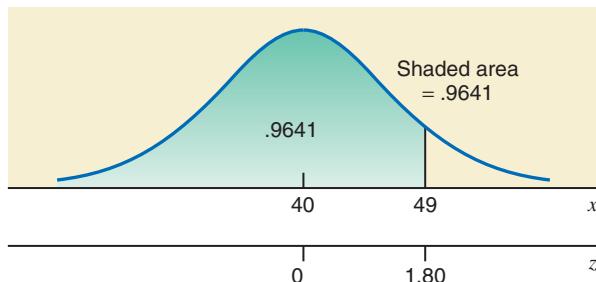
The required probability is given by the area to the right of  $z = 3.00$ . To find this area, first we find the area to the left of  $z = 3.00$ , which is .9987. Then we subtract this area from 1.0. Thus,

$$P(x > 55) = P(z > 3.00) = 1.0 - .9987 = .0013$$

*Calculating the probability to the left of a value of  $x$ .*

- (b) The probability that  $x$  will assume a value less than 49 is given by the area under the normal distribution curve to the left of 49, which is the shaded area in Figure 6.36. This area is obtained from Table IV as follows.

$$\text{For } x = 49: z = \frac{49 - 40}{5} = 1.80$$

**Figure 6.36** Finding  $P(x < 49)$ .

The required probability is given by the area to the left of  $z = 1.80$ . This area from Table IV is .9641. Therefore, the required probability is

$$P(x < 49) = P(z < 1.80) = \mathbf{.9641}$$

### ■ EXAMPLE 6–9

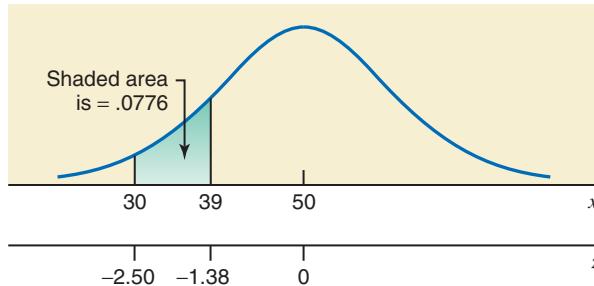
Let  $x$  be a continuous random variable that has a normal distribution with  $\mu = 50$  and  $\sigma = 8$ . Find the probability  $P(30 \leq x \leq 39)$ .

**Solution** For this normal distribution,  $\mu = 50$  and  $\sigma = 8$ . The probability  $P(30 \leq x \leq 39)$  is given by the area from  $x = 30$  to  $x = 39$  under the normal distribution curve. As shown in Figure 6.37, this area is given by the difference between the area to the left of  $x = 30$  and the area to the left of  $x = 39$ .

$$\text{For } x = 30: z = \frac{30 - 50}{8} = -2.50$$

$$\text{For } x = 39: z = \frac{39 - 50}{8} = -1.38$$

Finding the area between two  $x$  values that are less than the mean.



**Figure 6.37** Finding  $P(30 \leq x \leq 39)$ .

To find the required area, we first find the area to the left of  $z = -2.50$ , which is .0062. Then, we find the area to the left of  $z = -1.38$ , which is .0838. The difference between these two areas gives the required probability, which is

$$P(30 \leq x \leq 39) = P(-2.50 \leq z \leq -1.38) = .0838 - .0062 = \mathbf{.0776}$$

### ■ EXAMPLE 6–10

Let  $x$  be a continuous random variable that has a normal distribution with a mean of 80 and a standard deviation of 12. Find the area under the normal distribution curve

- (a) from  $x = 70$  to  $x = 135$     (b) to the left of 27

**Solution** For the given normal distribution,  $\mu = 80$  and  $\sigma = 12$ .

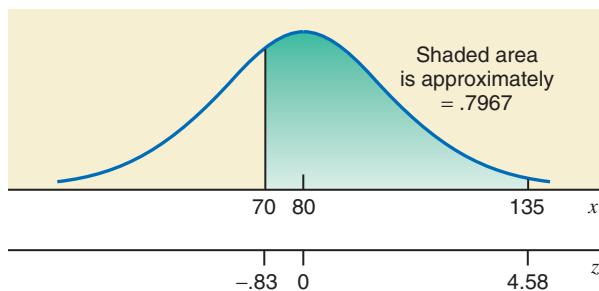
- (a) The  $z$  values for  $x = 70$  and  $x = 135$  are:

$$\text{For } x = 70: z = \frac{70 - 80}{12} = -.83$$

$$\text{For } x = 135: z = \frac{135 - 80}{12} = 4.58$$

Finding the area between two  $x$  values that are on different sides of the mean.

Thus, to find the required area we find the areas to the left of  $z = -.83$  and to the left of  $z = 4.58$  under the standard normal curve. From Table IV, the area to the left

**Figure 6.38** Area between  $x = 70$  and  $x = 135$ .

of  $z = -.83$  is .2033 and the area to the left of  $z = 4.58$  is approximately 1.0. Note that  $z = 4.58$  is not in Table IV.

Hence,

$$\begin{aligned} P(70 \leq x \leq 135) &= P(-.83 \leq z \leq 4.58) \\ &= 1.0 - .2033 = .7967 \text{ approximately} \end{aligned}$$

Figure 6.38 shows this area.

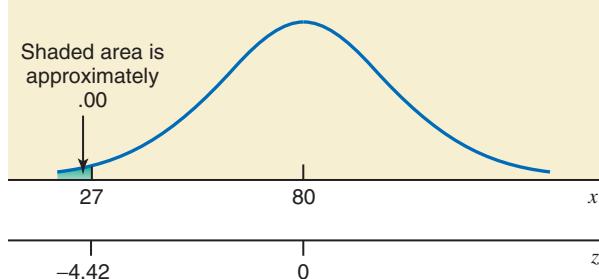
*Finding an area in the left tail*

- (b) First we find the  $z$  value for  $x = 27$ .

$$\text{For } x = 27: z = \frac{27 - 80}{12} = -4.42$$

As shown in Figure 6.39, the required area is given by the area under the standard normal distribution curve to the left of  $z = -4.42$ . This area is approximately zero.

$$P(x < 27) = P(z < -4.42) = .00 \text{ approximately}$$

**Figure 6.39** Area to the left of  $x = 27$ .

## EXERCISES

### CONCEPTS AND PROCEDURES



- 6.27** Find the  $z$  value for each of the following  $x$  values for a normal distribution with  $\mu = 30$  and  $\sigma = 5$ .
- $x = 39$
  - $x = 19$
  - $x = 24$
  - $x = 44$
- 6.28** Determine the  $z$  value for each of the following  $x$  values for a normal distribution with  $\mu = 16$  and  $\sigma = 3$ .
- $x = 12$
  - $x = 22$
  - $x = 19$
  - $x = 13$
- 6.29** Find the following areas under a normal distribution curve with  $\mu = 20$  and  $\sigma = 4$ .
- Area between  $x = 20$  and  $x = 27$
  - Area from  $x = 23$  to  $x = 26$
  - Area between  $x = 9.5$  and  $x = 17$
- 6.30** Find the following areas under a normal distribution curve with  $\mu = 12$  and  $\sigma = 2$ .
- Area between  $x = 7.76$  and  $x = 12$
  - Area between  $x = 14.48$  and  $x = 16.54$
  - Area from  $x = 8.22$  to  $x = 10.06$

- 6.31** Determine the area under a normal distribution curve with  $\mu = 55$  and  $\sigma = 7$
- to the right of  $x = 58$
  - to the right of  $x = 43$
  - to the left of  $x = 68$
  - to the left of  $x = 22$
- 6.32** Find the area under a normal distribution curve with  $\mu = 18.3$  and  $\sigma = 3.4$
- to the left of  $x = 10.9$
  - to the right of  $x = 14$
  - to the left of  $x = 22.7$
  - to the right of  $x = 29.2$
- 6.33** Let  $x$  be a continuous random variable that is normally distributed with a mean of 25 and a standard deviation of 6. Find the probability that  $x$  assumes a value
- between 29 and 36
  - between 22 and 35
- 6.34** Let  $x$  be a continuous random variable that has a normal distribution with a mean of 117.6 and a standard deviation of 14.6. Find the probability that  $x$  assumes a value
- between 77.9 and 98.3
  - between 85.3 and 142.6
- 6.35** Let  $x$  be a continuous random variable that is normally distributed with a mean of 80 and a standard deviation of 12. Find the probability that  $x$  assumes a value
- greater than 69
  - less than 73
  - greater than 101
  - less than 87
- 6.36** Let  $x$  be a continuous random variable that is normally distributed with a mean of 65 and a standard deviation of 15. Find the probability that  $x$  assumes a value
- less than 45
  - greater than 79
  - greater than 54
  - less than 70

## 6.3 Applications of the Normal Distribution

Sections 6.1 and 6.2 discussed the normal distribution, how to convert a normal distribution to the standard normal distribution, and how to find areas under a normal distribution curve. This section presents examples that illustrate the applications of the normal distribution.

### ■ EXAMPLE 6-11

According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of \$4129 for family coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose that the premiums for family coverage paid this year by all such workers are normally distributed with a mean of \$4129 and a standard deviation of \$600. Find the probability that such a premium paid this year by a randomly selected such worker is between \$3331 and \$4453.

*Using the normal distribution:  
the area between two points on  
different sides of the mean.*

**Solution** Let  $x$  denote the premium paid this year for family coverage by a randomly selected worker with employer-provided health insurance. Then,  $x$  is normally distributed with

$$\mu = \$4129 \quad \text{and} \quad \sigma = \$600$$

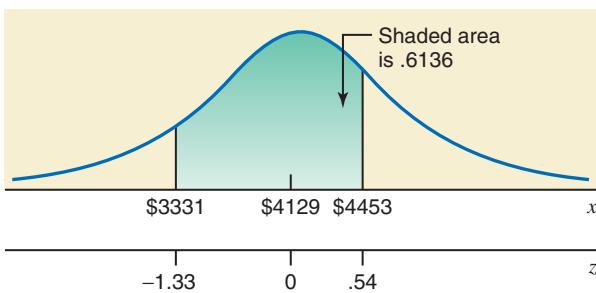
The probability that such a premium paid by a randomly selected such worker is between \$3331 and \$4453 is given by the area under the normal distribution curve of  $x$  that falls between  $x = \$3331$  and  $x = \$4453$  as shown in Figure 6.40. To find this area, first we find the areas to the left of  $x = \$3331$  and  $x = \$4453$ , respectively, and then take the difference between these two areas.

$$\text{For } x = \$3331: \quad z = \frac{3331 - 4129}{600} = -1.33$$

$$\text{For } x = \$4453: \quad z = \frac{4453 - 4129}{600} = .54$$

Thus, the required probability is given by the difference between the areas under the standard normal curve to the left of  $z = -1.33$  and to the left of  $z = .54$ . From Table IV in Appendix C,

**Figure 6.40** Area between  $x = \$3331$  and  $x = \$4453$ .



the area to the left of  $z = -1.33$  is .0918, and the area to the left of  $z = .54$  is .7054. Hence, the required probability is

$$P(\$3331 < x < \$4453) = P(-1.33 < z < .54) = .7054 - .0918 = .6136$$

Thus, the probability is .6136 that the premium paid this year for family coverage by a randomly selected worker with employer-provided health insurance is between \$3331 and \$4453. Converting this probability into a percentage, we can also state that (about) 61.36% of such workers paid premiums between \$3331 and \$4453 this year for family coverage. ■

*Using the normal distribution: probability that  $x$  is less than a value that is to the right of the mean.*



PhotoDisc, Inc./Getty Images

### ■ EXAMPLE 6-12

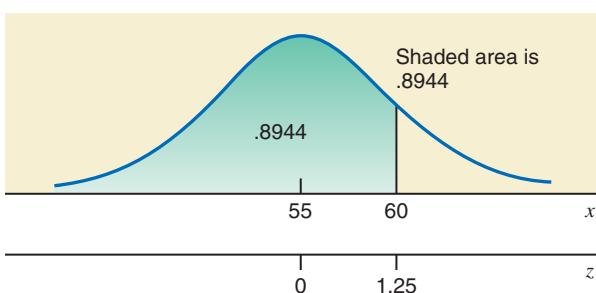
A racing car is one of the many toys manufactured by Mack Corporation. The assembly times for this toy follow a normal distribution with a mean of 55 minutes and a standard deviation of 4 minutes. The company closes at 5 P.M. every day. If one worker starts to assemble a racing car at 4 P.M., what is the probability that she will finish this job before the company closes for the day?

**Solution** Let  $x$  denote the time this worker takes to assemble a racing car. Then,  $x$  is normally distributed with

$$\mu = 55 \text{ minutes} \quad \text{and} \quad \sigma = 4 \text{ minutes}$$

We are to find the probability that this worker can assemble this car in 60 minutes or less (between 4 and 5 P.M.). This probability is given by the area under the normal curve to the left of  $x = 60$  minutes as shown in Figure 6.41.

**Figure 6.41** Area to the left of  $x = 60$ .



$$\text{For } x = 60: \quad z = \frac{60 - 55}{4} = 1.25$$

The required probability is given by the area under the standard normal curve to the left of  $z = 1.25$ , which is .8944 from Table IV of Appendix C. Thus, the required probability is

$$P(x \leq 60) = P(z \leq 1.25) = .8944$$

Thus, the probability is .8944 that this worker will finish assembling this racing car before the company closes for the day. ■

### ■ EXAMPLE 6-13

Hupper Corporation produces many types of soft drinks, including Orange Cola. The filling machines are adjusted to pour 12 ounces of soda into each 12-ounce can of Orange Cola. However, the actual amount of soda poured into each can is not exactly 12 ounces; it varies from can to can. It has been observed that the net amount of soda in such a can has a normal distribution with a mean of 12 ounces and a standard deviation of .015 ounce.

- (a) What is the probability that a randomly selected can of Orange Cola contains 11.97 to 11.99 ounces of soda?
- (b) What percentage of the Orange Cola cans contain 12.02 to 12.07 ounces of soda?

**Solution** Let  $x$  be the net amount of soda in a can of Orange Cola. Then,  $x$  has a normal distribution with  $\mu = 12$  ounces and  $\sigma = .015$  ounce.

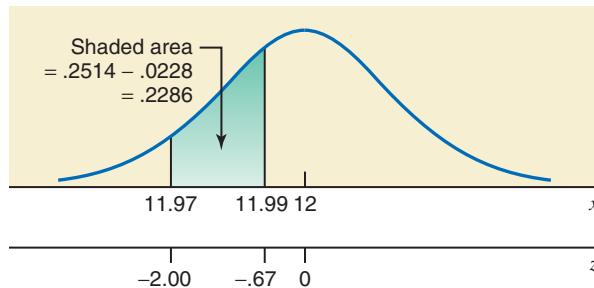
- (a) The probability that a randomly selected can contains 11.97 to 11.99 ounces of soda is given by the area under the normal distribution curve from  $x = 11.97$  to  $x = 11.99$ . This area is shown in Figure 6.42.

$$\text{For } x = 11.97: z = \frac{11.97 - 12}{.015} = -2.00$$

$$\text{For } x = 11.99: z = \frac{11.99 - 12}{.015} = -.67$$

*Using the normal distribution.*

*Calculating the probability between two points that are to the left of the mean.*



**Figure 6.42** Area between  $x = 11.97$  and  $x = 11.99$ .

The required probability is given by the area under the standard normal curve between  $z = -2.00$  and  $z = -.67$ . From Table IV of Appendix C, the area to the left of  $z = -2.00$  is .0228, and the area to the left of  $z = -.67$  is .2514. Hence, the required probability is

$$P(11.97 \leq x \leq 11.99) = P(-2.00 \leq z \leq -.67) = .2514 - .0228 = .2286$$

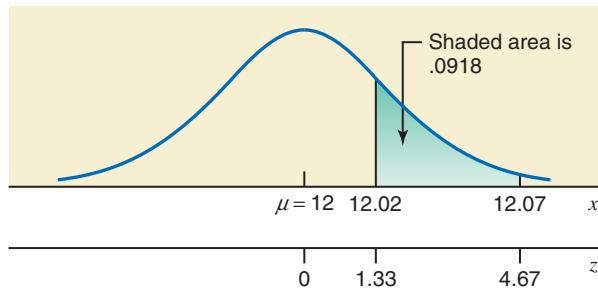
Thus, the probability is .2286 that any randomly selected can of Orange Cola will contain 11.97 to 11.99 ounces of soda. We can also state that about 22.86% of Orange Cola cans contain 11.97 to 11.99 ounces of soda.

- (b) The percentage of Orange Cola cans that contain 12.02 to 12.07 ounces of soda is given by the area under the normal distribution curve from  $x = 12.02$  to  $x = 12.07$ , as shown in Figure 6.43.

*Calculating the probability between two points that are to the right of the mean.*

$$\text{For } x = 12.02: z = \frac{12.02 - 12}{.015} = 1.33$$

$$\text{For } x = 12.07: z = \frac{12.07 - 12}{.015} = 4.67$$

**Figure 6.43** Area from  $x = 12.02$  to  $x = 12.07$ .

The required probability is given by the area under the standard normal curve between  $z = 1.33$  and  $z = 4.67$ . From Table IV of Appendix C, the area to the left of  $z = 1.33$  is .9082, and the area to the left of  $z = 4.67$  is approximately 1.0. Hence, the required probability is

$$P(12.02 \leq x \leq 12.07) = P(1.33 \leq z \leq 4.67) = 1.0 - .9082 = .0918$$

Converting this probability to a percentage, we can state that approximately 9.18% of all Orange Cola cans are expected to contain 12.02 to 12.07 ounces of soda. ■

### ■ EXAMPLE 6-14

*Finding the area to the left of x that is less than the mean.*

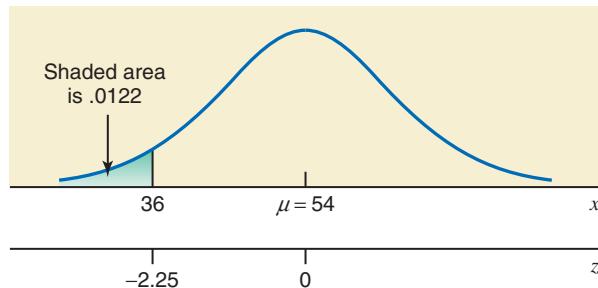


© jsemeniuk/iStockphoto

Suppose the life span of a calculator manufactured by Calculators Corporation has a normal distribution with a mean of 54 months and a standard deviation of 8 months. The company guarantees that any calculator that starts malfunctioning within 36 months of the purchase will be replaced by a new one. About what percentage of calculators made by this company are expected to be replaced?

**Solution** Let  $x$  be the life span of such a calculator. Then  $x$  has a normal distribution with  $\mu = 54$  and  $\sigma = 8$  months. The probability that a randomly selected calculator will start to malfunction within 36 months is given by the area under the normal distribution curve to the left of  $x = 36$ , as shown in Figure 6.44.

$$\text{For } x = 36: z = \frac{36 - 54}{8} = -2.25$$

**Figure 6.44** Area to the left of  $x = 36$ .

The required percentage is given by the area under the standard normal curve to the left of  $z = -2.25$ . From Table IV of Appendix C, this area is .0122. Hence, the required probability is

$$P(x < 36) = P(z < -2.25) = .0122$$

The probability that any randomly selected calculator manufactured by Calculators Corporation will start to malfunction within 36 months is .0122. Converting this probability to a percentage, we can state that approximately 1.22% of all calculators manufactured by this company are expected to start malfunctioning within 36 months. Hence, 1.22% of the calculators are expected to be replaced. ■

## EXERCISES

### APPLICATIONS

**6.37** Let  $x$  denote the time taken to run a road race. Suppose  $x$  is approximately normally distributed with a mean of 190 minutes and a standard deviation of 21 minutes. If one runner is selected at random, what is the probability that this runner will complete this road race

- a. in less than 160 minutes?
- b. in 215 to 245 minutes?

**6.38** According to the U.S. Employment and Training Administration, the average weekly unemployment benefit paid out in 2008 was \$297 (<http://www.ows.doleta.gov/unemploy/hb394.asp>). Suppose that the current distribution of weekly unemployment benefits paid out is approximately normally distributed with a mean of \$297 and a standard deviation of \$74.42. Find the probability that a randomly selected American who is receiving unemployment benefits is receiving

- a. more than \$400 per week
- b. between \$200 and \$340 per week

**6.39** According to the National Retail Federation's recent Back to College Consumer Intentions and Actions survey, families of college students spend an average of \$616.13 on new apparel, furniture for dorms or apartments, school supplies, and electronics ([www.nrf.com/modules.php?name=News&op=viewlive&sp\\_id=966](http://www.nrf.com/modules.php?name=News&op=viewlive&sp_id=966)). Suppose that the expenses on such Back to College items for the current year are approximately normally distributed with a mean of \$616.13 and a standard deviation of \$120. Find the probability that the amount of money spent on such items by a randomly selected family of a college student is

- a. less than \$450
- b. between \$500 and \$750

**6.40** Tommy Wait, a minor league baseball pitcher, is notorious for taking an excessive amount of time between pitches. In fact, his times between pitches are normally distributed with a mean of 36 seconds and a standard deviation of 2.5 seconds. What percentage of his times between pitches are

- a. longer than 39 seconds?
- b. between 29 and 34 seconds?

**6.41** A construction zone on a highway has a posted speed limit of 40 miles per hour. The speeds of vehicles passing through this construction zone are normally distributed with a mean of 46 miles per hour and a standard deviation of 4 miles per hour. Find the percentage of vehicles passing through this construction zone that are

- a. exceeding the posted speed limit
- b. traveling at speeds between 50 and 57 miles per hour

**6.42** The Bank of Connecticut issues Visa and MasterCard credit cards. It is estimated that the balances on all Visa credit cards issued by the Bank of Connecticut have a mean of \$845 and a standard deviation of \$270. Assume that the balances on all these Visa cards follow a normal distribution.

- a. What is the probability that a randomly selected Visa card issued by this bank has a balance between \$1000 and \$1440?
- b. What percentage of the Visa cards issued by this bank have a balance of \$730 or more?

**6.43** A 2011 analysis performed by ReadWrite Mobile revealed that the average number of apps downloaded per day per iOS device (such as iPhone, iPod, and iPad) exceeds 60 ([www.readwriteweb.com/mobile/2011/01/more-than-60-apps-downloaded-per-ios-device.php](http://www.readwriteweb.com/mobile/2011/01/more-than-60-apps-downloaded-per-ios-device.php)). Suppose that the current distribution of apps downloaded per day per iOS device is approximately normal with a mean of 65 and a standard deviation of 19.4. Find the probability that the number of apps downloaded on a randomly selected day by a randomly selected owner of an iOS device is

- a. 100 or more
- b. 45 or fewer

**6.44** The transmission on a model of a specific car has a warranty for 40,000 miles. It is known that the life of such a transmission has a normal distribution with a mean of 72,000 miles and a standard deviation of 13,000 miles.

- a. What percentage of the transmissions will fail before the end of the warranty period?
- b. What percentage of the transmissions will be good for more than 100,000 miles?

**6.45** According to the records of an electric company serving the Boston area, the mean electricity consumption for all households during winter is 1650 kilowatt-hours per month. Assume that the monthly electricity consumptions during winter by all households in this area have a normal distribution with a mean of 1650 kilowatt-hours and a standard deviation of 320 kilowatt-hours.

- a. Find the probability that the monthly electricity consumption during winter by a randomly selected household from this area is less than 1950 kilowatt-hours.
- b. What percentage of the households in this area have a monthly electricity consumption of 900 to 1300 kilowatt-hours?

**6.46** The management of a supermarket wants to adopt a new promotional policy of giving a free gift to every customer who spends more than a certain amount per visit at this supermarket. The expectation of the management is that after this promotional policy is advertised, the expenditures for all customers at this supermarket will be normally distributed with a mean of \$95 and a standard deviation of \$20. If the management decides to give free gifts to all those customers who spend more than \$130 at this supermarket during a visit, what percentage of the customers are expected to get free gifts?

**6.47** One of the cars sold by Walt's car dealership is a very popular subcompact car called Rhino. The final sale price of the basic model of this car varies from customer to customer depending on the negotiating skills and persistence of the customer. Assume that these sale prices of this car are normally distributed with a mean of \$19,800 and a standard deviation of \$350.

- Dolores paid \$19,445 for her Rhino. What percentage of Walt's customers paid less than Dolores for a Rhino?
- Cuthbert paid \$20,300 for a Rhino. What percentage of Walt's customers paid more than Cuthbert for a Rhino?

**6.48** A psychologist has devised a stress test for dental patients sitting in the waiting rooms. According to this test, the stress scores (on a scale of 1 to 10) for patients waiting for root canal treatments are found to be approximately normally distributed with a mean of 7.59 and a standard deviation of .73.

- What percentage of such patients have a stress score lower than 6.0?
- What is the probability that a randomly selected root canal patient sitting in the waiting room has a stress score between 7.0 and 8.0?
- The psychologist suggests that any patient with a stress score of 9.0 or higher should be given a sedative prior to treatment. What percentage of patients waiting for root canal treatments would need a sedative if this suggestion is accepted?

**6.49** According to the U.S. Department of Agriculture, the average American consumed 54.3 pounds (approximately seven gallons) of salad and cooking oils in 2008 ([www.ers.usda.gov/data/foodconsumption](http://www.ers.usda.gov/data/foodconsumption)). Suppose that the current distribution of salad and cooking oil consumption is approximately normally distributed with a mean of 54.3 pounds and a standard deviation of 14.5 pounds. What percentage of Americans' annual salad and cooking oil consumption is

- |  |  |
|--|--|
| <ol style="list-style-type: none"> <li>less than 10 pounds</li> <li>more than 90 pounds</li> </ol> | <ol style="list-style-type: none"> <li>between 40 and 60 pounds</li> <li>between 50 and 70 pounds</li> </ol> |
|--|--|

**6.50** Fast Auto Service guarantees that the maximum waiting time for its customers is 20 minutes for oil and lube service on their cars. It also guarantees that any customer who has to wait longer than 20 minutes for this service will receive a 50% discount on the charges. It is estimated that the mean time taken for oil and lube service at this garage is 15 minutes per car and the standard deviation is 2.4 minutes. Suppose the time taken for oil and lube service on a car follows a normal distribution.

- What percentage of customers will receive a 50% discount on their charges?
- Is it possible that it may take longer than 25 minutes for oil and lube service? Explain.

**6.51** The lengths of 3-inch nails manufactured on a machine are normally distributed with a mean of 3.0 inches and a standard deviation of .009 inch. The nails that are either shorter than 2.98 inches or longer than 3.02 inches are unusable. What percentage of all the nails produced by this machine are unusable?

**6.52** The pucks used by the National Hockey League for ice hockey must weigh between 5.5 and 6.0 ounces. Suppose the weights of pucks produced at a factory are normally distributed with a mean of 5.75 ounces and a standard deviation of .11 ounce. What percentage of the pucks produced at this factory cannot be used by the National Hockey League?

## 6.4 Determining the $z$ and $x$ Values When an Area Under the Normal Distribution Curve Is Known

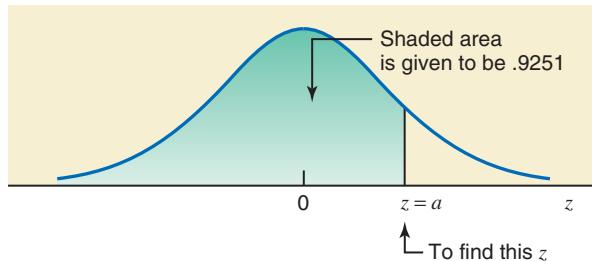
So far in this chapter we have discussed how to find the area under a normal distribution curve for an interval of  $z$  or  $x$ . Now we invert this procedure and learn how to find the corresponding value of  $z$  or  $x$  when an area under a normal distribution curve is known. Examples 6–15 through 6–17 describe this procedure for finding the  $z$  value.

### ■ EXAMPLE 6–15

Find the value of  $z$  such that the area under the standard normal curve to the left of  $z$  is .9251.

**Solution** As shown in Figure 6.45, we are to find the  $z$  value such that the area to the left of  $z$  is .9251. Since this area is greater than .50,  $z$  is positive and lies to the right of zero.

*Finding  $z$  when the area to the left of  $z$  is known.*



**Figure 6.45** Finding the  $z$  value.

To find the required value of  $z$ , we locate .9251 in the body of the normal distribution table, Table IV of Appendix C. The relevant portion of that table is reproduced as Table 6.4 here. Next we read the numbers in the column and row for  $z$  that correspond to .9251. As shown in Table 6.4, these numbers are 1.4 and .04, respectively. Combining these two numbers, we obtain the required value of  $z = 1.44$ .

**Table 6.4** Finding the  $z$  Value When Area Is Known

$z$	.00	.01	...	.04	...	.09
-3.4	.0003	.0003	...		...	.0002
-3.3	.0005	.0005	...		...	.0003
-3.2	.0007	.0007	...		...	.0005
•	•	•	...		...	•
•	•	•	...		...	•
•	•	•	...		...	•
1.4	1.4			.9251		
•	•	•	...	•	...	•
•	•	•	...	•	...	•
•	•	•	...	•	...	•
3.4	.9997	.9997	...	.9997	...	.9998

We locate this value in Table IV of Appendix C

### ■ EXAMPLE 6–16

Find the value of  $z$  such that the area under the standard normal curve in the right tail is .0050.

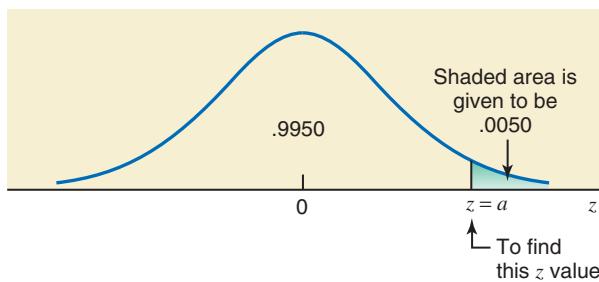
**Solution** To find the required value of  $z$ , we first find the area to the left of  $z$ . Hence,

$$\text{Area to the left of } z = 1.0 - .0050 = .9950$$

This area is shown in Figure 6.46.

Now we look for .9950 in the body of the normal distribution table. Table IV does not contain .9950. So we find the value closest to .9950, which is either .9949 or .9951. We can use either of these two values. If we choose .9951, the corresponding  $z$  value is 2.58. Hence, the required value of  $z$  is **2.58**, and the area to the right of  $z = 2.58$  is approximately .0050. Note that there is no apparent reason to choose .9951 and not to choose .9949. We can use either of the two values. If we choose .9949, the corresponding  $z$  value will be 2.57.

*Finding  $z$  when the area in the right tail is known.*

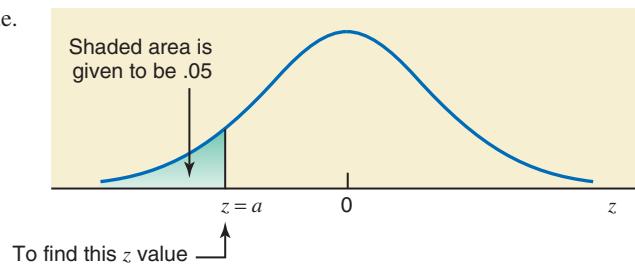
**Figure 6.46** Finding the  $z$  value.

*Finding  $z$  when the area in the left tail is known.*

### ■ EXAMPLE 6-17

Find the value of  $z$  such that the area under the standard normal curve in the left tail is .05.

**Solution** Because .05 is less than .5 and it is the area in the left tail, the value of  $z$  is negative. This area is shown in Figure 6.47.

**Figure 6.47** Finding the  $z$  value.

Next, we look for .0500 in the body of the normal distribution table. The value closest to .0500 in the normal distribution table is either .0505 or .0495. Suppose we use the value .0495. The corresponding  $z$  value is  $-1.65$ . Thus, the required value of  $z$  is  **$-1.65$**  and the area to the left of  $z = -1.65$  is approximately .05.

To find an  $x$  value when an area under a normal distribution curve is given, first we find the  $z$  value corresponding to that  $x$  value from the normal distribution table. Then, to find the  $x$  value, we substitute the values of  $\mu$ ,  $\sigma$ , and  $z$  in the following formula, which is obtained from  $z = (x - \mu)/\sigma$  by doing some algebraic manipulations. Also, if we know the values of  $x$ ,  $z$ , and  $\sigma$ , we can find  $\mu$  using this same formula. Exercises 6.63 and 6.64 present such cases.

**Finding an  $x$  Value for a Normal Distribution** For a normal curve, with known values of  $\mu$  and  $\sigma$  and for a given area under the curve to the left of  $x$ , the  $x$  value is calculated as

$$x = \mu + z\sigma$$

Examples 6-18 and 6-19 illustrate how to find an  $x$  value when an area under a normal distribution curve is known.

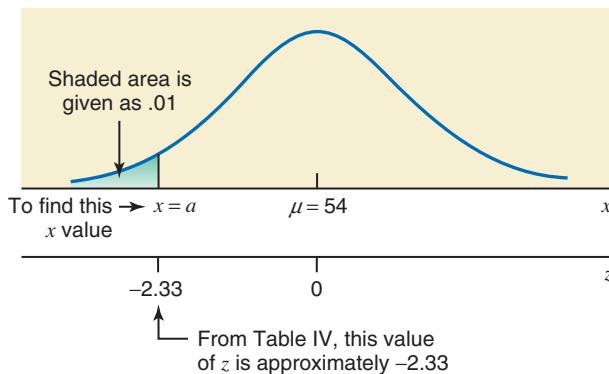
### ■ EXAMPLE 6-18

*Finding  $x$  when the area in the left tail is known.*

Recall Example 6-14. It is known that the life of a calculator manufactured by Calculators Corporation has a normal distribution with a mean of 54 months and a standard deviation of 8 months. What should the warranty period be to replace a malfunctioning calculator if the company does not want to replace more than 1% of all the calculators sold?

**Solution** Let  $x$  be the life of a calculator. Then,  $x$  follows a normal distribution with  $\mu = 54$  months and  $\sigma = 8$  months. The calculators that would be replaced are the ones that

start malfunctioning during the warranty period. The company's objective is to replace at most 1% of all the calculators sold. The shaded area in Figure 6.48 gives the proportion of calculators that are replaced. We are to find the value of  $x$  so that the area to the left of  $x$  under the normal curve is 1%, or .01.



**Figure 6.48** Finding an  $x$  value.

In the first step, we find the  $z$  value that corresponds to the required  $x$  value.

We find the  $z$  value from the normal distribution table for  $.0100$ . Table IV of Appendix C does not contain a value that is exactly  $.0100$ . The value closest to  $.0100$  in the table is  $.0099$ , and the  $z$  value for  $.0099$  is  $-2.33$ . Hence,

$$z = -2.33$$

Substituting the values of  $\mu$ ,  $\sigma$ , and  $z$  in the formula  $x = \mu + z\sigma$ , we obtain

$$x = \mu + z\sigma = 54 + (-2.33)(8) = 54 - 18.64 = 35.36$$

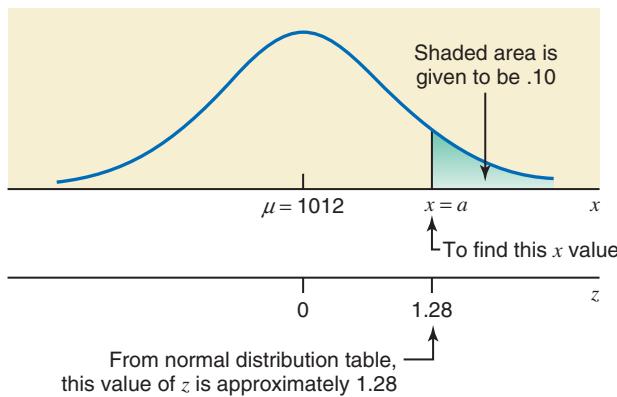
Thus, the company should replace all the calculators that start to malfunction within 35.36 months (which can be rounded to 35 months) of the date of purchase so that they will not have to replace more than 1% of the calculators. ■

### ■ EXAMPLE 6-19

According to the College Board, the mean combined (mathematics and critical reading) SAT score for all college-bound seniors was 1012 with a standard deviation of 213 in 2011 ([http://media.collegeboard.com/digitalServices/pdf/SAT-Percentile-Ranks-Composite-CR-M\\_2011.pdf](http://media.collegeboard.com/digitalServices/pdf/SAT-Percentile-Ranks-Composite-CR-M_2011.pdf)). Suppose that the current distribution of combined SAT scores for all college-bound seniors is approximately normal with a mean of 1012 and a standard deviation of 213. Jennifer is one of the college-bound seniors who took this test. It is found that 10% of all current college-bound seniors have SAT scores higher than Jennifer. What is Jennifer's SAT score?

Finding  $x$  when the area in the right tail is known.

**Solution** Let  $x$  represent the combined SAT scores of examinees. Then,  $x$  follows a normal distribution with  $\mu = 1012$  and  $\sigma = 213$ . We are to find the value of  $x$  such that the area under the normal distribution curve to the right of  $x$  is 10%, as shown in Figure 6.49.



**Figure 6.49** Finding an  $x$  value.

First, we find the area under the normal distribution curve to the left of the  $x$  value.

$$\text{Area to the left of the } x \text{ value} = 1.0 - .10 = .9000$$

To find the  $z$  value that corresponds to the required  $x$  value, we look for .9000 in the body of the normal distribution table. The value closest to .9000 in Table IV is .8997, and the corresponding  $z$  value is 1.28. Hence, the value of  $x$  is computed as

$$x = \mu + z\sigma = 1012 + 1.28(213) = 1012 + 272.64 = 1284.64 \approx \mathbf{1285}$$

Thus, Jennifer's combined SAT score is 1285. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**6.53** Find the value of  $z$  so that the area under the standard normal curve

- a. from 0 to  $z$  is .4772 and  $z$  is positive
- b. between 0 and  $z$  is (approximately) .4785 and  $z$  is negative
- c. in the left tail is (approximately) .3565
- d. in the right tail is (approximately) .1530

**6.54** Find the value of  $z$  so that the area under the standard normal curve

- a. from 0 to  $z$  is (approximately) .1965 and  $z$  is positive
- b. between 0 and  $z$  is (approximately) .2740 and  $z$  is negative
- c. in the left tail is (approximately) .2050
- d. in the right tail is (approximately) .1053

**6.55** Determine the value of  $z$  so that the area under the standard normal curve

- a. in the right tail is .0500      b. in the left tail is .0250
- c. in the left tail is .0100      d. in the right tail is .0050

**6.56** Determine the value of  $z$  so that the area under the standard normal curve

- a. in the right tail is .0250      b. in the left tail is .0500
- c. in the left tail is .0010      d. in the right tail is .0100

**6.57** Let  $x$  be a continuous random variable that follows a normal distribution with a mean of 200 and a standard deviation of 25.

- a. Find the value of  $x$  so that the area under the normal curve to the left of  $x$  is approximately .6330.
- b. Find the value of  $x$  so that the area under the normal curve to the right of  $x$  is approximately .05.
- c. Find the value of  $x$  so that the area under the normal curve to the right of  $x$  is .8051.
- d. Find the value of  $x$  so that the area under the normal curve to the left of  $x$  is .0150.
- e. Find the value of  $x$  so that the area under the normal curve between  $\mu$  and  $x$  is .4525 and the value of  $x$  is less than  $\mu$ .
- f. Find the value of  $x$  so that the area under the normal curve between  $\mu$  and  $x$  is approximately .4800 and the value of  $x$  is greater than  $\mu$ .

**6.58** Let  $x$  be a continuous random variable that follows a normal distribution with a mean of 550 and a standard deviation of 75.

- a. Find the value of  $x$  so that the area under the normal curve to the left of  $x$  is .0250.
- b. Find the value of  $x$  so that the area under the normal curve to the right of  $x$  is .9345.
- c. Find the value of  $x$  so that the area under the normal curve to the right of  $x$  is approximately .0275.
- d. Find the value of  $x$  so that the area under the normal curve to the left of  $x$  is approximately .9600.
- e. Find the value of  $x$  so that the area under the normal curve between  $\mu$  and  $x$  is approximately .4700 and the value of  $x$  is less than  $\mu$ .
- f. Find the value of  $x$  so that the area under the normal curve between  $\mu$  and  $x$  is approximately .4100 and the value of  $x$  is greater than  $\mu$ .

### APPLICATIONS

**6.59** Fast Auto Service provides oil and lube service for cars. It is known that the mean time taken for oil and lube service at this garage is 15 minutes per car and the standard deviation is 2.4 minutes. The management wants to promote the business by guaranteeing a maximum waiting time for its customers. If a customer's car

is not serviced within that period, the customer will receive a 50% discount on the charges. The company wants to limit this discount to at most 5% of the customers. What should the maximum guaranteed waiting time be? Assume that the times taken for oil and lube service for all cars have a normal distribution.

**6.60** The management of a supermarket wants to adopt a new promotional policy of giving a free gift to every customer who spends more than a certain amount per visit at this supermarket. The expectation of the management is that after this promotional policy is advertised, the expenditures for all customers at this supermarket will be normally distributed with a mean of \$95 and a standard deviation of \$20. If the management wants to give free gifts to at most 10% of the customers, what should the amount be above which a customer would receive a free gift?

**6.61** According to the records of an electric company serving the Boston area, the mean electricity consumption during winter for all households is 1650 kilowatt-hours per month. Assume that the monthly electric consumptions during winter by all households in this area have a normal distribution with a mean of 1650 kilowatt-hours and a standard deviation of 320 kilowatt-hours. The company sent a notice to Bill Johnson informing him that about 90% of the households use less electricity per month than he does. What is Bill Johnson's monthly electricity consumption?

**6.62** Rockingham Corporation makes electric shavers. The life (period during which a shaver does not need a major repair) of Model J795 of an electric shaver manufactured by this corporation has a normal distribution with a mean of 70 months and a standard deviation of 8 months. The company is to determine the warranty period for this shaver. Any shaver that needs a major repair during this warranty period will be replaced free by the company.

- What should the warranty period be if the company does not want to replace more than 1% of the shavers?
- What should the warranty period be if the company does not want to replace more than 5% of the shavers?

**\*6.63** A study has shown that 20% of all college textbooks have a price of \$184.52 or higher. It is known that the standard deviation of the prices of all college textbooks is \$36.35. Suppose the prices of all college textbooks have a normal distribution. What is the mean price of all college textbooks?

**\*6.64** A machine at Keats Corporation fills 64-ounce detergent jugs. The machine can be adjusted to pour, on average, any amount of detergent into these jugs. However, the machine does not pour exactly the same amount of detergent into each jug; it varies from jug to jug. It is known that the net amount of detergent poured into each jug has a normal distribution with a standard deviation of .35 ounce. The quality control inspector wants to adjust the machine such that at least 95% of the jugs have more than 64 ounces of detergent. What should the mean amount of detergent poured by this machine into these jugs be?

## 6.5 The Normal Approximation to the Binomial Distribution

Recall from Chapter 5 that:

- The binomial distribution is applied to a discrete random variable.
- Each repetition, called a trial, of a binomial experiment results in one of two possible outcomes (or events), either a success or a failure.
- The probabilities of the two (possible) outcomes (or events) remain the same for each repetition of the experiment.
- The trials are independent.

The binomial formula, which gives the probability of  $x$  successes in  $n$  trials, is

$$P(x) = {}_nC_x p^x q^{n-x}$$

The use of the binomial formula becomes very tedious when  $n$  is large. In such cases, the normal distribution can be used to approximate the binomial probability. Note that for a binomial problem, the exact probability is obtained by using the binomial formula. If we apply the normal distribution to solve a binomial problem, the probability that we obtain is an approximation to the exact probability. The approximation obtained by using the normal distribution is very close to the exact probability when  $n$  is large and  $p$  is very close to .50. However, this does not

mean that we should not use the normal approximation when  $p$  is not close to .50. The reason the approximation is closer to the exact probability when  $p$  is close to .50 is that the binomial distribution is symmetric when  $p = .50$ . The normal distribution is always symmetric. Hence, the two distributions are very close to each other when  $n$  is large and  $p$  is close to .50. However, this does not mean that whenever  $p = .50$ , the binomial distribution is the same as the normal distribution because not every symmetric bell-shaped curve is a normal distribution curve.

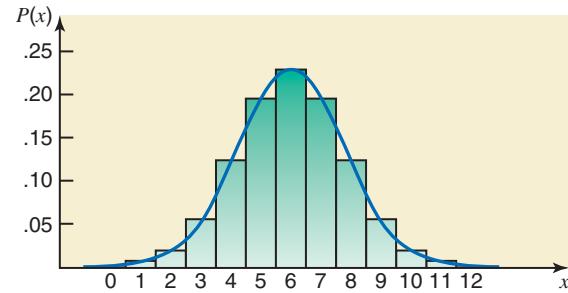
**Normal Distribution as an Approximation to Binomial Distribution** Usually, the normal distribution is used as an approximation to the binomial distribution when  $np$  and  $nq$  are both greater than 5, that is, when

$$np > 5 \quad \text{and} \quad nq > 5$$

Table 6.5 gives the binomial probability distribution of  $x$  for  $n = 12$  and  $p = .50$ . This table is written using Table I of Appendix C. Figure 6.50 shows the histogram and the smoothed polygon for the probability distribution of Table 6.5. As we can observe, the histogram in Figure 6.50 is symmetric, and the curve obtained by joining the upper midpoints of the rectangles is approximately bell shaped.

**Table 6.5** The Binomial Probability Distribution for  $n = 12$  and  $p = .50$

$x$	$P(x)$
0	.0002
1	.0029
2	.0161
3	.0537
4	.1208
5	.1934
6	.2256
7	.1934
8	.1208
9	.0537
10	.0161
11	.0029
12	.0002



**Figure 6.50** Histogram for the probability distribution of Table 6.5.

Examples 6–20 through 6–22 illustrate the application of the normal distribution as an approximation to the binomial distribution.

## ■ EXAMPLE 6–20

*Using the normal approximation to the binomial:  $x$  equals a specific value.*

According to an estimate, 50% of people in the United States have at least one credit card. If a random sample of 30 persons is selected, what is the probability that 19 of them will have at least one credit card?

**Solution** Let  $n$  be the total number of persons in the sample,  $x$  be the number of persons in the sample who have at least one credit card, and  $p$  be the probability that a person has at least one credit card. Then, this is a binomial problem with

$$n = 30, \quad p = .50, \quad q = 1 - p = .50,$$

$$x = 19, \quad n - x = 30 - 19 = 11$$

Using the binomial formula, the exact probability that 19 persons in a sample of 30 have at least one credit card is

$$P(19) = {}_{30}C_{19}(.50)^{19}(.50)^{11} = .0509$$

Now let us solve this problem using the normal distribution as an approximation to the binomial distribution. For this example,

$$np = 30(.50) = 15 \quad \text{and} \quad nq = 30(.50) = 15$$

Because  $np$  and  $nq$  are both greater than 5, we can use the normal distribution as an approximation to solve this binomial problem. We perform the following three steps.

**Step 1.** Compute  $\mu$  and  $\sigma$  for the binomial distribution.

To use the normal distribution, we need to know the mean and standard deviation of the distribution. Hence, the first step in using the normal approximation to the binomial distribution is to compute the mean and standard deviation of the binomial distribution. As we know from Chapter 5, the mean and standard deviation of a binomial distribution are given by  $np$  and  $\sqrt{npq}$ , respectively. Using these formulas, we obtain

$$\begin{aligned}\mu &= np = 30(.50) = 15 \\ \sigma &= \sqrt{npq} = \sqrt{30(.50)(.50)} = 2.73861279\end{aligned}$$

**Step 2.** Convert the discrete random variable into a continuous random variable.

The normal distribution applies to a continuous random variable, whereas the binomial distribution applies to a discrete random variable. The second step in applying the normal approximation to the binomial distribution is to convert the discrete random variable to a continuous random variable by making the **correction for continuity**.

### Definition

**Continuity Correction Factor** The addition of .5 and/or subtraction of .5 from the value(s) of  $x$  when the normal distribution is used as an approximation to the binomial distribution, where  $x$  is the number of successes in  $n$  trials, is called the *continuity correction factor*.

As shown in Figure 6.51, the probability of 19 successes in 30 trials is given by the area of the rectangle for  $x = 19$ . To make the correction for continuity, we use the interval 18.5 to 19.5 for 19 persons. This interval is actually given by the two boundaries of the rectangle for  $x = 19$ , which are obtained by subtracting .5 from 19 and by adding .5 to 19. Thus,  $P(x = 19)$  for the binomial problem will be approximately equal to  $P(18.5 \leq x \leq 19.5)$  for the normal distribution.

The area contained by the rectangle for  $x = 19$  is approximated by the area under the curve between 18.5 and 19.5.

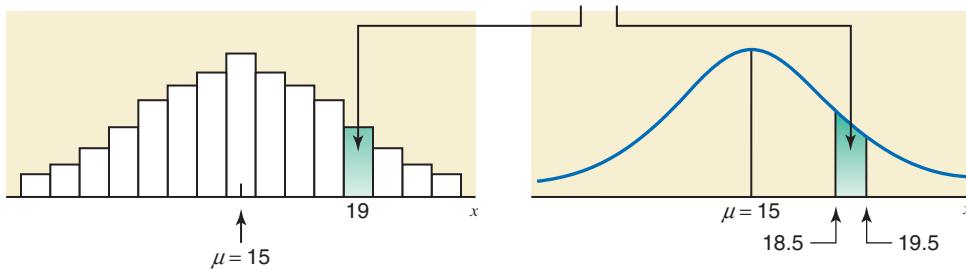


Figure 6.51

**Step 3.** Compute the required probability using the normal distribution.

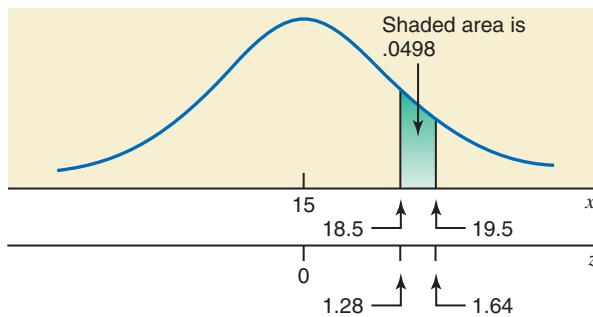
As shown in Figure 6.52, the area under the normal distribution curve between  $x = 18.5$  and  $x = 19.5$  will give us the (approximate) probability that 19 persons have at least one credit

card. We calculate this probability as follows:

$$\text{For } x = 18.5: z = \frac{18.5 - 15}{2.73861279} = 1.28$$

$$\text{For } x = 19.5: z = \frac{19.5 - 15}{2.73861279} = 1.64$$

**Figure 6.52** Area between  $x = 18.5$  and  $x = 19.5$ .



The required probability is given by the area under the standard normal curve between  $z = 1.28$  and  $z = 1.64$ . This area is obtained by subtracting the area to the left of  $z = 1.28$  from the area to the left of  $z = 1.64$ . From Table IV of Appendix C, the area to the left of  $z = 1.28$  is .8997 and the area to the left of  $z = 1.64$  is .9495. Hence, the required probability is

$$P(18.5 \leq x \leq 19.5) = P(1.28 \leq z \leq 1.64) = .9495 - .8997 = .0498$$

Thus, based on the normal approximation, the probability that 19 persons in a sample of 30 will have at least one credit card is approximately .0498. Earlier, using the binomial formula, we obtained the exact probability .0509. The error due to using the normal approximation is  $.0509 - .0498 = .0011$ . Thus, the exact probability is underestimated by .0011 if the normal approximation is used. ■

**Remember ►**

When applying the normal distribution as an approximation to the binomial distribution, always make a *correction for continuity*. The continuity correction is made by subtracting .5 from the lower limit of the interval and/or by adding .5 to the upper limit of the interval. For example, the binomial probability  $P(7 \leq x \leq 12)$  will be approximated by the probability  $P(6.5 \leq x \leq 12.5)$  for the normal distribution; the binomial probability  $P(x \geq 9)$  will be approximated by the probability  $P(x \geq 8.5)$  for the normal distribution; and the binomial probability  $P(x \leq 10)$  will be approximated by the probability  $P(x \leq 10.5)$  for the normal distribution. Note that the probability  $P(x \geq 9)$  has only the lower limit of 9 and no upper limit, and the probability  $P(x \leq 10)$  has only the upper limit of 10 and no lower limit.

### ■ EXAMPLE 6-21

**Using the normal approximation to the binomial:  $x$  assumes a value in an interval.**

According to an Arise Virtual Solutions Job survey, 32% of people working from home said that the biggest advantage of working from home is that there is no commute (*USA TODAY*, October 7, 2011). Suppose that this result is true for the current population of people who work from home. What is the probability that in a random sample of 400 people who work from home, 108 to 122 will say that the biggest advantage of working from home is that there is no commute?

**Solution** Let  $n$  be the number of people who work from home in the sample,  $x$  be the number of people in the sample who say that the biggest advantage of working from home is that there is no commute, and  $p$  be the probability that a person who works from home says that the biggest advantage of working from home is that there is no commute. Then, this is a binomial problem with

$$n = 400, \quad p = .32, \quad \text{and} \quad q = 1 - .32 = .68$$

We are to find the probability of 108 to 122 successes in 400 trials. Because  $n$  is large, it is easier to apply the normal approximation than to use the binomial formula. We can check that  $np$  and  $nq$  are both greater than 5. The mean and standard deviation of the binomial distribution are, respectively,

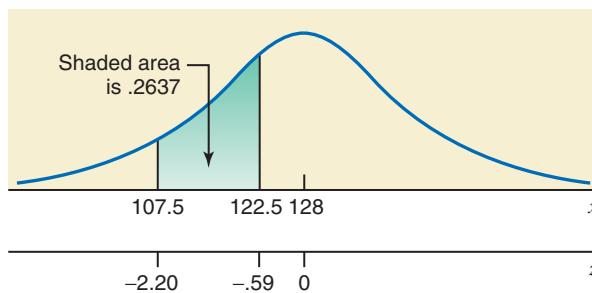
$$\mu = np = 400 (.32) = 128$$

$$\sigma = \sqrt{npq} = \sqrt{400(.32)(.68)} = 9.32952303$$

To make the continuity correction, we subtract .5 from 108 and add .5 to 122 to obtain the interval 107.5 to 122.5. Thus, the probability that 108 to 122 out of a sample of 400 people who work from home will say that the biggest advantage of working from home is that there is no commute is approximated by the area under the normal distribution curve from  $x = 107.5$  to  $x = 122.5$ . This area is shown in Figure 6.53. The  $z$  values for  $x = 107.5$  and  $x = 122.5$  are calculated as follows:

$$\text{For } x = 107.5: z = \frac{107.5 - 128}{9.32952303} = -2.20$$

$$\text{For } x = 122.5: z = \frac{122.5 - 128}{9.32952303} = -.59$$



**Figure 6.53** Area between  $x = 107.5$  and  $x = 122.5$ .

The required probability is given by the area under the standard normal curve between  $z = -2.20$  and  $z = -.59$ . This area is obtained by taking the difference between the areas under the standard normal curve to the left of  $z = -2.20$  and to the left of  $z = -.59$ . From Table IV of Appendix C, the area to the left of  $z = -2.20$  is .0139, and the area to the left of  $z = -.59$  is .2776. Hence, the required probability is

$$P(107.5 \leq x \leq 122.5) = P(-2.20 \leq z \leq -.59) = .2776 - .0139 = .2637$$

Thus, the probability that 108 to 122 people in a sample of 400 who work from home will say that the biggest advantage of working from home is that there is no commute is approximately .2637. ■

## EXAMPLE 6-22

According to a poll, 55% of American adults do not know that GOP stands for Grand Old Party (*Time*, October 17, 2011). Assume that this percentage is true for the current population of American adults. What is the probability that 397 or more American adults in a random sample of 700 do not know that GOP stands for Grand Old Party?

Using the normal approximation to the binomial:  $x$  is greater than or equal to a value.

**Solution** Let  $n$  be the sample size,  $x$  be the number of American adults in the sample who do not know that GOP stands for Grand Old Party, and  $p$  be the probability that a randomly selected American adult does not know that GOP stands for Grand Old Party. Then, this is a binomial problem with

$$n = 700, \quad p = .55, \quad \text{and} \quad q = 1 - .55 = .45$$

We are to find the probability of 397 or more successes in 700 trials. The mean and standard deviation of the binomial distribution are, respectively,

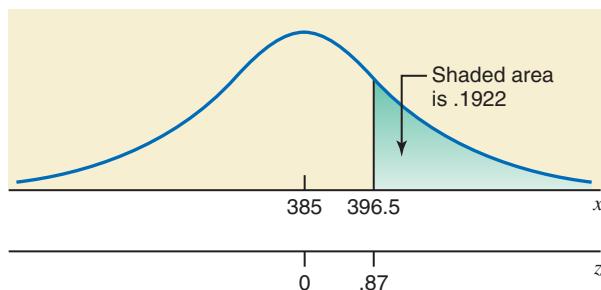
$$\mu = np = 700 (.55) = 385$$

$$\sigma = \sqrt{npq} = \sqrt{700(.55)(.45)} = 13.16244658$$

For the continuity correction, we subtract .5 from 397, which gives 396.5. Thus, the probability that 397 or more American adults in a random sample of 700 do not know that GOP stands for Grand Old Party is approximated by the area under the normal distribution curve to the right of  $x = 396.5$ , as shown in Figure 6.54. The  $z$  value for  $x = 396.5$  is calculated as follows.

$$\text{For } x = 396.5: \quad z = \frac{396.5 - 385}{13.16244658} = .87$$

**Figure 6.54** Area to the right of  $x = 396.5$ .



To find the required probability, we find the area to the left of  $z = .87$  and subtract this area from 1.0. From Table IV of Appendix C, the area to the left of  $z = .87$  is .8078. Hence,

$$P(x \geq 396.5) = P(z \geq .87) = 1.0 - .8078 = .1922$$

Thus, the probability that 397 or more American adults in a random sample of 700 will not know that GOP stands for Grand Old Party is approximately .1922. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

- 6.65** Under what conditions is the normal distribution usually used as an approximation to the binomial distribution?
- 6.66** For a binomial probability distribution,  $n = 20$  and  $p = .60$ .
- Find the probability  $P(x = 14)$  by using the table of binomial probabilities (Table I of Appendix C).
  - Find the probability  $P(x = 14)$  by using the normal distribution as an approximation to the binomial distribution. What is the difference between this approximation and the exact probability calculated in part a?
- 6.67** For a binomial probability distribution,  $n = 25$  and  $p = .40$ .
- Find the probability  $P(8 \leq x \leq 13)$  by using the table of binomial probabilities (Table I of Appendix C).
  - Find the probability  $P(8 \leq x \leq 13)$  by using the normal distribution as an approximation to the binomial distribution. What is the difference between this approximation and the exact probability calculated in part a?
- 6.68** For a binomial probability distribution,  $n = 80$  and  $p = .50$ . Let  $x$  be the number of successes in 80 trials.
- Find the mean and standard deviation of this binomial distribution.
  - Find  $P(x \geq 42)$  using the normal approximation.
  - Find  $P(41 \leq x \leq 48)$  using the normal approximation.
- 6.69** For a binomial probability distribution,  $n = 120$  and  $p = .60$ . Let  $x$  be the number of successes in 120 trials.
- Find the mean and standard deviation of this binomial distribution.
  - Find  $P(x \leq 69)$  using the normal approximation.
  - Find  $P(67 \leq x \leq 73)$  using the normal approximation.

**6.70** Find the following binomial probabilities using the normal approximation.

- a.  $n = 140, p = .45, P(x = 67)$
- b.  $n = 100, p = .55, P(52 \leq x \leq 60)$
- c.  $n = 90, p = .42, P(x \geq 40)$
- d.  $n = 104, p = .75, P(x \leq 72)$

**6.71** Find the following binomial probabilities using the normal approximation.

- a.  $n = 70, p = .30, P(x = 18)$
- b.  $n = 200, p = .70, P(133 \leq x \leq 145)$
- c.  $n = 85, p = .40, P(x \geq 30)$
- d.  $n = 150, p = .38, P(x \leq 62)$

## ■ APPLICATIONS

**6.72** According to the U.S. Census American Community Survey, 5.44% of workers in Portland, Oregon, commute to work on their bicycles. (Note: this is the highest percentage among all U.S. cities having at least 250,000 workers.) Find the probability that in a sample of 400 workers from Portland, Oregon, the number who commute to work on their bicycles is 23 to 27.

**6.73** According to an Allstate/National Journal poll, 39% of the U.S. adults polled said that it is *extremely or very likely* that “there will be a female president within 10–15 years” in the United States (*USA Today*, March 28, 2012). Suppose that this percentage is true for the current population of U.S. adults. Find the probability that in a random sample of 800 U.S. adults, more than 330 would hold the foregoing belief.

**6.74** The percentage of women in the work force has increased tremendously during the past few decades. Whereas only 35% of all employees in the United States in 1970 were women, this percentage is now 49% (*Bloomberg Businessweek*, January 9–January 15, 2012). Suppose that currently 49% of all employees in the United States are women. Find the probability that in a random sample of 400 employees, the number of women is

- a. exactly 205
- b. less than 190
- c. 210 to 220

**6.75** According to a November 8, 2010 report on [www.teleread.com](http://www.teleread.com), 7% of U.S. adults with online services currently read e-books. Assume that this percentage is true for the current population of U.S. adults with online services. Find the probability that in a random sample of 600 U.S. adults with online services, the number who read e-books is

- a. exactly 45
- b. at most 53
- c. 30 to 50

**6.76** According to a Gallup poll, 92% of Americans believe in God (*Time*, June 20, 2011). Suppose that this result is true for the current population of adult Americans. What is the probability that the number of adult Americans in a sample of 500 who believe in God is

- a. exactly 445
- b. at least 450
- c. 440 to 470

**6.77** An office supply company conducted a survey before marketing a new paper shredder designed for home use. In the survey, 80% of the people who tried the shredder were satisfied with it. Because of this high satisfaction rate, the company decided to market the new shredder. Assume that 80% of all people are satisfied with this shredder. During a certain month, 100 customers bought this shredder. Find the probability that of these 100 customers, the number who are satisfied is

- a. exactly 75
- b. 73 or fewer
- c. 74 to 85

**6.78** Johnson Electronics makes calculators. Consumer satisfaction is one of the top priorities of the company’s management. The company guarantees the refund of money or a replacement for any calculator that malfunctions within two years from the date of purchase. It is known from past data that despite all efforts, 5% of the calculators manufactured by this company malfunction within a 2-year period. The company recently mailed 500 such calculators to its customers.

- a. Find the probability that exactly 29 of the 500 calculators will be returned for refund or replacement within a 2-year period.
- b. What is the probability that 27 or more of the 500 calculators will be returned for refund or replacement within a 2-year period?
- c. What is the probability that 15 to 22 of the 500 calculators will be returned for refund or replacement within a 2-year period?

**6.79** Hurbert Corporation makes font cartridges for laser printers that it sells to Alpha Electronics Inc. The cartridges are shipped to Alpha Electronics in large volumes. The quality control department at

Alpha Electronics randomly selects 100 cartridges from each shipment and inspects them for being good or defective. If this sample contains 7 or more defective cartridges, the entire shipment is rejected. Hurbert Corporation promises that of all the cartridges, only 5% are defective.

- a. Find the probability that a given shipment of cartridges received by Alpha Electronics will be accepted.
- b. Find the probability that a given shipment of cartridges received by Alpha Electronics will not be accepted.

## USES AND MISUSES...

### (1) DON'T LOSE YOUR MEMORY

As discussed in the previous chapter, the Poisson distribution gives the probability of a specified number of events occurring in a time interval. The Poisson distribution provides a model for the number of emails a server might receive during a certain time period or the number of people arriving in line at a bank during lunch hour. These are nice to know for planning purposes, but sometimes we want to know the specific times at which emails or customers arrive. These times are governed by a special continuous probability distribution with certain unusual properties. This distribution is called the *exponential distribution*, and it is derived from the Poisson probability distribution.

Suppose you are a teller at a bank, and a customer has just arrived. You know that the customers arrive according to a Poisson process with a rate of  $\lambda$  customers per hour. Your boss might care how many customers arrive on average during a given time interval to ensure there are enough tellers available to handle the customers efficiently; you are more concerned with the time when the next customer will arrive. Remember that the probability that  $x$  customers arrive in an interval of length  $t$  is

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}.$$

The probability that at least one customer arrives within time  $t$  is 1 minus the probability that no customer arrives within time  $t$ . Hence,

$$\begin{aligned} P(\text{at least one customer arrives within time } t) &= 1 - P(0) \\ &= 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \\ &= 1 - e^{-\lambda t} \end{aligned}$$

If the bank receives an average of 15 customers per hour—an average of one every 4 minutes—and a customer has just arrived, the probability that a customer arrives within 4 minutes is  $1 - e^{-\lambda t} = 1 - e^{-(15/60)4} = .6321$ . In the same way, the probability that a customer arrives within 8 minutes is .8647.

Let us say that a customer arrived and went to your co-worker's window. No additional customer arrived within the next 2 minutes—an event with probability .6065—and you dozed off for 2 more minutes. When you open your eyes, you see that a customer has not arrived yet. What is the probability that a customer arrives within the next 4 minutes? From the calculation above, you might say that the answer is .8647. After all, you know that a customer arrived 8 minutes earlier. But .8647 is not the correct answer.

The exponential distribution, which governs the time between arrivals of a Poisson process, has a property called the *memoryless* property. For you as a bank teller, this means that if you know a customer

has not arrived during the past 4 minutes, then the clock is reset to zero, as if the previous customer had just arrived. So even after your nap, the probability that a customer arrives within 4 minutes is .6321. This interesting property reminds us again that we should be careful when we use mathematics to model real-world phenomena.

### (2) QUALITY IS JOB 1

During the early 1980s, Ford Motor Company adopted a new marketing slogan: "Quality Is Job 1." The new slogan coincided with Ford's release of the Taurus and the Mercury Sable, but the groundwork that resulted in Ford's sudden need to improve quality was laid some 30 years earlier by an American statistician—not in Detroit, but in Japan.

W. Edwards Deming is one of the most famous statisticians, if not the most famous statistician, in the field of statistical process control, a field that debunked the myth that you could not improve quality and lower costs simultaneously. After World War II, Deming was asked by the U.S. Armed Forces to assist with planning for the 1951 census in Japan. While there, he taught statistical process control and quality management to the managers and engineers at many of Japan's largest companies. After adopting Deming's principles, the quality of and the demand for Japanese products, including Japanese automobiles, increased tremendously.

Ford's interest in Japanese quality resulted from the fact that Ford was having a specific transmission produced simultaneously in Japan and the United States. Ford's U.S. customers were requesting cars with Japanese-produced transmissions, even if it required them to wait longer for the car. Despite the fact that the transmissions were made to the same specifications in the two countries, the parts used in the Japanese transmissions were much closer to the desired size than those used in the transmissions made in America. Knowing that Deming had done a great deal of work with Japanese companies, Ford hired him as a consultant. The result of Deming's work in statistical process control and proper management methods, along with Ford's willingness to implement his recommendations, was the production of Ford's Taurus and Mercury Sable lines of automobiles, which resulted in Ford earning a profit after numerous years of losses.

W. Edwards Deming died in 1993 at the age of 93 years, but he left a legacy. Japan introduced the Deming Prize in 1950. The Deming Prize is awarded annually to individuals and companies whose work has advanced knowledge in the statistical process control area. More information about Deming and The W. Edwards Deming Institute is available at [www.deming.org](http://www.deming.org).

*Source:* The W. Edwards Deming Institute ([www.deming.org](http://www.deming.org)), and [en.wikipedia.org/wiki/W.\\_Edwards\\_Deming](https://en.wikipedia.org/wiki/W._Edwards_Deming).

## Glossary

**Continuity correction factor** Addition of .5 and/or subtraction of .5 from the value(s) of  $x$  when the normal distribution is used as an approximation to the binomial distribution, where  $x$  is the number of successes in  $n$  trials.

**Continuous random variable** A random variable that can assume any value in one or more intervals.

**Normal probability distribution** The probability distribution of a continuous random variable that, when plotted, gives a specific

bell-shaped curve. The parameters of the normal distribution are the mean  $\mu$  and the standard deviation  $\sigma$ .

**Standard normal distribution** The normal distribution with  $\mu = 0$  and  $\sigma = 1$ . The units of the standard normal distribution are denoted by  $z$ .

**$z$  value or  $z$  score** The units of the standard normal distribution that are denoted by  $z$ .

## Supplementary Exercises

**6.80** The management at Ohio National Bank does not want its customers to wait in line for service for too long. The manager of a branch of this bank estimated that the customers currently have to wait an average of 8 minutes for service. Assume that the waiting times for all customers at this branch have a normal distribution with a mean of 8 minutes and a standard deviation of 2 minutes.

- Find the probability that a randomly selected customer will have to wait for less than 3 minutes.
- What percentage of the customers have to wait for 10 to 13 minutes?
- What percentage of the customers have to wait for 6 to 12 minutes?
- Is it possible that a customer may have to wait longer than 16 minutes for service? Explain.

**6.81** A company that has a large number of supermarket grocery stores claims that customers who pay by personal checks spend an average of \$87 on groceries at these stores with a standard deviation of \$22. Assume that the expenses incurred on groceries by all such customers at these stores are normally distributed.

- Find the probability that a randomly selected customer who pays by check spends more than \$114 on groceries.
- What percentage of customers paying by check spend between \$40 and \$60 on groceries?
- What percentage of customers paying by check spend between \$70 and \$105?
- Is it possible for a customer paying by check to spend more than \$185? Explain.

**6.82** At Jen and Perry Ice Cream Company, the machine that fills 1-pound cartons of Top Flavor ice cream is set to dispense 16 ounces of ice cream into every carton. However, some cartons contain slightly less than and some contain slightly more than 16 ounces of ice cream. The amounts of ice cream in all such cartons have a normal distribution with a mean of 16 ounces and a standard deviation of .18 ounce.

- Find the probability that a randomly selected carton contains 16.20 to 16.50 ounces of ice cream.
- What percentage of such cartons contain less than 15.70 ounces of ice cream?
- Is it possible for a carton to contain less than 15.20 ounces of ice cream? Explain.

**6.83** A machine at Kasem Steel Corporation makes iron rods that are supposed to be 50 inches long. However, the machine does not make all rods of exactly the same length. It is known that the probability distribution of the lengths of rods made on this machine is normal with a mean of 50 inches and a standard deviation of .06 inch. The rods that are either shorter than 49.85 inches or longer than 50.15 inches are discarded. What percentage of the rods made on this machine are discarded?

**6.84** Jenn Bard, who lives in the San Francisco Bay area, commutes by car from home to work. She knows that it takes her an average of 28 minutes for this commute in the morning. However, due to the variability in the traffic situation every morning, the standard deviation of these commutes is 5 minutes. Suppose the population of her morning commute times has a normal distribution with a mean of 28 minutes and a standard deviation of 5 minutes. Jenn has to be at work by 8:30 A.M. every morning. By what time must she leave home in the morning so that she is late for work at most 1% of the time?

**6.85** The print on the package of Sylvania CFL 65W replacement bulbs that use only 16W claims that these bulbs have an average life of 8000 hours. Assume that the distribution of lives of all such bulbs is normal with a mean of 8000 hours and a standard deviation of 400 hours. Let  $x$  be the life of a randomly selected such light bulb.

- Find  $x$  so that about 22.5% of such light bulbs have lives longer than this value.
- Find  $x$  so that about 63% of such light bulbs have lives shorter than this value.

**6.86** Major League Baseball rules require that the balls used in baseball games must have circumferences between 9 and 9.25 inches. Suppose the balls produced by the factory that supplies balls to Major League Baseball have circumferences normally distributed with a mean of 9.125 inches and a standard deviation of .06 inch. What percentage of these baseballs fail to meet the circumference requirement?

**6.87** According to an article on Yahoo.com on February 19, 2012, the average salary of actuaries in the U.S. is \$98,620 a year ([http://education.yahoo.net/articles/careers\\_for\\_shy\\_people\\_2.htm?kid=1KWO3](http://education.yahoo.net/articles/careers_for_shy_people_2.htm?kid=1KWO3)). Suppose that currently the distribution of annual salaries of all actuaries in the U.S. is approximately normal with a mean of \$98,620 and a standard deviation of \$18,000. How much would an actuary have to be paid in order to be in the highest-paid 10% of all actuaries?

**6.88** Mong Corporation makes auto batteries. The company claims that 80% of its LL70 batteries are good for 70 months or longer.

- What is the probability that in a sample of 100 such batteries, exactly 85 will be good for 70 months or longer?
- Find the probability that in a sample of 100 such batteries, at most 74 will be good for 70 months or longer.
- What is the probability that in a sample of 100 such batteries, 75 to 87 will be good for 70 months or longer?
- Find the probability that in a sample of 100 such batteries, 72 to 77 will be good for 70 months or longer.

**6.89** Stress on the job is a major concern of a large number of people who go into managerial positions. It is estimated that 80% of the managers of all companies suffer from job-related stress.

- What is the probability that in a sample of 200 managers of companies, exactly 150 suffer from job-related stress?
- Find the probability that in a sample of 200 managers of companies, at least 170 suffer from job-related stress.
- What is the probability that in a sample of 200 managers of companies, 165 or fewer suffer from job-related stress?
- Find the probability that in a sample of 200 managers of companies, 164 to 172 suffer from job-related stress.

## Advanced Exercises

**6.90** It is known that 15% of all homeowners pay a monthly mortgage of more than \$2500 and that the standard deviation of the monthly mortgage payments of all homeowners is \$350. Suppose that the monthly mortgage payments of all homeowners have a normal distribution. What is the mean monthly mortgage paid by all homeowners?

**6.91** At Jen and Perry Ice Cream Company, a machine fills 1-pound cartons of Top Flavor ice cream. The machine can be set to dispense, on average, any amount of ice cream into these cartons. However, the machine does not put exactly the same amount of ice cream into each carton; it varies from carton to carton. It is known that the amount of ice cream put into each such carton has a normal distribution with a standard deviation of .18 ounce. The quality control inspector wants to set the machine such that at least 90% of the cartons have more than 16 ounces of ice cream. What should be the mean amount of ice cream put into these cartons by this machine?

**6.92** Two companies, A and B, drill wells in a rural area. Company A charges a flat fee of \$3500 to drill a well regardless of its depth. Company B charges \$1000 plus \$12 per foot to drill a well. The depths of wells drilled in this area have a normal distribution with a mean of 250 feet and a standard deviation of 40 feet.

- What is the probability that Company B would charge more than Company A to drill a well?
- Find the mean amount charged by Company B to drill a well.

**6.93** Otto is trying out for the javelin throw to compete in the Olympics. The lengths of his javelin throws are normally distributed with a mean of 253 feet and a standard deviation of 8.4 feet. What is the probability that the longest of three of his throws is 270 feet or more?

**6.94** Lori just bought a new set of four tires for her car. The life of each tire is normally distributed with a mean of 45,000 miles and a standard deviation of 2000 miles. Find the probability that all four tires will last for at least 46,000 miles. Assume that the life of each of these tires is independent of the lives of other tires.

**6.95** The Jen and Perry Ice Cream company makes a gourmet ice cream. Although the law allows ice cream to contain up to 50% air, this product is designed to contain only 20% air. Because of variability inherent in the manufacturing process, management is satisfied if each pint contains between 18% and 22% air. Currently two of Jen and Perry's plants are making gourmet ice cream. At Plant A, the mean

amount of air per pint is 20% with a standard deviation of 2%. At Plant B, the mean amount of air per pint is 19% with a standard deviation of 1%. Assuming the amount of air is normally distributed at both plants, which plant is producing the greater proportion of pints that contain between 18% and 22% air?

**6.96** The highway police in a certain state are using aerial surveillance to control speeding on a highway with a posted speed limit of 55 miles per hour. Police officers watch cars from helicopters above a straight segment of this highway that has large marks painted on the pavement at 1-mile intervals. After the police officers observe how long a car takes to cover the mile, a computer estimates that car's speed. Assume that the errors of these estimates are normally distributed with a mean of 0 and a standard deviation of 2 miles per hour.

- a. The state police chief has directed his officers not to issue a speeding citation unless the aerial unit's estimate of speed is at least 65 miles per hour. What is the probability that a car traveling at 60 miles per hour or slower will be cited for speeding?
- b. Suppose the chief does not want his officers to cite a car for speeding unless they are 99% sure that it is traveling at 60 miles per hour or faster. What is the minimum estimate of speed at which a car should be cited for speeding?

**6.97** Ashley knows that the time it takes her to commute to work is approximately normally distributed with a mean of 45 minutes and a standard deviation of 3 minutes. What time must she leave home in the morning so that she is 95% sure of arriving at work by 9 A.M.?

**6.98** A soft-drink vending machine is supposed to pour 8 ounces of the drink into a paper cup. However, the actual amount poured into a cup varies. The amount poured into a cup follows a normal distribution with a mean that can be set to any desired amount by adjusting the machine. The standard deviation of the amount poured is always .07 ounce regardless of the mean amount. If the owner of the machine wants to be 99% sure that the amount in each cup is 8 ounces or more, to what level should she set the mean?

**6.99** According to the College Board (<http://professionals.collegeboard.com/gateway>), the mean SAT mathematics score for all college-bound seniors was 511 in 2011. Suppose that this is true for the current population of college-bound seniors. Furthermore, assume that 17% of college-bound seniors scored below 410 in this test. Assume that the distribution of SAT mathematics scores for college-bound seniors is approximately normal.

- a. Find the standard deviation of the mathematics SAT scores for college-bound seniors.
- b. Find the percentage of college-bound seniors whose mathematics SAT scores were above 660.

**6.100** Alpha Corporation is considering two suppliers to secure the large amounts of steel rods that it uses. Company A produces rods with a mean diameter of 8 mm and a standard deviation of .15 mm and sells 10,000 rods for \$400. Company B produces rods with a mean diameter of 8 mm and a standard deviation of .12 mm and sells 10,000 rods for \$460. A rod is usable only if its diameter is between 7.8 mm and 8.2 mm. Assume that the diameters of the rods produced by each company have a normal distribution. Which of the two companies should Alpha Corporation use as a supplier? Justify your answer with appropriate calculations.

**6.101** A gambler is planning to make a sequence of bets on a roulette wheel. Note that a roulette wheel has 38 numbers, of which 18 are red, 18 are black, and 2 are green. Each time the wheel is spun, each of the 38 numbers is equally likely to occur. The gambler will choose one of the following two sequences.  
*Single-number bet:* The gambler will bet \$5 on a particular number before each spin. He will win a net amount of \$175 if that number comes up and lose \$5 otherwise.

*Color bet:* The gambler will bet \$5 on the red color before each spin. He will win a net amount of \$5 if a red number comes up and lose \$5 otherwise.

- a. If the gambler makes a sequence of 25 bets, which of the two betting schemes offers him a better chance of coming out ahead (winning more money than losing) after the 25 bets?
- b. Now compute the probability of coming out ahead after 25 single-number bets of \$5 each and after 25 color bets of \$5 each. Do these results confirm your guess in part a? (Before using an approximation to find either probability, be sure to check whether it is appropriate.)

**6.102** A charter bus company is advertising a singles outing on a bus that holds 60 passengers. The company has found that, on average, 10% of ticket holders do not show up for such trips; hence, the company routinely overbooks such trips. Assume that passengers act independently of one another.

- a. If the company sells 65 tickets, what is the probability that the bus can hold all the ticket holders who actually show up? In other words, find the probability that 60 or fewer passengers show up.
- b. What is the largest number of tickets the company can sell and still be at least 95% sure that the bus can hold all the ticket holders who actually show up?

**6.103** The amount of time taken by a bank teller to serve a randomly selected customer has a normal distribution with a mean of 2 minutes and a standard deviation of .5 minute.

- a. What is the probability that both of two randomly selected customers will take less than 1 minute each to be served?

- b. What is the probability that at least one of four randomly selected customers will need more than 2.25 minutes to be served?

**6.104** Suppose you are conducting a binomial experiment that has 15 trials and the probability of success of .02. According to the sample size requirements, you cannot use the normal distribution to approximate the binomial distribution in this situation. Use the mean and standard deviation of this binomial distribution and the empirical rule to explain why there is a problem in this situation. (Note: Drawing the graph and marking the values that correspond to the empirical rule is a good way to start.)

**6.105** A variation of a roulette wheel has slots that are not of equal size. Instead, the width of any slot is proportional to the probability that a standard normal random variable  $z$  takes on a value between  $a$  and  $(a + .1)$ , where  $a = -3.0, -2.9, -2.8, \dots, 2.9, 3.0$ . In other words, there are slots for the intervals  $(-3.0, -2.9), (-2.9, -2.8), (-2.8, -2.7)$  through  $(2.9, 3.0)$ . There is one more slot that represents the probability that  $z$  falls outside the interval  $(-3.0, 3.0)$ . Find the following probabilities.

- a. The ball lands in the slot representing  $(.3, .4)$ .
- b. The ball lands in any of the slots representing  $(-.1, .4)$ .
- c. In at least one out of five games, the ball lands in the slot representing  $(-.1, .4)$ .
- d. In at least 100 out of 500 games, the ball lands in the slot representing  $(.4, .5)$ .

**6.106** Refer to Exercise 6.98. In that exercise, suppose the mean is set to be 8 ounces, but the standard deviation is unknown. The cups used in the machine can hold up to 8.2 ounces, but these cups will overflow if more than 8.2 ounces is dispensed by the machine. What is the smallest possible standard deviation that will result in overflows occurring 3% of the time?

## APPENDIX 6.1 NORMAL QUANTILE PLOTS

Many of the methods that are used in statistics require that the sampled data come from a normal distribution. While it is impossible to determine if this holds true without taking a census (i.e., looking at the population data), there are statistical tools that can be used to determine if this is a reasonable assumption. One of the simplest tools to use is called a normal quantile plot. The idea of the plot is to compare the values in a data set with the corresponding values one would predict for a standard normal distribution.

Although normal quantile plots are typically created using technology, it is helpful to see an example to understand how they are created and what the various numbers represent. To demonstrate, consider the data in the following table, which contains the 2001 salaries of the mayors of 10 large cities.

City	Mayor's Salary (\$)	City	Mayor's Salary (\$)
Chicago, IL	170,000	Newark, NJ	147,000
New York, NY	165,000	San Francisco, CA	146,891
Houston, TX	160,500	Jacksonville, FL	127,230
Detroit, MI	157,300	Baltimore, MD	125,000
Los Angeles, CA	147,390	Boston, MA	125,000

Each data point represents 1/10 of the distribution, with the smallest value representing the smallest 10%, the next representing the 10%–20% interval, and so on. In each case, we estimate that the data points fall in the middle of their respective intervals. For these 10 data points, these midpoints would be at the 5%, 15%, 25%, and so on, locations, while if we have 20 data points, these locations would be at 2.5%, 7.5%, 12.5%, and so on. Next we determine the  $z$  scores for these locations. The following table shows the  $z$  scores for the 10-data point scenario and for the 20-data point scenario.

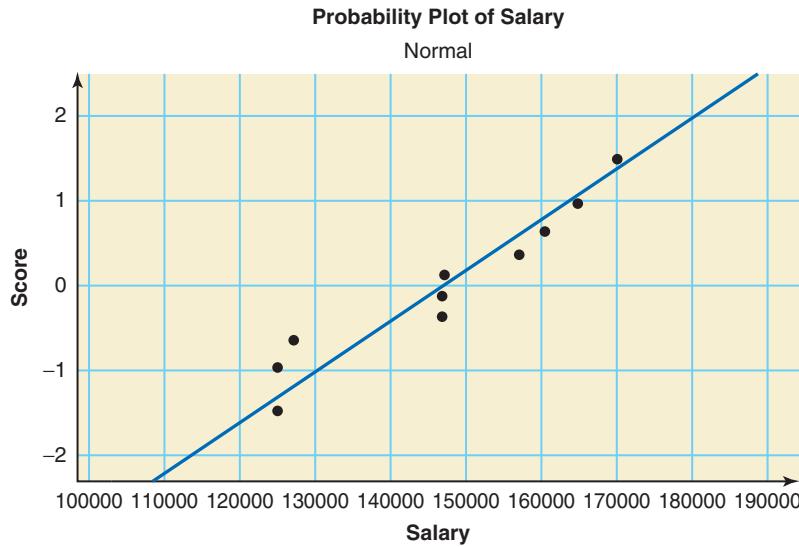
### Ten data points

Location (%)	5	15	25	35	45	55	65	75	85	95
$z$ Score	-1.645	-1.036	-0.674	-0.385	-0.126	0.126	0.385	0.674	1.036	1.645

### Twenty data points

Location (%)	2.5	7.5	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5
$z$ Score	-1.960	-1.440	-1.150	-0.935	-0.755	-0.598	-0.454	-0.319	-0.189	-0.063
Location (%)	52.5	57.5	62.5	67.5	72.5	77.5	82.5	87.5	92.5	97.5
$z$ Score	0.063	0.189	0.319	0.454	0.598	0.755	0.935	1.150	1.440	1.960

Next we make a two-dimensional plot that places the data on the horizontal axis and the  $z$  scores on the vertical axis.

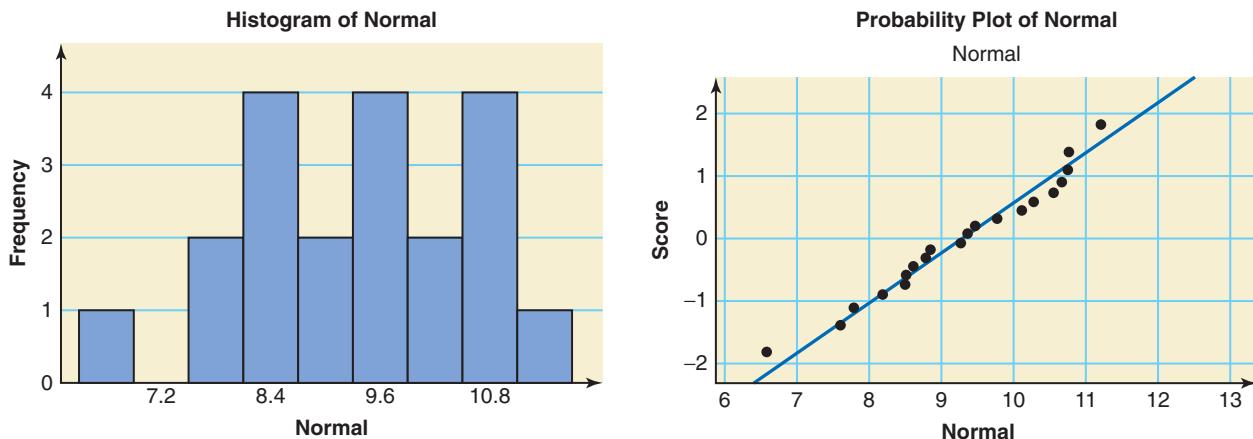


If the data are in complete agreement with a normal distribution, the points will lie on the line displayed in the graph. As the likelihood that the given data come from a normal distribution decreases, the plot of data points will become less linear.

So, how do we interpret the plot of 10 salaries in the graph? There are a couple of features that we can point out. There are two groups of points that are stacked almost vertically (near \$125,000 and \$147,000). Depending on the software, multiple data points of the same value will be stacked or will appear as one point. In addition, there is a fairly big gap between these two groups. This is not unusual with small data sets, even if the data come from a normal distribution. Most times, in order to state that a very small data set does not come from a normal distribution, many people will be able to see that the data are very strongly skewed or have an outlier simply by looking at a sorted list of the data.

To understand the correspondence between the shape of a data set and its normal quantile plot, it is useful to look at a dotplot or histogram side by side with the normal quantile plot. We will consider a few common cases here.

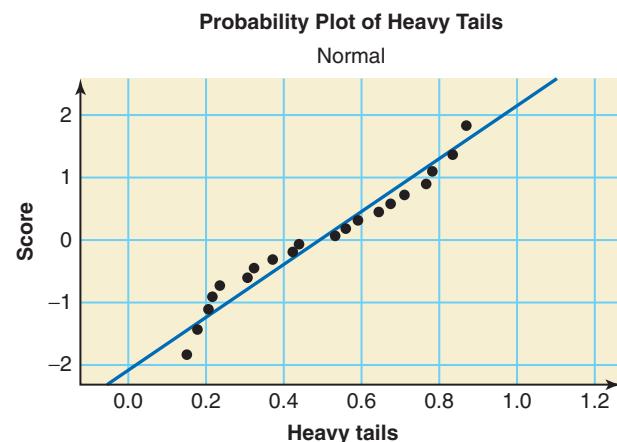
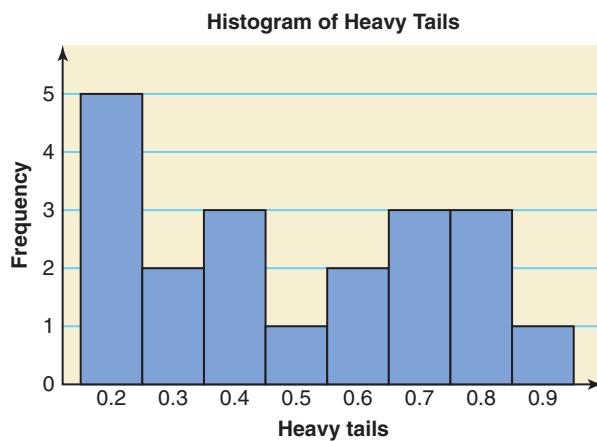
(1). First, following are the two graphs for 20 data points randomly selected from a normal distribution.



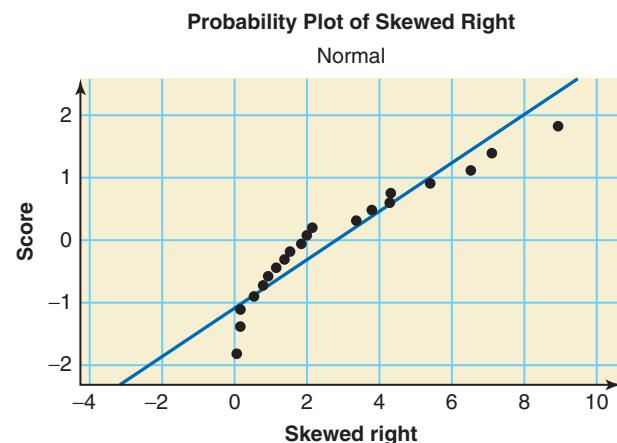
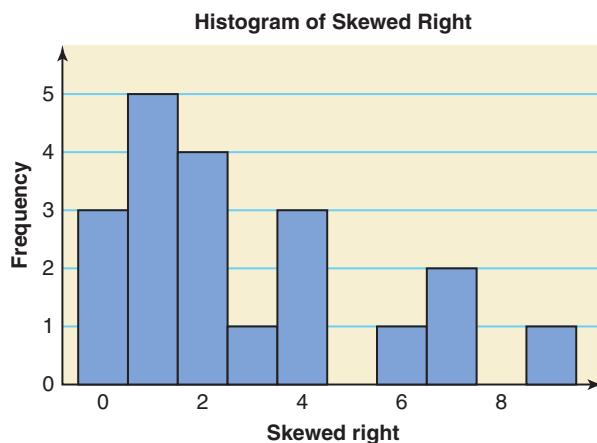
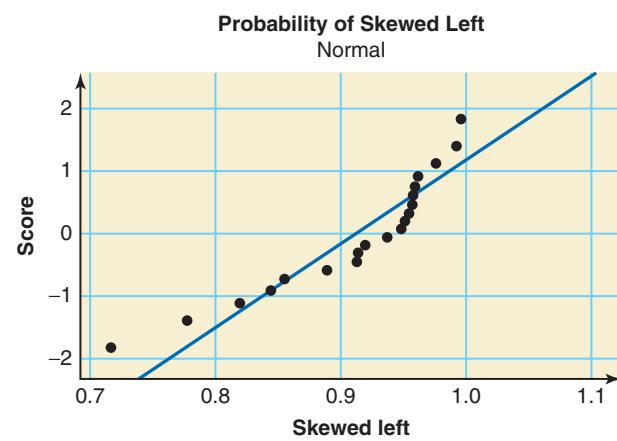
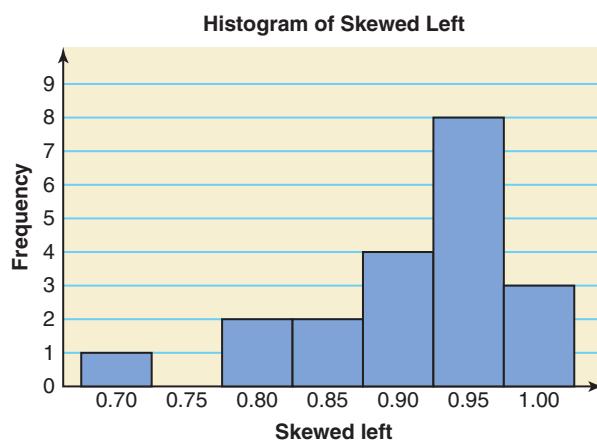
As you can see, just because data come from a normal distribution does not imply that they will be perfectly linear. In general, the points are close to the line, but small patterns such as in the upper right or the gap in the lower left can occur without invalidating the normality assumption.

(2). In the next case, we consider data that come from a distribution that is *heavy tailed*, which means that the distribution has higher percentages of values in its tails than one would expect in a normal distribution. For example, the Empirical Rule states that a normal distribution has approximately 2.5% of the observations below  $\mu - 2\sigma$  and 2.5% of the observations above  $\mu + 2\sigma$ . If a data set has 10%

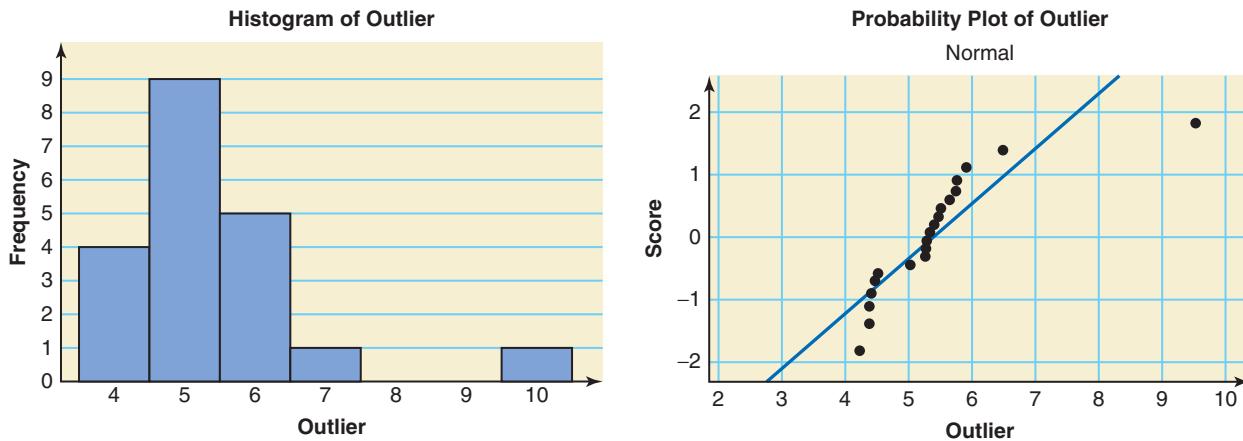
of the observations below  $\mu - 2\sigma$  and 10% of the observations above  $\mu + 2\sigma$ , it would be classified as being heavy tailed. The resulting shape of the normal quantile plot will look somewhat like a playground slide. As the tails get heavier, the ends of the plot become steeper and the middle gets flatter.



- (3). Skewed distributions have normal quantile plots that are shaped somewhat like a boomerang, which has a rounded V shape, with one end of the boomerang stretched out more than the other side. Just like all other graphs of skewed distributions, the side that is stretched out identifies the direction of the skew. Again, as the distribution becomes more skewed, the bend in the quantile plot will become more severe.



- (4). Our last example involves an outlier. As in other graphs, potential outliers are fairly easy to identify, basically by finding a large horizontal jump in the left or right tail. However, you need to be careful when distinguishing a skewed distribution from one that has an outlier. In our skewed-to-the-right example, there is an approximate difference of 2 between the two largest values, yet the largest data value is still fairly close to the line in a vertical direction. In our outlier example, there is a substantial vertical distance between the largest data value and the line. Moreover, we do not see the *bow* shape in the latter plot.



It is important to remember that these plots contain examples of a variety of common features. However, it is also important to remember that some of these features are not mutually exclusive. As one example, it is possible for a distribution to have heavy tails and an outlier. Identifying issues that would reject the notion of normality will be important in determining the types of inference procedures that can be used, which we will begin examining in Chapter 8.

## Self-Review Test

- The normal probability distribution is applied to
  - a continuous random variable
  - a discrete random variable
  - any random variable
- For a continuous random variable, the probability of a single value of  $x$  is always
  - zero
  - 1.0
  - between 0 and 1
- Which of the following is not a characteristic of the normal distribution?
  - The total area under the curve is 1.0.
  - The curve is symmetric about the mean.
  - The two tails of the curve extend indefinitely.
  - The value of the mean is always greater than the value of the standard deviation.
- The parameters of a normal distribution are
  - $\mu$ ,  $z$ , and  $\sigma$
  - $\mu$  and  $\sigma$
  - $\mu$ ,  $x$ , and  $\sigma$
- For the standard normal distribution,
  - $\mu = 0$  and  $\sigma = 1$
  - $\mu = 1$  and  $\sigma = 0$
  - $\mu = 100$  and  $\sigma = 10$
- The  $z$  value for  $\mu$  for a normal distribution curve is always
  - positive
  - negative
  - 0
- For a normal distribution curve, the  $z$  value for an  $x$  value that is less than  $\mu$  is always
  - positive
  - negative
  - 0
- Usually the normal distribution is used as an approximation to the binomial distribution when
  - $n \geq 30$
  - $np > 5$  and  $nq > 5$
  - $n > 20$  and  $p = .50$
- Find the following probabilities for the standard normal distribution.
  - $P(.85 \leq z \leq 2.33)$
  - $P(-2.97 \leq z \leq 1.49)$
  - $P(z \leq -1.29)$
  - $P(z > -.74)$
- Find the value of  $z$  for the standard normal curve such that the area
  - in the left tail is .1000
  - between 0 and  $z$  is .2291 and  $z$  is positive
  - in the right tail is .0500
  - between 0 and  $z$  is .3571 and  $z$  is negative

11. In a National Highway Traffic Safety Administration (NHTSA) report, data provided to the NHTSA by Goodyear stated that the average tread life of properly inflated automobile tires is 45,000 miles (*Source*: [http://www.safercar.gov/cars/rules/rulings/TPMS\\_FMVSS\\_No138/part5.5.html](http://www.safercar.gov/cars/rules/rulings/TPMS_FMVSS_No138/part5.5.html)). Suppose that the current distribution of tread life of properly inflated automobile tires is normally distributed with a mean of 45,000 miles and a standard deviation of 2360 miles.
- Find the probability that a randomly selected automobile tire has a tread life between 42,000 and 46,000 miles.
  - What is the probability that a randomly selected automobile tire has a tread life of less than 38,000 miles?
  - What is the probability that a randomly selected automobile tire has a tread life of more than 50,000 miles?
  - Find the probability that a randomly selected automobile tire has a tread life between 46,500 and 47,500 miles.
12. Refer to Problem 11.
- Suppose that 6% of all automobile tires with the longest tread life have a tread life of at least  $x$  miles. Find the value of  $x$ .
  - Suppose that 2% of all automobile tires with the shortest tread life have a tread life of at most  $x$  miles. Find the value of  $x$ .
13. Gluten sensitivity, which is also known as wheat intolerance, affects approximately 15% of people. The condition involves great difficulty in digesting wheat, but is not the same as wheat allergy, which has much more severe reactions (*Source*: <http://www.foodintol.com/wheat.asp>). A random sample of 800 individuals is selected.
- Find the probability that the number of individuals in this sample who have wheat intolerance is
    - exactly 115
    - 103 to 142
    - at least 107
    - at most 100
    - between 111 and 123
  - Find the probability that at least 675 of the individuals in this sample do *not* have wheat intolerance.
  - Find the probability that 682 to 697 of the individuals in this sample do *not* have wheat intolerance.

## Mini-Projects

### ■ MINI-PROJECT 6-1

Consider the data on heights of NFL players that accompany this text (see Appendix B).

- Use statistical software to obtain a histogram. Do these heights appear to be symmetrically distributed? If not, in which direction do they seem to be skewed?
- Compute  $\mu$  and  $\sigma$  for heights of all players.
- What percentage of these heights lie in the interval  $\mu - \sigma$  to  $\mu + \sigma$ ? What about in the interval  $\mu - 2\sigma$  to  $\mu + 2\sigma$ ? In the interval  $\mu - 3\sigma$  to  $\mu + 3\sigma$ ?
- How do the percentages in part c compare to the corresponding percentages for a normal distribution (68.26%, 95.44%, and 99.74%, respectively)?
- Based on the percentages in the Empirical Rule, approximately 34.13% of data values should fall in each of the intervals  $\mu - \sigma$  to  $\mu$  and  $\mu$  to  $\mu + \sigma$ . Similarly, approximately 13.59% of the data values should fall in each of the intervals  $\mu - 2\sigma$  to  $\mu - \sigma$  and  $\mu + \sigma$  to  $\mu + 2\sigma$ , and approximately 2.15% of the data values should fall in each of the intervals  $\mu - 3\sigma$  to  $\mu - 2\sigma$  and  $\mu + 2\sigma$  to  $\mu + 3\sigma$ . Calculate the percentages of the values in the NFL data that fall in each of these intervals. How do they compare to the values given by the Empirical Rule?
- Use statistical software to select three random samples of 20 players each. Create a histogram and a dotplot of heights for each sample, and calculate the mean and standard deviation of heights for each sample. How well do your graphs and summary statistics match up with the corresponding population graphs and parameter values obtained in earlier parts? Does it seem reasonable that they might not match up very well?

### ■ MINI-PROJECT 6-2

Consider the data on weights of NFL players (see Appendix B).

- Use statistical software to obtain a histogram. Do these weights appear to be symmetrically distributed? If not, in which direction do they seem to be skewed?

- b. Compute  $\mu$  and  $\sigma$  for weights of all players.
- c. What percentage of these weights lie in the interval  $\mu - \sigma$  to  $\mu + \sigma$ ? What about in the interval  $\mu - 2\sigma$  to  $\mu + 2\sigma$ ? In the interval  $\mu - 3\sigma$  to  $\mu + 3\sigma$ ?
- d. How do the percentages in part c compare to the corresponding percentages for a normal distribution (68.26%, 95.44%, and 99.74%, respectively)?
- e. Based on the percentages in the Empirical Rule, approximately 34.13% of data values should fall in each of the intervals  $\mu - \sigma$  to  $\mu$  and  $\mu$  to  $\mu + \sigma$ . Similarly, approximately 13.59% of the data values should fall in each of the intervals  $\mu - 2\sigma$  to  $\mu - \sigma$  and  $\mu + \sigma$  to  $\mu + 2\sigma$ , and approximately 2.15% of the data values should fall in each of the intervals  $\mu - 3\sigma$  to  $\mu - 2\sigma$  and  $\mu + 2\sigma$  to  $\mu + 3\sigma$ . Calculate the percentages of the values in the NFL data that fall in each of these intervals. How do they compare to the values given by the Empirical Rule?
- f. Use statistical software to select three random samples of 20 players each. Create a histogram and a dotplot of weights for each sample, and calculate the mean and standard deviation of weights for each sample. How well do your graphs and summary statistics match up with the corresponding population graphs and parameter values obtained in earlier parts? Does it seem reasonable that they might not match up very well?

### ■ MINI-PROJECT 6-3

The National Oceanic and Atmospheric Administration (NOAA) Web site has daily historical data of precipitation amounts, as well as the minimum and maximum temperatures available for a large number of weather stations throughout the United States. For the purpose of this Mini-Project, you will need to download 2 consecutive months of data, 1 month at a time. To obtain the data, go to <http://www7.ncdc.noaa.gov/IPS/coop/coop.html> and choose your location and month of interest. Answer the following questions with regard to the maximum daily temperature.

- a. Use statistical software to obtain a histogram and a dotplot for your data. Comment on the shape of the distribution as observed from these graphs.
- b. Calculate  $\bar{x}$  and  $s$ .
- c. What percentage of the temperatures are in the interval  $\bar{x} - s$  to  $\bar{x} + s$ ?
- d. What percentage are in the interval  $\bar{x} - 2s$  to  $\bar{x} + 2s$ ?
- e. How do these percentages compare to the corresponding percentages for a normal distribution (68.26% and 95.44%, respectively)?
- f. Now find the minimum temperatures in your town for 60 days by using the same source that you used to find the maximum temperatures or by using a different source. Then repeat parts a through e for this data set.

## DECIDE FOR YOURSELF

## DECIDING ABOUT THE SHAPE OF A DISTRIBUTION

Reporting summary measures such as the mean, median, and standard deviation has become very common in modern life. Many companies, government agencies, and so forth will report the mean and standard deviation of a variable, but they will very rarely provide information on the shape of the distribution of that variable. In Chapters 5 and 6, you have learned some basic properties of some distributions that can help you to decide if a specific type of distribution is a good fit for a set of data.

According to the *National Diet and Nutrition Survey: Adults Aged 19 to 64*, British men spend an average of 2.15 hours per day in moderate- or high-intensity physical activity. The standard deviation of these activity times for this sample was 3.59 hours. (*Source:* <http://www.food.gov.uk/multimedia/pdfs/ndnsfour.pdf>.) Can we infer that these activity times could follow a normal distribution? The following questions may provide an answer.

1. Sketch a normal curve marking the points representing 1, 2, and 3 standard deviations above and below the mean, and calculate the values at these points using a mean of 2.15 hours and a standard deviation of 3.59 hours.
2. Examine the curve with your calculations. Explain why it is impossible for this distribution to be normal based on your graph and calculations.
3. Considering the variable being measured, is it more likely that the distribution is skewed to the left or that it is skewed to the right? Explain why.
4. Suppose that the standard deviation for this sample was .70 hour instead of 3.59 hours, which makes it numerically possible for the distribution to be normal. Again, considering the variable being measured, explain why the normal distribution is still not a logical choice for this distribution.

# TECHNOLOGY INSTRUCTION

## Normal and Inverse Normal Probabilities

**TI-84**

- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies below  $b$ , select **2nd >VARS >normalcdf**. In the **normalcdf(** menu, enter **-E99** at the **lower:** prompt,  $b$  at the **upper:** prompt,  $\mu$  at the  **$\mu$ :** prompt,  $\sigma$  at the  **$\sigma$ :** prompt, and then highlight **Paste** and press **ENTER** twice. (See **Screen 6.1**.)

```
normalcdf
lower:-e99
upper:125
μ:100
σ:15
Paste
```

Screen 6.1

Note: To type **E99**, press **2nd >comma key** (which is the key just above the **7** key). The function is labeled **EE**, but only **E** is displayed on the screen. Then type **9** twice. For **-E99**, press the negative **(-)** key (which is to the right of the decimal key) before **E99**.

- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies above  $a$ , select **2nd >VARS >normalcdf**. In the **normalcdf(** menu, enter  $a$  at the **lower:** prompt, **E99** at the **upper:** prompt,  $\mu$  at the  **$\mu$ :** prompt,  $\sigma$  at the  **$\sigma$ :** prompt, and then highlight **Paste** and press **ENTER** twice. (See **Screen 6.2**.)
- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies between  $a$  and  $b$ , select **2nd >VARS >normalcdf**. In the **normalcdf(** menu, enter  $a$  at the **lower:** prompt,  $b$  at the **upper:** prompt,  $\mu$  at the  **$\mu$ :** prompt,  $\sigma$  at the  **$\sigma$ :** prompt, and then highlight **Paste** and press **ENTER** twice. (See **Screen 6.3**.)
- To find a value of  $a$  for a normal random variable  $x$  with mean  $\mu$  and standard deviation  $\sigma$  such that the probability of  $x$  being less than  $a$  is  $p$ , select **2nd >VARS >invNorm**. In the **invNorm(** menu, enter  $p$  at the **area:** prompt,  $\mu$  at the  **$\mu$ :** prompt,  $\sigma$  at the  **$\sigma$ :** prompt, and then highlight **Paste** and press **ENTER** twice. (See **Screen 6.4**.)

```
normalcdf
lower:90
upper:e99
μ:100
σ:15
Paste
```

Screen 6.2

- To create a normal quantile plot for a list of data, press **STAT PLOT**, which you access by pressing **2nd > Y=**. The **Y=** key is located at the top left of the calculator buttons. Make sure that only one plot is turned on. If more than one plot is turned on, you can turn off the unwanted plots by using the following steps. Press the number corresponding to the plot you wish to turn off. A screen similar to **Screen 6.5** will appear. Use the arrow keys to move the cursor to the **Off** button, then press **ENTER**. Now use the arrow keys to move to the row with **Plot1**, **Plot2**, and **Plot3**. If there is another plot that you need to turn off, select that plot by moving the cursor to that plot, press **ENTER**, and repeat the previous procedure. If you do not need to turn off any graphs, move the cursor to the plot you wish to use and press **ENTER**. Make sure that this plot is turned **On**. At the **Type:** prompt use the right arrow to move to the third column in the second row, and press **ENTER**. Move to the **Xlist:** prompt to enter the name of the list where the data are located. Press **2nd > STAT**, then use the up and down arrows to move through the list names until you find the list you want to use. Press **ENTER**. (Note: If you are using one of the lists named **L1**, **L2**, **L3**, **L4**, **L5**, or **L6**, you can enter the list name by pressing **2nd** followed by one of the numbers **1** through **6**, as they correspond to the list names **L1** through **L6**.) At the **Data Axis:** prompt, select **X**. At the **Mark:** prompt, choose any symbol you wish to use. To see the graph, select **ZOOM > 9** (the **ZOOMSTAT** function), where **ZOOM** is the third key in the top row. This sets the window settings to display your graph. (See **Screen 6.6**.)

```
normalcdf
lower:90
upper:125
μ:100
σ:15
Paste
```

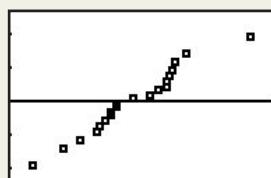
Screen 6.3

```
invNorm
area:.99
μ:100
σ:15
Paste
```

Screen 6.4

```
Plot1 Plot2 Plot3
On Off
Type: L1 L2 L3
Data List:L1
Data Axis:X Y
Mark: □ +
```

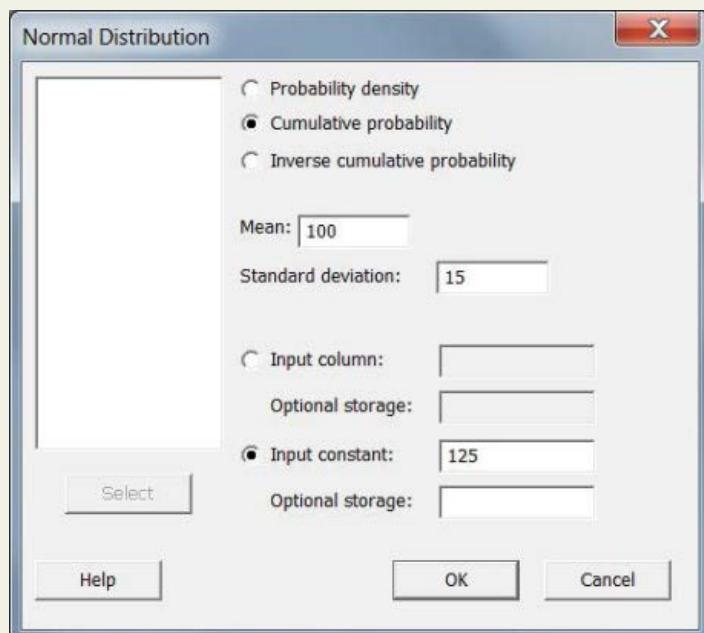
Screen 6.5



Screen 6.6

**Minitab**

1. For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies below  $a$ , select **Calc >Probability Distributions >Normal**. Select **Cumulative probability**, and enter the mean  $\mu$  and the standard deviation  $\sigma$ . Select **Input constant** and enter  $a$ , then select **OK**. (See Screens 6.7 and 6.8.)
2. To find a value of  $a$  for a normal random variable  $x$  with mean  $\mu$  and standard deviation  $\sigma$  such that the probability of  $x$  being less than  $a$  is  $p$ , select **Calc >Probability Distributions >Normal**. Select **Inverse cumulative probability** and enter the mean  $\mu$  and the standard deviation  $\sigma$ . Select **Input constant** and enter  $a$ , then select **OK**. (See Screens 6.7 and 6.8.)



Screen 6.7

```
Session
```

**Cumulative Distribution Function**

Normal with mean = 100 and standard deviation = 15

x	P( X <= x )
125	0.952210

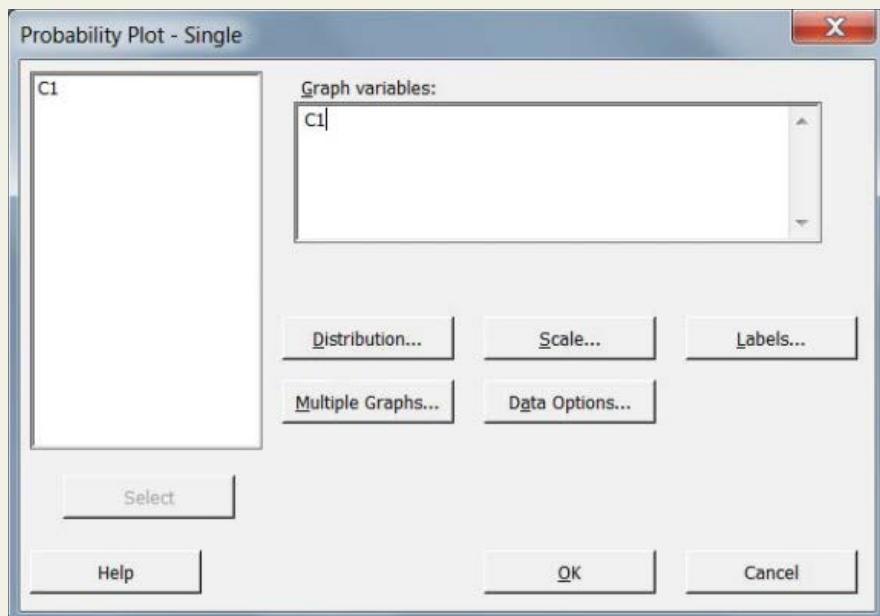
**Inverse Cumulative Distribution Function**

Normal with mean = 100 and standard deviation = 15

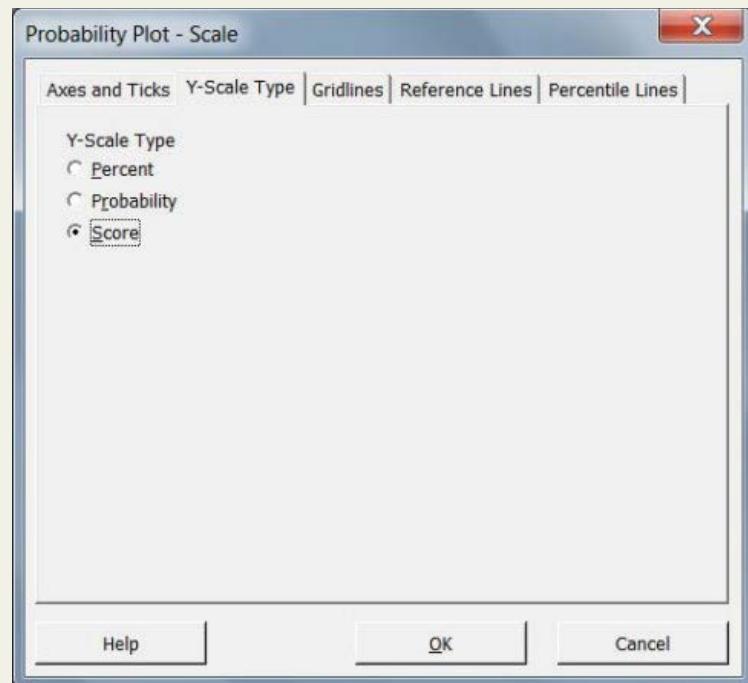
P( X <= x )	x
0.99	134.895

Screen 6.8

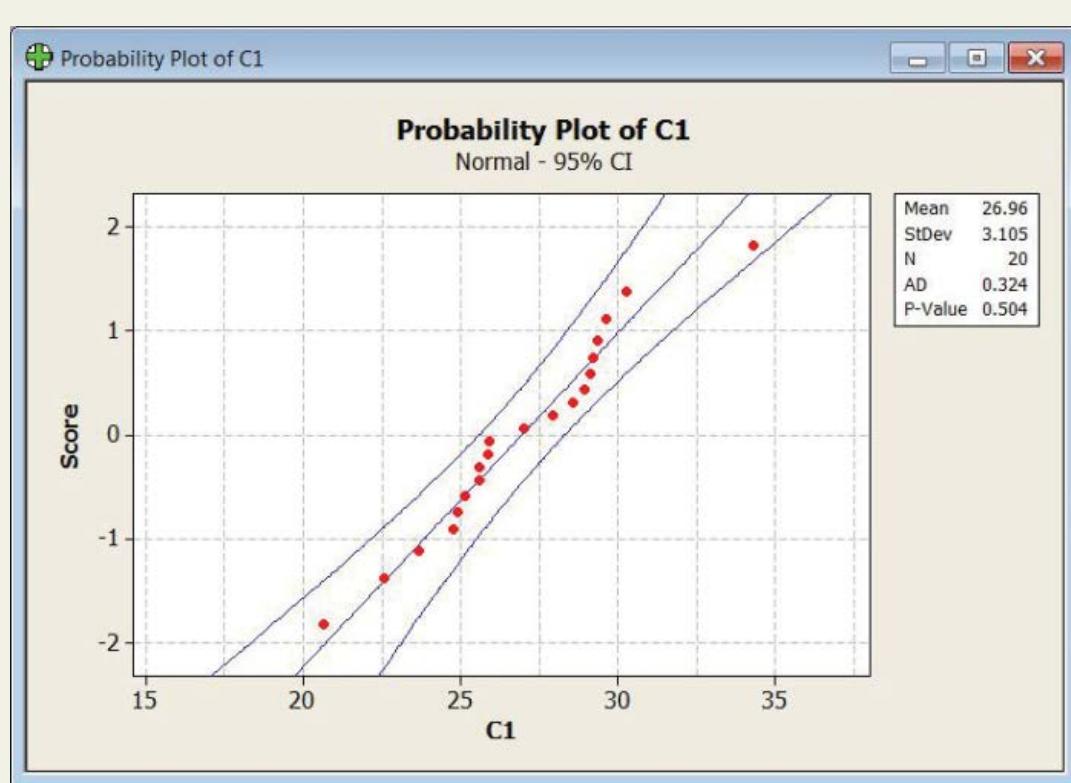
3. To create a normal quantile plot for quantitative data values entered in column C1, select **Graph > Probability Plot**, select **Simple**, and click **OK**. In the resulting dialog box, type C1 in the box below **Graph Variables**. (See Screen 6.9.) Click the **Scale** button, and then click the **Y-Scale** tab. Select **Score**. (See Screen 6.10.) Click **OK** to close the **Scale** box, and then click **OK** in the **Probability Plot** box to create the plot. (See Screen 6.11.)



Screen 6.9



Screen 6.10



Screen 6.11

**Excel**

- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies below  $b$ , type =NORM.DIST( $b, \mu, \sigma, 1$ ). (See Screen 6.12.)
- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies above  $a$ , type =1-NORM.DIST( $a, \mu, \sigma, 1$ ).
- For a given mean  $\mu$  and standard deviation  $\sigma$ , to find the probability that a normal random variable  $x$  lies between  $a$  and  $b$ , type =NORM.DIST( $b, \mu, \sigma, 1$ ) - NORM.DIST( $a, \mu, \sigma, 1$ ).
- To find a value of  $a$  for a normal random variable  $x$  with mean  $\mu$  and standard deviation  $\sigma$  such that the probability of  $x$  being less than  $a$  is  $p$ , type =NORM.INV( $p, \mu, \sigma$ ). (See Screen 6.13.)

Note: If you are using office 2007 or earlier versions, the function names do not contain a dot (·).

	A	B	C	D
1	Mean	100		
2	Std. Dev.	15		
3				
4	P(X<125)	=NORM.DIST(125,100,15,1)		

Screen 6.12

	A	B	C	D
1	Mean	100		
2	Std. Dev.	15		
3				
4	P(X<a)=.99	=NORM.INV(.99,100,15)		

Screen 6.13

## TECHNOLOGY ASSIGNMENTS

**TA6.1** Find the area under the standard normal curve

- a. to the left of  $z = -1.94$
- b. to the left of  $z = .83$
- c. to the right of  $z = 1.45$
- d. to the right of  $z = -1.65$
- e. between  $z = .75$  and  $z = 1.90$
- f. between  $z = -1.20$  and  $z = 1.55$

**TA6.2** Find the following areas under a normal curve with  $\mu = 86$  and  $\sigma = 14$ .

- a. Area to the left of  $x = 71$
- b. Area to the left of  $x = 96$
- c. Area to the right of  $x = 90$
- d. Area to the right of  $x = 75$
- e. Area between  $x = 65$  and  $x = 75$
- f. Area between  $x = 72$  and  $x = 95$

**TA6.3** The transmission on a particular model of car has a warranty for 40,000 miles. It is known that the life of such a transmission has a normal distribution with a mean of 72,000 miles and a standard deviation of 12,000 miles. Answer the following questions.

- a. What percentage of the transmissions will fail before the end of the warranty period?
- b. What percentage of the transmissions will be good for more than 100,000 miles?
- c. What percentage of the transmissions will be good for 80,000 to 100,000 miles?

**TA6.4** Refer to Exercise 6.38. Assume that the distribution of weekly unemployment benefits in the United States is approximately normal with a mean of \$297 and a standard deviation of \$74.42.

- a. Find the probability that the weekly unemployment benefit received by a randomly selected person who is currently receiving an unemployment benefit is
  - i. more than \$200
  - ii. \$275 to \$375
  - iii. \$0 or less (theoretically the normal distribution extends from negative infinity to positive infinity; realistically, unemployment benefits must be positive, so this answer provides an idea of the level of approximation used in modeling this variable)
  - iv. more than \$689, which is the maximum weekly unemployment benefit that a person can receive in the United States (this is similar to part c, except that we are looking at the upper tail of the distribution)
- b. What is the amount of the weekly unemployment benefit that will place someone in the highest 6.5% of all weekly unemployment benefits received?

**TA6.5** Refer to Exercise 6.39. Suppose that the current amounts spent by families of college students on new apparel, dorm furniture, supplies, and electronics follow a normal distribution with a mean of \$616.13 and a standard deviation of \$120.

- a. Find the proportion of such families that spend on the aforementioned list of items:
  - i. less than \$825
  - ii. between \$400 and \$500
- b. How much would the family of a college student have to spend to be among the 2.8% of families with the lowest expenditures on the aforementioned list of items?

**TA6.6** In Appendix 6.1, we learned how to create a normal quantile plot, which can be used to determine how well a data set matches a normal distribution. Use technology to create a normal quantile plot for the following data set.

Does the data plot appear to be approximately linear? If not, how does it differ? What does this imply for the possibility that the data come from a normal distribution?

**TA6.7** In Appendix 6.1, we learned how to create a normal quantile plot, which can be used to determine how well a data set matches a normal distribution. Use your technology of choice to create a normal quantile plot and a histogram or dotplot for the data sets in each of the following exercises. After creating these plots, describe how the various shapes of these plots correspond to the characteristics (symmetry, skewness, outliers) of a distribution.

- a. Exercise 3.105
- b. Exercise 3.110
- c. Exercise 3.136 without the two largest values
- d. Exercise 3.136 without the three largest values
- e. Exercise 3.140 (for weights)



© Steve Cole/Stockphoto

## Sampling Distributions

- 7.1 Sampling Distribution, Sampling Error, and Nonsampling Errors
- 7.2 Mean and Standard Deviation of  $\bar{x}$
- 7.3 Shape of the Sampling Distribution of  $\bar{x}$
- 7.4 Applications of the Sampling Distribution of  $\bar{x}$
- 7.5 Population and Sample Proportions; and Mean, Standard Deviation, and Shape of the Sampling Distribution of  $\hat{p}$
- 7.6 Applications of the Sampling Distribution of  $\hat{p}$

You read about opinion polls in newspapers, magazines, and on the Web every day. These polls are based on sample surveys. Have you heard of sampling and nonsampling errors? It is good to be aware of such errors while reading these opinion poll results. Sound sampling methods are essential for opinion poll results to be valid and to lower the effects of such errors.

Chapters 5 and 6 discussed probability distributions of discrete and continuous random variables. This chapter extends the concept of probability distribution to that of a sample statistic. As we discussed in Chapter 3, a sample statistic is a numerical summary measure calculated for sample data. The mean, median, mode, and standard deviation calculated for sample data are called *sample statistics*. On the other hand, the same numerical summary measures calculated for population data are called *population parameters*. A population parameter is always a constant (at a given point in time), whereas a sample statistic is always a random variable. Because every random variable must possess a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is more commonly called its *sampling distribution*. This chapter discusses the sampling distributions of the sample mean and the sample proportion. The concepts covered in this chapter are the foundation of the inferential statistics discussed in succeeding chapters.

## 7.1 Sampling Distribution, Sampling Error, and Nonsampling Errors

This section introduces the concepts of sampling distribution, sampling error, and nonsampling errors. Before we discuss these concepts, we will briefly describe the concept of a population distribution.

The **population distribution** is the probability distribution derived from the information on all elements of a population.

### Definition

**Population Distribution** The *population distribution* is the probability distribution of the population data.

Suppose there are only five students in an advanced statistics class and the midterm scores of these five students are

70      78      80      80      95

Let  $x$  denote the score of a student. Using single-valued classes (because there are only five data values, there is no need to group them), we can write the frequency distribution of scores as in Table 7.1 along with the relative frequencies of classes, which are obtained by dividing the frequencies of classes by the population size. Table 7.2, which lists the probabilities of various  $x$  values, presents the probability distribution of the population. Note that these probabilities are the same as the relative frequencies.

**Table 7.1** Population Frequency and Relative Frequency Distributions

$x$	$f$	Relative Frequency
70	1	$1/5 = .20$
78	1	$1/5 = .20$
80	2	$2/5 = .40$
95	1	$1/5 = .20$
	$N = 5$	Sum = 1.00

**Table 7.2** Population Probability Distribution

$x$	$P(x)$
70	.20
78	.20
80	.40
95	.20
	$\Sigma P(x) = 1.00$

The values of the mean and standard deviation calculated for the probability distribution of Table 7.2 give the values of the population parameters  $\mu$  and  $\sigma$ . These values are  $\mu = 80.60$  and  $\sigma = 8.09$ . The values of  $\mu$  and  $\sigma$  for the probability distribution of Table 7.2 can be calculated using the formulas given in Section 5.3 of Chapter 5 (see Exercise 7.6).

### 7.1.1 Sampling Distribution

As mentioned at the beginning of this chapter, the value of a population parameter is always constant. For example, for any population data set, there is only one value of the population

mean,  $\mu$ . However, we cannot say the same about the sample mean,  $\bar{x}$ . We would expect different samples of the same size drawn from the same population to yield different values of the sample mean,  $\bar{x}$ . The value of the sample mean for any one sample will depend on the elements included in that sample. Consequently, *the sample mean,  $\bar{x}$ , is a random variable*. Therefore, like other random variables, the sample mean possesses a probability distribution, which is more commonly called the **sampling distribution of  $\bar{x}$** . Other sample statistics, such as the median, mode, and standard deviation, also possess sampling distributions.

### Definition

**Sampling Distribution of  $\bar{x}$**  The probability distribution of  $\bar{x}$  is called its sampling distribution. It lists the various values that  $\bar{x}$  can assume and the probability of each value of  $\bar{x}$ .

In general, the probability distribution of a sample statistic is called its *sampling distribution*.

Reconsider the population of midterm scores of five students given in Table 7.1. Consider all possible samples of three scores each that can be selected, without replacement, from that population. The total number of possible samples, given by the combinations formula discussed in Chapter 4, is 10; that is,

$$\text{Total number of samples} = {}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 10$$

Suppose we assign the letters A, B, C, D, and E to the scores of the five students, so that

$$A = 70, \quad B = 78, \quad C = 80, \quad D = 80, \quad E = 95$$

Then, the 10 possible samples of three scores each are

$$ABC, \quad ABD, \quad ABE, \quad ACD, \quad ACE, \quad ADE, \quad BCD, \quad BCE, \quad BDE, \quad CDE$$

These 10 samples and their respective means are listed in Table 7.3. Note that the first two samples have the same three scores. The reason for this is that two of the students (C and D) have the same score, and, hence, the samples ABC and ABD contain the same values. The mean of each sample is obtained by dividing the sum of the three scores included in that sample by 3. For instance, the mean of the first sample is  $(70 + 78 + 80)/3 = 76$ . Note that the values of the means of samples in Table 7.3 are rounded to two decimal places.

By using the values of  $\bar{x}$  given in Table 7.3, we record the frequency distribution of  $\bar{x}$  in Table 7.4. By dividing the frequencies of the various values of  $\bar{x}$  by the sum of all frequencies, we obtain the relative frequencies of classes, which are listed in the third column of Table 7.4. These relative frequencies are used as probabilities and listed in Table 7.5. This table gives the sampling distribution of  $\bar{x}$ .

If we select just one sample of three scores from the population of five scores, we may draw any of the 10 possible samples. Hence, the sample mean,  $\bar{x}$ , can assume any of the values listed in Table 7.5 with the corresponding probability. For instance, the probability that the mean of a randomly selected sample of three scores is 81.67 is .20. This probability can be written as

$$P(\bar{x} = 81.67) = .20$$

**Table 7.3** All Possible Samples and Their Means When the Sample Size Is 3

Sample	Scores in the Sample	$\bar{x}$
ABC	70, 78, 80	76.00
ABD	70, 78, 80	76.00
ABE	70, 78, 95	81.00
ACD	70, 80, 80	76.67
ACE	70, 80, 95	81.67
ADE	70, 80, 95	81.67
BCD	78, 80, 80	79.33
BCE	78, 80, 95	84.33
BDE	78, 80, 95	84.33
CDE	80, 80, 95	85.00

**Table 7.4** Frequency and Relative Frequency Distributions of  $\bar{x}$  When the Sample Size Is 3

$\bar{x}$	$f$	Relative Frequency
76.00	2	2/10 = .20
76.67	1	1/10 = .10
79.33	1	1/10 = .10
81.00	1	1/10 = .10
81.67	2	2/10 = .20
84.33	2	2/10 = .20
85.00	1	1/10 = .10

$\Sigma f = 10$       Sum = 1.00

**Table 7.5** Sampling Distribution of  $\bar{x}$  When the Sample Size Is 3

$\bar{x}$	$P(\bar{x})$
76.00	.20
76.67	.10
79.33	.10
81.00	.10
81.67	.20
84.33	.20
85.00	.10

$\Sigma P(\bar{x}) = 1.00$

## 7.1.2 Sampling and Nonsampling Errors

Usually, different samples selected from the same population will give different results because they contain different elements. This is obvious from Table 7.3, which shows that the mean of a sample of three scores depends on which three of the five scores are included in the sample. The result obtained from any one sample will generally be different from the result obtained from the corresponding population. The difference between the value of a sample statistic obtained from a sample and the value of the corresponding population parameter obtained from the population is called the **sampling error**. Note that this difference represents the sampling error only if the sample is random and no nonsampling error has been made. Otherwise, only a part of this difference will be due to the sampling error.

### Definition

**Sampling Error** *Sampling error* is the difference between the value of a sample statistic and the value of the corresponding population parameter. In the case of the mean,

$$\text{Sampling error} = \bar{x} - \mu$$

assuming that the sample is random and no nonsampling error has been made.

It is important to remember that *a sampling error occurs because of chance*. The errors that occur for other reasons, such as errors made during collection, recording, and tabulation of data, are called **nonsampling errors**. These errors occur because of human mistakes, and not chance. Note that there is only one kind of sampling error—the error that occurs due to chance. However, there is not just one nonsampling error, but there are many nonsampling errors that may occur for different reasons.

### Definition

**Nonsampling Errors** The errors that occur in the collection, recording, and tabulation of data are called *nonsampling errors*.

The following paragraph, reproduced from the *Current Population Reports* of the U.S. Bureau of the Census, explains how nonsampling errors can occur.

Nonsampling errors can be attributed to many sources, e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, errors made in collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, biases resulting from the differing recall periods caused by the interviewing pattern used, and failure of all units in the universe to have some probability of being selected for the sample (undercoverage).

The following are the main reasons for the occurrence of nonsampling errors.

1. If a sample is nonrandom (and, hence, most likely nonrepresentative), the sample results may be too different from the census results. Even a randomly selected sample can become nonrandom if some of the members included in the sample cannot be contacted. A very good example of this comes from an article published in a magazine in 1988. As reported in a July 11, 1988, article in *U.S. News & World Report* ("The Numbers Racket: How Polls and Statistics Lie"), during the 1984 presidential election a test poll was conducted in which the only subjects interviewed were those who could be reached on the first try. The results of this poll indicated that Ronald Reagan had a 3 percentage point lead over Walter Mondale. However, when interviewers made an effort to contact everyone on their lists (calling some households up to 30 times before reaching someone), this lead increased to 13%. It turned out that this 13% lead was much closer to the actual election results. Apparently, people who planned to vote Republican spent less time at home.
2. The questions may be phrased in such a way that they are not fully understood by the members of the sample or population. As a result, the answers obtained are not accurate.
3. The respondents may intentionally give false information in response to some sensitive questions. For example, people may not tell the truth about their drinking habits, incomes, or opinions about minorities. Sometimes the respondents may give wrong answers because of ignorance. For example, a person may not remember the exact amount he or she spent on clothes last year. If asked in a survey, he or she may give an inaccurate answer.
4. The poll taker may make a mistake and enter a wrong number in the records or make an error while entering the data on a computer.

Note that nonsampling errors can occur both in a sample survey and in a census, whereas sampling error occurs only when a sample survey is conducted. Nonsampling errors can be minimized by preparing the survey questionnaire carefully and handling the data cautiously. However, it is impossible to avoid sampling error.

Example 7–1 illustrates the sampling and nonsampling errors using the mean.

### ■ EXAMPLE 7–1

*Illustrating sampling and nonsampling errors.*

Reconsider the population of five scores given in Table 7.1. Suppose one sample of three scores is selected from this population, and this sample includes the scores 70, 80, and 95. Find the sampling error.

**Solution** The scores of the five students are 70, 78, 80, 80, and 95. The population mean is

$$\mu = \frac{70 + 78 + 80 + 80 + 95}{5} = 80.60$$

Now a random sample of three scores from this population is taken and this sample includes the scores 70, 80, and 95. The mean for this sample is

$$\bar{x} = \frac{70 + 80 + 95}{3} = 81.67$$

Consequently,

$$\text{Sampling error} = \bar{x} - \mu = 81.67 - 80.60 = \mathbf{1.07}$$

That is, the mean score estimated from the sample is 1.07 higher than the mean score of the population. Note that this difference occurred due to chance—that is, because we used a sample instead of the population. ■

Now suppose, when we select the sample of three scores, we mistakenly record the second score as 82 instead of 80. As a result, we calculate the sample mean as

$$\bar{x} = \frac{70 + 82 + 95}{3} = 82.33$$

Consequently, the difference between this sample mean and the population mean is

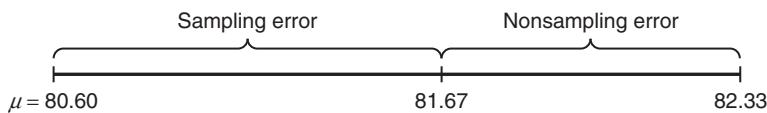
$$\bar{x} - \mu = 82.33 - 80.60 = 1.73$$

However, this difference between the sample mean and the population mean does not represent the sampling error. As we calculated earlier, only 1.07 of this difference is due to the sampling error. The remaining portion, which is equal to  $1.73 - 1.07 = .66$ , represents the nonsampling error because it occurred due to the error we made in recording the second score in the sample. Thus, in this case,

$$\text{Sampling error} = \mathbf{1.07}$$

$$\text{Nonsampling error} = \mathbf{.66}$$

Figure 7.1 shows the sampling and nonsampling errors for these calculations.



**Figure 7.1** Sampling and nonsampling errors.

Thus, the sampling error is the difference between the correct value of  $\bar{x}$  and the value of  $\mu$ , where the correct value of  $\bar{x}$  is the value of  $\bar{x}$  that does not contain any nonsampling errors. In contrast, the nonsampling error(s) is (are) obtained by subtracting the correct value of  $\bar{x}$  from the incorrect value of  $\bar{x}$ , where the incorrect value of  $\bar{x}$  is the value that contains the nonsampling error(s). For our example,

$$\text{Sampling error} = \bar{x} - \mu = 81.67 - 80.60 = 1.07$$

$$\text{Nonsampling error} = \text{Incorrect } \bar{x} - \text{Correct } \bar{x} = 82.33 - 81.67 = .66$$

Note that in the real world we do not know the mean of a population. Hence, we select a sample to use the sample mean as an estimate of the population mean. Consequently, we never know the size of the sampling error.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**7.1** Briefly explain the meaning of a population distribution and a sampling distribution. Give an example of each.

**7.2** Explain briefly the meaning of sampling error. Give an example. Does such an error occur only in a sample survey, or can it occur in both a sample survey and a census?

**7.3** Explain briefly the meaning of nonsampling errors. Give an example. Do such errors occur only in a sample survey, or can they occur in both a sample survey and a census?

- 7.4** Consider the following population of six numbers.

15      13      8      17      9      12

- Find the population mean.
- Liza selected one sample of four numbers from this population. The sample included the numbers 13, 8, 9, and 12. Calculate the sample mean and sampling error for this sample.
- Refer to part b. When Liza calculated the sample mean, she mistakenly used the numbers 13, 8, 6, and 12 to calculate the sample mean. Find the sampling and nonsampling errors in this case.
- List all samples of four numbers (without replacement) that can be selected from this population. Calculate the sample mean and sampling error for each of these samples.

- 7.5** Consider the following population of 10 numbers.

20      25      13      19      9      15      11      7      17      30

- Find the population mean.
- Rich selected one sample of nine numbers from this population. The sample included the numbers 20, 25, 13, 9, 15, 11, 7, 17, and 30. Calculate the sample mean and sampling error for this sample.
- Refer to part b. When Rich calculated the sample mean, he mistakenly used the numbers 20, 25, 13, 9, 15, 11, 17, 17, and 30 to calculate the sample mean. Find the sampling and nonsampling errors in this case.
- List all samples of nine numbers (without replacement) that can be selected from this population. Calculate the sample mean and sampling error for each of these samples.

## ■ APPLICATIONS

- 7.6** Using the formulas of Section 5.3 of Chapter 5 for the mean and standard deviation of a discrete random variable, verify that the mean and standard deviation for the population probability distribution of Table 7.2 are 80.60 and 8.09, respectively.

- 7.7** The following data give the ages (in years) of all six members of a family.

55      53      28      25      21      15

- Let  $x$  denote the age of a member of this family. Write the population distribution of  $x$ .
- List all the possible samples of size four (without replacement) that can be selected from this population. Calculate the mean for each of these samples. Write the sampling distribution of  $\bar{x}$ .
- Calculate the mean for the population data. Select one random sample of size four and calculate the sample mean  $\bar{x}$ . Compute the sampling error.

- 7.8** The following data give the years of teaching experience for all five faculty members of a department at a university.

7      8      14      7      20

- Let  $x$  denote the years of teaching experience for a faculty member of this department. Write the population distribution of  $x$ .
- List all the possible samples of size three (without replacement) that can be selected from this population. Calculate the mean for each of these samples. Write the sampling distribution of  $\bar{x}$ .
- Calculate the mean for the population data. Select one random sample of size three and calculate the sample mean  $\bar{x}$ . Compute the sampling error.

## 7.2 Mean and Standard Deviation of $\bar{x}$

The mean and standard deviation calculated for the sampling distribution of  $\bar{x}$  are called the **mean and standard deviation of  $\bar{x}$** . Actually, the mean and standard deviation of  $\bar{x}$  are, respectively, the mean and standard deviation of the means of all samples of the same size selected from a population. The standard deviation of  $\bar{x}$  is also called the **standard error of  $\bar{x}$** .

### Definition

**Mean and Standard Deviation of  $\bar{x}$**  The mean and standard deviation of the sampling distribution of  $\bar{x}$  are called the *mean and standard deviation of  $\bar{x}$*  and are denoted by  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$ , respectively.

If we calculate the mean and standard deviation of the 10 values of  $\bar{x}$  listed in Table 7.3, we obtain the mean,  $\mu_{\bar{x}}$ , and the standard deviation,  $\sigma_{\bar{x}}$ , of  $\bar{x}$ . Alternatively, we can calculate the mean and standard deviation of the sampling distribution of  $\bar{x}$  listed in Table 7.5. These will also be the values of  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$ . From these calculations, we will obtain  $\mu_{\bar{x}} = 80.60$  and  $\sigma_{\bar{x}} = 3.30$  (see Exercise 7.25 at the end of this section).

The mean of the sampling distribution of  $\bar{x}$  is always equal to the mean of the population.

**Mean of the Sampling Distribution of  $\bar{x}$**  The *mean of the sampling distribution of  $\bar{x}$*  is always equal to the mean of the population. Thus,

$$\mu_{\bar{x}} = \mu$$

Thus, if we select all possible samples (of the same size) from a population and calculate their means, the mean ( $\mu_{\bar{x}}$ ) of all these sample means will be the same as the mean ( $\mu$ ) of the population. If we calculate the mean for the population probability distribution of Table 7.2 and the mean for the sampling distribution of Table 7.5 by using the formula learned in Section 5.3 of Chapter 5, we get the same value of 80.60 for  $\mu$  and  $\mu_{\bar{x}}$  (see Exercise 7.25).

The sample mean,  $\bar{x}$ , is called an **estimator** of the population mean,  $\mu$ . When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that sample statistic is said to be an **unbiased estimator**. For the sample mean  $\bar{x}$ ,  $\mu_{\bar{x}} = \mu$ . Hence,  $\bar{x}$  is an unbiased estimator of  $\mu$ . This is a very important property that an estimator should possess.

However, the standard deviation,  $\sigma_{\bar{x}}$ , of  $\bar{x}$  is not equal to the standard deviation,  $\sigma$ , of the population distribution (unless  $n = 1$ ). The standard deviation of  $\bar{x}$  is equal to the standard deviation of the population divided by the square root of the sample size; that is,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This formula for the standard deviation of  $\bar{x}$  holds true only when the sampling is done either with replacement from a finite population or with or without replacement from an infinite population. These two conditions can be replaced by the condition that the above formula holds true if the sample size is small in comparison to the population size. The sample size is considered to be small compared to the population size if the sample size is equal to or less than 5% of the population size; that is, if

$$\frac{n}{N} \leq .05$$

If this condition is not satisfied, we use the following formula to calculate  $\sigma_{\bar{x}}$ :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

where the factor  $\sqrt{\frac{N-n}{N-1}}$  is called the finite population correction factor.

In most practical applications, the sample size is small compared to the population size. Consequently, in most cases, the formula used to calculate  $\sigma_{\bar{x}}$  is  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ .

**Standard Deviation of the Sampling Distribution of  $\bar{x}$**  The *standard deviation of the sampling distribution of  $\bar{x}$*  is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the standard deviation of the population and  $n$  is the sample size. This formula is used when  $n/N \leq .05$ , where  $N$  is the population size.

Following are two important observations regarding the sampling distribution of  $\bar{x}$ .

1. *The spread of the sampling distribution of  $\bar{x}$  is smaller than the spread of the corresponding population distribution.* In other words,  $\sigma_{\bar{x}} < \sigma$ . This is obvious from the formula for  $\sigma_{\bar{x}}$ . When  $n$  is greater than 1, which is usually true, the denominator in  $\sigma/\sqrt{n}$  is greater than 1. Hence,  $\sigma_{\bar{x}}$  is smaller than  $\sigma$ .
2. *The standard deviation of the sampling distribution of  $\bar{x}$  decreases as the sample size increases.* This feature of the sampling distribution of  $\bar{x}$  is also obvious from the formula

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is said to be a **consistent estimator**. This is another important property that an estimator should possess. It is obvious from the above formula for  $\sigma_{\bar{x}}$  that as  $n$  increases, the value of  $\sqrt{n}$  also increases and, consequently, the value of  $\sigma/\sqrt{n}$  decreases. Thus, the sample mean  $\bar{x}$  is a consistent estimator of the population mean  $\mu$ . Example 7–2 illustrates this feature.

## ■ EXAMPLE 7–2

*Finding the mean and standard deviation of  $\bar{x}$ .*



Image Source/GettyImages, Inc.

The mean wage per hour for all 5000 employees who work at a large company is \$27.50, and the standard deviation is \$3.70. Let  $\bar{x}$  be the mean wage per hour for a random sample of certain employees selected from this company. Find the mean and standard deviation of  $\bar{x}$  for a sample size of

- (a) 30      (b) 75      (c) 200

**Solution** From the given information, for the population of all employees,

$$N = 5000, \quad \mu = \$27.50, \quad \text{and} \quad \sigma = \$3.70$$

- (a) The mean,  $\mu_{\bar{x}}$ , of the sampling distribution of  $\bar{x}$  is

$$\mu_{\bar{x}} = \mu = \$27.50$$

In this case,  $n = 30$ ,  $N = 5000$ , and  $n/N = 30/5000 = .006$ . Because  $n/N$  is less than .05, the standard deviation of  $\bar{x}$  is obtained by using the formula  $\sigma/\sqrt{n}$ . Hence,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.70}{\sqrt{30}} = \$0.676$$

Thus, we can state that if we take all possible samples of size 30 from the population of all employees of this company and prepare the sampling distribution of  $\bar{x}$ , the mean and standard deviation of this sampling distribution of  $\bar{x}$  will be \$27.50 and \$.676, respectively.

- (b) In this case,  $n = 75$  and  $n/N = 75/5000 = .015$ , which is less than .05. The mean and standard deviation of  $\bar{x}$  are

$$\mu_{\bar{x}} = \mu = \$27.50 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.70}{\sqrt{75}} = \$0.427$$

- (c) In this case,  $n = 200$  and  $n/N = 200/5000 = .04$ , which is less than .05. Therefore, the mean and standard deviation of  $\bar{x}$  are

$$\mu_{\bar{x}} = \mu = \$27.50 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.70}{\sqrt{200}} = \$0.262$$

From the preceding calculations we observe that the mean of the sampling distribution of  $\bar{x}$  is always equal to the mean of the population whatever the size of the sample. However, the value of the standard deviation of  $\bar{x}$  decreases from \$.676 to \$.427 and then to \$.262 as the sample size increases from 30 to 75 and then to 200. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

- 7.9** Let  $\bar{x}$  be the mean of a sample selected from a population.
- What is the mean of the sampling distribution of  $\bar{x}$  equal to?
  - What is the standard deviation of the sampling distribution of  $\bar{x}$  equal to? Assume  $n/N \leq .05$ .
- 7.10** What is an estimator? When is an estimator unbiased? Is the sample mean,  $\bar{x}$ , an unbiased estimator of  $\mu$ ? Explain.
- 7.11** When is an estimator said to be consistent? Is the sample mean,  $\bar{x}$ , a consistent estimator of  $\mu$ ? Explain.
- 7.12** How does the value of  $\sigma_{\bar{x}}$  change as the sample size increases? Explain.
- 7.13** Consider a large population with  $\mu = 60$  and  $\sigma = 10$ . Assuming  $n/N \leq .05$ , find the mean and standard deviation of the sample mean,  $\bar{x}$ , for a sample size of
- 18
  - 90
- 7.14** Consider a large population with  $\mu = 90$  and  $\sigma = 18$ . Assuming  $n/N \leq .05$ , find the mean and standard deviation of the sample mean,  $\bar{x}$ , for a sample size of
- 10
  - 35
- 7.15** A population of  $N = 5000$  has  $\sigma = 25$ . In each of the following cases, which formula will you use to calculate  $\sigma_{\bar{x}}$  and why? Using the appropriate formula, calculate  $\sigma_{\bar{x}}$  for each of these cases.
- $n = 300$
  - $n = 100$
- 7.16** A population of  $N = 100,000$  has  $\sigma = 40$ . In each of the following cases, which formula will you use to calculate  $\sigma_{\bar{x}}$  and why? Using the appropriate formula, calculate  $\sigma_{\bar{x}}$  for each of these cases.
- $n = 2500$
  - $n = 7000$
- \*7.17** For a population,  $\mu = 125$  and  $\sigma = 36$ .
- For a sample selected from this population,  $\mu_{\bar{x}} = 125$  and  $\sigma_{\bar{x}} = 3.6$ . Find the sample size. Assume  $n/N \leq .05$ .
  - For a sample selected from this population,  $\mu_{\bar{x}} = 125$  and  $\sigma_{\bar{x}} = 2.25$ . Find the sample size. Assume  $n/N \leq .05$ .
- \*7.18** For a population,  $\mu = 46$  and  $\sigma = 10$ .
- For a sample selected from this population,  $\mu_{\bar{x}} = 46$  and  $\sigma_{\bar{x}} = 2.0$ . Find the sample size. Assume  $n/N \leq .05$ .
  - For a sample selected from this population,  $\mu_{\bar{x}} = 46$  and  $\sigma_{\bar{x}} = 1.6$ . Find the sample size. Assume  $n/N \leq .05$ .

### APPLICATIONS

- 7.19** According to the Project on Student Debt, the average student loan for college graduates of the class of 2010 was \$25,000 (*USA TODAY*, April 24, 2012). Suppose that the student loans for all college graduates of the class of 2010 have a mean of \$25,000 and a standard deviation of \$6280. Let  $\bar{x}$  be the average student loan of a random sample of 400 college graduates from the class of 2010. Find the mean and standard deviation of the sampling distribution of  $\bar{x}$ .
- 7.20** The living spaces of all homes in a city have a mean of 2300 square feet and a standard deviation of 500 square feet. Let  $\bar{x}$  be the mean living space for a random sample of 25 homes selected from this city. Find the mean and standard deviation of the sampling distribution of  $\bar{x}$ .
- 7.21** According to a report in *The New York Times*, bank tellers in the United States earn an average of \$25,510 a year (Jessica Silver-Greenberg, *The New York Times*, April 22, 2012). Suppose that the current distribution of salaries of all bank tellers in the United States has a mean of \$25,510 and a standard deviation of \$4550. Let  $\bar{x}$  be the average salary of a random sample of 200 such tellers. Find the mean and standard deviation of the sampling distribution of  $\bar{x}$ .
- 7.22** According to the American Automobile Association's 2012 annual report *Your Driving Costs*, the cost of owning and operating a four-wheel drive SUV is \$11,350 per year (*USA TODAY*, April 27, 2012). Note that this cost includes expenses for gasoline, maintenance, insurance, and financing for a vehicle that is driven 15,000 miles a year. Suppose that the distribution of such costs of owning and operating all four-wheel drive SUVs has a mean of \$11,350 with a standard deviation of \$2390. Let  $\bar{x}$  be the average of such costs of owning and operating a four-wheel drive SUV based on a random sample of 400 four-wheel drive SUVs. Find the mean and standard deviation of the sampling distribution of  $\bar{x}$ .

**\*7.23** Suppose the standard deviation of recruiting costs per player for all female basketball players recruited by all public universities in the Midwest is \$2000. Let  $\bar{x}$  be the mean recruiting cost for a sample of a certain number of such players. What sample size will give the standard deviation of  $\bar{x}$  equal to \$125? Assume  $n/N \leq .05$ .

**\*7.24** The standard deviation of the 2011 gross sales of all corporations is known to be \$139.50 million. Let  $\bar{x}$  be the mean of the 2011 gross sales of a sample of corporations. What sample size will produce the standard deviation of  $\bar{x}$  equal to \$15.50 million? Assume  $n/N \leq .05$ .

**\*7.25** Consider the sampling distribution of  $\bar{x}$  given in Table 7.5.

- Calculate the value of  $\mu_{\bar{x}}$  using the formula  $\mu_{\bar{x}} = \sum \bar{x} P(\bar{x})$ . Is the value of  $\mu$  calculated in Exercise 7.6 the same as the value of  $\mu_{\bar{x}}$  calculated here?
- Calculate the value of  $\sigma_{\bar{x}}$  by using the formula

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - (\mu_{\bar{x}})^2}$$

- From Exercise 7.6,  $\sigma = 8.09$ . Also, our sample size is 3, so that  $n = 3$ . Therefore,  $\sigma/\sqrt{n} = 8.09/\sqrt{3} = 4.67$ . From part b, you should get  $\sigma_{\bar{x}} = 3.30$ . Why does  $\sigma/\sqrt{n}$  not equal  $\sigma_{\bar{x}}$  in this case?
- In our example (given in the beginning of Section 7.1) on scores,  $N = 5$  and  $n = 3$ . Hence,  $n/N = 3/5 = .60$ . Because  $n/N$  is greater than .05, the appropriate formula to find  $\sigma_{\bar{x}}$  is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Show that the value of  $\sigma_{\bar{x}}$  calculated by using this formula gives the same value as the one calculated in part b above.

## 7.3 Shape of the Sampling Distribution of $\bar{x}$

The shape of the sampling distribution of  $\bar{x}$  relates to the following two cases:

- The population from which samples are drawn has a normal distribution.
- The population from which samples are drawn does not have a normal distribution.

### 7.3.1 Sampling from a Normally Distributed Population

When the population from which samples are drawn is normally distributed with its mean equal to  $\mu$  and standard deviation equal to  $\sigma$ , then:

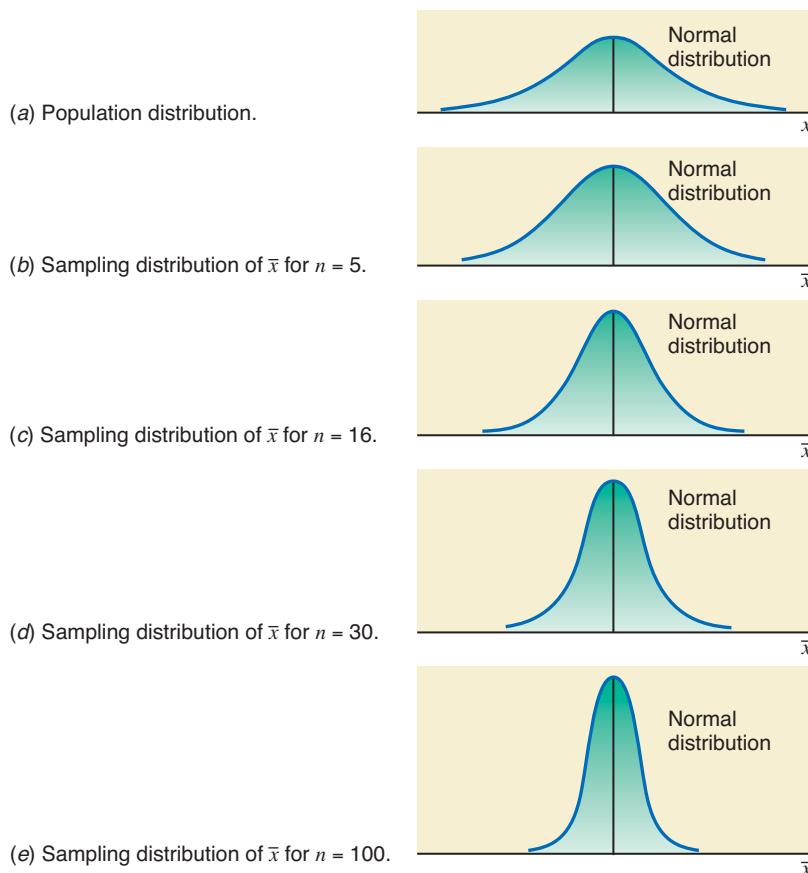
- The mean of  $\bar{x}$ ,  $\mu_{\bar{x}}$ , is equal to the mean of the population,  $\mu$ .
- The standard deviation of  $\bar{x}$ ,  $\sigma_{\bar{x}}$ , is equal to  $\sigma/\sqrt{n}$ , assuming  $n/N \leq .05$ .
- The shape of the sampling distribution of  $\bar{x}$  is normal, whatever the value of  $n$ .

**Sampling Distribution of  $\bar{x}$  When the Population Has a Normal Distribution** If the population from which the samples are drawn is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , then the sampling distribution of the sample mean,  $\bar{x}$ , will also be normally distributed with the following mean and standard deviation, irrespective of the sample size:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

**Remember ▶** For  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$  to be true,  $n/N$  must be less than or equal to .05.

Figure 7.2a shows the probability distribution curve for a population. The distribution curves in Figure 7.2b through Figure 7.2e show the sampling distributions of  $\bar{x}$  for different sample sizes taken from the population of Figure 7.2a. As we can observe, the population has a normal distribution. Because of this, the sampling distribution of  $\bar{x}$  is normal for each of



**Figure 7.2** Population distribution and sampling distributions of  $\bar{x}$ .

the four cases illustrated in Figure 7.2b through Figure 7.2e. Also notice from Figure 7.2b through Figure 7.2e that the spread of the sampling distribution of  $\bar{x}$  decreases as the sample size increases.

Example 7–3 illustrates the calculation of the mean and standard deviation of  $\bar{x}$  and the description of the shape of its sampling distribution.

### ■ EXAMPLE 7–3

In a recent SAT, the mean score for all examinees was 1020. Assume that the distribution of SAT scores of all examinees is normal with a mean of 1020 and a standard deviation of 153. Let  $\bar{x}$  be the mean SAT score of a random sample of certain examinees. Calculate the mean and standard deviation of  $\bar{x}$  and describe the shape of its sampling distribution when the sample size is

- (a) 16      (b) 50      (c) 1000

*Finding the mean, standard deviation, and sampling distribution of  $\bar{x}$ : normal population.*

**Solution** Let  $\mu$  and  $\sigma$  be the mean and standard deviation of SAT scores of all examinees, and let  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$  be the mean and standard deviation of the sampling distribution of  $\bar{x}$ , respectively. Then, from the given information,

$$\mu = 1020 \quad \text{and} \quad \sigma = 153$$

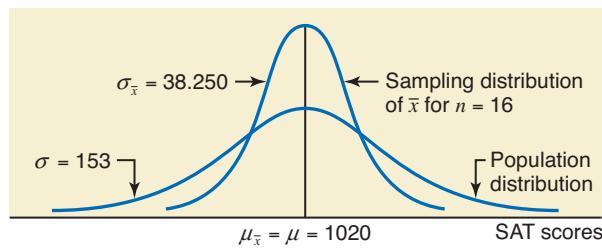
- (a) The mean and standard deviation of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu = 1020 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{16}} = 38.250$$

Because the SAT scores of all examinees are assumed to be normally distributed, the sampling distribution of  $\bar{x}$  for samples of 16 examinees is also normal. Figure 7.3

shows the population distribution and the sampling distribution of  $\bar{x}$ . Note that because  $\sigma$  is greater than  $\sigma_{\bar{x}}$ , the population distribution has a wider spread but smaller height than the sampling distribution of  $\bar{x}$  in Figure 7.3.

Figure 7.3

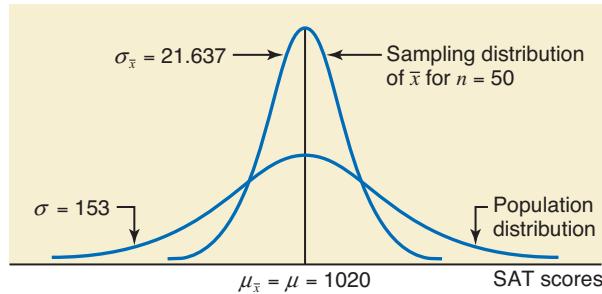


- (b) The mean and standard deviation of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu = 1020 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{50}} = 21.637$$

Again, because the SAT scores of all examinees are assumed to be normally distributed, the sampling distribution of  $\bar{x}$  for samples of 50 examinees is also normal. The population distribution and the sampling distribution of  $\bar{x}$  are shown in Figure 7.4.

Figure 7.4

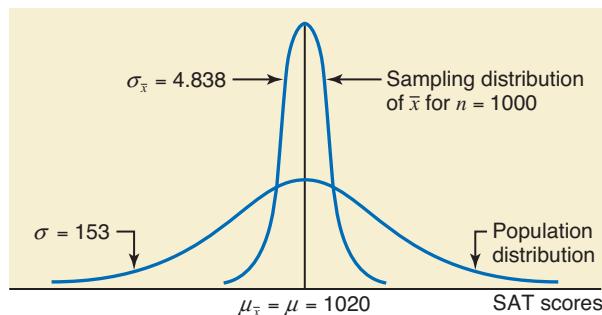


- (c) The mean and standard deviation of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu = 1020 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{1000}} = 4.838$$

Again, because the SAT scores of all examinees are assumed to be normally distributed, the sampling distribution of  $\bar{x}$  for samples of 1000 examinees is also normal. The two distributions are shown in Figure 7.5.

Figure 7.5



Thus, whatever the sample size, the sampling distribution of  $\bar{x}$  is normal when the population from which the samples are drawn is normally distributed. ■

### 7.3.2 Sampling from a Population That Is Not Normally Distributed

Most of the time the population from which the samples are selected is not normally distributed. In such cases, the shape of the sampling distribution of  $\bar{x}$  is inferred from a very important theorem called the **central limit theorem**.

**Central Limit Theorem** According to the *central limit theorem*, for a large sample size, the sampling distribution of  $\bar{x}$  is approximately normal, irrespective of the shape of the population distribution. The mean and standard deviation of the sampling distribution of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

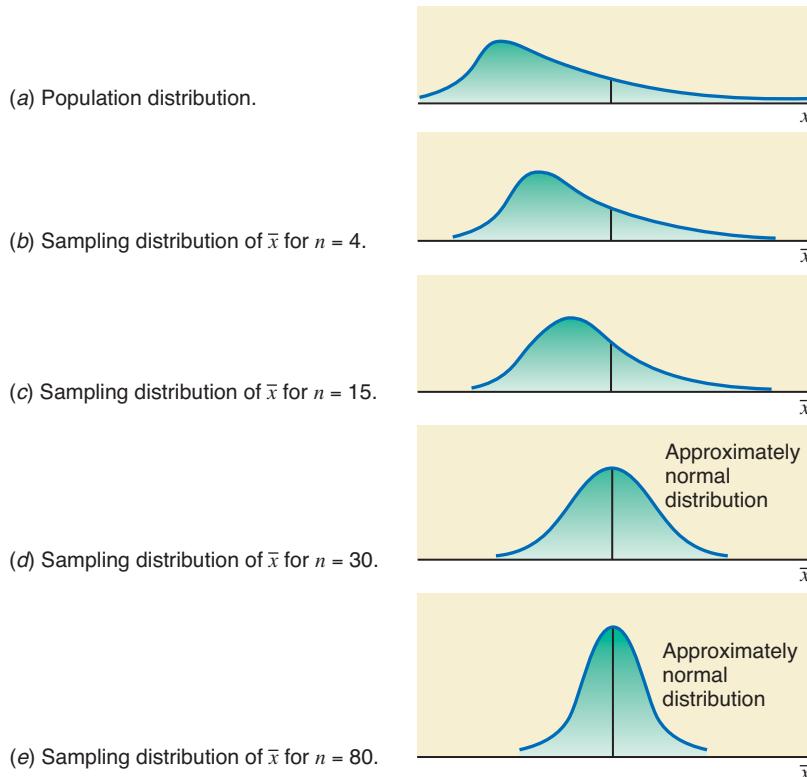
The sample size is usually considered to be large if  $n \geq 30$ .

Note that when the population does not have a normal distribution, the shape of the sampling distribution is not exactly normal, but it is approximately normal for a large sample size. The approximation becomes more accurate as the sample size increases. Another point to remember is that the central limit theorem applies to *large* samples only. Usually, if the sample size is 30 or larger, it is considered sufficiently large so that the central limit theorem can be applied to the sampling distribution of  $\bar{x}$ . Thus:

1. When  $n \geq 30$ , the shape of the sampling distribution of  $\bar{x}$  is approximately normal irrespective of the shape of the population distribution. This is so due to the central limit theorem.
2. The mean of  $\bar{x}$ ,  $\mu_{\bar{x}}$ , is equal to the mean of the population,  $\mu$ .
3. The standard deviation of  $\bar{x}$ ,  $\sigma_{\bar{x}}$ , is equal to  $\sigma/\sqrt{n}$  if  $n/N \leq .05$ .

Again, remember that for  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$  to apply,  $n/N$  must be less than or equal to .05, otherwise we multiply  $\sigma/\sqrt{n}$  by the finite population correction factor explained earlier in this chapter.

Figure 7.6a shows the probability distribution curve for a population. The distribution curves in Figure 7.6b through Figure 7.6e show the sampling distributions of  $\bar{x}$  for different sample



**Figure 7.6** Population distribution and sampling distributions of  $\bar{x}$ .

sizes taken from the population of Figure 7.6a. As we can observe, the population is not normally distributed. The sampling distributions of  $\bar{x}$  shown in parts b and c, when  $n < 30$ , are not normal. However, the sampling distributions of  $\bar{x}$  shown in parts d and e, when  $n \geq 30$ , are (approximately) normal. Also notice that the spread of the sampling distribution of  $\bar{x}$  decreases as the sample size increases.

Example 7–4 illustrates the calculation of the mean and standard deviation of  $\bar{x}$  and describes the shape of the sampling distribution of  $\bar{x}$  when the sample size is large.

### ■ EXAMPLE 7–4

*Finding the mean, standard deviation, and sampling distribution of  $\bar{x}$ : nonnormal population.*

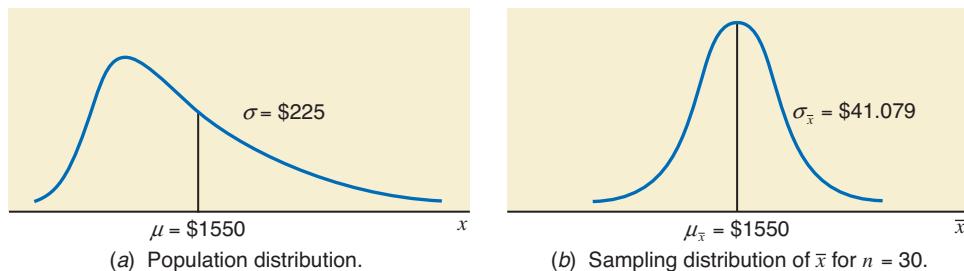
- (a) 30      (b) 100

**Solution** Although the population distribution of rents paid by all tenants is not normal, in each case the sample size is large ( $n \geq 30$ ). Hence, the central limit theorem can be applied to infer the shape of the sampling distribution of  $\bar{x}$ .

- (a) Let  $\bar{x}$  be the mean rent paid by a sample of 30 tenants. Then, the sampling distribution of  $\bar{x}$  is approximately normal with the values of the mean and standard deviation given as

$$\mu_{\bar{x}} = \mu = \$1550 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{225}{\sqrt{30}} = \$41.079$$

Figure 7.7 shows the population distribution and the sampling distribution of  $\bar{x}$ .

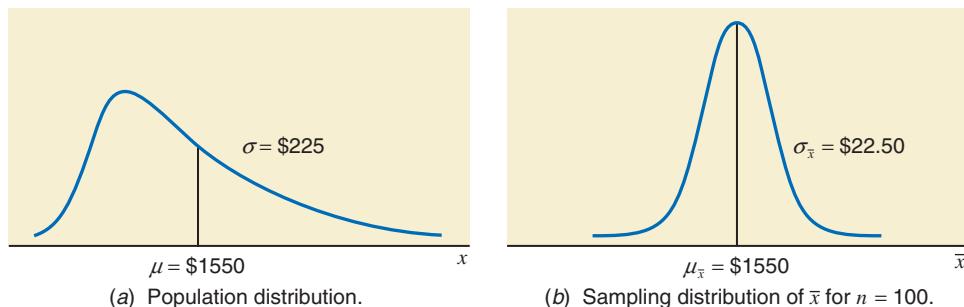


**Figure 7.7**

- (b) Let  $\bar{x}$  be the mean rent paid by a sample of 100 tenants. Then, the sampling distribution of  $\bar{x}$  is approximately normal with the values of the mean and standard deviation given as

$$\mu_{\bar{x}} = \mu = \$1550 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{225}{\sqrt{100}} = \$22.50$$

Figure 7.8 shows the population distribution and the sampling distribution of  $\bar{x}$ .



**Figure 7.8**

## EXERCISES

### CONCEPTS AND PROCEDURES



**7.26** What condition or conditions must hold true for the sampling distribution of the sample mean to be normal when the sample size is less than 30?

**7.27** Explain the central limit theorem.

**7.28** A population has a distribution that is skewed to the left. Indicate in which of the following cases the central limit theorem will apply to describe the sampling distribution of the sample mean.

- a.  $n = 400$       b.  $n = 25$       c.  $n = 36$

**7.29** A population has a distribution that is skewed to the right. A sample of size  $n$  is selected from this population. Describe the shape of the sampling distribution of the sample mean for each of the following cases.

- a.  $n = 25$       b.  $n = 80$       c.  $n = 29$

**7.30** A population has a normal distribution. A sample of size  $n$  is selected from this population. Describe the shape of the sampling distribution of the sample mean for each of the following cases.

- a.  $n = 94$       b.  $n = 11$

**7.31** A population has a normal distribution. A sample of size  $n$  is selected from this population. Describe the shape of the sampling distribution of the sample mean for each of the following cases.

- a.  $n = 23$       b.  $n = 450$

### APPLICATIONS

**7.32** The delivery times for all food orders at a fast-food restaurant during the lunch hour are normally distributed with a mean of 7.7 minutes and a standard deviation of 2.1 minutes. Let  $\bar{x}$  be the mean delivery time for a random sample of 16 orders at this restaurant. Calculate the mean and standard deviation of  $\bar{x}$ , and describe the shape of its sampling distribution.

**7.33** Among college students who hold part-time jobs during the school year, the distribution of the time spent working per week is approximately normally distributed with a mean of 20.20 hours and a standard deviation of 2.60 hours. Let  $\bar{x}$  be the average time spent working per week for 18 randomly selected college students who hold part-time jobs during the school year. Calculate the mean and the standard deviation of the sampling distribution of  $\bar{x}$ , and describe the shape of this sampling distribution.

**7.34** The amounts of electricity bills for all households in a particular city have an approximately normal distribution with a mean of \$140 and a standard deviation of \$30. Let  $\bar{x}$  be the mean amount of electricity bills for a random sample of 25 households selected from this city. Find the mean and standard deviation of  $\bar{x}$ , and comment on the shape of its sampling distribution.

**7.35** The GPAs of all 5540 students enrolled at a university have an approximately normal distribution with a mean of 3.02 and a standard deviation of .29. Let  $\bar{x}$  be the mean GPA of a random sample of 48 students selected from this university. Find the mean and standard deviation of  $\bar{x}$ , and comment on the shape of its sampling distribution.

**7.36** The weights of all people living in a particular town have a distribution that is skewed to the right with a mean of 133 pounds and a standard deviation of 24 pounds. Let  $\bar{x}$  be the mean weight of a random sample of 45 persons selected from this town. Find the mean and standard deviation of  $\bar{x}$  and comment on the shape of its sampling distribution.

**7.37** According to an estimate, the average age at first marriage for men in the United States was 28.2 years in 2010 (*Time*, March 21, 2011). Suppose that currently the mean age for all U.S. men at the time of first marriage is 28.2 years with a standard deviation of 6 years and that this distribution is strongly skewed to the right. Let  $\bar{x}$  be the average age at the time of first marriage for 25 randomly selected U.S. men. Find the mean and the standard deviation of the sampling distribution of  $\bar{x}$ . What if the sample size is 100? How do the shapes of the sampling distributions differ for the two sample sizes?

**7.38** Suppose that the incomes of all people in the United States who own hybrid (gas and electric) automobiles are normally distributed with a mean of \$78,000 and a standard deviation of \$8300. Let  $\bar{x}$  be the mean income of a random sample of 50 owners of such automobiles. Calculate the mean and standard deviation of  $\bar{x}$  and describe the shape of its sampling distribution.

**7.39** According to the American Time Use Survey, Americans watch television on weekdays for an average of 151 minutes per day (*Time*, July 11, 2011). Suppose that the current distribution of times spent

watching television per weekday by all Americans has a mean of 151 minutes and a standard deviation of 20 minutes. Let  $\bar{x}$  be the average time spent watching television on a weekday by 200 randomly selected Americans. Find the mean and the standard deviation of the sampling distribution of  $\bar{x}$ . What is the shape of the sampling distribution of  $\bar{x}$ ? Do you need to know the shape of the population distribution in order to make this conclusion? Explain why or why not.

## 7.4 Applications of the Sampling Distribution of $\bar{x}$

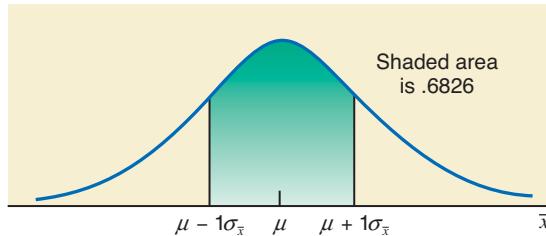
From the central limit theorem, for large samples, the sampling distribution of  $\bar{x}$  is approximately normal with mean  $\mu$  and standard deviation  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Based on this result, we can make the following statements about  $\bar{x}$  for large samples. The areas under the curve of  $\bar{x}$  mentioned in these statements are found from the normal distribution table.

- If we take all possible samples of the same (large) size from a population and calculate the mean for each of these samples, then about 68.26% of the sample means will be within one standard deviation ( $\sigma_{\bar{x}}$ ) of the population mean.* Alternatively, we can state that if we take one sample (of  $n \geq 30$ ) from a population and calculate the mean for this sample, the probability that this sample mean will be within one standard deviation ( $\sigma_{\bar{x}}$ ) of the population mean is .6826. That is,

$$P(\mu - 1\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 1\sigma_{\bar{x}}) = .8413 - .1587 = .6826$$

This probability is shown in Figure 7.9.

**Figure 7.9**  $P(\mu - 1\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 1\sigma_{\bar{x}})$

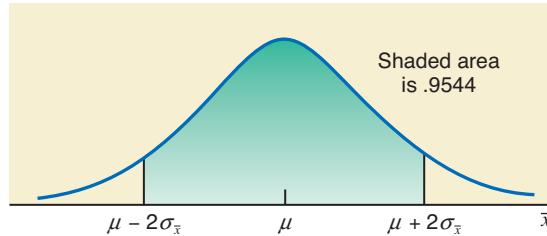


- If we take all possible samples of the same (large) size from a population and calculate the mean for each of these samples, then about 95.44% of the sample means will be within two standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean.* Alternatively, we can state that if we take one sample (of  $n \geq 30$ ) from a population and calculate the mean for this sample, the probability that this sample mean will be within two standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean is .9544. That is,

$$P(\mu - 2\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 2\sigma_{\bar{x}}) = .9772 - .0228 = .9544$$

This probability is shown in Figure 7.10.

**Figure 7.10**  $P(\mu - 2\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 2\sigma_{\bar{x}})$ .

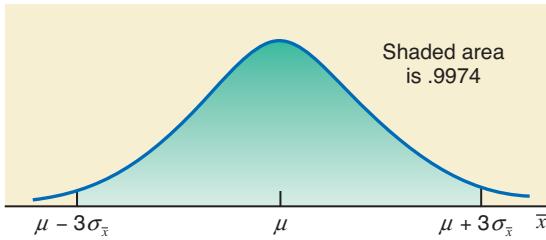


- If we take all possible samples of the same (large) size from a population and calculate the mean for each of these samples, then about 99.74% of the sample means will be within*

three standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean. Alternatively, we can state that if we take one sample (of  $n \geq 30$ ) from a population and calculate the mean for this sample, the probability that this sample mean will be within three standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean is .9974. That is,

$$P(\mu - 3\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 3\sigma_{\bar{x}}) = .9987 - .0013 = .9974$$

This probability is shown in Figure 7.11.



**Figure 7.11**  $P(\mu - 3\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 3\sigma_{\bar{x}})$

When conducting a survey, we usually select one sample and compute the value of  $\bar{x}$  based on that sample. We never select all possible samples of the same size and then prepare the sampling distribution of  $\bar{x}$ . Rather, we are more interested in finding the probability that the value of  $\bar{x}$  computed from one sample falls within a given interval. Examples 7–5 and 7–6 illustrate this procedure.

### ■ EXAMPLE 7–5

Assume that the weights of all packages of a certain brand of cookies are normally distributed with a mean of 32 ounces and a standard deviation of .3 ounce. Find the probability that the mean weight,  $\bar{x}$ , of a random sample of 20 packages of this brand of cookies will be between 31.8 and 31.9 ounces.

*Calculating the probability of  $\bar{x}$  in an interval: normal population.*

**Solution** Although the sample size is small ( $n < 30$ ), the shape of the sampling distribution of  $\bar{x}$  is normal because the population is normally distributed. The mean and standard deviation of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu = 32 \text{ ounces} \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.3}{\sqrt{20}} = .06708204 \text{ ounce}$$

We are to compute the probability that the value of  $\bar{x}$  calculated for one randomly drawn sample of 20 packages is between 31.8 and 31.9 ounces; that is,

$$P(31.8 < \bar{x} < 31.9)$$

This probability is given by the area under the normal distribution curve for  $\bar{x}$  between the points  $\bar{x} = 31.8$  and  $\bar{x} = 31.9$ . The first step in finding this area is to convert the two  $\bar{x}$  values to their respective  $z$  values.



© Burwell and BurwellPhotography/  
iStockphoto

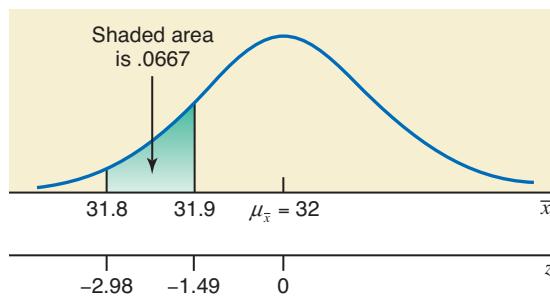
**z Value for a Value of  $\bar{x}$**  The  $z$  value for a value of  $\bar{x}$  is calculated as

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

The  $z$  values for  $\bar{x} = 31.8$  and  $\bar{x} = 31.9$  are computed below, and they are shown on the  $z$  scale below the normal distribution curve for  $\bar{x}$  in Figure 7.12.

$$\text{For } \bar{x} = 31.8: \quad z = \frac{31.8 - 32}{.06708204} = -2.98$$

$$\text{For } \bar{x} = 31.9: \quad z = \frac{31.9 - 32}{.06708204} = -1.49$$

Figure 7.12  $P(31.8 < \bar{x} < 31.9)$ 

The probability that  $\bar{x}$  is between 31.8 and 31.9 is given by the area under the standard normal curve between  $z = -2.98$  and  $z = -1.49$ . Thus, the required probability is

$$\begin{aligned} P(31.8 < \bar{x} < 31.9) &= P(-2.98 < z < -1.49) \\ &= P(z < -1.49) - P(z < -2.98) \\ &= .0681 - .0014 = .0667 \end{aligned}$$

Therefore, the probability is .0667 that the mean weight of a sample of 20 packages will be between 31.8 and 31.9 ounces. ■

### ■ EXAMPLE 7–6

*Calculating the probability of  $\bar{x}$  in an interval:  $n \geq 30$ .*

According to Moebs Services Inc., an individual checking account at major U.S. banks costs the banks between \$350 and \$450 per year (*Time*, November 21, 2011). Suppose that the current average cost of all checking accounts at major U.S. banks is \$400 per year with a standard deviation of \$30. Let  $\bar{x}$  be the current average annual cost of a random sample of 225 individual checking accounts at major banks in America.

- (a) What is the probability that the average annual cost of the checking accounts in this sample is within \$4 of the population mean?
- (b) What is the probability that the average annual cost of the checking accounts in this sample is less than the population mean by \$2.70 or more?

**Solution** From the given information, for the annual costs of all individual checking accounts at major banks in America,

$$\mu = \$400 \quad \text{and} \quad \sigma = \$30$$

Although the shape of the probability distribution of the population (annual costs of all individual checking accounts at major U.S. banks) is unknown, the sampling distribution of  $\bar{x}$  is approximately normal because the sample is large ( $n \geq 30$ ). Remember that when the sample is large, the central limit theorem applies. The mean and standard deviation of the sampling distribution of  $\bar{x}$  are, respectively,

$$\mu_{\bar{x}} = \mu = \$400 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{225}} = \$2.00$$

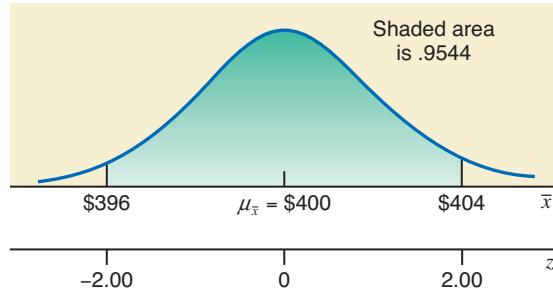
- (a) The probability that the mean of the annual costs of checking accounts in this sample is within \$4 of the population mean is written as

$$P(396 \leq \bar{x} \leq 404)$$

This probability is given by the area under the normal curve for  $\bar{x}$  between  $\bar{x} = \$396$  and  $\bar{x} = \$404$ , as shown in Figure 7.13. We find this area as follows:

$$\text{For } \bar{x} = \$396: \quad z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{396 - 400}{2.00} = -2.00$$

$$\text{For } \bar{x} = \$404: \quad z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{404 - 400}{2.00} = 2.00$$



**Figure 7.13**  $P(\$396 \leq \bar{x} \leq \$400)$

Hence, the required probability is

$$\begin{aligned} P(\$396 \leq \bar{x} \leq \$404) &= P(-2.00 \leq z \leq 2.00) \\ &= P(z \leq 2.00) - P(z \leq -2.00) \\ &= .9772 - .0228 = .9544 \end{aligned}$$

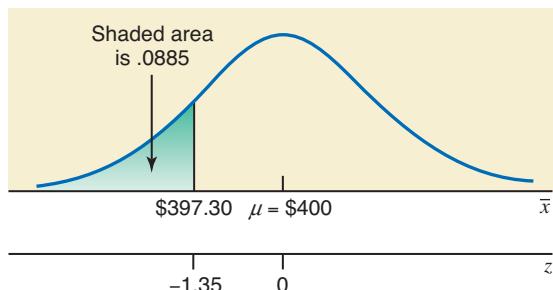
Therefore, the probability that the average annual cost of the 225 checking accounts in this sample is within \$4 of the population mean is .9544.

- (b) The probability that the average annual cost of the checking accounts in this sample is less than the population mean by \$2.70 or more is written as

$$P(\bar{x} \leq 397.30)$$

This probability is given by the area under the normal curve for  $\bar{x}$  to the left of  $\bar{x} = \$397.30$ , as shown in Figure 7.14. We find this area as follows:

$$\text{For } \bar{x} = \$397.30: \quad z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{397.30 - 400}{2.00} = -1.35$$



**Figure 7.14**  $P(\bar{x} \leq \$397.30)$

Hence, the required probability is

$$P(\bar{x} \leq 397.30) = P(z \leq -1.35) = .0885$$

Thus, the probability that the average annual cost of the checking accounts in this sample is less than the population mean by \$2.70 or more is .0885. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**7.40** If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.5 standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean?

**7.41** If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 1.5 standard deviations ( $\sigma_{\bar{x}}$ ) of the population mean?

**7.42** For a population,  $N = 10,000$ ,  $\mu = 124$ , and  $\sigma = 18$ . Find the  $z$  value for each of the following for  $n = 36$ .

- a.  $\bar{x} = 128.60$       b.  $\bar{x} = 119.30$       c.  $\bar{x} = 116.88$       d.  $\bar{x} = 132.05$

**7.43** For a population,  $N = 205,000$ ,  $\mu = 66$ , and  $\sigma = 7$ . Find the  $z$  value for each of the following for  $n = 49$ .

- a.  $\bar{x} = 68.44$       b.  $\bar{x} = 58.75$       c.  $\bar{x} = 62.35$       d.  $\bar{x} = 71.82$

**7.44** Let  $x$  be a continuous random variable that has a normal distribution with  $\mu = 75$  and  $\sigma = 14$ . Assuming  $n/N \leq .05$ , find the probability that the sample mean,  $\bar{x}$ , for a random sample of 20 taken from this population will be

- a. between 68.5 and 77.3      b. less than 72.4

**7.45** Let  $x$  be a continuous random variable that has a normal distribution with  $\mu = 48$  and  $\sigma = 8$ . Assuming  $n/N \leq .05$ , find the probability that the sample mean,  $\bar{x}$ , for a random sample of 16 taken from this population will be

- a. between 49.6 and 52.2      b. more than 45.7

**7.46** Let  $x$  be a continuous random variable that has a distribution skewed to the right with  $\mu = 60$  and  $\sigma = 10$ . Assuming  $n/N \leq .05$ , find the probability that the sample mean,  $\bar{x}$ , for a random sample of 40 taken from this population will be

- a. less than 62.20      b. between 61.4 and 64.2

**7.47** Let  $x$  be a continuous random variable that follows a distribution skewed to the left with  $\mu = 90$  and  $\sigma = 18$ . Assuming  $n/N \leq .05$ , find the probability that the sample mean,  $\bar{x}$ , for a random sample of 64 taken from this population will be

- a. less than 82.3      b. greater than 86.7

### ■ APPLICATIONS

**7.48** According to Moebs Services Inc., an individual checking account at U.S. community banks costs these banks between \$175 and \$200 per year (*Time*, November 21, 2011). Suppose that the average annual cost of all such checking accounts at U.S. community banks is \$190 with a standard deviation of \$20. Find the probability that the average annual cost of a random sample of 100 such checking accounts at U.S. community banks is

- a. less than \$187      b. more than \$193.5      c. \$191.70 to 194.5

**7.49** The GPAs of all students enrolled at a large university have an approximately normal distribution with a mean of 3.02 and a standard deviation of .29. Find the probability that the mean GPA of a random sample of 20 students selected from this university is

- a. 3.10 or higher      b. 2.90 or lower      c. 2.95 to 3.11

**7.50** The delivery times for all food orders at a fast-food restaurant during the lunch hour are normally distributed with a mean of 7.7 minutes and a standard deviation of 2.1 minutes. Find the probability that the mean delivery time for a random sample of 16 such orders at this restaurant is

- a. between 7 and 8 minutes  
b. within 1 minute of the population mean  
c. less than the population mean by 1 minute or more

**7.51** As mentioned in Exercise 7.22, according to the American Automobile Association's 2012 annual report *Your Driving Costs*, the cost of owning and operating a four-wheel drive SUV is \$11,350 per year (*USA TODAY*, April 27, 2012). Note that this cost includes expenses for gasoline, maintenance, insurance, and financing for a vehicle that is driven 15,000 miles a year. Suppose that the distribution of such costs of owning and operating all four-wheel drive SUVs has a mean of \$11,350 with a standard deviation of \$2390. Find the probability that for a random sample of 400 four-wheel drive SUVs, the average cost of owning and operating is

- a. more than \$11,540      b. less than \$11,110      c. \$11,250 to \$11,600

**7.52** The times that college students spend studying per week have a distribution that is skewed to the right with a mean of 8.4 hours and a standard deviation of 2.7 hours. Find the probability that the mean time spent studying per week for a random sample of 45 students would be

- a. between 8 and 9 hours
- b. less than 8 hours

**7.53** The credit card debts of all college students have a distribution that is skewed to the right with a mean of \$2840 and a standard deviation of \$672. Find the probability that the mean credit card debt for a random sample of 36 college students would be

- a. between \$2600 and \$2950
- b. less than \$3060

**7.54** As mentioned in Exercise 7.39, according to the American Time Use Survey, Americans watch television each weekday for an average of 151 minutes (*Time*, July 11, 2011). Suppose that the current distribution of times spent watching television every weekday by all Americans has a mean of 151 minutes and a standard deviation of 20 minutes. Find the probability that the average time spent watching television on a weekday by 200 randomly selected Americans is

- a. 148.70 to 150 minutes
- b. more than 153 minutes
- c. at most 146 minutes

**7.55** The amounts of electricity bills for all households in a city have a skewed probability distribution with a mean of \$140 and a standard deviation of \$30. Find the probability that the mean amount of electric bills for a random sample of 75 households selected from this city will be

- a. between \$132 and \$136
- b. within \$6 of the population mean
- c. more than the population mean by at least \$4

**7.56** According to a PNC Financial Independence Survey released in March 2012, today's U.S. adults in their 20s "hold an average debt of about \$45,000, which includes everything from cars to credit cards to student loans to mortgages" (*USA TODAY*, April 24, 2012). Suppose that the current distribution of debts of all U.S. adults in their 20s has a mean of \$45,000 and a standard deviation of \$12,720. Find the probability that the average debt of a random sample of 144 U.S. adults in their 20s is

- a. less than \$42,600
- b. more than \$46,240
- c. \$43,190 to \$46,980

**7.57** As mentioned in Exercise 7.33, among college students who hold part-time jobs during the school year, the distribution of the time spent working per week is approximately normally distributed with a mean of 20.20 hours and a standard deviation of 2.60 hours. Find the probability that the average time spent working per week for 18 randomly selected college students who hold part-time jobs during the school year is

- a. not within 1 hour of the population mean
- b. 20 to 20.50 hours
- c. at least 22 hours
- d. no more than 21 hours

**7.58** Johnson Electronics Corporation makes electric tubes. It is known that the standard deviation of the lives of these tubes is 150 hours. The company's research department takes a sample of 100 such tubes and finds that the mean life of these tubes is 2250 hours. What is the probability that this sample mean is within 25 hours of the mean life of all tubes produced by this company?

**7.59** A machine at Katz Steel Corporation makes 3-inch-long nails. The probability distribution of the lengths of these nails is normal with a mean of 3 inches and a standard deviation of .1 inch. The quality control inspector takes a sample of 25 nails once a week and calculates the mean length of these nails. If the mean of this sample is either less than 2.95 inches or greater than 3.05 inches, the inspector concludes that the machine needs an adjustment. What is the probability that based on a sample of 25 nails, the inspector will conclude that the machine needs an adjustment?

## 7.5

## Population and Sample Proportions; and Mean, Standard Deviation, and Shape of the Sampling Distribution of $\hat{p}$

The concept of proportion is the same as the concept of relative frequency discussed in Chapter 2 and the concept of probability of success in a binomial experiment. The relative frequency of a category or class gives the proportion of the sample or population that belongs to that category or class. Similarly, the probability of success in a binomial experiment represents the proportion of the sample or population that possesses a given characteristic.

In this section, first we will learn about the population and sample proportions. Then we will discuss the mean, standard deviation and shape of the sampling distribution of  $\hat{p}$ .

### 7.5.1 Population and Sample Proportions

The **population proportion**, denoted by  $p$ , is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population. The **sample proportion**, denoted by  $\hat{p}$  (pronounced *p hat*), gives a similar ratio for a sample.

**Population and Sample Proportions** The *population and sample proportions*, denoted by  $p$  and  $\hat{p}$ , respectively, are calculated as

$$p = \frac{X}{N} \quad \text{and} \quad \hat{p} = \frac{x}{n}$$

where

$N$  = total number of elements in the population

$n$  = total number of elements in the sample

$X$  = number of elements in the population that possess a specific characteristic

$x$  = number of elements in the sample that possess a specific characteristic

Example 7–7 illustrates the calculation of the population and sample proportions.

#### ■ EXAMPLE 7–7

*Calculating the population and sample proportions.*

Suppose a total of 789,654 families live in a particular city and 563,282 of them own homes. A sample of 240 families is selected from this city, and 158 of them own homes. Find the proportion of families who own homes in the population and in the sample.

**Solution** For the population of this city,

$$N = \text{population size} = 789,654$$

$$X = \text{families in the population who own homes} = 563,282$$

The proportion of all families in this city who own homes is

$$p = \frac{X}{N} = \frac{563,282}{789,654} = .71$$

Now, a sample of 240 families is taken from this city, and 158 of them are home-owners. Then,

$$n = \text{sample size} = 240$$

$$x = \text{families in the sample who own homes} = 158$$

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{158}{240} = .66$$

As in the case of the mean, the difference between the sample proportion and the corresponding population proportion gives the sampling error, assuming that the sample is random and no nonsampling error has been made. Thus, in the case of the proportion,

$$\text{Sampling error} = \hat{p} - p$$

For instance, for Example 7–7,

$$\text{Sampling error} = \hat{p} - p = .66 - .71 = -.05$$

### 7.5.2 Sampling Distribution of $\hat{p}$

Just like the sample mean  $\bar{x}$ , the sample proportion  $\hat{p}$  is a random variable. In other words, the population proportion  $p$  is a constant as it assumes one and only one value. However, the sample proportion  $\hat{p}$  can assume one of a large number of possible values depending on which sample is selected. Hence,  $\hat{p}$  is a random variable and it possesses a probability distribution, which is called its **sampling distribution**.

#### Definition

**Sampling Distribution of the Sample Proportion,  $\hat{p}$**  The probability distribution of the sample proportion,  $\hat{p}$ , is called its *sampling distribution*. It gives the various values that  $\hat{p}$  can assume and their probabilities.

The value of  $\hat{p}$  calculated for a particular sample depends on what elements of the population are included in that sample. Example 7–8 illustrates the concept of the sampling distribution of  $\hat{p}$ .

#### ■ EXAMPLE 7–8

Boe Consultant Associates has five employees. Table 7.6 gives the names of these five employees and information concerning their knowledge of statistics.

Illustrating the sampling distribution of  $\hat{p}$ .

**Table 7.6** Information on the Five Employees of Boe Consultant Associates

Name	Knows Statistics
Ally	Yes
John	No
Susan	No
Lee	Yes
Tom	Yes

If we define the population proportion,  $p$ , as the proportion of employees who know statistics, then

$$p = 3/5 = .60$$

Now, suppose we draw all possible samples of three employees each and compute the proportion of employees, for each sample, who know statistics. The total number of samples of size three that can be drawn from the population of five employees is

$$\text{Total number of samples} = {}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 10$$

Table 7.7 lists these 10 possible samples and the proportion of employees who know statistics for each of those samples. Note that we have rounded the values of  $\hat{p}$  to two decimal places.

**Table 7.7** All Possible Samples of Size 3 and the Value of  $\hat{p}$  for Each Sample

Sample	Proportion Who Know Statistics $\hat{p}$
Ally, John, Susan	$1/3 = .33$
Ally, John, Lee	$2/3 = .67$
Ally, John, Tom	$2/3 = .67$
Ally, Susan, Lee	$2/3 = .67$
Ally, Susan, Tom	$2/3 = .67$
Ally, Lee, Tom	$3/3 = 1.00$
John, Susan, Lee	$1/3 = .33$
John, Susan, Tom	$1/3 = .33$
John, Lee, Tom	$2/3 = .67$
Susan, Lee, Tom	$2/3 = .67$

Using Table 7.7, we prepare the frequency distribution of  $\hat{p}$  as recorded in Table 7.8, along with the relative frequencies of classes, which are obtained by dividing the frequencies of classes by the population size. The relative frequencies are used as probabilities and listed in Table 7.9. This table gives the sampling distribution of  $\hat{p}$ .

**Table 7.8** Frequency and Relative Frequency Distributions of  $\hat{p}$  When the Sample Size Is 3

$\hat{p}$	f	Relative Frequency
.33	3	$3/10 = .30$
.67	6	$6/10 = .60$
1.00	1	$1/10 = .10$
$\Sigma f = 10$		Sum = 1.00

**Table 7.9** Sampling Distribution of  $\hat{p}$  When the Sample Size Is 3

$\hat{p}$	$P(\hat{p})$
.33	.30
.67	.60
1.00	.10
$\Sigma P(\hat{p}) = 1.00$	

### 7.5.3 Mean and Standard Deviation of $\hat{p}$

The **mean** of  $\hat{p}$ , which is the same as the mean of the sampling distribution of  $\hat{p}$ , is always equal to the population proportion,  $p$ , just as the mean of the sampling distribution of  $\bar{x}$  is always equal to the population mean,  $\mu$ .

**Mean of the Sample Proportion** The *mean of the sample proportion*,  $\hat{p}$ , is denoted by  $\mu_{\hat{p}}$  and is equal to the population proportion,  $p$ . Thus,

$$\mu_{\hat{p}} = p$$

The sample proportion,  $\hat{p}$ , is called an **estimator** of the population proportion,  $p$ . As mentioned earlier, when the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that sample statistic is said to be an **unbiased estimator**. Since for the sample proportion  $\mu_{\hat{p}} = p$ ,  $\hat{p}$  is an unbiased estimator of  $p$ .

The **standard deviation** of  $\hat{p}$ , denoted by  $\sigma_{\hat{p}}$ , is given by the following formula. This formula is true only when the sample size is small compared to the population size. As we know from Section 7.2, the sample size is said to be small compared to the population size if  $n/N \leq .05$ .

**Standard Deviation of the Sample Proportion** The *standard deviation of the sample proportion*,  $\hat{p}$ , is denoted by  $\sigma_{\hat{p}}$  and is given by the formula

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

where  $p$  is the population proportion,  $q = 1 - p$ , and  $n$  is the sample size. This formula is used when  $n/N \leq .05$ , where  $N$  is the population size.

However, if  $n/N$  is greater than .05, then  $\sigma_{\hat{p}}$  is calculated as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}$$

where the factor

$$\sqrt{\frac{N-n}{N-1}}$$

is called the finite population correction factor.

In almost all cases, the sample size is small compared to the population size and, consequently, the formula used to calculate  $\sigma_{\hat{p}}$  is  $\sqrt{pq/n}$ .

As mentioned earlier, if the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is said to be a **consistent estimator**. It is obvious from the above formula for  $\sigma_{\hat{p}}$  that as  $n$  increases, the value of  $\sqrt{pq/n}$  decreases. Thus, the sample proportion,  $\hat{p}$ , is a consistent estimator of the population proportion,  $p$ .

## 7.5.4 Shape of the Sampling Distribution of $\hat{p}$

The shape of the sampling distribution of  $\hat{p}$  is inferred from the central limit theorem.

**Central Limit Theorem for Sample Proportion** According to the central limit theorem, the *sampling distribution of  $\hat{p}$*  is approximately normal for a sufficiently large sample size. In the case of proportion, the sample size is considered to be sufficiently large if  $np$  and  $nq$  are both greater than 5; that is, if

$$np > 5 \quad \text{and} \quad nq > 5$$

Note that the sampling distribution of  $\hat{p}$  will be approximately normal if  $np > 5$  and  $nq > 5$ . This is the same condition that was required for the application of the normal approximation to the binomial probability distribution in Chapter 6.

Example 7–9 shows the calculation of the mean and standard deviation of  $\hat{p}$  and describes the shape of its sampling distribution.

### ■ EXAMPLE 7–9

According to a *New York Times/CBS News* poll conducted during June 24–28, 2011, 55% of adults polled said that owning a home is a *very important part of the American Dream* (*The New York Times*, June 30, 2011). Assume that this result is true for the current population of American adults. Let  $\hat{p}$  be the proportion of American adults in a random sample of 2000 who will say that owning a home is a *very important part of the American Dream*. Find the mean and standard deviation of  $\hat{p}$  and describe the shape of its sampling distribution.

*Finding the mean and standard deviation, and describing the shape of the sampling distribution of  $\hat{p}$ .*

**Solution** Let  $p$  be the proportion of all American adults who will say that owning a home is a *very important part of the American Dream*. Then,

$$p = .55, \quad q = 1 - p = 1 - .55 = .45, \quad \text{and} \quad n = 2000$$

The mean of the sampling distribution of  $\hat{p}$  is

$$\mu_{\hat{p}} = p = .55$$

The standard deviation of  $\hat{p}$  is

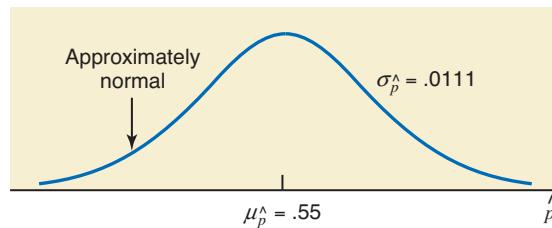
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.55)(.45)}{2000}} = .0111$$

The values of  $np$  and  $nq$  are

$$np = 2000(.55) = 1100 \quad \text{and} \quad nq = 2000(.45) = 900$$

Because  $np$  and  $nq$  are both greater than 5, we can apply the central limit theorem to make an inference about the shape of the sampling distribution of  $\hat{p}$ . Thus, the sampling distribution of  $\hat{p}$  is approximately normal with a mean of .55 and a standard deviation of .0111, as shown in Figure 7.15.

**Figure 7.15** Sampling distribution of  $\hat{p}$ .



■

## EXERCISES

### CONCEPTS AND PROCEDURES

**7.60** In a population of 1000 subjects, 640 possess a certain characteristic. In a sample of 40 subjects selected from this population, 24 possess the same characteristic. What are the values of the population and sample proportions?

**7.61** In a population of 5000 subjects, 600 possess a certain characteristic. In a sample of 120 subjects selected from this population, 18 possess the same characteristic. What are the values of the population and sample proportions?

**7.62** In a population of 18,700 subjects, 30% possess a certain characteristic. In a sample of 250 subjects selected from this population, 25% possess the same characteristic. How many subjects in the population and sample, respectively, possess this characteristic?

**7.63** In a population of 9500 subjects, 75% possess a certain characteristic. In a sample of 400 subjects selected from this population, 78% possess the same characteristic. How many subjects in the population and sample, respectively, possess this characteristic?

**7.64** Let  $\hat{p}$  be the proportion of elements in a sample that possess a characteristic.

- a. What is the mean of  $\hat{p}$ ?
- b. What is the formula to calculate the standard deviation of  $\hat{p}$ ? Assume  $n/N \leq .05$ .
- c. What condition(s) must hold true for the sampling distribution of  $\hat{p}$  to be approximately normal?

**7.65** For a population,  $N = 12,000$  and  $p = .71$ . A random sample of 900 elements selected from this population gave  $\hat{p} = .66$ . Find the sampling error.

**7.66** For a population,  $N = 2800$  and  $p = .29$ . A random sample of 80 elements selected from this population gave  $\hat{p} = .33$ . Find the sampling error.

**7.67** What is the estimator of the population proportion? Is this estimator an unbiased estimator of  $p$ ? Explain why or why not.

**7.68** Is the sample proportion a consistent estimator of the population proportion? Explain why or why not.

**7.69** How does the value of  $\sigma_{\hat{p}}$  change as the sample size increases? Explain. Assume  $n/N \leq .05$ .

**7.70** Consider a large population with  $p = .63$ . Assuming  $n/N \leq .05$ , find the mean and standard deviation of the sample proportion  $\hat{p}$  for a sample size of

- a. 100      b. 900

**7.71** Consider a large population with  $p = .21$ . Assuming  $n/N \leq .05$ , find the mean and standard deviation of the sample proportion  $\hat{p}$  for a sample size of

- a. 400      b. 750

**7.72** A population of  $N = 4000$  has a population proportion equal to .12. In each of the following cases, which formula will you use to calculate  $\sigma_{\hat{p}}$  and why? Using the appropriate formula, calculate  $\sigma_{\hat{p}}$  for each of these cases.

- a.  $n = 800$       b.  $n = 30$

**7.73** A population of  $N = 1400$  has a population proportion equal to .47. In each of the following cases, which formula will you use to calculate  $\sigma_{\hat{p}}$  and why? Using the appropriate formula, calculate  $\sigma_{\hat{p}}$  for each of these cases.

- a.  $n = 90$       b.  $n = 50$

**7.74** According to the central limit theorem, the sampling distribution of  $\hat{p}$  is approximately normal when the sample is large. What is considered a large sample in the case of the proportion? Briefly explain.

**7.75** Indicate in which of the following cases the central limit theorem will apply to describe the sampling distribution of the sample proportion.

- a.  $n = 400$  and  $p = .28$       b.  $n = 80$  and  $p = .05$   
c.  $n = 60$  and  $p = .12$       d.  $n = 100$  and  $p = .035$

**7.76** Indicate in which of the following cases the central limit theorem will apply to describe the sampling distribution of the sample proportion.

- a.  $n = 20$  and  $p = .45$       b.  $n = 75$  and  $p = .22$   
c.  $n = 350$  and  $p = .01$       d.  $n = 200$  and  $p = .022$

## ■ APPLICATIONS

**7.77** A company manufactured six television sets on a given day, and these TV sets were inspected for being good or defective. The results of the inspection follow.

Good      Good      Defective      Defective      Good      Good

- a. What proportion of these TV sets are good?  
b. How many total samples (without replacement) of size five can be selected from this population?  
c. List all the possible samples of size five that can be selected from this population and calculate the sample proportion,  $\hat{p}$ , of television sets that are good for each sample. Prepare the sampling distribution of  $\hat{p}$ .  
d. For each sample listed in part c, calculate the sampling error.

**7.78** Investigation of all five major fires in a western desert during one of the recent summers found the following causes.

Arson      Accident      Accident      Arson      Accident

- a. What proportion of those fires were due to arson?  
b. How many total samples (without replacement) of size three can be selected from this population?  
c. List all the possible samples of size three that can be selected from this population and calculate the sample proportion  $\hat{p}$  of the fires due to arson for each sample. Prepare the table that gives the sampling distribution of  $\hat{p}$ .  
d. For each sample listed in part c, calculate the sampling error.

**7.79** Beginning in the second half of 2011, there were widespread protests in many American cities that were primarily against Wall Street corruption and the gap between the rich and the poor in America. According to a *Time Magazine/ABT SRBI* poll conducted by telephone during October 9–10, 2011, 86% of adults who were familiar with those protests agreed that Wall Street and lobbyists have too much influence in Washington (*The New York Times*, October 22, 2011). Assume that this percentage is true for the current population of American adults. Let  $\hat{p}$  be the proportion in a random sample of 400 American adults who hold the opinion that Wall Street and lobbyists have too much influence in Washington. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape.

**7.80** According to a poll, 55% of Americans do not know that GOP stands for Grand Old Party (*Time*, October 17, 2011). Assume that this percentage is true for the current population of Americans. Let  $\hat{p}$  be the proportion in a random sample of 900 Americans who do not know that GOP stands for Grand Old Party. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape.

**7.81** In a *Time/Money Magazine* poll of Americans age 18 years and older, 65% agreed with the statement, “We are less sure our children will achieve the American Dream” (*Time*, October 10, 2011). Assume that this result is true for the current population of Americans age 18 years and older. Let  $\hat{p}$  be the proportion in a random sample of 600 Americans age 18 years and older who agree with the above statement. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape.

**7.82** In a *Time Magazine/Aspen* poll of American adults conducted by the strategic research firm Penn Schoen Berland, these adults were asked, “In your opinion, what is more important for the U.S. to focus on in the next decade?” Eighty-three percent of the adults polled said *domestic issues* (*Time*, July 11, 2011). Assume that this percentage is true for the current population of American adults. Let  $\hat{p}$  be the proportion in a random sample of 1000 American adults who hold the above opinion. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape.

## 7.6 Applications of the Sampling Distribution of $\hat{p}$

As mentioned in Section 7.4, when we conduct a study, we usually take only one sample and make all decisions or inferences on the basis of the results of that one sample. We use the concepts of the mean, standard deviation, and shape of the sampling distribution of  $\hat{p}$  to determine the probability that the value of  $\hat{p}$  computed from one sample falls within a given interval. Examples 7–10 and 7–11 illustrate this application.

### ■ EXAMPLE 7–10

*Calculating the probability that  $\hat{p}$  is in an interval.*

According to a Pew Research Center nationwide telephone survey of American adults conducted by phone between March 15 and April 24, 2011, 75% of adults said that college education has become too expensive for most people and they cannot afford it (*Time*, May 30, 2011). Suppose that this result is true for the current population of American adults. Let  $\hat{p}$  be the proportion in a random sample of 1400 adult Americans who will hold the said opinion. Find the probability that 76.5% to 78% of adults in this sample will hold this opinion.

**Solution** From the given information,

$$n = 1400, \quad p = .75, \quad \text{and} \quad q = 1 - p = 1 - .75 = .25$$

where  $p$  is the proportion of all adult Americans who hold the said opinion.

The mean of the sample proportion  $\hat{p}$  is

$$\mu_{\hat{p}} = p = .75$$

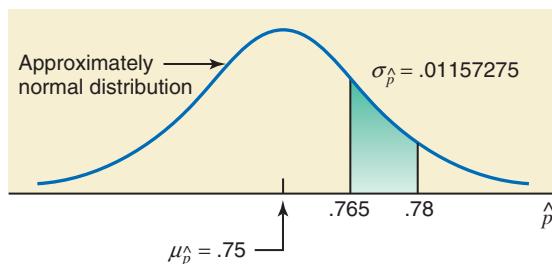
The standard deviation of  $\hat{p}$  is

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.75)(.25)}{1400}} = .01157275$$

The values of  $np$  and  $nq$  are

$$np = 1400 (.75) = 1050 \quad \text{and} \quad nq = 1400 (.25) = 350$$

Because  $np$  and  $nq$  are both greater than 5, we can infer from the central limit theorem that the sampling distribution of  $\hat{p}$  is approximately normal. The probability that  $\hat{p}$  is between .765 and .78 is given by the area under the normal curve for  $\hat{p}$  between  $\hat{p} = .765$  and  $\hat{p} = .78$ , as shown in Figure 7.16.

Figure 7.16  $P(.765 < \hat{p} < .78)$ 

The first step in finding the area under the normal curve between  $\hat{p} = .765$  and  $\hat{p} = .78$  is to convert these two values to their respective  $z$  values. The  $z$  value for  $\hat{p}$  is computed using the following formula

**z Value for a Value of  $\hat{p}$**  The  $z$  value for a value of  $\hat{p}$  is calculated as

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

The two values of  $\hat{p}$  are converted to their respective  $z$  values, and then the area under the normal curve between these two points is found using the normal distribution table.

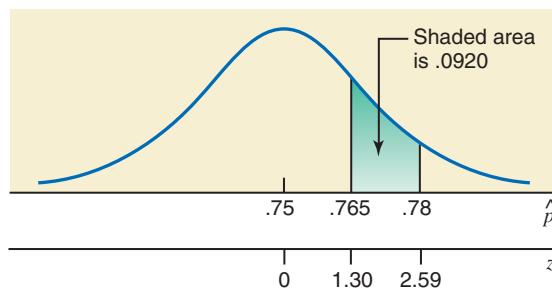
$$\text{For } \hat{p} = .765: \quad z = \frac{.765 - .75}{.01157275} = 1.30$$

$$\text{For } \hat{p} = .78: \quad z = \frac{.78 - .75}{.01157275} = 2.59$$

Thus, the probability that  $\hat{p}$  is between  $.765$  and  $.78$  is given by the area under the standard normal curve between  $z = 1.30$  and  $z = 2.59$ . This area is shown in Figure 7.17. The required probability is

$$P(.765 < \hat{p} < .78) = P(1.30 < z < 2.59) = P(z < 2.59) - P(z < 1.30)$$

$$= .9952 - .9032 = .0920$$

Figure 7.17  $P(.765 < \hat{p} < .78)$ 

Thus, the probability is  $.0920$  that  $76.5\%$  to  $78\%$  of American adults in a random sample of 1400 will say that college education has become too expensive for most people and they cannot afford it.

### ■ EXAMPLE 7-11

Maureen Webster, who is running for mayor in a large city, claims that she is favored by  $53\%$  of all eligible voters of that city. Assume that this claim is true. What is the probability that in a random sample of 400 registered voters taken from this city, less than  $49\%$  will favor Maureen Webster?

Calculating the probability that  $\hat{p}$  is less than a certain value.

**Solution** Let  $p$  be the proportion of all eligible voters who favor Maureen Webster. Then,

$$p = .53 \quad \text{and} \quad q = 1 - p = 1 - .53 = .47$$

The mean of the sampling distribution of the sample proportion  $\hat{p}$  is

$$\mu_{\hat{p}} = p = .53$$

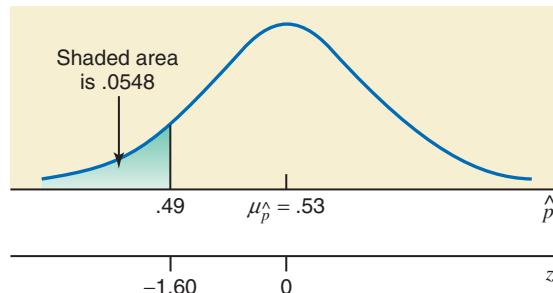
The population of all voters is large (because the city is large), and the sample size is small compared to the population. Consequently, we can assume that  $n/N \leq .05$ . Hence, the standard deviation of  $\hat{p}$  is calculated as

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.53)(.47)}{400}} = .02495496$$

From the central limit theorem, the shape of the sampling distribution of  $\hat{p}$  is approximately normal. (The reader should check that  $np > 5$  and  $nq > 5$  and, hence, the sample size is large.) The probability that  $\hat{p}$  is less than .49 is given by the area under the normal distribution curve for  $\hat{p}$  to the left of  $\hat{p} = .49$ , as shown in Figure 7.18. The  $z$  value for  $\hat{p} = .49$  is

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.49 - .53}{.02495496} = -1.60$$

**Figure 7.18**  $P(\hat{p} < .49)$



Thus, the required probability from Table IV is

$$\begin{aligned} P(\hat{p} < .49) &= P(z < -1.60) \\ &= \mathbf{.0548} \end{aligned}$$

Hence, the probability that less than 49% of the voters in a random sample of 400 will favor Maureen Webster is .0548. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**7.83** If all possible samples of the same (large) size are selected from a population, what percentage of all sample proportions will be within 2.0 standard deviations ( $\sigma_{\hat{p}}$ ) of the population proportion?

**7.84** If all possible samples of the same (large) size are selected from a population, what percentage of all sample proportions will be within 3.0 standard deviations ( $\sigma_{\hat{p}}$ ) of the population proportion?

**7.85** For a population,  $N = 30,000$  and  $p = .59$ . Find the  $z$  value for each of the following for  $n = 100$ .

- a.  $\hat{p} = .56$
- b.  $\hat{p} = .68$
- c.  $\hat{p} = .53$
- d.  $\hat{p} = .65$

**7.86** For a population,  $N = 18,000$  and  $p = .25$ . Find the  $z$  value for each of the following for  $n = 70$ .

- a.  $\hat{p} = .26$
- b.  $\hat{p} = .32$
- c.  $\hat{p} = .17$
- d.  $\hat{p} = .20$

### ■ APPLICATIONS

**7.87** Refer to Exercise 7.79. Beginning in the second half of 2011, there were widespread protests in many American cities that were primarily against Wall Street corruption and the gap between the rich and the

poor in America. According to a *Time Magazine*/ABT SRBI poll conducted by telephone during October 9–10, 2011, 86% of adults who were familiar with those protests agreed that Wall Street and lobbyists have too much influence in Washington (*The New York Times*, October 22, 2011). Assume that this percentage is true for the current population of American adults. Let  $\hat{p}$  be the proportion in a random sample of 400 American adults who hold the opinion that Wall Street and lobbyists have too much influence in Washington. Find the probability that the value of  $\hat{p}$  will be

- a. greater than .88      b. between .82 and .84

**7.88** A survey of all medium- and large-sized corporations showed that 64% of them offer retirement plans to their employees. Let  $\hat{p}$  be the proportion in a random sample of 50 such corporations that offer retirement plans to their employees. Find the probability that the value of  $\hat{p}$  will be

- a. between .54 and .61      b. greater than .71

**7.89** According to a poll, 55% of Americans do not know that GOP stands for Grand Old Party (*Time*, October 17, 2011). Assume that this percentage is true for the current population of Americans. Let  $\hat{p}$  be the proportion in a random sample of 900 Americans who do not know that GOP stands for Grand Old Party. Find the probability that the value of  $\hat{p}$  will be

- a. less than .58      b. between .53 and .59

**7.90** Dartmouth Distribution Warehouse makes deliveries of a large number of products to its customers. It is known that 85% of all the orders it receives from its customers are delivered on time. Let  $\hat{p}$  be the proportion of orders in a random sample of 100 that are delivered on time. Find the probability that the value of  $\hat{p}$  will be

- a. between .81 and .88      b. less than .87

**7.91** Brooklyn Corporation manufactures DVDs. The machine that is used to make these DVDs is known to produce 6% defective DVDs. The quality control inspector selects a sample of 150 DVDs every week and inspects them for being good or defective. If 8% or more of the DVDs in the sample are defective, the process is stopped and the machine is readjusted. What is the probability that based on a sample of 150 DVDs, the process will be stopped to readjust the machine?

**7.92** Mong Corporation makes auto batteries. The company claims that 80% of its LL70 batteries are good for 70 months or longer. Assume that this claim is true. Let  $\hat{p}$  be the proportion in a sample of 100 such batteries that are good for 70 months or longer.

- a. What is the probability that this sample proportion is within .05 of the population proportion?  
b. What is the probability that this sample proportion is less than the population proportion by .06 or more?  
c. What is the probability that this sample proportion is greater than the population proportion by .07 or more?

## USES AND MISUSES... BEWARE OF BIAS

Mathematics tells us that the sample mean,  $\bar{x}$ , is an unbiased and consistent estimator for the population mean,  $\mu$ . This is great news because it allows us to estimate properties of a population based on those of a sample; this is the essence of statistics. But statistics always makes a number of assumptions about the sample from which the mean and standard deviation are calculated. Failure to respect these assumptions can introduce bias in your calculations. In statistics, *bias* means a deviation of the expected value of a statistical estimator from the parameter it estimates.

Let's say you are a quality control manager for a refrigerator parts company. One of the parts that you manufacture has a specification that the length of the part be 2.0 centimeters plus or minus .025 centimeter. The manufacturer expects that the parts it receives have a mean length of 2.0 centimeters and a small variation around that mean. The manufacturing process is to mold the part to something a little bit bigger than necessary—say, 2.1 centimeters—and finish the process by hand. Because the action of cutting material is irreversible, the machinists tend to miss their target by

approximately .01 centimeter, so the mean length of the parts is not 2.0 centimeters, but rather 2.01 centimeters. It is your job to catch this.

One of your quality control procedures is to select completed parts randomly and test them against specification. Unfortunately, your measurement device is also subject to variation and might consistently underestimate the length of the parts. If your measurements are consistently .01 centimeter too short, your sample mean will not catch the manufacturing error in the population of parts.

The solution to the manufacturing problem is relatively straightforward: Be certain to calibrate your measurement instrument. Calibration becomes very difficult when working with people. It is known that people tend to overestimate the number of times that they vote and underestimate the time it takes to complete a project. Basing statistical results on this type of data can result in distorted estimates of the properties of your population. It is very important to be careful to weed out bias in your data because once it gets into your calculations, it is very hard to get it out.

## Glossary

**Central limit theorem** The theorem from which it is inferred that for a large sample size ( $n \geq 30$ ), the shape of the sampling distribution of  $\bar{x}$  is approximately normal. Also, by the same theorem, the shape of the sampling distribution of  $\hat{p}$  is approximately normal for a sample for which  $np > 5$  and  $nq > 5$ .

**Consistent estimator** A sample statistic with a standard deviation that decreases as the sample size increases.

**Estimator** The sample statistic that is used to estimate a population parameter.

**Mean of  $\hat{p}$**  The mean of the sampling distribution of  $\hat{p}$ , denoted by  $\mu_{\hat{p}}$ , is equal to the population proportion  $p$ .

**Mean of  $\bar{x}$**  The mean of the sampling distribution of  $\bar{x}$ , denoted by  $\mu_{\bar{x}}$ , is equal to the population mean  $\mu$ .

**Nonsampling errors** The errors that occur during the collection, recording, and tabulation of data.

**Population distribution** The probability distribution of the population data.

**Population proportion  $p$**  The ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

**Sample proportion  $\hat{p}$**  The ratio of the number of elements in a sample with a specific characteristic to the total number of elements in that sample.

**Sampling distribution of  $\hat{p}$**  The probability distribution of all the values of  $\hat{p}$  calculated from all possible samples of the same size selected from a population.

**Sampling distribution of  $\bar{x}$**  The probability distribution of all the values of  $\bar{x}$  calculated from all possible samples of the same size selected from a population.

**Sampling error** The difference between the value of a sample statistic calculated from a random sample and the value of the corresponding population parameter. This type of error occurs due to chance.

**Standard deviation of  $\hat{p}$**  The standard deviation of the sampling distribution of  $\hat{p}$ , denoted by  $\sigma_{\hat{p}}$ , is equal to  $\sqrt{pq/n}$  when  $n/N \leq .05$ .

**Standard deviation of  $\bar{x}$**  The standard deviation of the sampling distribution of  $\bar{x}$ , denoted by  $\sigma_{\bar{x}}$ , is equal to  $\sigma/\sqrt{n}$  when  $n/N \leq .05$ .

**Unbiased estimator** An estimator with an expected value (or mean) that is equal to the value of the corresponding population parameter.

## Supplementary Exercises

**7.93** The package of Sylvania CFL 65-watt replacement bulbs that use only 16 watts claims that these bulbs have an average life of 8000 hours. Assume that the lives of all such bulbs have a normal distribution with a mean of 8000 hours and a standard deviation of 400 hours. Let  $\bar{x}$  be the average life of 25 randomly selected such bulbs. Find the mean and standard deviation of  $\bar{x}$ , and comment on the shape of its sampling distribution.

**7.94** A January 2010 article on money.cnn.com reported that the average monthly cable bill in the United States was \$75. The article also stated that the annual percentage increase in the average monthly cable bill is 5% ([http://money.cnn.com/2010/01/06/news/companies/cable\\_bill\\_cost\\_increase/index.htm](http://money.cnn.com/2010/01/06/news/companies/cable_bill_cost_increase/index.htm)). Suppose that the current distribution of all monthly cable bills in the United States is approximately normal with a mean of \$82.69 and a standard deviation of \$11.17. Let  $\bar{x}$  be the average monthly cable bill for 23 randomly selected U.S. households with cable. Find the mean and standard deviation of  $\bar{x}$ , and comment on the shape of its sampling distribution.

**7.95** Refer to Exercise 7.93. The package of Sylvania CFL 65-watt replacement bulbs that use only 16 watts claims that these bulbs have an average life of 8000 hours. Assume that the lives of all such bulbs have a normal distribution with a mean of 8000 hours and a standard deviation of 400 hours. Find the probability that the mean life of a random sample of 25 such bulbs is

- a. less than 7890 hours
- b. between 7850 and 7910 hours
- c. within 130 hours of the population mean
- d. less than the population mean by 150 hours or more

**7.96** Refer to Exercise 7.94. The current distribution of all monthly cable bills in the United States is approximately normal with a mean of \$82.69 and a standard deviation of \$11.17. Find the probability that the average monthly cable bill for 23 U.S. households with cable is

- a. less than \$80
- b. between \$75 and \$85
- c. within \$5 of the population mean
- d. more than \$90

**7.97** The Toyota Prius hybrid car is estimated to get an average of 50 miles per gallon (mpg) of gas. However, the gas mileage varies from car to car due to a variety of conditions, driving styles, and other factors and has been reported to be as high as 70 mpg. Suppose that the distribution of miles per gallon for Toyota Prius hybrid cars has a mean of 50 mpg and a standard deviation of 5.9 mpg. Find the probability that the average miles per gallon for 38 randomly selected Prius hybrid cars is

- a. more than 51.5
- b. between 48 and 51
- c. less than 53
- d. greater than the population mean by 2.5 or more

**7.98** A machine at Keats Corporation fills 64-ounce detergent jugs. The probability distribution of the amount of detergent in these jugs is normal with a mean of 64 ounces and a standard deviation of .4 ounce. The quality control inspector takes a sample of 16 jugs once a week and measures the amount of detergent in these jugs. If the mean of this sample is either less than 63.75 ounces or greater than 64.25 ounces, the inspector concludes that the machine needs an adjustment. What is the probability that based on a sample of 16 jugs, the inspector will conclude that the machine needs an adjustment when actually it does not?

**7.99** In a large city, 88% of the cases of car burglar alarms that go off are false. Let  $\hat{p}$  be the proportion of false alarms in a random sample of 80 cases of car burglar alarms that go off in this city. Calculate the mean and standard deviation of  $\hat{p}$ , and describe the shape of its sampling distribution.

**7.100** According to a *Time Magazine*/ABT SRBI poll conducted by telephone during October 9–10, 2011, 73% of adults age 18 years and older said that they are in favor of raising taxes on those with annual incomes of \$1 million or more to help cut the federal deficit (*Time*, October 24, 2011). Assume that this percentage is true for the current population of all American adults age 18 years and older. Let  $\hat{p}$  be the proportion of American adults age 18 years and older in a random sample of 900 who will hold the above opinion. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape.

**7.101** Refer to Exercise 7.100. Assume that 73% of adults age 18 years and older are in favor of raising taxes on those with annual incomes of \$1 million or more to help cut the federal deficit. A random sample of 900 American adults age 18 years and older is selected.

- a. Find the probability that the sample proportion is
  - i. less than .76
  - ii. between .70 and .75
- b. What is the probability that the sample proportion is within .025 of the population proportion?
- c. What is the probability that the sample proportion is greater than the population proportion by .03 or more?

**7.102** According to a Pew Research Center nationwide telephone survey of American adults conducted by phone between March 15 and April 24, 2011, 25% of American college graduates said that their student loans make it harder for them to buy a home (*Time*, May 30, 2011). Suppose that this result is true for the current population of American college graduates. Let  $\hat{p}$  be the proportion in a random sample of 1000 American college graduates who will say that their student loans make it harder for them to buy a home. Find the probability that the value of  $\hat{p}$  is

- a. within .02 of the population proportion
- b. not within .02 of the population proportion
- c. greater than the population proportion by .025 or more
- d. less than the population proportion by .03 or more

## Advanced Exercises

**7.103** Let  $\mu$  be the mean annual salary of Major League Baseball players for 2012. Assume that the standard deviation of the salaries of these players is \$2,845,000. What is the probability that the 2012 mean salary of a random sample of 32 baseball players was within \$500,000 of the population mean,  $\mu$ ? Assume that  $n/N \leq .05$ .

**7.104** The test scores for 300 students were entered into a computer, analyzed, and stored in a file. Unfortunately, someone accidentally erased a major portion of this file from the computer. The only information that is available is that 30% of the scores were below 65 and 15% of the scores were above 90. Assuming the scores are normally distributed, find their mean and standard deviation.

**7.105** A chemist has a 10-gallon sample of river water taken just downstream from the outflow of a chemical plant. He is concerned about the concentration,  $c$  (in parts per million), of a certain toxic substance in the water. He wants to take several measurements, find the mean concentration of the toxic substance for this sample, and have a 95% chance of being within .5 part per million of the true mean value of  $c$ . If the concentration of the toxic substance in all measurements is normally distributed with  $\sigma = .8$  part per million, how many measurements are necessary to achieve this goal?

**7.106** A television reporter is covering the election for mayor of a large city and will conduct an exit poll (interviews with voters immediately after they vote) to make an early prediction of the outcome. Assume that the eventual winner of the election will get 60% of the votes.

- What is the probability that a prediction based on an exit poll of a random sample of 25 voters will be correct? In other words, what is the probability that 13 or more of the 25 voters in the sample will have voted for the eventual winner?
- How large a sample would the reporter have to take so that the probability of correctly predicting the outcome would be .95 or higher?

**7.107** A city is planning to build a hydroelectric power plant. A local newspaper found that 53% of the voters in this city favor the construction of this plant. Assume that this result holds true for the population of all voters in this city.

- What is the probability that more than 50% of the voters in a random sample of 200 voters selected from this city will favor the construction of this plant?
- A politician would like to take a random sample of voters in which more than 50% would favor the plant construction. How large a sample should be selected so that the politician is 95% sure of this outcome?

**7.108** Refer to Exercise 6.93. Otto is trying out for the javelin throw to compete in the Olympics. The lengths of his javelin throws are normally distributed with a mean of 253 feet and a standard deviation of 8.4 feet. What is the probability that the total length of three of his throws will exceed 885 feet?

**7.109** A certain elevator has a maximum legal carrying capacity of 6000 pounds. Suppose that the population of all people who ride this elevator have a mean weight of 160 pounds with a standard deviation of 25 pounds. If 35 of these people board the elevator, what is the probability that their combined weight will exceed 6000 pounds? Assume that the 35 people constitute a random sample from the population.

**7.110** A Census Bureau report revealed that 43.7% of Americans who moved between 2009 and 2010 did so for housing-related reasons, such as the desire to live in a new or better home or apartment ([http://www.census.gov/newsroom/releases/archives/mobility\\_of\\_the\\_population/cb11-91.html](http://www.census.gov/newsroom/releases/archives/mobility_of_the_population/cb11-91.html)). Suppose that this percentage is true for the current population of Americans.

- Suppose that 49% of the people in a random sample of 100 Americans who moved recently did so for housing-related reasons. How likely is it for the sample proportion in a sample of 100 to be .49 or more when the population proportion is .437?
- Refer to part a. How likely is it for the sample proportion in a random sample of 200 to be .49 or more when the population proportion is .437?
- What is the smallest sample size that will produce a sample proportion of .49 or more in no more than 5% of all sample surveys of that size?

**7.111** Refer to the sampling distribution discussed in Section 7.1. Calculate and replace the sample means in Table 7.3 with the sample medians, and then calculate the average of these sample medians. Does this average of the medians equal the population mean? If yes, why does this make sense? If no, how could you change exactly two of the five data values in this example so that the average of the sample medians equals the population mean?

**7.112** Suppose you want to calculate  $P(a \leq \bar{x} \leq b)$ , where  $a$  and  $b$  are two numbers and  $x$  has a distribution with mean  $\mu$  and standard deviation  $\sigma$ . If  $a < \mu < b$  (i.e.,  $\mu$  lies in the interval  $a$  to  $b$ ), what happens to the probability  $P(a \leq \bar{x} \leq b)$  as the sample size becomes larger?

## Self-Review Test

- A sampling distribution is the probability distribution of
  - a population parameter
  - a sample statistic
  - any random variable
- Nonsampling errors are
  - the errors that occur because the sample size is too large in relation to the population size
  - the errors made while collecting, recording, and tabulating data
  - the errors that occur because an untrained person conducts the survey
- A sampling error is
  - the difference between the value of a sample statistic based on a random sample and the value of the corresponding population parameter
  - the error made while collecting, recording, and tabulating data
  - the error that occurs because the sample is too small

4. The mean of the sampling distribution of  $\bar{x}$  is always equal to  
a.  $\mu$       b.  $\mu - 5$       c.  $\sigma/\sqrt{n}$
5. The condition for the standard deviation of the sample mean to be  $\sigma/\sqrt{n}$  is that  
a.  $np > 5$       b.  $n/N \leq .05$       c.  $n > 30$
6. The standard deviation of the sampling distribution of the sample mean decreases when  
a.  $x$  increases      b.  $n$  increases      c.  $n$  decreases
7. When samples are selected from a normally distributed population, the sampling distribution of the sample mean has a normal distribution  
a. if  $n \geq 30$       b. if  $n/N \leq .05$       c. all the time
8. When samples are selected from a nonnormally distributed population, the sampling distribution of the sample mean has an approximately normal distribution  
a. if  $n \geq 30$       b. if  $n/N \leq .05$       c. always
9. In a sample of 200 customers of a mail-order company, 174 are found to be satisfied with the service they receive from the company. The proportion of customers in this sample who are satisfied with the company's service is  
a. .87      b. .174      c. .148
10. The mean of the sampling distribution of  $\hat{p}$  is always equal to  
a.  $p$       b.  $\mu$       c.  $\hat{p}$
11. The condition for the standard deviation of the sampling distribution of the sample proportion to be  $\sqrt{pq/n}$  is  
a.  $np > 5$  and  $nq > 5$       b.  $n > 30$       c.  $n/N \leq .05$
12. The sampling distribution of  $\hat{p}$  is (approximately) normal if  
a.  $np > 5$  and  $nq > 5$       b.  $n > 30$       c.  $n/N \leq .05$
13. Briefly state and explain the central limit theorem.
14. The weights of all students at a large university have an approximately normal distribution with a mean of 145 pounds and a standard deviation of 18 pounds. Let  $\bar{x}$  be the mean weight of a random sample of certain students selected from this university. Calculate the mean and standard deviation of  $\bar{x}$  and describe the shape of its sampling distribution for a sample size of  
a. 25      b. 100
15. According to an estimate, the average price of homes in Martha's Vineyard, Massachusetts, was \$650,000 in 2011 (*USA Today*, August 11, 2011). Suppose that the current population distribution of home prices in Martha's Vineyard has a mean of \$650,000 and a standard deviation of \$140,000, but the shape of this distribution is unknown. Let  $\bar{x}$  be the average price of a random sample of certain homes selected from Martha's Vineyard. Calculate the mean and the standard deviation of the sampling distribution of  $\bar{x}$  and describe its shape for a sample size of  
a. 20      b. 100      c. 400
16. Refer to Problem 15 above. The current population distribution of home prices in Martha's Vineyard has a mean of \$650,000 and a standard deviation of \$140,000, but the shape of this distribution is unknown. Find the probability that the average price of a random sample of 100 homes selected from Martha's Vineyard is  
a. between \$620,000 and \$635,000      b. within \$24,000 of the population mean  
c. \$630,000 or more      d. not within \$20,000 of the population mean  
e. less than \$640,000      f. less than \$660,000  
g. more than \$670,000      h. between \$640,000 and \$665,000
17. At Jen and Perry Ice Cream Company, the machine that fills one-pound cartons of Top Flavor ice cream is set to dispense 16 ounces of ice cream into every carton. However, some cartons contain slightly less than and some contain slightly more than 16 ounces of ice cream. The amounts of ice cream in all such cartons have a normal distribution with a mean of 16 ounces and a standard deviation of .18 ounce.  
a. Find the probability that the mean amount of ice cream in a random sample of 16 such cartons will be  
i. between 15.90 and 15.95 ounces  
ii. less than 15.95 ounces  
iii. more than 15.97 ounces

- b. What is the probability that the mean amount of ice cream in a random sample of 16 such cartons will be within .10 ounce of the population mean?
  - c. What is the probability that the mean amount of ice cream in a random sample of 16 such cartons will be less than the population mean by .135 ounce or more?
18. In a September 2011 CNN/ORC International Poll, 15% of Americans said that they trust the (Federal) Government in Washington to do what is right *always or most of the time* ([http://caffertyfile.blogs.cnn.com/2011/09/28/our-government-is-more-badly-divided-than-maybe-it-has-ever-been-whats-the-answer/?hpt=hp\\_t2](http://caffertyfile.blogs.cnn.com/2011/09/28/our-government-is-more-badly-divided-than-maybe-it-has-ever-been-whats-the-answer/?hpt=hp_t2)). Let  $\hat{p}$  be the proportion of Americans in a random sample who hold the aforementioned opinion. Find the mean and standard deviation of the sampling distribution of  $\hat{p}$  and describe its shape when the sample size is
- a. 30      b. 300      c. 3000
19. In a *Time Magazine*/Aspen Ideas Festival poll conducted by Penn Schoen Berland during June 1–8, 2011, Americans age 18 years and older were asked, “Overall, do you think the past decade has been one of progress or decline for the U.S. as a country?” Of the respondents, 68% said *decline* (*Time*, July 11, 2011). Assume that this result is true for the current population of American adults.
- a. Find the probability that in a random sample of 1000 American adults, the proportion who will say *decline* in response to the aforementioned question is
    - i. greater than .70
    - ii. between .66 and .71
    - iii. less than .65
    - iv. between .695 and .715
    - v. less than .69
    - vi. more than .67
  - b. What is the probability that in a random sample of 1000 American adults, the proportion who will say *decline* in response to the aforementioned question is within .025 of the population proportion?
  - c. What is the probability that in a random sample of 1000 American adults, the proportion who will say *decline* in response to the aforementioned question is not within .03 of the population proportion?
  - d. What is the probability that in a random sample of 1000 American adults, the proportion who will say *decline* in response to the aforementioned question is greater than the population proportion by .02 or more?

## Mini-Projects

### ■ MINI-PROJECT 7-1

Consider the data on weights of NFL players as given in Data Set III on the Web site for this book.

- a. Compute  $\mu$  and  $\sigma$  for this data set.
- b. Take 20 random samples of five players each, and find  $\bar{x}$  for each sample.
- c. Compute the mean and standard deviation of the 20 sample means obtained in part b.
- d. Using the formulas given in Section 7.2, find  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$  for  $n = 5$ .
- e. How do your values of  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$  in part d compare with those in part c?
- f. What percentage of the 20 sample means found in part b lie in the interval  $\mu_{\bar{x}} - \sigma_{\bar{x}}$  to  $\mu_{\bar{x}} + \sigma_{\bar{x}}$ ? In the interval  $\mu_{\bar{x}} - 2\sigma_{\bar{x}}$  to  $\mu_{\bar{x}} + 2\sigma_{\bar{x}}$ ? In the interval  $\mu_{\bar{x}} - 3\sigma_{\bar{x}}$  to  $\mu_{\bar{x}} + 3\sigma_{\bar{x}}$ ?
- g. How do the percentages in part f compare to the corresponding percentages for a normal distribution (68%, 95%, and 99.7%, respectively)?
- h. Repeat parts b through g using 20 samples of 10 players each.

### ■ MINI-PROJECT 7-2

Consider Data Set II, Data on States, that accompanies this text. Let  $p$  denote the proportion of the 50 states that have a per capita income of less than \$35,000.

- a. Find  $p$ .
- b. Select 20 random samples of 5 states each, and find the sample proportion  $\hat{p}$  for each sample.
- c. Compute the mean and standard deviation of the 20 sample proportions obtained in part b.
- d. Using the formulas given in Section 7.5.3, compute  $\mu_{\hat{p}}$  and  $\sigma_{\hat{p}}$ . Is the finite population correction factor required here?
- e. Compare your mean and standard deviation of  $\hat{p}$  from part c with the values calculated in part d.
- f. Repeat parts b through e using 20 samples of 10 states each.

### ■ MINI-PROJECT 7-3

You are to conduct the experiment of sampling 10 times (with replacement) from the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. You can do this in a variety of ways. One way is to write each digit on a separate piece

of paper, place all the slips in a hat, and select 10 times from the hat, returning each selected slip before the next pick. As alternatives, you can use a 10-sided die, statistical software, or a calculator that generates random numbers. Perform the experiment using any of these methods, and compute the sample mean  $\bar{x}$  for the 10 numbers obtained. Now repeat the procedure 49 more times. When you are done, you will have 50 sample means.

- a. Make a table of the population distribution for the 10 digits, and display it using a graph.
- b. Make a stem-and-leaf display of your 50 sample means. What shape does it have?
- c. What does the central limit theorem say about the shape of the sampling distribution of  $\bar{x}$ ? What mean and standard deviation does the sampling distribution of  $\bar{x}$  have in this problem?

### ■ MINI-PROJECT 7-4

Reconsider Mini-Project 7-3. Now repeat that project and parts a through c, but this time use a skewed distribution (as explained below), instead of a symmetric distribution, to take samples. This project is more easily performed using a computer or graphing calculator, but it can be done using a hat or a random numbers table. In this project, sample 10 times from a population of digits that contains twenty 0s, fifteen 1s, ten 2s, seven 3s, four 4s, three 5s, two 6s, and one each of the numbers 7, 8, and 9. Select 50 samples of size 10, and repeat parts a through c of Mini-Project 7-3. How do the parts a and b of this project compare to parts a and b of Mini-Project 7-3 in regard to the shapes of the distributions? How does this relate to what the central limit theorem says?

## DECIDE FOR YOURSELF DECIDING ABOUT ELECTIONS

In the first week of November during an election year, you are very likely to hear the following statement on TV news, “We are now able to make a projection. In (*insert the name of the state where you live*), we project that the winner will be (*insert one of the elected officials from your state*).” Many people are aware that news agencies conduct exit polls on election day. A commonly asked question is, “How can an agency call a race based on the results from a sample of only 1200 voters or so, and do this with a high (although not perfect) accuracy level?” Although the actual methods used to make projections based on exit polls are above the level of this book, we will examine a similar but simpler version of the question here. The concepts and logic involved in this process will help you understand the statistical inference concepts discussed in subsequent chapters.

Consider a simple election where there are only two candidates, named A and B. Suppose  $p$  and  $q$  are the proportions of votes received by candidates A and B, respectively. Suppose we conduct an exit poll based on a simple random sample of 800 voters and

determine the mean, standard deviation, and shape of the sampling distribution of  $\hat{p}$ , where  $\hat{p}$  is the proportion of voters in the sample who voted for candidate A.

1. Suppose that 440 of the 800 voters included in the exit poll voted for candidate A, which gives  $\hat{p} = .55$ . Assuming that each candidate received 50% of the votes (i.e.,  $p = .50$  and  $q = .50$ , where  $p$  and  $q$  are the proportions of votes received by candidates A and B, respectively), what is the probability that at least 440 out of 800 voters in a sample would vote for candidate A?
2. Based on your answer to the above question, the results of the poll make it reasonable to conclude that the proportion of all voters who voted for candidate A is actually higher than .5. Explain why.
3. What implications do the above answers have for the result of the election? Will you make a projection about this election based on the results of this exit poll?



### Sampling Distribution of Means

TI-84

To create a sampling distribution of a sample mean using the TI-84 requires a good deal of programming, which we will not do here. However, it is quite easy to create a sampling distribution for a sample proportion using the TI-84. Let  $n$  and  $p$  represent the number of trials and the probability of a success, respectively, for a binomial experiment. On the TI-84, press **2nd STAT OPS** **seq(**. In the **seq(** menu, select **MATH > PRB > randBin( $n,p$ )/ $n$**  at the **Expr:** prompt, X

```

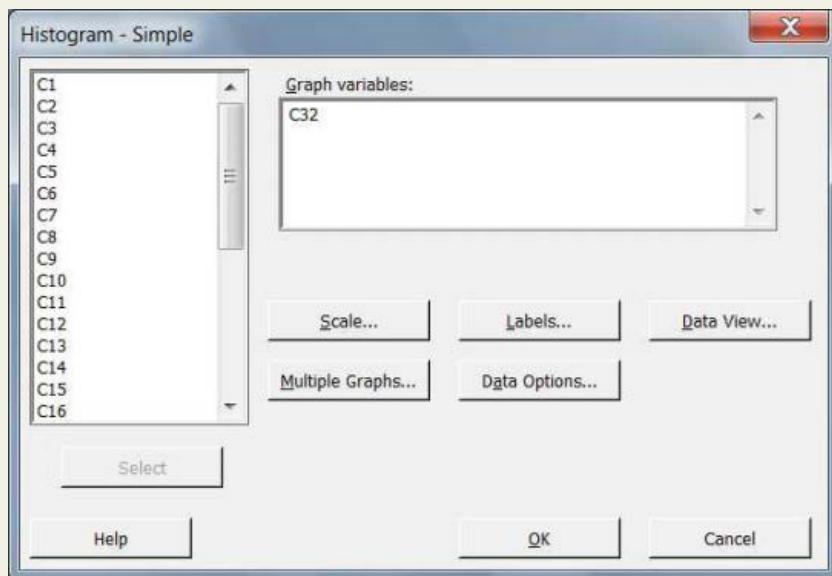
Expr:randBin(50
Variable:X
start:1
end:100
step:1
Paste
Expr:...50,.4)/50
Variable:X
start:1
end:100
step:1
Paste

```

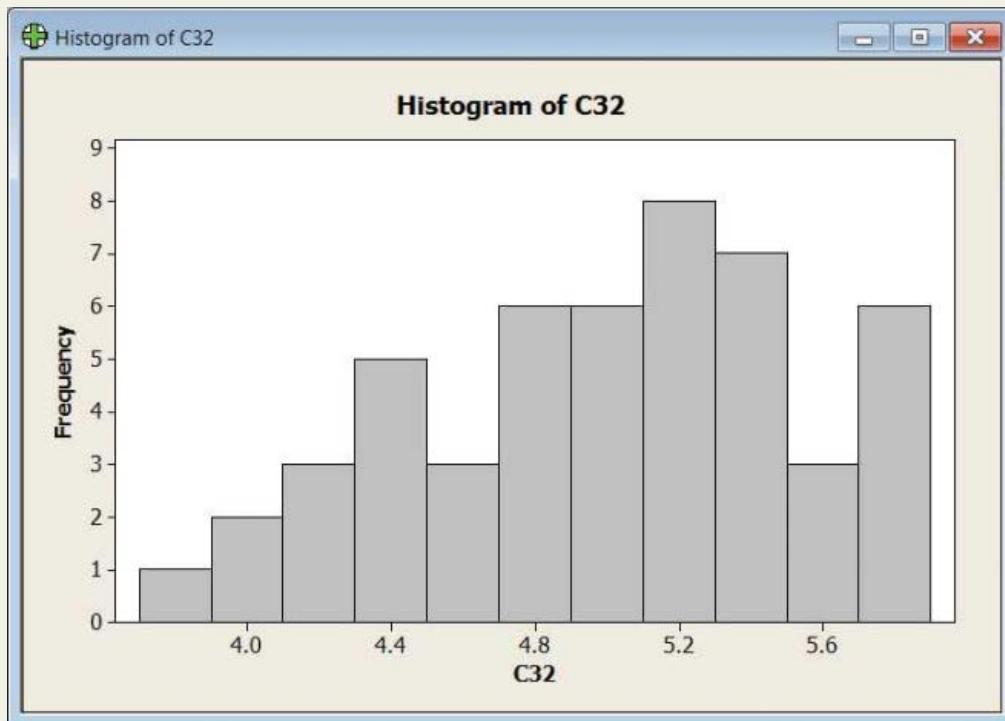
Screen 7.1

at the **X:** prompt, 1 at the **start:** prompt, 100 at the **end:** prompt, 1 at the **step:** prompt, and then highlight **Paste** and press **ENTER**. (See Screen 7.1) Now type **STO > L1 > ENTER**. This will produce 100 values of  $\hat{p}$  and store them in **L1**. If you want more or fewer values of  $\hat{p}$ , change 100 in the above command to any desired number. Then, you can create a histogram of the data using the technology instructions of Chapter 2.

## Minitab



Screen 7.2



Screen 7.3

1. To see an example of the sampling distribution of a sample mean, select **Calc >Random Data >Integer**. We will create 50 samples of size 30, each value a random integer between 0 and 10. Each sample will be a row, so that when we find the mean of each row, the result will go in a column.
2. Enter 50 for **Number of Rows of Data to Generate**.
3. Enter **c1-c30** for **Store in columns**.
4. Enter 0 for **Minimum value** and 10 for **Maximum value**.
5. Select **OK**.
6. Select **Calc >Row Statistics**. Select **Mean** and enter **c1-c30** for **Input variables**. Enter **c32** for **Store results in**.
7. Select **OK**.
8. Select **Graph >Histogram**. Enter **C32** in the **Graph variables:** box (see Screen 7.2). Click **OK** to obtain the histogram that will appear in the graph window (see Screen 7.3). Is this histogram bell shaped? What do you think is the center of this histogram?
9. To see an example of a sampling distribution of a sample proportion, select **Calc >Random Data > Binomial**. Suppose you want to create the number of successes for 100 binomial experiments, each consisting of 80 trials and a probability of success of .40. Each row will contain the number of successes for a set of 80 trials. We will then use these values

to calculate the sample proportions for the 100 experiments, placing the values of the proportions in a different column. In the dialog box you obtain in response to the above commands, follow the following steps:

Enter **100** for **Number of Rows of Data to Generate**.

Enter **C1** for **Store in columns**.

Enter **80** for **Number of trials** and **.40** for **Event probability**.

Click **OK**.

Select **Calc >Calculator**. Enter **C2** in **Store result in variable**.

Enter **C1/80** in **Expression**.

Select **OK**.

Select **Graph >Histogram**.

Enter **C2** in the **Graph variables:** box.

Click **OK** to obtain the histogram that will appear in the graph window. Is this histogram approximately bell shaped? What do you think is the center of this histogram?

## Excel

	A	B	C
1	7.884601	=average(a1:a2)	
2	5.291139		
3			

Screen 7.4

1. To see an example of the sampling distribution of means, use the **rand** function described in Chapter 4 to create a sample of two random numbers between 0 and 10 in column A.
2. Use the **average** function to find their mean B. (See **Screen 7.4**.)
3. Cut and paste the pair of random numbers and their mean 30 times.
4. Use the **frequency** function described in Chapter 2 to find the frequency counts between 0 and 1, 1 and 2, 2 and 3, and so forth through 9 and 10.
5. Use the **Chart wizard** to plot a frequency histogram. Is the histogram bell shaped? Where is it centered?

## TECHNOLOGY ASSIGNMENTS

**TA7.1** Create 200 samples, each containing the results of 30 rolls of a die. Calculate the means of these 200 samples. Construct the histogram, and calculate the mean and standard deviation of these 200 sample means.

**TA7.2** Create 150 samples each containing the results of selecting 35 numbers from 1 through 100. Calculate the means of these 150 samples. Construct the histogram, and calculate the mean and standard deviation of these 150 sample means.

**TA7.3** Refer to Self-Review Test Problem 18. In this assignment, we will explore properties of the sampling distribution of a sample proportion for different sample sizes, as well as by looking at the sample proportion of *failures* instead of *successes*. Self-Review Test Problem 18 stated that 15% of Americans say they trust the government in Washington to do what is right *always or most of the time*.

- a. Using technology, simulate 1000 binomial experiments with 25 trials and probability of success of .15. Calculate the sample proportion of *successes* for each of the 1000 experiments. In another column or list, calculate the sample proportion of *failures* by subtracting the sample proportion of successes from 1.0.
- b. Create two histograms, one of the sample proportions of successes and one of the sample proportions of failures. Calculate the mean and standard deviation for each of the sets of 1000 sample proportions. What are the similarities and differences in the histograms and the summary statistics? Are the histograms approximately bell-shaped?
- c. Repeat parts a and b with 250 trials. In addition, compare the similarities and differences for 25 and 250 trials.



© Steve Debenport/iStockphoto

## Estimation of the Mean and Proportion

### 8.1 Estimation, Point Estimate, and Interval Estimate

### 8.2 Estimation of a Population Mean: $\sigma$ Known

#### Case Study 8-1 How Much Did Registered Nurses Earn in 2011?

### 8.3 Estimation of a Population Mean: $\sigma$ Not Known

### 8.4 Estimation of a Population Proportion: Large Samples

#### Case Study 8-2 Do You Bring Your Lunch From Home?

Do you plan to become a registered nurse? If you do, do you know how much registered nurses earn a year? According to the U.S. Bureau of Labor Statistics, registered nurses earned an average of \$69,110 in 2011. The earnings of registered nurses varied greatly from state to state. Whereas the 2011 average earnings of registered nurses was \$90,860 in California, it was \$64,020 in Florida. (See Case Study 8-1.)

Now we are entering that part of statistics called *inferential statistics*. In Chapter 1 inferential statistics was defined as the part of statistics that helps us make decisions about some characteristics of a population based on sample information. In other words, inferential statistics uses the sample results to make decisions and draw conclusions about the population from which the sample is drawn. Estimation is the first topic to be considered in our discussion of inferential statistics. Estimation and hypothesis testing (discussed in Chapter 9) taken together are usually referred to as inference making. This chapter explains how to estimate the population mean and population proportion for a single population.

## 8.1 Estimation, Point Estimate, and Interval Estimate

In this section, first we discuss the concept of estimation and then the concepts of point and interval estimates.

### 8.1.1 Estimation: An Introduction

**Estimation** is a procedure by which a numerical value or values are assigned to a population parameter based on the information collected from a sample.

#### Definition

**Estimation** The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic is called *estimation*.

In inferential statistics,  $\mu$  is called the *true population mean* and  $p$  is called the *true population proportion*. There are many other population parameters, such as the median, mode, variance, and standard deviation.

The following are a few examples of estimation: an auto company may want to estimate the mean fuel consumption for a particular model of a car; a manager may want to estimate the average time taken by new employees to learn a job; the U.S. Census Bureau may want to find the mean housing expenditure per month incurred by households; and the AWAH (Association of Wives of Alcoholic Husbands) may want to find the proportion (or percentage) of all husbands who are alcoholic.

The examples about estimating the mean fuel consumption, estimating the average time taken to learn a job by new employees, and estimating the mean housing expenditure per month incurred by households are illustrations of estimating the *true population mean*,  $\mu$ . The example about estimating the proportion (or percentage) of all husbands who are alcoholic is an illustration of estimating the *true population proportion*,  $p$ .

If we can conduct a *census* (a survey that includes the entire population) each time we want to find the value of a population parameter, then the estimation procedures explained in this and subsequent chapters are not needed. For example, if the U.S. Census Bureau can contact every household in the United States to find the mean housing expenditure incurred by households, the result of the survey (which will actually be a census) will give the value of  $\mu$ , and the procedures learned in this chapter will not be needed. However, it is too expensive, very time consuming, or virtually impossible to contact every member of a population to collect information to find the true value of a population parameter. Therefore, we usually take a sample from the population and calculate the value of the appropriate sample statistic. Then we assign a value or values to the corresponding population parameter based on the value of the sample statistic. This chapter (and subsequent chapters) explains how to assign values to population parameters based on the values of sample statistics.

For example, to estimate the mean time taken to learn a certain job by new employees, the manager will take a sample of new employees and record the time taken by each of these employees to learn the job. Using this information, he or she will calculate the sample mean,  $\bar{x}$ . Then, based on the value of  $\bar{x}$ , he or she will assign certain values to  $\mu$ . As another example, to estimate the mean housing expenditure per month incurred by all households in the United States, the Census Bureau will take a sample of certain households, collect the information on the housing expenditure that each of these households incurs per month, and compute the value of the sample mean,  $\bar{x}$ . Based on this value of  $\bar{x}$ , the bureau will then assign values to the population mean,  $\mu$ . Similarly, the AWAH will take a sample of husbands and determine the value of the sample proportion,  $\hat{p}$ , which represents the proportion of husbands in the sample who are alcoholic. Using this value of the sample proportion,  $\hat{p}$ , AWAH will assign values to the population proportion,  $p$ .

The value(s) assigned to a population parameter based on the value of a sample statistic is called an **estimate** of the population parameter. For example, suppose the manager takes a

sample of 40 new employees and finds that the mean time,  $\bar{x}$ , taken to learn this job for these employees is 5.5 hours. If he or she assigns this value to the population mean, then 5.5 hours is called an estimate of  $\mu$ . The sample statistic used to estimate a population parameter is called an **estimator**. Thus, the sample mean,  $\bar{x}$ , is an estimator of the population mean,  $\mu$ ; and the sample proportion,  $\hat{p}$ , is an estimator of the population proportion,  $p$ .

### Definition

**Estimate and Estimator** The value(s) assigned to a population parameter based on the value of a sample statistic is called an *estimate*. The sample statistic used to estimate a population parameter is called an *estimator*.

The estimation procedure involves the following steps.

1. Select a sample.
2. Collect the required information from the members of the sample.
3. Calculate the value of the sample statistic.
4. Assign value(s) to the corresponding population parameter.

Remember, **the procedures to be learned in this chapter assume that the sample taken is a simple random sample**. If the sample is not a simple random sample (see Appendix A for a few other kinds of samples), then the procedures to be used to estimate a population mean or proportion become more complex. These procedures are outside the scope of this book.

## 8.1.2 Point and Interval Estimates

An estimate may be a point estimate or an interval estimate. These two types of estimates are described in this section.

### A Point Estimate

If we select a sample and compute the value of the sample statistic for this sample, then this value gives the **point estimate** of the corresponding population parameter.

### Definition

**Point Estimate** The value of a sample statistic that is used to estimate a population parameter is called a *point estimate*.

Thus, the value computed for the sample mean,  $\bar{x}$ , from a sample is a point estimate of the corresponding population mean,  $\mu$ . For the example mentioned earlier, suppose the Census Bureau takes a sample of 10,000 households and determines that the mean housing expenditure per month,  $\bar{x}$ , for this sample is \$1970. Then, using  $\bar{x}$  as a point estimate of  $\mu$ , the Bureau can state that the mean housing expenditure per month,  $\mu$ , for all households is about \$1970. Thus,

$$\text{Point estimate of a population parameter} = \text{Value of the corresponding sample statistic}$$

Each sample selected from a population is expected to yield a different value of the sample statistic. Thus, the value assigned to a population mean,  $\mu$ , based on a point estimate depends on which of the samples is drawn. Consequently, the point estimate assigns a value to  $\mu$  that almost always differs from the true value of the population mean.

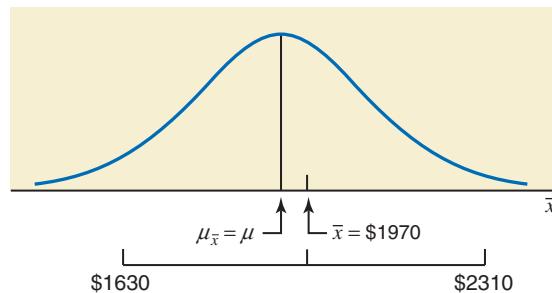
### An Interval Estimate

In the case of **interval estimation**, instead of assigning a single value to a population parameter, an interval is constructed around the point estimate, and then a probabilistic statement that this interval contains the corresponding population parameter is made.

### Definition

**Interval Estimation** In *interval estimation*, an interval is constructed around the point estimate, and it is stated that this interval is likely to contain the corresponding population parameter.

For the example about the mean housing expenditure, instead of saying that the mean housing expenditure per month for all households is \$1970, we may obtain an interval by subtracting a number from \$1970 and adding the same number to \$1970. Then we state that this interval contains the population mean,  $\mu$ . For purposes of illustration, suppose we subtract \$340 from \$1970 and add \$340 to \$1970. Consequently, we obtain the interval (\$1970 - \$340) to (\$1970 + \$340), or \$1630 to \$2310. Then we state that the interval \$1630 to \$2310 is likely to contain the population mean,  $\mu$ , and that the mean housing expenditure per month for all households in the United States is between \$1630 and \$2310. This procedure is called *interval estimation*. The value \$1630 is called the *lower limit* of the interval, and \$2310 is called the *upper limit* of the interval. The number we add to and subtract from the point estimate is called the **margin of error**. Figure 8.1 illustrates the concept of interval estimation.



**Figure 8.1** Interval estimation.

The question arises: What number should we subtract from and add to a point estimate to obtain an interval estimate? The answer to this question depends on two considerations:

1. The standard deviation  $\sigma_{\bar{x}}$  of the sample mean,  $\bar{x}$
2. The level of confidence to be attached to the interval

First, the larger the standard deviation of  $\bar{x}$ , the greater is the number subtracted from and added to the point estimate. Thus, it is obvious that if the range over which  $\bar{x}$  can assume values is larger, then the interval constructed around  $\bar{x}$  must be wider to include  $\mu$ .

Second, the quantity subtracted and added must be larger if we want to have a higher confidence in our interval. We always attach a probabilistic statement to the interval estimation. This probabilistic statement is given by the **confidence level**. An interval constructed based on this confidence level is called a **confidence interval**.

### Definition

**Confidence Level and Confidence Interval** Each interval is constructed with regard to a given *confidence level* and is called a *confidence interval*. The confidence interval is given as

$$\text{Point estimate} \pm \text{Margin of error}$$

The confidence level associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by  $(1 - \alpha)100\%$ .

The confidence level is denoted by  $(1 - \alpha)100\%$ , where  $\alpha$  is the Greek letter *alpha*. When expressed as probability, it is called the *confidence coefficient* and is denoted by  $1 - \alpha$ . In passing, note that  $\alpha$  is called the *significance level*, which will be explained in detail in Chapter 9.

Although any value of the confidence level can be chosen to construct a confidence interval, the more common values are 90%, 95%, and 99%. The corresponding confidence coefficients are .90, .95, and .99, respectively. The next section describes how to construct a confidence interval for the population mean when the population standard deviation,  $\sigma$ , is known.

Sections 8.2 and 8.3 discuss the procedures that are used to estimate a population mean  $\mu$ . In Section 8.2 we assume that the population standard deviation  $\sigma$  is known, and in Section 8.3 we do not assume that the population standard deviation  $\sigma$  is known. In the latter situation, we use the sample standard deviation  $s$  instead of  $\sigma$ . In the real world, the population standard deviation  $\sigma$  is almost never known. Consequently, we (almost) always use the sample standard deviation  $s$ .

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 8.1 Briefly explain the meaning of an estimator and an estimate.
- 8.2 Explain the meaning of a point estimate and an interval estimate.

## 8.2 Estimation of a Population Mean: $\sigma$ Known

This section explains how to construct a confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known. Here, there are three possible cases, as follows.

**Case I.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is normally distributed,

then we use the normal distribution to make the confidence interval for  $\mu$  because from Section 7.3.1 of Chapter 7 the sampling distribution of  $\bar{x}$  is normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , assuming that  $n/N \leq .05$ .

**Case II.** If the following two conditions are fulfilled:

1. The population standard deviation  $\sigma$  is known
2. The sample size is large (i.e.,  $n \geq 30$ ),

then, again, we use the normal distribution to make the confidence interval for  $\mu$  because from Section 7.3.2 of Chapter 7, due to the central limit theorem, the sampling distribution of  $\bar{x}$  is (approximately) normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , assuming that  $n/N \leq .05$ .

**Case III.** If the following three conditions are fulfilled:

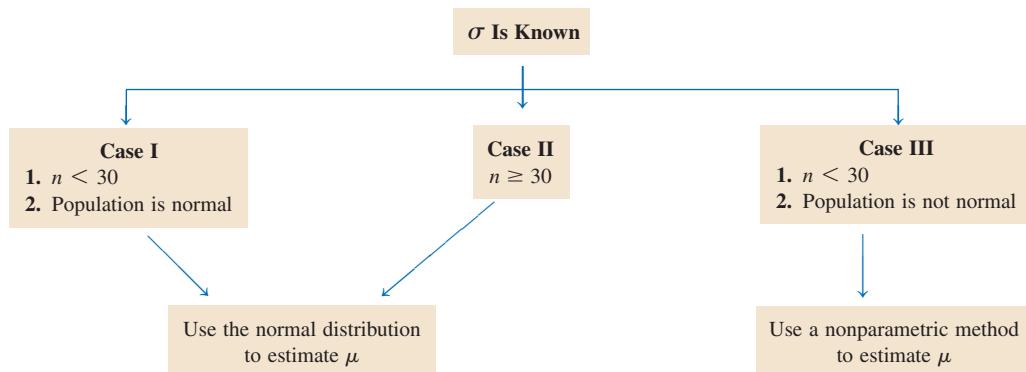
1. The population standard deviation  $\sigma$  is known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is not normally distributed (or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for  $\mu$ . Such procedures are covered in Chapter 15 that is on the Web site of the text.

This section will cover the first two cases. The procedure for making a confidence interval for  $\mu$  is the same in both these cases. Note that in Case I, the population does not have to be exactly normally distributed. As long as it is close to the normal distribution without any outliers, we can use the normal distribution procedure. In Case II, although 30 is considered a large sample, if the population distribution is very different from the normal distribution, then 30 may

not be a large enough sample size for the sampling distribution of  $\bar{x}$  to be normal and, hence, to use the normal distribution.

The following chart summarizes the above three cases.



**Confidence Interval for  $\mu$**  The  $(1 - \alpha)100\%$  confidence interval for  $\mu$  under Cases I and II is

$$\bar{x} \pm z\sigma_{\bar{x}}$$

where

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

The value of  $z$  used here is obtained from the standard normal distribution table (Table IV of Appendix C) for the given confidence level.

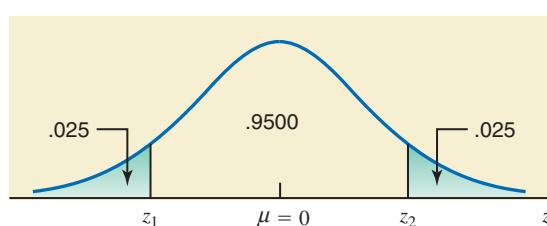
The quantity  $z\sigma_{\bar{x}}$  in the confidence interval formula is called the **margin of error** and is denoted by  $E$ .

### Definition

**Margin of Error** The margin of error for the estimate for  $\mu$ , denoted by  $E$ , is the quantity that is subtracted from and added to the value of  $\bar{x}$  to obtain a confidence interval for  $\mu$ . Thus,

$$E = z\sigma_{\bar{x}}$$

The value of  $z$  in the confidence interval formula is obtained from the standard normal distribution table (Table IV of Appendix C) for the given confidence level. To illustrate, suppose we want to construct a 95% confidence interval for  $\mu$ . A 95% confidence level means that the total area under the normal curve for  $\bar{x}$  between two points (at the same distance) on different sides of  $\mu$  is 95%, or .95, as shown in Figure 8.2. Note that we have denoted these two points by  $z_1$  and  $z_2$  in Figure 8.2. To find the value of  $z$  for a 95% confidence level, we first find the areas to the left of these two points,  $z_1$  and  $z_2$ . Then we find the  $z$  values for these two areas from the normal distribution table. Note that these two values of  $z$  will be the same but with opposite signs. To find these values of  $z$ , we perform the following two steps:



**Figure 8.2** Finding  $z$  for a 95% confidence level.

- The first step is to find the areas to the left of  $z_1$  and  $z_2$ , respectively. Note that the area between  $z_1$  and  $z_2$  is denoted by  $1 - \alpha$ . Hence, the total area in the two tails is  $\alpha$  because the total area under the curve is 1.0. Therefore, the area in each tail, as shown in Figure 8.3, is  $\alpha/2$ . In our example,  $1 - \alpha = .95$ . Hence, the total area in both tails is  $\alpha = 1 - .95 = .05$ . Consequently, the area in each tail is  $\alpha/2 = .05/2 = .025$ . Then, the area to the left of  $z_1$  is .0250, and the area to the left of  $z_2$  is  $.0250 + .95 = .9750$ .
- Now find the  $z$  values from Table IV of Appendix C such that the areas to the left of  $z_1$  and  $z_2$  are .0250 and .9750, respectively. These  $z$  values are  $-1.96$  and  $1.96$ , respectively.

Thus, for a confidence level of 95%, we will use  $z = 1.96$  in the confidence interval formula.

**Figure 8.3** Area in the tails.

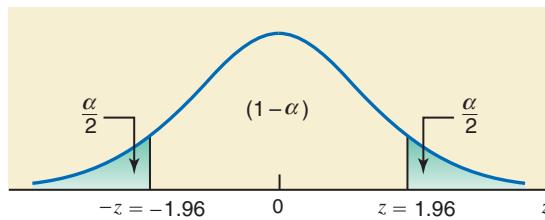


Table 8.1 lists the  $z$  values for some of the most commonly used confidence levels. Note that we always use the positive value of  $z$  in the formula.

**Table 8.1**  $z$  Values for Commonly Used Confidence Levels

Confidence Level	Areas to Look for in Table IV	$z$ Value
90%	.0500 and .9500	1.64 or 1.65
95%	.0250 and .9750	1.96
96%	.0200 and .9800	2.05
97%	.0150 and .9850	2.17
98%	.0100 and .9900	2.33
99%	.0050 and .9950	2.57 or 2.58

Example 8–1 describes the procedure used to construct a confidence interval for  $\mu$  when  $\sigma$  is known, the sample size is small, but the population from which the sample is drawn is normally distributed.

### EXAMPLE 8–1

Finding the point estimate and confidence interval for  $\mu$ :  $\sigma$  known,  $n < 30$ , and population normal.

A publishing company has just published a new college textbook. Before the company decides the price at which to sell this textbook, it wants to know the average price of all such textbooks in the market. The research department at the company took a sample of 25 comparable textbooks and collected information on their prices. This information produced a mean price of \$145 for this sample. It is known that the standard deviation of the prices of all such textbooks is \$35 and the population of such prices is normal.

- (a) What is the point estimate of the mean price of all such college textbooks?
- (b) Construct a 90% confidence interval for the mean price of all such college textbooks.

**Solution** Here,  $\sigma$  is known and, although  $n < 30$ , the population is normally distributed. Hence, we can use the normal distribution. From the given information,

$$n = 25, \bar{x} = \$145, \text{ and } \sigma = \$35$$

The standard deviation of  $\bar{x}$  is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{35}{\sqrt{25}} = \$7.00$$

- (a) The point estimate of the mean price of all such college textbooks is \$145; that is,

Point estimate of  $\mu = \bar{x} = \$145$

- (b) The confidence level is 90%, or .90. First we find the  $z$  value for a 90% confidence level. Here, the area in each tail of the normal distribution curve is  $\alpha/2 = (1 - .90)/2 = .05$ . Now in Table IV, look for the areas .0500 and .9500 and find the corresponding values of  $z$ . These values are  $z = -1.65$  and  $z = 1.65$ .<sup>1</sup>

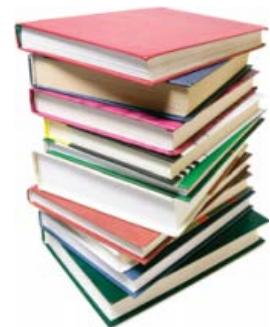
Next, we substitute all the values in the confidence interval formula for  $\mu$ . The 90% confidence interval for  $\mu$  is

$$\begin{aligned}\bar{x} \pm z\sigma_{\bar{x}} &= 145 \pm 1.65(7.00) = 145 \pm 11.55 \\ &= (145 - 11.55) \text{ to } (145 + 11.55) = \$133.45 \text{ to } \$156.55\end{aligned}$$

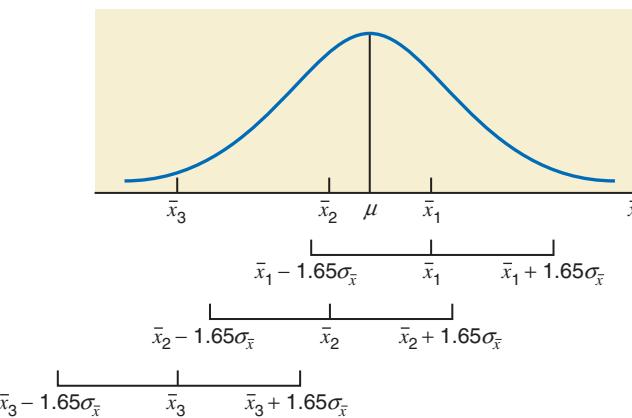
Thus, we are 90% confident that the mean price of all such college textbooks is between \$133.45 and \$156.55. Note that we cannot say for sure whether the interval \$133.45 to \$156.55 contains the true population mean or not. Since  $\mu$  is a constant, we cannot say that the probability is .90 that this interval contains  $\mu$  because either it contains  $\mu$  or it does not. Consequently, the probability is either 1.0 or 0 that this interval contains  $\mu$ . All we can say is that we are 90% confident that the mean price of all such college textbooks is between \$133.45 and \$156.55.

In the above estimate, \$11.55 is called the margin of error or give-and-take figure. ■

How do we interpret a 90% confidence level? In terms of Example 8–1, if we take all possible samples of 25 such college textbooks each and construct a 90% confidence interval for  $\mu$  around each sample mean, we can expect that 90% of these intervals will include  $\mu$  and 10% will not. In Figure 8.4 we show means  $\bar{x}_1$ ,  $\bar{x}_2$ , and  $\bar{x}_3$  of three different samples of the same size drawn from the same population. Also shown in this figure are the 90% confidence intervals constructed around these three sample means. As we observe, the 90% confidence intervals constructed around  $\bar{x}_1$  and  $\bar{x}_2$  include  $\mu$ , but the one constructed around  $\bar{x}_3$  does not. We can state for a 90% confidence level that if we take many samples of the same size from a population and construct 90% confidence intervals around the means of these samples, then 90% of these confidence intervals will be like the ones around  $\bar{x}_1$  and  $\bar{x}_2$  in Figure 8.4, which include  $\mu$ , and 10% will be like the one around  $\bar{x}_3$ , which does not include  $\mu$ .



© Oleg Prikhodko/iStockphoto



**Figure 8.4** Confidence intervals.

<sup>1</sup>Note that there is no apparent reason for choosing .0495 and .9505 and not choosing .0505 and .9495 in Table IV. If we choose .0505 and .9495, the  $z$  values will be  $-1.64$  and  $1.64$ . An alternative is to use the average of  $1.64$  and  $1.65$ , which is  $1.645$ , which we will not do in this text.

Example 8–2 illustrates how to obtain a confidence interval for  $\mu$  when  $\sigma$  is known and the sample size is large ( $n \geq 30$ ).

### ■ EXAMPLE 8–2

*Constructing a confidence interval for  $\mu$ :  $\sigma$  known and  $n \geq 30$ .*

According to Moebs Services Inc., an individual checking account at major U.S. banks costs the banks between \$350 and \$450 per year (*Time*, November 21, 2011). A recent random sample of 600 such checking accounts produced a mean annual cost of \$500 to major U.S. banks. Assume that the standard deviation of annual costs to major U.S. banks of all such checking accounts is \$40. Make a 99% confidence interval for the current mean annual cost to major U.S. banks of all such checking accounts.

**Solution** From the given information,

$$n = 600, \quad \bar{x} = \$500, \quad \sigma = \$40,$$

$$\text{Confidence level} = 99\% \text{ or } .99$$

In this example, although the shape of the population distribution is unknown, the population standard deviation is known, and the sample size is large ( $n \geq 30$ ). Hence, we can use the normal distribution to make a confidence interval for  $\mu$ . To make this confidence interval, first we find the standard deviation of  $\bar{x}$ . The value of  $\sigma_{\bar{x}}$  is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{600}} = 1.63299316$$

To find  $z$  for a 99% confidence level, first we find the area in each of the two tails of the normal distribution curve, which is  $(1 - .99)/2 = .0050$ . Then, we look for .0050 and  $.0050 + .99 = .9950$  areas in the normal distribution table to find the two  $z$  values. These two  $z$  values are (approximately)  $-2.58$  and  $2.58$ . Thus, we will use  $z = 2.58$  in the confidence interval formula. Substituting all the values in the formula, we obtain the 99% confidence interval for  $\mu$ ,

$$\bar{x} \pm z\sigma_{\bar{x}} = 500 \pm 2.58(1.63299316) = 500 \pm 4.21 = \$495.79 \text{ to } \$504.21$$

Thus, we can state with 99% confidence that the current mean annual cost to major U.S. banks of all individual checking accounts is between \$495.79 and \$504.21. ■

The **width of a confidence interval** depends on the size of the margin of error,  $z\sigma_{\bar{x}}$ , which depends on the values of  $z$ ,  $\sigma$ , and  $n$  because  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . However, the value of  $\sigma$  is not under the control of the investigator. Hence, the width of a confidence interval can be controlled using

1. The value of  $z$ , which depends on the confidence level
2. The sample size  $n$

The confidence level determines the value of  $z$ , which in turn determines the size of the margin of error. The value of  $z$  increases as the confidence level increases, and it decreases as the confidence level decreases. For example, the value of  $z$  is approximately 1.65 for a 90% confidence level, 1.96 for a 95% confidence level, and approximately 2.58 for a 99% confidence level. Hence, the higher the confidence level, the larger the width of the confidence interval, other things remaining the same.

For the same value of  $\sigma$ , an increase in the sample size decreases the value of  $\sigma_{\bar{x}}$ , which in turn decreases the size of the margin of error when the confidence level remains unchanged. Therefore, an increase in the sample size decreases the width of the confidence interval.

Thus, if we want to decrease the width of a confidence interval, we have two choices:

1. Lower the confidence level
2. Increase the sample size

Lowering the confidence level is not a good choice, however, because a lower confidence level may give less reliable results. Therefore, we should always prefer to increase the sample size if we want to decrease the width of a confidence interval. Next we illustrate, using Example 8–2,

how either a decrease in the confidence level or an increase in the sample size decreases the width of the confidence interval.

## ① Confidence Level and the Width of the Confidence Interval

Reconsider Example 8–2. Suppose all the information given in that example remains the same. First, let us decrease the confidence level to 95%. From the normal distribution table,  $z = 1.96$  for a 95% confidence level. Then, using  $z = 1.96$  in the confidence interval for Example 8–2, we obtain

$$\bar{x} \pm z\sigma_{\bar{x}} = 500 \pm 1.96(1.63299316) = 500 \pm 3.20 = \$496.80 \text{ to } \$503.20$$

Comparing this confidence interval to the one obtained in Example 8–2, we observe that the width of the confidence interval for a 95% confidence level is smaller than the one for a 99% confidence level.

## ② Sample Size and the Width of the Confidence Interval

Consider Example 8–2 again. Now suppose the information given in that example is based on a sample size of 1000. Further assume that all other information given in that example, including the confidence level, remains the same. First, we calculate the standard deviation of the sample mean using  $n = 1000$ :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{1000}} = 1.26491106$$

Then, the 99% confidence interval for  $\mu$  is

$$\bar{x} \pm z\sigma_{\bar{x}} = 500 \pm 2.58(1.26491106) = 500 \pm 3.26 = \$496.74 \text{ to } \$503.26$$

Comparing this confidence interval to the one obtained in Example 8–2, we observe that the width of the 99% confidence interval for  $n = 1000$  is smaller than the 99% confidence interval for  $n = 600$ .

### 8.2.1 Determining the Sample Size for the Estimation of Mean

One reason we usually conduct a sample survey and not a census is that almost always we have limited resources at our disposal. In light of this, if a smaller sample can serve our purpose, then we will be wasting our resources by taking a larger sample. For instance, suppose we want to estimate the mean life of a certain auto battery. If a sample of 40 batteries can give us the confidence interval we are looking for, then we will be wasting money and time if we take a sample of a much larger size—say, 500 batteries. In such cases, if we know the confidence level and the width of the confidence interval that we want, then we can find the (approximate) size of the sample that will produce the required result.

From earlier discussion, we learned that  $E = z\sigma_{\bar{x}}$  is called the margin of error of estimate for  $\mu$ . As we know, the standard deviation of the sample mean is equal to  $\sigma/\sqrt{n}$ . Therefore, we can write the margin of error of estimate for  $\mu$  as

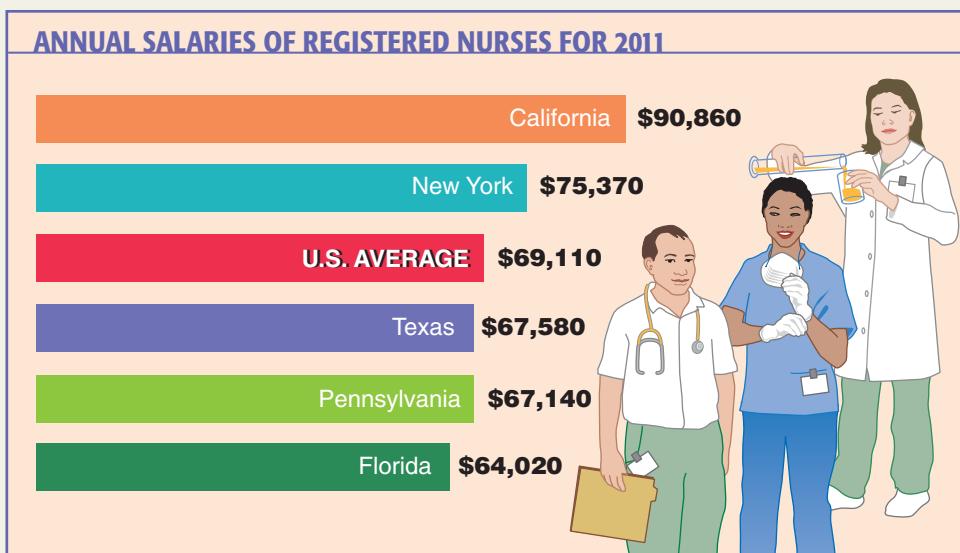
$$E = z \cdot \frac{\sigma}{\sqrt{n}}$$

Suppose we predetermine the size of the margin of error,  $E$ , and want to find the size of the sample that will yield this margin of error. From the above expression, the following formula is obtained that determines the required sample size  $n$ .

**Determining the Sample Size for the Estimation of  $\mu$**  Given the confidence level and the standard deviation of the population, the sample size that will produce a predetermined margin of error  $E$  of the confidence interval estimate of  $\mu$  is

$$n = \frac{z^2 \sigma^2}{E^2}$$

## HOW MUCH DID REGISTERED NURSES EARN IN 2011?



Data source: U.S. Bureau of Labor Statistics, March 2012.

As shown in the accompanying chart, according to a survey by the U.S. Bureau of Labor Statistics, registered nurses in the United States earned an average of \$69,110 in 2011 (<http://www.bls.gov/oes/current/oes291111.htm>). The average earnings of registered nurses varied greatly from state to state. Whereas the 2011 average earnings of registered nurses was \$90,860 in California, it was only \$55,710 in South Dakota (which is not shown in the graph). The 2011 earnings of registered nurses also varied greatly among different metropolitan areas. Whereas such average earnings was \$120,540 in the Vallejo–Fairfield (California) metropolitan area, it was \$85,340 in the Los Angeles–Long Beach–Glendale metropolitan division. This average was \$60,260 for the Greenville, North Carolina, metropolitan area. (Note that these numbers for metropolitan areas are not shown in the accompanying graph.) As we know, such estimates are based on sample surveys. If we know the sample size and the population standard deviation for any state or metropolitan area, we can find a confidence interval for the 2011 average earnings of registered nurses for that state or metropolitan area. For example, if we know the sample size and the population standard deviation of the 2011 earnings of registered nurses in Texas, we can make a confidence interval for the 2011 average earnings of all registered nurses in Texas using the following formula:

$$\bar{x} \pm z\sigma_{\bar{x}}$$

In this formula, we can substitute the values of  $\bar{x}$ ,  $z$ , and  $\sigma_{\bar{x}}$  to obtain the confidence interval. Remember that  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Suppose we want to find a 98% confidence interval for the 2011 average earnings of registered nurses in Texas. Suppose that the 2011 average earnings of registered nurses in Texas (given in the graph) is based on a random sample of 1600 registered nurses and that the population standard deviation for such 2011 earnings is \$6240. Then the 98% confidence interval for the corresponding population mean is calculated as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6240}{\sqrt{1600}} = \$156.00$$

$$\bar{x} \pm z\sigma_{\bar{x}} = 67,580 \pm 2.33(156.00) = 67,580 \pm 363.48 = \$67,216.52 \text{ to } \$67,943.48$$

Thus, we can state with 98% confidence that the 2011 average earnings of all registered nurses in Texas was in the interval \$67,216.52 to \$67,943.48. We can find the confidence intervals for the other states mentioned in the graph the same way. Note that the sample means given in the graph are the point estimates of the corresponding population means.

In practice, we typically do not know the value of the population standard deviation, but we do know the value of the sample standard deviation, which is calculated from the sample data. In this case, we will find a confidence interval for the population mean using the  $t$  distribution procedure, which is explained in the next section.

If we do not know  $\sigma$ , we can take a preliminary sample (of any arbitrarily determined size) and find the sample standard deviation,  $s$ . Then we can use  $s$  for  $\sigma$  in the formula. However, note that using  $s$  for  $\sigma$  may give a sample size that eventually may produce an error much larger (or smaller) than the predetermined margin of error. This will depend on how close  $s$  and  $\sigma$  are.

Example 8–3 illustrates how we determine the sample size that will produce the margin of error of estimate for  $\mu$  within a certain limit.

### ■ EXAMPLE 8–3

An alumni association wants to estimate the mean debt of this year's college graduates. It is known that the population standard deviation of the debts of this year's college graduates is \$11,800. How large a sample should be selected so that the estimate with a 99% confidence level is within \$800 of the population mean?

*Determining the sample size for the estimation of  $\mu$ .*

**Solution** The alumni association wants the 99% confidence interval for the mean debt of this year's college graduates to be

$$\bar{x} \pm 800$$

Hence, the maximum size of the margin of error of estimate is to be \$800; that is,

$$E = \$800$$

The value of  $z$  for a 99% confidence level is 2.58. The value of  $\sigma$  is given to be \$11,800. Therefore, substituting all values in the formula and simplifying, we obtain

$$n = \frac{z^2 \sigma^2}{E^2} = \frac{(2.58)^2 (11,800)^2}{(800)^2} = 1448.18 \approx 1449$$

Thus, the required sample size is 1449. If the alumni association takes a sample of 1449 of this year's college graduates, computes the mean debt for this sample, and then makes a 99% confidence interval around this sample mean, the margin of error of estimate will be approximately \$800. Note that we have rounded the final answer for the sample size to the next higher integer. This is always the case when determining the sample size. ■



© Christopher Futcher/iStockphoto

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**8.3** What is the point estimator of the population mean,  $\mu$ ? How would you calculate the margin of error for an estimate of  $\mu$ ?

**8.4** Explain the various alternatives for decreasing the width of a confidence interval. Which is the best alternative?

**8.5** Briefly explain how the width of a confidence interval decreases with an increase in the sample size. Give an example.

**8.6** Briefly explain how the width of a confidence interval decreases with a decrease in the confidence level. Give an example.

**8.7** Briefly explain the difference between a confidence level and a confidence interval.

**8.8** What is the margin of error of estimate for  $\mu$  when  $\sigma$  is known? How is it calculated?

**8.9** How will you interpret a 99% confidence interval for  $\mu$ ? Explain.

**8.10** Find  $z$  for each of the following confidence levels.

- a. 90%      b. 95%      c. 96%      d. 97%      e. 98%      f. 99%

**8.11** For a data set obtained from a sample,  $n = 20$  and  $\bar{x} = 24.5$ . It is known that  $\sigma = 3.1$ . The population is normally distributed.

- a. What is the point estimate of  $\mu$ ?
- b. Make a 99% confidence interval for  $\mu$ .
- c. What is the margin of error of estimate for part b?

**8.12** For a data set obtained from a sample,  $n = 81$  and  $\bar{x} = 48.25$ . It is known that  $\sigma = 4.8$ .

- What is the point estimate of  $\mu$ ?
- Make a 95% confidence interval for  $\mu$ .
- What is the margin of error of estimate for part b?

**8.13** The standard deviation for a population is  $\sigma = 15.3$ . A sample of 36 observations selected from this population gave a mean equal to 74.8.

- Make a 90% confidence interval for  $\mu$ .
- Construct a 95% confidence interval for  $\mu$ .
- Determine a 99% confidence interval for  $\mu$ .
- Does the width of the confidence intervals constructed in parts a through c increase as the confidence level increases? Explain your answer.

**8.14** The standard deviation for a population is  $\sigma = 14.8$ . A sample of 25 observations selected from this population gave a mean equal to 143.72. The population is known to have a normal distribution.

- Make a 99% confidence interval for  $\mu$ .
- Construct a 95% confidence interval for  $\mu$ .
- Determine a 90% confidence interval for  $\mu$ .
- Does the width of the confidence intervals constructed in parts a through c decrease as the confidence level decreases? Explain your answer.

**8.15** The standard deviation for a population is  $\sigma = 6.30$ . A random sample selected from this population gave a mean equal to 81.90. The population is known to be normally distributed.

- Make a 99% confidence interval for  $\mu$  assuming  $n = 16$ .
- Construct a 99% confidence interval for  $\mu$  assuming  $n = 20$ .
- Determine a 99% confidence interval for  $\mu$  assuming  $n = 25$ .
- Does the width of the confidence intervals constructed in parts a through c decrease as the sample size increases? Explain.

**8.16** The standard deviation for a population is  $\sigma = 7.14$ . A random sample selected from this population gave a mean equal to 48.52.

- Make a 95% confidence interval for  $\mu$  assuming  $n = 196$ .
- Construct a 95% confidence interval for  $\mu$  assuming  $n = 100$ .
- Determine a 95% confidence interval for  $\mu$  assuming  $n = 49$ .
- Does the width of the confidence intervals constructed in parts a through c increase as the sample size decreases? Explain.

**8.17** For a population, the value of the standard deviation is 2.65. A sample of 35 observations taken from this population produced the following data.

42	51	42	31	28	36	49
29	46	37	32	27	33	41
47	41	28	46	34	39	48
26	35	37	38	46	48	39
29	31	44	41	37	38	46

- What is the point estimate of  $\mu$ ?
- Make a 98% confidence interval for  $\mu$ .
- What is the margin of error of estimate for part b?

**8.18** For a population, the value of the standard deviation is 4.96. A sample of 32 observations taken from this population produced the following data.

74	85	72	73	86	81	77	60
83	78	79	88	76	73	84	78
81	72	82	81	79	83	88	86
78	83	87	82	80	84	76	74

- What is the point estimate of  $\mu$ ?
- Make a 99% confidence interval for  $\mu$ .
- What is the margin of error of estimate for part b?

**8.19** For a population data set,  $\sigma = 12.5$ .

- How large a sample should be selected so that the margin of error of estimate for a 99% confidence interval for  $\mu$  is 2.50?
- How large a sample should be selected so that the margin of error of estimate for a 96% confidence interval for  $\mu$  is 3.20?

**8.20** For a population data set,  $\sigma = 14.50$ .

- What should the sample size be for a 98% confidence interval for  $\mu$  to have a margin of error of estimate equal to 5.50?
- What should the sample size be for a 95% confidence interval for  $\mu$  to have a margin of error of estimate equal to 4.25?

**8.21** Determine the sample size for the estimate of  $\mu$  for the following.

- $E = 2.3$ ,  $\sigma = 15.40$ , confidence level = 99%
- $E = 4.1$ ,  $\sigma = 23.45$ , confidence level = 95%
- $E = 25.9$ ,  $\sigma = 122.25$ , confidence level = 90%

**8.22** Determine the sample size for the estimate of  $\mu$  for the following.

- $E = .17$ ,  $\sigma = .90$ , confidence level = 99%
- $E = 1.45$ ,  $\sigma = 5.82$ , confidence level = 95%
- $E = 5.65$ ,  $\sigma = 18.20$ , confidence level = 90%

## ■ APPLICATIONS

**8.23** A travel agent wants to gather information on the per-night cost at hotels in Caribbean countries. She took a random sample of 52 rooms from various hotels in those countries. The sample produced a mean cost for the 52 rooms to be \$208.35 per night. If the population standard deviation of costs for a one-night stay in Caribbean hotels is \$47.45, find a 99% confidence interval for the average cost per night in Caribbean hotels.

**8.24** A city planner wants to estimate the average monthly residential water usage in the city. He selected a random sample of 40 households from the city, which gave the mean water usage to be 3415.70 gallons over a 1-month period. Based on earlier data, the population standard deviation of the monthly residential water usage in this city is 389.60 gallons. Make a 95% confidence interval for the average monthly residential water usage for all households in this city.

**8.25** An entertainment company is in the planning stages of producing a new computer-animated movie for national release, so they need to determine the production time (labor-hours necessary) to produce the movie. The mean production time for a random sample of 14 big-screen computer-animated movies is found to be 53,550 labor-hours. Suppose that the population standard deviation is known to be 7462 labor-hours and the distribution of production times is normal.

- Construct a 98% confidence interval for the mean production time to produce a big-screen computer-animated movie.
- Explain why we need to make the confidence interval. Why is it not correct to say that the average production time needed to produce all big-screen computer-animated movies is 53,550 labor-hours?

**8.26** Lazurus Steel Corporation produces iron rods that are supposed to be 36 inches long. The machine that makes these rods does not produce each rod exactly 36 inches long. The lengths of the rods vary slightly. It is known that when the machine is working properly, the mean length of the rods made on this machine is 36 inches. The standard deviation of the lengths of all rods produced on this machine is always equal to .10 inch. The quality control department takes a sample of 20 such rods every week, calculates the mean length of these rods, and makes a 99% confidence interval for the population mean. If either the upper limit of this confidence interval is greater than 36.05 inches or the lower limit of this confidence interval is less than 35.95 inches, the machine is stopped and adjusted. A recent sample of 20 rods produced a mean length of 36.02 inches. Based on this sample, will you conclude that the machine needs an adjustment? Assume that the lengths of all such rods have a normal distribution.

**8.27** At Farmer's Dairy, a machine is set to fill 32-ounce milk cartons. However, this machine does not put exactly 32 ounces of milk into each carton; the amount varies slightly from carton to carton. It is known that when the machine is working properly, the mean net weight of these cartons is 32 ounces. The standard deviation of the amounts of milk in all such cartons is always equal to .15 ounce. The quality control department takes a sample of 25 such cartons every week, calculates the mean net weight of these cartons, and makes a 99% confidence interval for the population mean. If either the upper limit of this confidence interval is greater than 32.15 ounces or the lower limit of this confidence interval is less than 31.85 ounces, the machine is stopped and adjusted. A recent sample of 25 such cartons produced a mean net weight of 31.94 ounces. Based on this sample, will you conclude that the machine needs an adjustment? Assume that the amounts of milk put in all such cartons have a normal distribution.

**8.28** A consumer agency that proposes that lawyers' rates are too high wanted to estimate the mean hourly rate for all lawyers in New York City. A sample of 70 lawyers taken from New York City showed that the

mean hourly rate charged by them is \$570. The population standard deviation of hourly charges for all lawyers in New York City is \$110.

- Construct a 99% confidence interval for the mean hourly charges for all lawyers in New York City.
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which alternative is the best?

**8.29** A bank manager wants to know the mean amount of mortgage paid per month by homeowners in an area. A random sample of 120 homeowners selected from this area showed that they pay an average of \$1575 per month for their mortgages. The population standard deviation of such mortgages is \$215.

- Find a 97% confidence interval for the mean amount of mortgage paid per month by all homeowners in this area.
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which alternative is the best?

**8.30** A marketing researcher wants to find a 95% confidence interval for the mean amount that visitors to a theme park spend per person per day. She knows that the standard deviation of the amounts spent per person per day by all visitors to this park is \$11. How large a sample should the researcher select so that the estimate will be within \$2 of the population mean?

**8.31** A company that produces detergents wants to estimate the mean amount of detergent in 64-ounce jugs at a 99% confidence level. The company knows that the standard deviation of the amounts of detergent in all such jugs is .20 ounce. How large a sample should the company select so that the estimate is within .04 ounce of the population mean?

**8.32** A department store manager wants to estimate at a 98% confidence level the mean amount spent by all customers at this store. The manager knows that the standard deviation of amounts spent by all customers at this store is \$31. What sample size should he choose so that the estimate is within \$3 of the population mean?

**8.33** Refer to Exercise 8.24. A city planner wants to estimate, with a 97% confidence level, the average monthly residential water usage in the city. Based on earlier data, the population standard deviation of the monthly residential water usage in this city is 389.60 gallons. How large a sample should be selected so that the estimate for the average monthly residential water usage in this city is within 100 gallons of the population mean?

**\*8.34** You are interested in estimating the mean commuting time from home to school for all commuter students at your school. Briefly explain the procedure you will follow to conduct this study. Collect the required data from a sample of 30 or more such students and then estimate the population mean at a 99% confidence level. Assume that the population standard deviation for such times is 5.5 minutes.

**\*8.35** You are interested in estimating the mean age of cars owned by all people in the United States. Briefly explain the procedure you will follow to conduct this study. Collect the required data on a sample of 30 or more cars and then estimate the population mean at a 95% confidence level. Assume that the population standard deviation is 2.4 years.

## 8.3 Estimation of a Population Mean: $\sigma$ Not Known

This section explains how to construct a confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known. Here, again, there are three possible cases:

**Case I.** If the following three conditions are fulfilled:

- The population standard deviation  $\sigma$  is not known
- The sample size is small (i.e.,  $n < 30$ )
- The population from which the sample is selected is normally distributed,

then we use the  $t$  distribution (explained in Section 8.3.1) to make the confidence interval for  $\mu$ .

**Case II.** If the following two conditions are fulfilled:

1. The population standard deviation  $\sigma$  is not known
2. The sample size is large (i.e.,  $n \geq 30$ ),

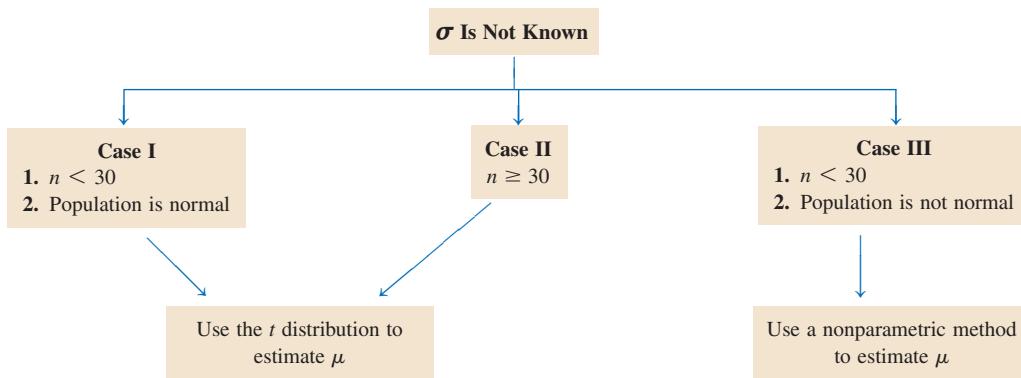
then again we use the  $t$  distribution to make the confidence interval for  $\mu$ .

**Case III.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is not known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is not normally distributed (or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for  $\mu$ . Such procedures are covered in Chapter 15, which is on the Web site for this text.

The following chart summarizes the above three cases.



In the next subsection, we discuss the  $t$  distribution, and then in Section 8.3.2 we show how to use the  $t$  distribution to make a confidence interval for  $\mu$  when  $\sigma$  is not known and conditions of Cases I or II are satisfied.

### 8.3.1 The $t$ Distribution

The  **$t$  distribution** was developed by W. S. Gosset in 1908 and published under the pseudonym *Student*. As a result, the  $t$  distribution is also called *Student's t distribution*. The  $t$  distribution is similar to the normal distribution in some respects. Like the normal distribution curve, the  $t$  distribution curve is symmetric (bell shaped) about the mean and never meets the horizontal axis. The total area under a  $t$  distribution curve is 1.0, or 100%. However, the  $t$  distribution curve is flatter than the standard normal distribution curve. In other words, the  $t$  distribution curve has a lower height and a wider spread (or, we can say, a larger standard deviation) than the standard normal distribution. However, as the sample size increases, the  $t$  distribution approaches the standard normal distribution. The units of a  $t$  distribution are denoted by  $t$ .

The shape of a particular  $t$  distribution curve depends on the number of **degrees of freedom (df)**. For the purpose of this chapter and Chapter 9, the number of degrees of freedom for a  $t$  distribution is equal to the sample size minus one, that is,

$$df = n - 1$$

The number of degrees of freedom is the only parameter of the  $t$  distribution. There is a different  $t$  distribution for each number of degrees of freedom. Like the standard normal distribution, the mean of the  $t$  distribution is 0. But unlike the standard normal distribution, whose standard

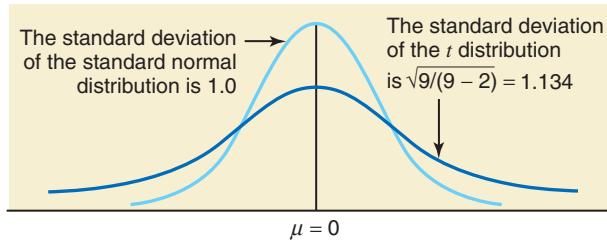
deviation is 1, the standard deviation of a  $t$  distribution is  $\sqrt{df/(df - 2)}$  for  $df > 2$ . Thus, the standard deviation of a  $t$  distribution is always greater than 1 and, hence, it is larger than the standard deviation of the standard normal distribution.

### Definition

**The  $t$  Distribution** The  $t$  distribution is a specific type of bell-shaped distribution with a lower height and a wider spread than the standard normal distribution. As the sample size becomes larger, the  $t$  distribution approaches the standard normal distribution. The  $t$  distribution has only one parameter, called the degrees of freedom ( $df$ ). The mean of the  $t$  distribution is equal to 0, and its standard deviation is  $\sqrt{df/(df - 2)}$ .

Figure 8.5 shows the standard normal distribution and the  $t$  distribution for 9 degrees of freedom. The standard deviation of the standard normal distribution is 1.0, and the standard deviation of the  $t$  distribution is  $\sqrt{9/(9 - 2)} = \sqrt{9/7} = 1.134$ .

**Figure 8.5** The  $t$  distribution for  $df = 9$  and the standard normal distribution.



As stated earlier, the number of degrees of freedom for a  $t$  distribution for the purpose of this chapter is  $n - 1$ . *The number of degrees of freedom is defined as the number of observations that can be chosen freely.* As an example, suppose we know that the mean of four values is 20. Consequently, the sum of these four values is  $20(4) = 80$ . Now, how many values out of four can we choose freely so that the sum of these four values is 80? The answer is that we can freely choose  $4 - 1 = 3$  values. Suppose we choose 27, 8, and 19 as the three values. Given these three values and the information that the mean of the four values is 20, the fourth value is  $80 - 27 - 8 - 19 = 26$ . Thus, once we have chosen three values, the fourth value is automatically determined. Consequently, the number of degrees of freedom for this example is

$$df = n - 1 = 4 - 1 = 3$$

We subtract 1 from  $n$  because we lose 1 degree of freedom to calculate the mean.

Table V of Appendix C lists the values of  $t$  for the given number of degrees of freedom and areas in the right tail of a  $t$  distribution. Because the  $t$  distribution is symmetric, these are also the values of  $-t$  for the same number of degrees of freedom and the same areas in the left tail of the  $t$  distribution. Example 8–4 describes how to read Table V of Appendix C.

### ■ EXAMPLE 8–4

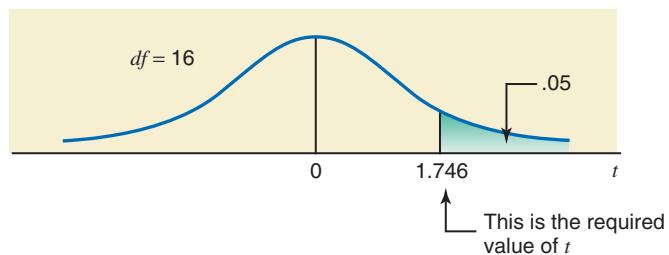
Find the value of  $t$  for 16 degrees of freedom and .05 area in the right tail of a  $t$  distribution curve.

**Solution** In Table V of Appendix C, we locate 16 in the column of degrees of freedom (labeled  $df$ ) and .05 in the row of *Area in the right tail under the  $t$  distribution curve* at the top of the table. The entry at the intersection of the row of 16 and the column of .05, which is 1.746, gives the required value of  $t$ . The relevant portion of Table V of Appendix C is shown here as Table 8.2. The value of  $t$  read from the  $t$  distribution table is shown in Figure 8.6.

**Table 8.2** Determining  $t$  for 16 df and .05 Area in the Right Tail

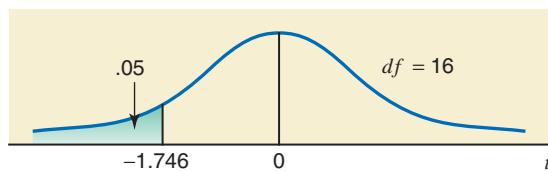
$df$	Area in the right tail				
	.10	.05	.025	...	.001
1	3.078	6.314	12.706	...	318.309
2	1.886	2.920	4.303	...	22.327
3	1.638	2.353	3.182	...	10.215
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
$df \rightarrow 16$	1.337	1.746	2.120	...	3.686
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
75	1.293	1.665	1.992	...	3.202
$\infty$	1.282	1.645	1.960	...	3.090

The required value of  $t$  for 16 df and .05 area in the right tail



**Figure 8.6** The value of  $t$  for 16 df and .05 area in the right tail.

Because of the symmetric shape of the  $t$  distribution curve, the value of  $t$  for 16 degrees of freedom and .05 area in the left tail is  $-1.746$ . Figure 8.7 illustrates this case.



**Figure 8.7** The value of  $t$  for 16 df and .05 area in the left tail.

### 8.3.2 Confidence Interval for $\mu$ Using the $t$ Distribution

To reiterate, when the conditions mentioned under Cases I or II in the beginning of this section hold true, we use the  $t$  distribution to construct a confidence interval for the population mean,  $\mu$ .

When the population standard deviation  $\sigma$  is not known, then we replace it by the sample standard deviation  $s$ , which is its estimator. Consequently, for the standard deviation of  $\bar{x}$ , we use

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

for  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Note that the value of  $s_{\bar{x}}$  is a point estimate of  $\sigma_{\bar{x}}$ .

**Confidence Interval for  $\mu$  Using the  $t$  Distribution** The  $(1 - \alpha)100\%$  confidence interval for  $\mu$  is

$$\bar{x} \pm ts_{\bar{x}}$$

where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The value of  $t$  is obtained from the  $t$  distribution table for  $n - 1$  degrees of freedom and the given confidence level. Here  $ts_{\bar{x}}$  is the margin of error of the estimate; that is,

$$E = ts_{\bar{x}}$$

Examples 8–5 and 8–6 describe the procedure of constructing a confidence interval for  $\mu$  using the  $t$  distribution.

### ■ EXAMPLE 8–5

Constructing a 95% confidence interval for  $\mu$  using the  $t$  distribution.

According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance coverage paid an average premium of \$4129 for family health insurance coverage during 2011 (*USA TODAY*, October 10, 2011). A recent random sample of 25 workers from New York City who have employer-provided health insurance coverage paid an average premium of \$6600 for family health insurance coverage with a standard deviation of \$800. Make a 95% confidence interval for the current average premium paid for family health insurance coverage by all workers in New York City who have employer-provided health insurance coverage. Assume that the distribution of premiums paid for family health insurance coverage by all workers in New York City who have employer-provided health insurance coverage is normally distributed.

**Solution** Here,  $\sigma$  is not known,  $n < 30$ , and the population is normally distributed. All conditions mentioned in Case I of the chart given in the beginning of this section are satisfied. Therefore, we will use the  $t$  distribution to make a confidence interval for  $\mu$ . From the given information,

$$n = 25, \quad \bar{x} = \$6600, \quad s = \$800, \quad \text{Confidence level} = 95\% \text{ or } .95$$

The value of  $s_{\bar{x}}$  is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{800}{\sqrt{25}} = \$160$$

To find the value of  $t$ , we need to know the degrees of freedom and the area under the  $t$  distribution curve in each tail.

$$\text{Degrees of freedom} = n - 1 = 25 - 1 = 24$$

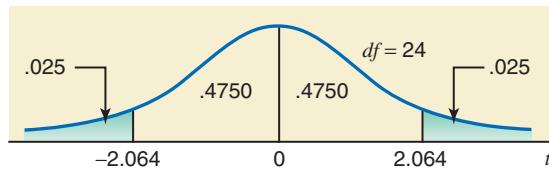
To find the area in each tail, we divide the confidence level by 2 and subtract the number obtained from .5. Thus,

$$\text{Area in each tail} = .5 - (.95/2) = .5 - .4750 = .025$$

From the  $t$  distribution table, Table V of Appendix C, the value of  $t$  for  $df = 24$  and .025 area in the right tail is 2.064. The value of  $t$  is shown in Figure 8.8.

By substituting all values in the formula for the confidence interval for  $\mu$ , we obtain the 95% confidence interval as

$$\bar{x} \pm ts_{\bar{x}} = 6600 \pm 2.064(160) = 6600 \pm 330.24 = \$6269.76 \text{ to } \$6930.24$$

Figure 8.8 The value of  $t$ .

Thus, we can state with 95% confidence that the current mean premium paid for family health insurance coverage by all workers in New York City who have employer-provided health insurance coverage is between \$6269.76 and \$6930.24.

Note that  $\bar{x} = \$6600$  is a point estimate of  $\mu$  in this example, and \$330.24 is the margin of error.

### ■ EXAMPLE 8-6

Sixty-four randomly selected adults who buy books for general reading were asked how much they usually spend on books per year. The sample produced a mean of \$1450 and a standard deviation of \$300 for such annual expenses. Determine a 99% confidence interval for the corresponding population mean.

Constructing a 99% confidence interval for  $\mu$  using the  $t$  distribution.

**Solution** From the given information,

$$n = 64, \quad \bar{x} = \$1450, \quad s = \$300,$$

and

Confidence level = 99%, or .99

Here  $\sigma$  is not known, but the sample size is large ( $n \geq 30$ ). Hence, we will use the  $t$  distribution to make a confidence interval for  $\mu$ . First we calculate the standard deviation of  $\bar{x}$ , the number of degrees of freedom, and the area in each tail of the  $t$  distribution.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{300}{\sqrt{64}} = \$37.50$$

$$df = n - 1 = 64 - 1 = 63$$

$$\text{Area in each tail} = .5 - (.99/2) = .5 - .4950 = .005$$

From the  $t$  distribution table,  $t = 2.656$  for 63 degrees of freedom and .005 area in the right tail. The 99% confidence interval for  $\mu$  is

$$\begin{aligned} \bar{x} \pm ts_{\bar{x}} &= \$1450 \pm 2.656(37.50) \\ &= \$1450 \pm \$99.60 = \$1350.40 \text{ to } \$1549.60 \end{aligned}$$

Thus, we can state with 99% confidence that based on this sample the mean annual expenditure on books by all adults who buy books for general reading is between \$1350.40 and \$1549.60.



PhotoDisc, Inc./Getty Images

Again, we can decrease the width of a confidence interval for  $\mu$  either by lowering the confidence level or by increasing the sample size, as was done in Section 8.2. However, increasing the sample size is the better alternative.

#### Note: What If the Sample Size Is Large and the Number of $df$ Is Not in the $t$ Distribution Table?

In the above section, when  $\sigma$  is not known, we used the  $t$  distribution to make a confidence interval for  $\mu$  in Cases I and II. Note that in Case II, the sample size is large. If we have access to technology, it does not matter how large (greater than 30) the sample size is; we can use the  $t$  distribution. However, if we are using the  $t$  distribution table (Table V of Appendix C), this may pose a problem. Usually such a table goes only up to a certain number of degrees of freedom. For example, Table V in Appendix C goes only up to 75 degrees of freedom. Thus, if the

sample size is larger than 76, we cannot use Table V to find the  $t$  value for the given confidence level to use in the confidence interval in this section. In such a situation when  $n$  is large (for example, 500) and the number of  $df$  is not included in the  $t$  distribution table, there are two options:

1. Use the  $t$  value from the last row (the row of  $\infty$ ) in Table V.
2. Use the normal distribution as an approximation to the  $t$  distribution.

Note that the  $t$  values you will obtain from the last row of the  $t$  distribution table are the same as obtained from the normal distribution table for the same confidence levels, the only difference being the decimal places. To use the normal distribution as an approximation to the  $t$  distribution to make a confidence interval for  $\mu$ , the procedure is exactly like the one in Section 8.2, except that now we replace  $\sigma$  by  $s$ , and  $\sigma_{\bar{x}}$  by  $s_{\bar{x}}$ .

Again, note that here we can use the normal distribution as a convenience and as an approximation, but if we can, we should use the  $t$  distribution by using technology. Exercises 8.50, 8.51, and 8.56 at the end of this section present such situations.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**8.36** Briefly explain the similarities and the differences between the standard normal distribution and the  $t$  distribution.

**8.37** What are the parameters of a normal distribution and a  $t$  distribution? Explain.

**8.38** Briefly explain the meaning of the degrees of freedom for a  $t$  distribution. Give one example.

**8.39** What assumptions must hold true to use the  $t$  distribution to make a confidence interval for  $\mu$ ?

**8.40** Find the value of  $t$  for the  $t$  distribution for each of the following.

- |   |   |
|---|---|
| a. Area in the right tail = .05 and $df = 12$ | b. Area in the left tail = .025 and $n = 66$  |
| c. Area in the left tail = .001 and $df = 49$ | d. Area in the right tail = .005 and $n = 24$ |

**8.41** a. Find the value of  $t$  for the  $t$  distribution with a sample size of 21 and area in the left tail equal to .10.

b. Find the value of  $t$  for the  $t$  distribution with a sample size of 14 and area in the right tail equal to .025.

c. Find the value of  $t$  for the  $t$  distribution with 45 degrees of freedom and .001 area in the right tail.

d. Find the value of  $t$  for the  $t$  distribution with 37 degrees of freedom and .005 area in the left tail.

**8.42** For each of the following, find the area in the appropriate tail of the  $t$  distribution.

- |                              |                               |
|------------------------------|-------------------------------|
| a. $t = 2.467$ and $df = 28$ | b. $t = -1.672$ and $df = 58$ |
| c. $t = -2.670$ and $n = 55$ | d. $t = 2.819$ and $n = 23$   |

**8.43** For each of the following, find the area in the appropriate tail of the  $t$  distribution.

- |                               |                               |
|-------------------------------|-------------------------------|
| a. $t = -1.302$ and $df = 42$ | b. $t = 2.797$ and $n = 25$   |
| c. $t = 1.397$ and $n = 9$    | d. $t = -2.383$ and $df = 67$ |

**8.44** Find the value of  $t$  from the  $t$  distribution table for each of the following.

- a. Confidence level = 99% and  $df = 13$
- b. Confidence level = 95% and  $n = 36$
- c. Confidence level = 90% and  $df = 16$

**8.45** a. Find the value of  $t$  from the  $t$  distribution table for a sample size of 22 and a confidence level of 95%.

b. Find the value of  $t$  from the  $t$  distribution table for 60 degrees of freedom and a 90% confidence level.

c. Find the value of  $t$  from the  $t$  distribution table for a sample size of 24 and a confidence level of 99%.

**8.46** A sample of 18 observations taken from a normally distributed population produced the following data:

28.4	27.3	25.5	25.5	31.1	23.0	26.3	24.6	28.4
37.2	23.9	28.7	27.9	25.1	27.2	25.3	22.6	22.7

- a. What is the point estimate of  $\mu$ ?
- b. Make a 99% confidence interval for  $\mu$ .
- c. What is the margin of error of estimate for  $\mu$  in part b?

**8.47** A sample of 11 observations taken from a normally distributed population produced the following data:

-7.1    10.3    8.7    -3.6    -6.0    -7.5    5.2    3.7    9.8    -4.4    6.4

- What is the point estimate of  $\mu$ ?
- Make a 95% confidence interval for  $\mu$ .
- What is the margin of error of estimate for  $\mu$  in part b?

**8.48** Suppose, for a sample selected from a normally distributed population,  $\bar{x} = 68.50$  and  $s = 8.9$ .

- Construct a 95% confidence interval for  $\mu$  assuming  $n = 16$ .
- Construct a 90% confidence interval for  $\mu$  assuming  $n = 16$ . Is the width of the 90% confidence interval smaller than the width of the 95% confidence interval calculated in part a? If yes, explain why.
- Find a 95% confidence interval for  $\mu$  assuming  $n = 25$ . Is the width of the 95% confidence interval for  $\mu$  with  $n = 25$  smaller than the width of the 95% confidence interval for  $\mu$  with  $n = 16$  calculated in part a? If so, why? Explain.

**8.49** Suppose, for a sample selected from a population,  $\bar{x} = 25.5$  and  $s = 4.9$ .

- Construct a 95% confidence interval for  $\mu$  assuming  $n = 47$ .
- Construct a 99% confidence interval for  $\mu$  assuming  $n = 47$ . Is the width of the 99% confidence interval larger than the width of the 95% confidence interval calculated in part a? If yes, explain why.
- Find a 95% confidence interval for  $\mu$  assuming  $n = 32$ . Is the width of the 95% confidence interval for  $\mu$  with  $n = 32$  larger than the width of the 95% confidence interval for  $\mu$  with  $n = 47$  calculated in part a? If so, why? Explain.

**8.50** a. A sample of 100 observations taken from a population produced a sample mean equal to 55.32 and a standard deviation equal to 8.4. Make a 90% confidence interval for  $\mu$ .

- Another sample of 100 observations taken from the same population produced a sample mean equal to 57.40 and a standard deviation equal to 7.5. Make a 90% confidence interval for  $\mu$ .
- A third sample of 100 observations taken from the same population produced a sample mean equal to 56.25 and a standard deviation equal to 7.9. Make a 90% confidence interval for  $\mu$ .
- The true population mean for this population is 55.80. Which of the confidence intervals constructed in parts a through c cover this population mean and which do not?

**8.51** a. A sample of 400 observations taken from a population produced a sample mean equal to 92.45 and a standard deviation equal to 12.20. Make a 98% confidence interval for  $\mu$ .

- Another sample of 400 observations taken from the same population produced a sample mean equal to 91.75 and a standard deviation equal to 14.50. Make a 98% confidence interval for  $\mu$ .
- A third sample of 400 observations taken from the same population produced a sample mean equal to 89.63 and a standard deviation equal to 13.40. Make a 98% confidence interval for  $\mu$ .
- The true population mean for this population is 90.65. Which of the confidence intervals constructed in parts a through c cover this population mean and which do not?

## ■ APPLICATIONS

**8.52** A random sample of 16 airline passengers at the Bay City airport showed that the mean time spent waiting in line to check in at the ticket counters was 31 minutes with a standard deviation of 7 minutes. Construct a 99% confidence interval for the mean time spent waiting in line by all passengers at this airport. Assume that such waiting times for all passengers are normally distributed.

**8.53** A random sample of 20 acres gave a mean yield of wheat equal to 41.2 bushels per acre with a standard deviation of 3 bushels. Assuming that the yield of wheat per acre is normally distributed, construct a 90% confidence interval for the population mean  $\mu$ .

**8.54** Almost all employees working for financial companies in New York City receive large bonuses at the end of the year. A sample of 65 employees selected from financial companies in New York City showed that they received an average bonus of \$55,000 last year with a standard deviation of \$18,000. Construct a 95% confidence interval for the average bonus that all employees working for financial companies in New York City received last year.

**8.55** According to the 2010 Time Use Survey conducted by the U.S. Bureau of Labor Statistics, Americans of age 15 years and older spent an average of 164 minutes per day watching TV in 2010 (*USA TODAY*, June 23, 2011). Suppose a recent sample of 25 people of age 15 years and older selected from a city showed that they spend an average of 172 minutes per day watching TV with a standard deviation of

28 minutes. Make a 90% confidence interval for the average time that all people of age 15 years and older in this city spend per day watching TV. Assume that the times spent by all people of age 15 years and older in this city watching TV have a normal distribution.

**8.56** The high price of medicines is a source of major expense for those seniors in the United States who have to pay for these medicines themselves. A random sample of 2000 seniors who pay for their medicines showed that they spent an average of \$4600 last year on medicines with a standard deviation of \$800. Make a 98% confidence interval for the corresponding population mean.

**8.57** Jack's Auto Insurance Company customers sometimes have to wait a long time to speak to a customer service representative when they call regarding disputed claims. A random sample of 25 such calls yielded a mean waiting time of 22 minutes with a standard deviation of 6 minutes. Construct a 99% confidence interval for the population mean of such waiting times. Assume that such waiting times for the population follow a normal distribution.

**8.58** A random sample of 36 mid-sized cars tested for fuel consumption gave a mean of 26.4 miles per gallon with a standard deviation of 2.3 miles per gallon.

- Find a 99% confidence interval for the population mean,  $\mu$ .
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Describe all possible alternatives. Which alternative is the best and why?

**8.59** The mean time taken to design a house plan by 40 architects was found to be 23 hours with a standard deviation of 3.75 hours.

- Construct a 98% confidence interval for the population mean  $\mu$ .
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Describe all possible alternatives. Which alternative is the best and why?

**8.60** The following data give the speeds (in miles per hour), as measured by radar, of 10 cars traveling on Interstate I-15:

76      72      80      68      76      74      71      78      82      65

Assuming that the speeds of all cars traveling on this highway have a normal distribution, construct a 90% confidence interval for the mean speed of all cars traveling on this highway.

**8.61** A company randomly selected nine office employees and secretly monitored their computers for one month. The times (in hours) spent by these employees using their computers for non-job-related activities (playing games, personal communications, etc.) during this month are as follows:

7      12      9      8      11      4      14      1      6

Assuming that such times for all employees are normally distributed, make a 95% confidence interval for the corresponding population mean for all employees of this company.

**8.62** A dentist wants to find the average time taken by one of her hygienists to take X-rays and clean teeth for patients. She recorded the time to serve 24 randomly selected patients by this hygienist. The data (in minutes) are as follows:

36.80    39.80    38.60    38.30    34.30    32.60    38.70    34.50    37.00    36.80    40.90    33.80  
37.10    33.00    35.10    38.20    36.60    38.80    39.60    39.70    35.10    38.20    32.70    39.50

Assume that such times for this hygienist for all patients are approximately normal.

- What is the point estimate of the corresponding population mean.
- Construct a 99% confidence interval for the average time taken by this hygienist to take X-rays and to clean teeth for all patients.

**8.63** A businesswoman is considering whether to open a coffee shop in a local shopping center. Before making this decision, she wants to know how much money people spend per week at coffee shops in that area. She took a random sample of 26 customers from the area who visit coffee shops and asked them to record the amount of money (in dollars) they would spend during the next week at coffee shops. At the end of the week, she obtained the following data (in dollars) from these 26 customers:

16.96	38.83	15.28	14.84	5.99	64.50	12.15	14.68	33.37
37.10	18.15	67.89	12.17	40.13	5.51	8.80	34.53	35.54
8.51	37.18	41.52	13.83	12.96	22.78	5.29	9.09	

Assume that the distribution of weekly expenditures at coffee shops by all customers who visit coffee shops in this area is approximately normal.

- What is the point estimate of the corresponding population mean?
- Make a 95% confidence interval for the average amount of money spent per week at coffee shops by all customers who visit coffee shops in this area.

- 8.64** A random sample of 34 participants in a Zumba dance class had their heart rates measured before and after a moderate 10-minute workout. The following data correspond to the increase in each individual's heart rate (in beats per minute):

59	70	57	42	57	59	41	54	44	36	59	61
52	42	41	32	60	54	52	53	51	47	62	62
44	69	50	37	50	54	48	52	61	45		

- What is the point estimate of the corresponding population mean?
- Make a 98% confidence interval for the average increase in a person's heart rate after a moderate 10-minute Zumba workout.

- 8.65** The following data give the number of pitches thrown by both teams in each of a random sample of 24 Major League Baseball games played between the beginning of the 2012 season and May 16, 2012.

234	281	264	251	284	266	337	291
309	245	331	284	239	282	226	286
361	278	317	306	325	256	295	276

- Create a histogram of these data using the class intervals 210 to less than 230, 230 to less than 250, 250 to less than 270, and so on. Based on the histogram, does it seem reasonable to assume that these data are approximately normally distributed?
- Calculate the value of the point estimate of the corresponding population mean.
- Assuming that the distribution of total number of pitches thrown by both teams in Major League Baseball games is approximately normal, construct a 99% confidence interval for the average number of pitches thrown by both teams in a Major League Baseball game.

- \*8.66** You are working for a supermarket. The manager has asked you to estimate the mean time taken by a cashier to serve customers at this supermarket. Briefly explain how you will conduct this study. Collect data on the time taken by any supermarket cashier to serve 40 customers. Then estimate the population mean. Choose your own confidence level.

- \*8.67** You are working for a bank. The bank manager wants to know the mean waiting time for all customers who visit this bank. She has asked you to estimate this mean by taking a sample. Briefly explain how you will conduct this study. Collect data on the waiting times for 45 customers who visit a bank. Then estimate the population mean. Choose your own confidence level.

## 8.4 Estimation of a Population Proportion: Large Samples

Often we want to estimate the population proportion or percentage. (Recall that a percentage is obtained by multiplying the proportion by 100.) For example, the production manager of a company may want to estimate the proportion of defective items produced on a machine. A bank manager may want to find the percentage of customers who are satisfied with the service provided by the bank.

Again, if we can conduct a census each time we want to find the value of a population proportion, there is no need to learn the procedures discussed in this section. However, we usually derive our results from sample surveys. Hence, to take into account the variability in the results obtained from different sample surveys, we need to know the procedures for estimating a population proportion.

Recall from Chapter 7 that the population proportion is denoted by  $p$ , and the sample proportion is denoted by  $\hat{p}$ . This section explains how to estimate the population proportion,  $p$ , using the sample proportion,  $\hat{p}$ . The sample proportion,  $\hat{p}$ , is a sample statistic, and it possesses a sampling distribution. From Chapter 7, we know that for large samples:

- The sampling distribution of the sample proportion,  $\hat{p}$ , is (approximately) normal.
- The mean,  $\mu_{\hat{p}}$ , of the sampling distribution of  $\hat{p}$  is equal to the population proportion,  $p$ .
- The standard deviation,  $\sigma_{\hat{p}}$ , of the sampling distribution of the sample proportion,  $\hat{p}$ , is  $\sqrt{pq/n}$ , where  $q = 1 - p$ .

In the case of a proportion, a sample is considered to be large if  $np$  and  $nq$  are both greater than 5. If  $p$  and  $q$  are not known, then  $n\hat{p}$  and  $n\hat{q}$  should each be greater than 5 for the sample to be large.

◀ Remember

When estimating the value of a population proportion, we do not know the values of  $p$  and  $q$ . Consequently, we cannot compute  $\sigma_{\hat{p}}$ . Therefore, in the estimation of a population proportion, we use the value of  $s_{\hat{p}}$  as an estimate of  $\sigma_{\hat{p}}$ . The value of  $s_{\hat{p}}$  is calculated using the following formula.

**Estimator of the Standard Deviation of  $\hat{p}$**  The value of  $s_{\hat{p}}$ , which gives a point estimate of  $\sigma_{\hat{p}}$ , is calculated as follows. Here,  $s_{\hat{p}}$  is an estimator of  $\sigma_{\hat{p}}$ .

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The sample proportion,  $\hat{p}$ , is the point estimator of the corresponding population proportion,  $p$ . Then to find the confidence interval for  $p$ , we add to and subtract from  $\hat{p}$  a number that is called the **margin of error**,  $E$ .

**Confidence Interval for the Population Proportion,  $p$**  The  $(1 - \alpha)100\%$  confidence interval for the population proportion,  $p$ , is

$$\hat{p} \pm z s_{\hat{p}}$$

The value of  $z$  used here is obtained from the standard normal distribution table for the given confidence level, and  $s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$ . The term  $z s_{\hat{p}}$  is called the *margin of error*,  $E$ .

Examples 8–7 and 8–8 illustrate the procedure for constructing a confidence interval for  $p$ .

### ■ EXAMPLE 8–7

Finding the point estimate and 99% confidence interval for  $p$ : large sample.

According to a *New York Times/CBS News* poll conducted during June 24–28, 2011, 55% of American adults polled said that owning a home is a *very important part of the American Dream* (*The New York Times*, June 30, 2011). This poll was based on a sample of 979 American adults.

- (a) What is the point estimate of the corresponding population proportion?
- (b) Find, with a 99% confidence level, the percentage of all American adults who will say that owning a home is a *very important part of the American Dream*. What is the margin of error of this estimate?

**Solution** Let  $p$  be the proportion of all American adults who will say that owning a home is a *very important part of the American Dream*, and let  $\hat{p}$  be the corresponding sample proportion. From the given information,

$$n = 979, \quad \hat{p} = .55, \quad \text{and} \quad \hat{q} = 1 - \hat{p} = 1 - .55 = .45$$

First, we calculate the value of the standard deviation of the sample proportion as follows:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(.55)(.45)}{979}} = .01589997$$

Note that  $n\hat{p}$  and  $n\hat{q}$  are both greater than 5. (The reader should check this condition.)

Consequently, the sampling distribution of  $\hat{p}$  is approximately normal, and we will use the normal distribution to make a confidence interval about  $p$ .

- (a) The point estimate of the proportion of all American adults who will say that owning a home is a *very important part of the American Dream* is equal to .55; that is,

$$\text{Point estimate of } p = \hat{p} = .55$$

- (b) The confidence level is 99%, or .99. To find  $z$  for a 99% confidence level, first we find the area in each of the two tails of the normal distribution curve, which is  $(1 - .99)/2 = .0050$ . Then, we look for .0050 and  $.0050 + .99 = .9950$  areas in the normal distribution table to find the two values of  $z$ . These two  $z$  values are (approximately)  $-2.58$  and  $2.58$ . Thus, we will use  $z = 2.58$  in the confidence interval formula. Substituting all the values in the confidence interval formula for  $p$ , we obtain

$$\begin{aligned}\hat{p} \pm z s_{\hat{p}} &= .55 \pm 2.58(.01589997) = .55 \pm .041 \\ &= \mathbf{.509 \text{ to } .591 \text{ or } 50.9\% \text{ to } 59.1\%}\end{aligned}$$

Thus, we can state with 99% confidence that .509 to .591, or 50.9% to 59.1%, of all American adults will say that owning a home is a *very* important part of the American Dream.

The margin of error associated with this estimate of  $p$  is .041 or 4.1%, that is,

$$\text{Margin of error} = z s_{\hat{p}} = \pm .041 \text{ or } \pm 4.1\% \quad \blacksquare$$

### ■ EXAMPLE 8–8

According to a Pew Research Center nationwide telephone survey of adults conducted March 15 to April 24, 2011, 86% of college graduates said that college education was a good investment (*Time*, May 30, 2011). Suppose that this survey included 1450 college graduates. Construct a 97% confidence interval for the corresponding population proportion.

*Constructing a 97% confidence interval for  $p$ : large sample.*

**Solution** Let  $p$  be the proportion of all college graduates who would say that college education is a good investment, and let  $\hat{p}$  be the corresponding sample proportion. From the given information,

$$n = 1450, \quad \hat{p} = .86, \quad \hat{q} = 1 - \hat{p} = 1 - .86 = .14; \quad \text{Confidence level} = 97\%$$

The standard deviation of the sample proportion is

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(.86)(.14)}{1450}} = .00911233$$

Note that if we check  $n\hat{p}$  and  $n\hat{q}$ , both are greater than 5. Consequently, we can use the normal distribution to make a confidence interval for  $p$ .

From the normal distribution table, the value of  $z$  for a 97% confidence level is 2.17. Note that to find this  $z$  value, you will look for the areas .0150 and .9850 in Table IV. Substituting all the values in the formula, we find that the 97% confidence interval for  $p$  is

$$\begin{aligned}\hat{p} \pm z s_{\hat{p}} &= .86 \pm 2.17(.00911233) = .86 \pm .02 \\ &= \mathbf{.84 \text{ to } .88, \text{ or } 84\% \text{ to } 88\%}\end{aligned}$$

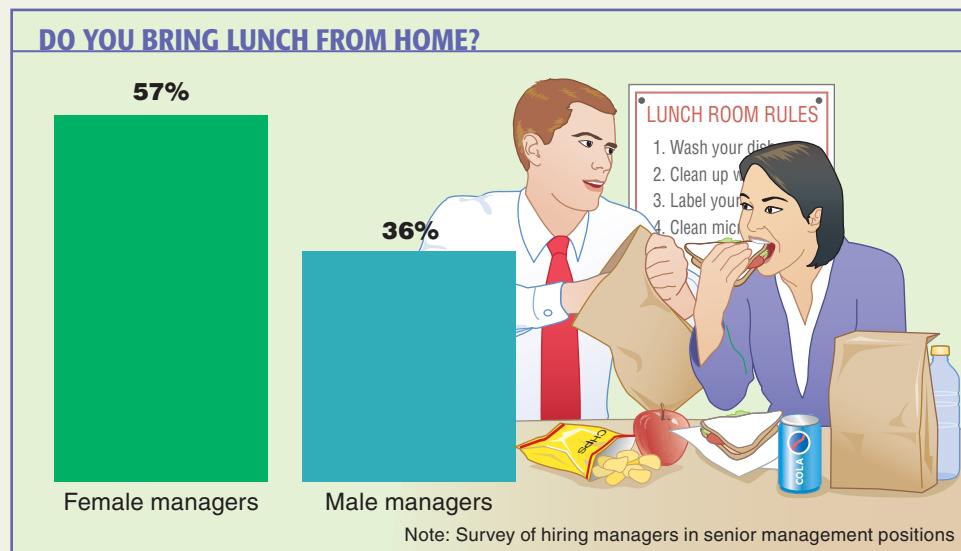
Thus, we can state with 97% confidence that the proportion of all college graduates who would say that college education is a good investment is between .84 and .88. This confidence interval can be converted into a percentage interval as 84% to 88%. ■

Again, we can decrease the width of a confidence interval for  $p$  either by lowering the confidence level or by increasing the sample size. However, lowering the confidence level is not a good choice because it simply decreases the likelihood that the confidence interval contains  $p$ . Hence, to decrease the width of a confidence interval for  $p$ , we should always increase the sample size.

#### 8.4.1 Determining the Sample Size for the Estimation of Proportion

Just as we did with the mean, we can also determine the sample size for estimating the population proportion,  $p$ . This sample size will yield an error of estimate that may not be larger than a predetermined margin of error. By knowing the sample size that can give us the required

## DO YOU BRING YOUR LUNCH FROM HOME?



Data source: Harris Interactive survey on behalf of CareerBuilder conducted between August 16 and September 8, 2011.

The above chart shows the percentage of male and female managers who bring their lunches to work from home. These percentages are based on a survey of 561 hiring managers who held senior management positions and were employed full-time, were not self-employed, and held non-government jobs. The survey was conducted by Harris Interactive on behalf of CareerBuilder between August 16 and September 8, 2011. According to the survey, 41% of the managers included in the survey said that they bring their lunch from home. As we can observe from the graph, when classified by gender, 36% of the male managers and 57% of the female managers in the survey said that they bring lunch from home. Using the procedure learned in this section, we can make a confidence interval for each of the two population proportions as shown in the table below.

Category	Sample Proportion	Confidence Interval
Men who bring lunch from home	.36	.36 ± $z s_{\hat{p}}$
Women who bring lunch from home	.57	.57 ± $z s_{\hat{p}}$

For each of the two confidence intervals listed in the table, we can substitute the value of  $z$  and the value of  $s_{\hat{p}}$ , which is calculated as  $\sqrt{(\hat{p}\hat{q})/n}$ . For example, suppose we want to find a 96% confidence interval for the proportion of all female managers who bring lunch from home. The calculations show that of the 561 managers surveyed, there were 133 female managers. Then this confidence interval is determined as follows:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.57)(0.43)}{133}} = .04292851$$

$$\hat{p} \pm z s_{\hat{p}} = .57 \pm 2.05(.04292851) = .57 \pm .088 = .482 \text{ to } .658$$

Thus, we can state with a 96% confidence that 48.2% to 65.8% of all female managers bring lunch to work from home.

We can find the confidence interval for the population proportion of all male managers who bring lunch from home in the same way.

*Source:* <http://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?id=pr669&sd=11/16/2011&ed=11/16/2011>.

results, we can save our scarce resources by not taking an unnecessarily large sample. From Section 8.4, the margin of error,  $E$ , of the interval estimation of the population proportion is

$$E = z s_{\hat{p}} = z \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

By manipulating this expression algebraically, we obtain the following formula to find the required sample size given  $E$ ,  $\hat{p}$ ,  $\hat{q}$ , and  $z$ .

**Determining the Sample Size for the Estimation of  $p$**  Given the confidence level and the values of  $\hat{p}$  and  $\hat{q}$ , the sample size that will produce a predetermined margin of error  $E$  of the confidence interval *estimate of  $p$*  is

$$n = \frac{z^2 \hat{p} \hat{q}}{E^2}$$

We can observe from this formula that to find  $n$ , we need to know the values of  $\hat{p}$  and  $\hat{q}$ . However, the values of  $\hat{p}$  and  $\hat{q}$  are not known to us. In such a situation, we can choose one of the following alternatives.

1. We make the *most conservative estimate* of the sample size  $n$  by using  $\hat{p} = .50$  and  $\hat{q} = .50$ . For a given  $E$ , these values of  $\hat{p}$  and  $\hat{q}$  will give us the largest sample size in comparison to any other pair of values of  $\hat{p}$  and  $\hat{q}$  because the product of  $\hat{p} = .50$  and  $\hat{q} = .50$  is greater than the product of any other pair of values for  $\hat{p}$  and  $\hat{q}$ .
2. We take a *preliminary sample* (of arbitrarily determined size) and calculate  $\hat{p}$  and  $\hat{q}$  for this sample. Then, we use these values of  $\hat{p}$  and  $\hat{q}$  to find  $n$ .

Examples 8–9 and 8–10 illustrate how to determine the sample size that will produce the error of estimation for the population proportion within a predetermined margin of error value. Example 8–9 gives the most conservative estimate of  $n$ , and Example 8–10 uses the results from a preliminary sample to determine the required sample size.

## ■ EXAMPLE 8–9

Lombard Electronics Company has just installed a new machine that makes a part that is used in clocks. The company wants to estimate the proportion of these parts produced by this machine that are defective. The company manager wants this estimate to be within .02 of the population proportion for a 95% confidence level. What is the most conservative estimate of the sample size that will limit the margin of error to within .02 of the population proportion?

*Determining the most conservative estimate of  $n$  for the estimation of  $p$ .*

**Solution** The company manager wants the 95% confidence interval to be

$$\hat{p} \pm .02$$

Therefore,

$$E = .02$$

The value of  $z$  for a 95% confidence level is 1.96. For the most conservative estimate of the sample size, we will use  $\hat{p} = .50$  and  $\hat{q} = .50$ . Hence, the required sample size is

$$n = \frac{z^2 \hat{p} \hat{q}}{E^2} = \frac{(1.96)^2 (.50)(.50)}{(.02)^2} = 2401$$

Thus, if the company takes a sample of 2401 parts, there is a 95% chance that the estimate of  $p$  will be within .02 of the population proportion. ■

## ■ EXAMPLE 8–10

Consider Example 8–9 again. Suppose a preliminary sample of 200 parts produced by this machine showed that 7% of them are defective. How large a sample should the company select so that the 95% confidence interval for  $p$  is within .02 of the population proportion?

*Determining  $n$  for the estimation of  $p$  using preliminary sample results.*

**Solution** Again, the company wants the 95% confidence interval for  $p$  to be

$$\hat{p} \pm .02$$

Hence,

$$E = .02$$

The value of  $z$  for a 95% confidence level is 1.96. From the preliminary sample,

$$\hat{p} = .07 \quad \text{and} \quad \hat{q} = 1 - .07 = .93$$

Using these values of  $\hat{p}$  and  $\hat{q}$ , we obtain

$$n = \frac{z^2 \hat{p} \hat{q}}{E^2} = \frac{(1.96)^2(0.07)(0.93)}{(0.02)^2} = \frac{(3.8416)(0.07)(0.93)}{0.0004} = 625.22 \approx \mathbf{626}$$

Note that if the value of  $n$  is not an integer, we always round it up here.

Thus, if the company takes a sample of 626 items, there is a 95% chance that the estimate of  $p$  will be within .02 of the population proportion. However, we should note that this sample size will produce the margin of error within .02 only if  $\hat{p}$  is .07 or less for the new sample. If  $\hat{p}$  for the new sample happens to be much higher than .07, the margin of error will not be within .02. Therefore, to avoid such a situation, we may be more conservative and take a much larger sample than 626 items. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**8.68** What assumption(s) must hold true to use the normal distribution to make a confidence interval for the population proportion,  $p$ ?

**8.69** What is the point estimator of the population proportion,  $p$ ?

**8.70** Check if the sample size is large enough to use the normal distribution to make a confidence interval for  $p$  for each of the following cases.

- a.  $n = 50$  and  $\hat{p} = .25$
- b.  $n = 160$  and  $\hat{p} = .03$
- c.  $n = 400$  and  $\hat{p} = .65$
- d.  $n = 75$  and  $\hat{p} = .06$

**8.71** Check if the sample size is large enough to use the normal distribution to make a confidence interval for  $p$  for each of the following cases.

- a.  $n = 80$  and  $\hat{p} = .85$
- b.  $n = 110$  and  $\hat{p} = .98$
- c.  $n = 35$  and  $\hat{p} = .40$
- d.  $n = 200$  and  $\hat{p} = .08$

**8.72** a. A sample of 300 observations taken from a population produced a sample proportion of .63. Make a 95% confidence interval for  $p$ .

b. Another sample of 300 observations taken from the same population produced a sample proportion of .59. Make a 95% confidence interval for  $p$ .

c. A third sample of 300 observations taken from the same population produced a sample proportion of .67. Make a 95% confidence interval for  $p$ .

d. The true population proportion for this population is .65. Which of the confidence intervals constructed in parts a through c cover this population proportion and which do not?

**8.73** a. A sample of 1100 observations taken from a population produced a sample proportion of .32. Make a 90% confidence interval for  $p$ .

b. Another sample of 1100 observations taken from the same population produced a sample proportion of .36. Make a 90% confidence interval for  $p$ .

c. A third sample of 1100 observations taken from the same population produced a sample proportion of .30. Make a 90% confidence interval for  $p$ .

d. The true population proportion for this population is .34. Which of the confidence intervals constructed in parts a through c cover this population proportion and which do not?

**8.74** A sample of 200 observations selected from a population produced a sample proportion equal to .91.

- a. Make a 90% confidence interval for  $p$ .
- b. Construct a 95% confidence interval for  $p$ .

- c. Make a 99% confidence interval for  $p$ .
  - d. Does the width of the confidence intervals constructed in parts a through c increase as the confidence level increases? If yes, explain why.
- 8.75** A sample of 200 observations selected from a population gave a sample proportion equal to .27.
- a. Make a 99% confidence interval for  $p$ .
  - b. Construct a 97% confidence interval for  $p$ .
  - c. Make a 90% confidence interval for  $p$ .
  - d. Does the width of the confidence intervals constructed in parts a through c decrease as the confidence level decreases? If yes, explain why.
- 8.76** A sample selected from a population gave a sample proportion equal to .73.
- a. Make a 99% confidence interval for  $p$  assuming  $n = 100$ .
  - b. Construct a 99% confidence interval for  $p$  assuming  $n = 600$ .
  - c. Make a 99% confidence interval for  $p$  assuming  $n = 1500$ .
  - d. Does the width of the confidence intervals constructed in parts a through c decrease as the sample size increases? If yes, explain why.
- 8.77** A sample selected from a population gave a sample proportion equal to .31.
- a. Make a 95% confidence interval for  $p$  assuming  $n = 1200$ .
  - b. Construct a 95% confidence interval for  $p$  assuming  $n = 500$ .
  - c. Make a 95% confidence interval for  $p$  assuming  $n = 80$ .
  - d. Does the width of the confidence intervals constructed in parts a through c increase as the sample size decreases? If yes, explain why.
- 8.78** a. How large a sample should be selected so that the margin of error of estimate for a 99% confidence interval for  $p$  is .035 when the value of the sample proportion obtained from a preliminary sample is .29?
- b. Find the most conservative sample size that will produce the margin of error for a 99% confidence interval for  $p$  equal to .035.
- 8.79** a. How large a sample should be selected so that the margin of error of estimate for a 98% confidence interval for  $p$  is .045 when the value of the sample proportion obtained from a preliminary sample is .53?
- b. Find the most conservative sample size that will produce the margin of error for a 98% confidence interval for  $p$  equal to .045.
- 8.80** Determine the most conservative sample size for the estimation of the population proportion for the following.
- a.  $E = .025$ , confidence level = 95%
  - b.  $E = .05$ , confidence level = 90%
  - c.  $E = .015$ , confidence level = 99%
- 8.81** Determine the sample size for the estimation of the population proportion for the following, where  $\hat{p}$  is the sample proportion based on a preliminary sample.
- a.  $E = .025$ ,  $\hat{p} = .16$ , confidence level = 99%
  - b.  $E = .05$ ,  $\hat{p} = .85$ , confidence level = 95%
  - c.  $E = .015$ ,  $\hat{p} = .97$ , confidence level = 90%

## ■ APPLICATIONS

- 8.82** According to a Pew Research Center nationwide telephone survey of adults conducted March 15 to April 24, 2011, 55% of college graduates said that their college education prepared them for a job (*Time*, May 30, 2011). Suppose that this survey included 1450 college graduates.
- a. What is the point estimate of the corresponding population proportion?
  - b. Construct a 98% confidence interval for the proportion of all college graduates who will say that their college education prepared them for a job. What is the margin of error for this estimate?
- 8.83** The express check-out lanes at Wally's Supermarket are limited to customers purchasing 12 or fewer items. Cashiers at this supermarket have complained that many customers who use the express lanes have more than 12 items. A recently taken random sample of 200 customers entering express lanes at this supermarket found that 74 of them had more than 12 items.
- a. Construct a 98% confidence interval for the percentage of all customers at this supermarket who enter express lanes with more than 12 items.
  - b. Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which alternative is the best?

**8.84** According to a Pew Research Center nationwide telephone survey of adults conducted March 15 to April 24, 2011, 69% of college graduates said that their college education gave them maturity (*Time*, May 30, 2011). Suppose that this survey included 1450 college graduates.

- What is the point estimate of the corresponding population proportion?
- Construct a 95% confidence interval for the proportion of all college graduates who will say that their college education gave them maturity. What is the margin of error for this estimate?

**8.85** It is said that happy and healthy workers are efficient and productive. A company that manufactures exercising machines wanted to know the percentage of large companies that provide on-site health club facilities. A sample of 240 such companies showed that 96 of them provide such facilities on site.

- What is the point estimate of the percentage of all such companies that provide such facilities on site?
- Construct a 97% confidence interval for the percentage of all such companies that provide such facilities on site. What is the margin of error for this estimate?

**8.86** A mail-order company promises its customers that the products ordered will be mailed within 72 hours after an order is placed. The quality control department at the company checks from time to time to see if this promise is fulfilled. Recently the quality control department took a sample of 50 orders and found that 35 of them were mailed within 72 hours of the placement of the orders.

- Construct a 98% confidence interval for the percentage of all orders that are mailed within 72 hours of their placement.
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which alternative is the best?

**8.87** In a random sample of 50 homeowners selected from a large suburban area, 19 said that they had serious problems with excessive noise from their neighbors.

- Make a 99% confidence interval for the percentage of all homeowners in this suburban area who have such problems.
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which option is best?

**8.88** An Accountemps survey asked workers to identify what behavior of coworkers irritates them the most. Forty-one percent of the workers surveyed said that *sloppy work* is the most irritating behavior. Suppose that this percentage is based on a random sample of 500 workers.

- Construct a 95% confidence interval for the proportion of all workers who will say that *sloppy work* is the most irritating behavior of their coworkers.
- Suppose the confidence interval obtained in part a is too wide. How can the width of this interval be reduced? Discuss all possible alternatives. Which alternative is the best?

**8.89** In a *Time/Money Magazine* poll of Americans of age 18 years and older, 65% agreed with the statement, “We are less sure our children will achieve the American Dream” (*Time*, October 10, 2011). Assume that this poll was based on a random sample of 1600 Americans.

- Construct a 95% confidence interval for the proportion of all Americans of age 18 years and older who will agree with the aforementioned statement.
- Explain why we need to construct a confidence interval. Why can we not simply say that 65% of all Americans of age 18 years and older agree with the aforementioned statement?

**8.90** A researcher wanted to know the percentage of judges who are in favor of the death penalty. He took a random sample of 15 judges and asked them whether or not they favor the death penalty. The responses of these judges are given here.

Yes	No	Yes	Yes	No	No	No	Yes
Yes	No	Yes	Yes	Yes	No	Yes	

- What is the point estimate of the population proportion?
- Make a 95% confidence interval for the percentage of all judges who are in favor of the death penalty.

**8.91** The management of a health insurance company wants to know the percentage of its policyholders who have tried alternative treatments (such as acupuncture, herbal therapy, etc.). A random sample of 24 of the company’s policyholders were asked whether or not they have ever tried such treatments. The following are their responses.

Yes	No	No	Yes	No	Yes	No	No
No	Yes	No	No	Yes	No	Yes	No
No	No	Yes	No	No	No	Yes	No

- What is the point estimate of the corresponding population proportion?
- Construct a 99% confidence interval for the percentage of this company’s policyholders who have tried alternative treatments.

**8.92** Tony's Pizza guarantees all pizza deliveries within 30 minutes of the placement of orders. An agency wants to estimate the proportion of all pizzas that are delivered within 30 minutes by Tony's. What is the most conservative estimate of the sample size that would limit the margin of error to within .02 of the population proportion for a 99% confidence interval?

**8.93** Refer to Exercise 8.92. Assume that a preliminary study has shown that 93% of all Tony's pizzas are delivered within 30 minutes. How large should the sample size be so that the 99% confidence interval for the population proportion has a margin of error of .02?

**8.94** A consumer agency wants to estimate the proportion of all drivers who wear seat belts while driving. Assume that a preliminary study has shown that 76% of drivers wear seat belts while driving. How large should the sample size be so that the 99% confidence interval for the population proportion has a margin of error of .03?

**8.95** Refer to Exercise 8.94. What is the most conservative estimate of the sample size that would limit the margin of error to within .03 of the population proportion for a 99% confidence interval?

**\*8.96** You want to estimate the proportion of students at your college who hold off-campus (part-time or full-time) jobs. Briefly explain how you will make such an estimate. Collect data from 40 students at your college on whether or not they hold off-campus jobs. Then calculate the proportion of students in this sample who hold off-campus jobs. Using this information, estimate the population proportion. Select your own confidence level.

**\*8.97** You want to estimate the percentage of students at your college or university who are satisfied with the campus food services. Briefly explain how you will make such an estimate. Select a sample of 30 students and ask them whether or not they are satisfied with the campus food services. Then calculate the percentage of students in the sample who are satisfied. Using this information, find the confidence interval for the corresponding population percentage. Select your own confidence level.

## USES AND MISUSES... NATIONAL VERSUS LOCAL UNEMPLOYMENT RATE

Reading a newspaper article, you learn that the national unemployment rate is 8.1%. The next month you read another article that states that a recent survey in your area, based on a random sample of the labor force, estimates that the local unemployment rate is 7.7% with a margin of error of .5%. Thus, you conclude that the unemployment rate in your area is somewhere between 7.2% and 8.2%.

So, what does this say about the local unemployment picture in your area versus the national unemployment situation? Since a major portion of the interval for the local unemployment rate is below 8.1%, is it reasonable to conclude that the local unemployment rate is below the national unemployment rate? Not really. When looking at the confidence interval, you have some degree of confidence, usually between 90% and 99%. If we use  $z = 1.96$  to calculate the margin of error, which is the  $z$  value for a 95% confidence level, we

can state that we are 95% confident that the local unemployment rate falls in the interval we obtain by using the margin of error. However, since 8.1% is in the interval for the local unemployment rate, the one thing that you can say is that it appears reasonable to conclude that the local and national unemployment rates are not different. However, if the national rate was 8.3%, then a conclusion that the two rates differ is reasonable because we are confident that the local unemployment rate falls between 7.2% and 8.2%.

When making conclusions based on the types of confidence intervals you have learned and will learn in this course, you will only be able to conclude that either there is a difference or there is not a difference. However, the methods you will learn in Chapter 9 will also allow you to determine the validity of a conclusion that states that the local rate is lower (or higher) than the national rate.

## Glossary

**Confidence interval** An interval constructed around the value of a sample statistic to estimate the corresponding population parameter.

**Confidence level** Confidence level, denoted by  $(1 - \alpha)100\%$ , that states how much confidence we have that a confidence interval contains the true population parameter.

**Degrees of freedom ( $df$ )** The number of observations that can be chosen freely. For the estimation of  $\mu$  using the  $t$  distribution, the degrees of freedom are  $n - 1$ .

**Estimate** The value of a sample statistic that is used to find the corresponding population parameter.

**Estimation** A procedure by which a numerical value or values are assigned to a population parameter based on the information collected from a sample.

**Estimator** The sample statistic that is used to estimate a population parameter.

**Interval estimate** An interval constructed around the point estimate that is likely to contain the corresponding population parameter. Each interval estimate has a confidence level.

**Margin of error** The quantity that is subtracted from and added to the value of a sample statistic to obtain a confidence interval for the corresponding population parameter.

**Point estimate** The value of a sample statistic assigned to the corresponding population parameter.

**t distribution** A continuous distribution with a specific type of bell-shaped curve with its mean equal to 0 and standard deviation equal to  $\sqrt{df}/(df - 2)$  for  $df > 2$ .

## Supplementary Exercises



**8.98** Because of inadequate public school budgets and lack of money available to teachers for classroom materials, many teachers often use their own money to buy materials used in the classrooms. A random sample of 100 public school teachers selected from an eastern state showed that they spent an average of \$290 of their own money on such materials during the 2011–2012 school year. The population standard deviation was \$70.

- a. What is the point estimate of the mean of such expenses incurred during the 2011–2012 school year by all public school teachers in this state?
- b. Make a 95% confidence interval for the corresponding population mean.

**8.99** A bank manager wants to know the mean amount owed on credit card accounts that become delinquent. A random sample of 100 delinquent credit card accounts taken by the manager produced a mean amount owed on these accounts equal to \$2640. The population standard deviation was \$578.

- a. What is the point estimate of the mean amount owed on all delinquent credit card accounts at this bank?
- b. Construct a 97% confidence interval for the mean amount owed on all delinquent credit card accounts for this bank.

**8.100** York Steel Corporation produces iron rings that are supplied to other companies. These rings are supposed to have a diameter of 24 inches. The machine that makes these rings does not produce each ring with a diameter of exactly 24 inches. The diameter of each of the rings varies slightly. It is known that when the machine is working properly, the rings made on this machine have a mean diameter of 24 inches. The standard deviation of the diameters of all rings produced on this machine is always equal to .06 inch. The quality control department takes a sample of 25 such rings every week, calculates the mean of the diameters for these rings, and makes a 99% confidence interval for the population mean. If either the lower limit of this confidence interval is less than 23.975 inches or the upper limit of this confidence interval is greater than 24.025 inches, the machine is stopped and adjusted. A recent such sample of 25 rings produced a mean diameter of 24.015 inches. Based on this sample, can you conclude that the machine needs an adjustment? Explain. Assume that the population distribution is normal.

**8.101** Yunan Corporation produces bolts that are supplied to other companies. These bolts are supposed to be 4 inches long. The machine that makes these bolts does not produce each bolt exactly 4 inches long. It is known that when the machine is working properly, the mean length of the bolts made on this machine is 4 inches. The standard deviation of the lengths of all bolts produced on this machine is always equal to .04 inch. The quality control department takes a sample of 20 such bolts every week, calculates the mean length of these bolts, and makes a 98% confidence interval for the population mean. If either the upper limit of this confidence interval is greater than 4.02 inches or the lower limit of this confidence interval is less than 3.98 inches, the machine is stopped and adjusted. A recent such sample of 20 bolts produced a mean length of 3.99 inches. Based on this sample, will you conclude that the machine needs an adjustment? Assume that the population distribution is normal.

**8.102** A hospital administration wants to estimate the mean time spent by patients waiting for treatment at the emergency room. The waiting times (in minutes) recorded for a random sample of 35 such patients are given below. The population standard deviation is not known.

30	7	68	76	47	60	51
64	25	35	29	30	35	62
96	104	58	32	32	102	27
45	11	64	62	72	39	92
84	47	12	33	55	84	36

Construct a 99% confidence interval for the corresponding population mean.

- 8.103** A local gasoline dealership in a small town wants to estimate the average amount of gasoline that people in that town use in a 1-week period. The dealer asked 44 randomly selected customers to keep a diary of their gasoline usage, and this information produced the following data on gas used (in gallons) by these people during a 1-week period. The population standard deviation is not known.

23.1	13.6	25.8	10.0	7.6	18.9	26.6	23.8	12.3	15.8	21.0
26.9	22.9	18.3	23.5	21.6	15.5	23.5	11.8	15.3	11.9	19.2
14.5	9.6	12.1	18.0	20.6	14.2	7.1	13.2	5.3	13.1	10.9
10.5	5.1	5.2	6.5	8.3	10.5	7.4	7.4	5.3	10.6	13.0

Construct a 95% confidence interval for the average weekly gas usage by people in this town.

- 8.104** A random sample of 25 life insurance policyholders showed that the average premium they pay on their life insurance policies is \$685 per year with a standard deviation of \$74. Assuming that the life insurance policy premiums for all life insurance policyholders have a normal distribution, make a 99% confidence interval for the population mean,  $\mu$ .

- 8.105** A drug that provides relief from headaches was tried on 18 randomly selected patients. The experiment showed that the mean time to get relief from headaches for these patients after taking this drug was 24 minutes with a standard deviation of 4.5 minutes. Assuming that the time taken to get relief from a headache after taking this drug is (approximately) normally distributed, determine a 95% confidence interval for the mean relief time for this drug for all patients.

- 8.106** A survey of 500 randomly selected adult men showed that the mean time they spend per week watching sports on television is 9.75 hours with a standard deviation of 2.2 hours. Construct a 90% confidence interval for the population mean,  $\mu$ .

- 8.107** A random sample of 300 female members of health clubs in Los Angeles showed that they spend, on average, 4.5 hours per week doing physical exercise with a standard deviation of .75 hour. Find a 98% confidence interval for the population mean.

- 8.108** A computer company that recently developed a new software product wanted to estimate the mean time taken to learn how to use this software by people who are somewhat familiar with computers. A random sample of 12 such persons was selected. The following data give the times taken (in hours) by these persons to learn how to use this software.

1.75	2.25	2.40	1.90	1.50	2.75
2.15	2.25	1.80	2.20	3.25	2.60

Construct a 95% confidence interval for the population mean. Assume that the times taken by all persons who are somewhat familiar with computers to learn how to use this software are approximately normally distributed.

- 8.109** A company that produces 8-ounce low-fat yogurt cups wanted to estimate the mean number of calories for such cups. A random sample of 10 such cups produced the following numbers of calories.

147	159	153	146	144	148	163	153	143	158
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Construct a 99% confidence interval for the population mean. Assume that the numbers of calories for such cups of yogurt produced by this company have an approximately normal distribution.

- 8.110** An insurance company selected a sample of 50 auto claims filed with it and investigated those claims carefully. The company found that 12% of those claims were fraudulent.

- What is the point estimate of the percentage of all auto claims filed with this company that are fraudulent?
- Make a 99% confidence interval for the percentage of all auto claims filed with this company that are fraudulent.

- 8.111** A casino player has grown suspicious about a specific roulette wheel. Specifically, this player believes that the slots for the numbers 0 and 00, which can lead to larger payoffs, are slightly smaller than the rest of 36 slots, which means that the ball would land in these two slots less often than it would if all of the slots were of the same size. This player watched 430 spins on this roulette wheel, and found that the ball landed in 0 or 00 slot 14 times.

- What is the value of the point estimate of the proportion of all roulette spins on this wheel in which the ball would land in 0 or 00 slot?
- Construct a 95% confidence interval for the proportion of all roulette spins on this wheel in which the ball would land in 0 or 00 slot.

- c. If all of the slots on this wheel are of the same size, the ball should land in 0 or 00 slot 5.26% of the time. Based on the confidence interval you calculated in part b, does the player's suspicion seem reasonable?

**8.112** A sample of 20 managers was taken, and they were asked whether or not they usually take work home. The responses of these managers are given below, where *yes* indicates they usually take work home and *no* means they do not.

Yes	Yes	No	No	No	Yes	No	No	No	No
Yes	Yes	No	Yes	Yes	No	No	No	No	Yes

Make a 99% confidence interval for the percentage of all managers who take work home.

**8.113** Salaried workers at a large corporation receive 2 weeks' paid vacation per year. Sixteen randomly selected workers from this corporation were asked whether or not they would be willing to take a 3% reduction in their annual salaries in return for 2 additional weeks of paid vacation. The following are the responses of these workers.

No	Yes	No	No	Yes	No	No	Yes
Yes	No	No	No	Yes	No	No	No

Construct a 97% confidence interval for the percentage of all salaried workers at this corporation who would accept a 3% pay cut in return for 2 additional weeks of paid vacation.

**8.114** A researcher wants to determine a 99% confidence interval for the mean number of hours that adults spend per week doing community service. How large a sample should the researcher select so that the estimate is within 1.2 hours of the population mean? Assume that the standard deviation for time spent per week doing community service by all adults is 3 hours.

**8.115** An economist wants to find a 90% confidence interval for the mean sale price of houses in a state. How large a sample should she select so that the estimate is within \$3500 of the population mean? Assume that the standard deviation for the sale prices of all houses in this state is \$31,500.

**8.116** A large city with chronic economic problems is considering legalizing casino gambling. The city council wants to estimate the proportion of all adults in the city who favor legalized casino gambling. What is the most conservative estimate of the sample size that would limit the margin of error to be within .05 of the population proportion for a 95% confidence interval?

**8.117** Refer to Exercise 8.116. Assume that a preliminary sample has shown that 63% of the adults in this city favor legalized casino gambling. How large should the sample size be so that the 95% confidence interval for the population proportion has a margin of error of .05?

## Advanced Exercises

**8.118** Let  $\mu$  be the hourly wage (excluding tips) for workers who provide hotel room service in a large city. A random sample of a number (more than 30) of such workers yielded a 95% confidence interval for  $\mu$  of \$8.46 to \$9.86 using the normal distribution with a known population standard deviation.

- a. Find the value of  $\bar{x}$  for this sample.
- b. Find a 99% confidence interval for  $\mu$  based on this sample.

**8.119** In April 2012, N3L Optics conducted a telephone poll of 1080 adult Americans aged 18 years and older. One of the questions asked respondents to identify which outdoor activities and sports they favor for fitness. Respondents could choose more than one activity/sport. Of the respondents, 76% said walking, 35% mentioned hiking, and 27% said team sports ([http://n3loptics.com/news\\_items/57](http://n3loptics.com/news_items/57)). Using these results, find a 98% confidence interval for the population percentage that corresponds to each response. Write a one-page report to present your results to a group of college students who have not taken statistics. Your report should answer questions such as the following: (1) What is a confidence interval? (2) Why is a range of values (interval) more informative than a single percentage (point estimate)? (3) What does 98% confidence mean in this context? (4) What assumptions, if any, are you making when you construct each confidence interval?

**8.120** A group of veterinarians wants to test a new canine vaccine for Lyme disease. (Lyme disease is transmitted by the bite of an infected deer tick.) In an area that has a high incidence of Lyme disease, 100 dogs are randomly selected (with their owners' permission) to receive the vaccine. Over a 12-month period, these dogs are periodically examined by veterinarians for symptoms of Lyme disease. At the end of 12 months, 10 of these 100 dogs are diagnosed with the disease. During the same 12-month period, 18% of the unvaccinated dogs in the area have been found to have Lyme disease. Let  $p$  be the proportion of all potential vaccinated dogs who would contract Lyme disease in this area.

- a. Find a 95% confidence interval for  $p$ .
- b. Does 18% lie within your confidence interval of part a? Does this suggest the vaccine might or might not be effective to some degree?
- c. Write a brief critique of this experiment, pointing out anything that may have distorted the results or conclusions.

**8.121** When one is attempting to determine the required sample size for estimating a population mean, and the information on the population standard deviation is not available, it may be feasible to take a small preliminary sample and use the sample standard deviation to estimate the required sample size,  $n$ . Suppose that we want to estimate  $\mu$ , the mean commuting distance for students at a community college, to a margin of error within 1 mile with a confidence level of 95%. A random sample of 20 students yields a standard deviation of 4.1 miles. Use this value of the sample standard deviation,  $s$ , to estimate the required sample size,  $n$ . Assume that the corresponding population has a normal distribution.

**8.122** A gas station attendant would like to estimate  $p$ , the proportion of all households that own more than two vehicles. To obtain an estimate, the attendant decides to ask the next 200 gasoline customers how many vehicles their households own. To obtain an estimate of  $p$ , the attendant counts the number of customers who say there are more than two vehicles in their households and then divides this number by 200. How would you critique this estimation procedure? Is there anything wrong with this procedure that would result in sampling and/or nonsampling errors? If so, can you suggest a procedure that would reduce this error?

**8.123** A couple considering the purchase of a new home would like to estimate the average number of cars that go past the location per day. The couple guesses that the number of cars passing this location per day has a population standard deviation of 170.

- a. On how many randomly selected days should the number of cars passing the location be observed so that the couple can be 99% certain the estimate will be within 100 cars of the true average?
- b. Suppose the couple finds out that the population standard deviation of the number of cars passing the location per day is not 170 but is actually 272. If they have already taken a sample of the size computed in part a, what confidence does the couple have that their point estimate is within 100 cars of the true average?
- c. If the couple has already taken a sample of the size computed in part a and later finds out that the population standard deviation of the number of cars passing the location per day is actually 130, they can be 99% confident their point estimate is within how many cars of the true average?

**8.124** The U.S. Senate just passed a bill by a vote of 55–45 (with all 100 senators voting). A student who took an elementary statistics course last semester says, “We can use these data to make a confidence interval about  $p$ . We have  $n = 100$  and  $\hat{p} = 55/100 = .55$ .” Hence, according to him, a 95% confidence interval for  $p$  is

$$\hat{p} \pm z\sigma_{\hat{p}} = .55 \pm 1.96 \sqrt{\frac{(.55)(.45)}{100}} = .55 \pm .098 = .452 \text{ to } .648$$

Does this make sense? If not, what is wrong with the student’s reasoning?

**8.125** When calculating a confidence interval for the population mean  $\mu$  with a known population standard deviation  $\sigma$ , describe the effects of the following two changes on the confidence interval: (1) doubling the sample size, (2) quadrupling (multiplying by 4) the sample size. Give two reasons why this relationship does not hold true if you are calculating a confidence interval for the population mean  $\mu$  with an unknown population standard deviation.

**8.126** At the end of Section 8.2, we noted that we always round up when calculating the minimum sample size for a confidence interval for  $\mu$  with a specified margin of error and confidence level. Using the formula for the margin of error, explain why we must always round up in this situation.

**8.127** Calculating a confidence interval for the proportion requires a minimum sample size. Calculate a confidence interval, using any confidence level of 90% or higher, for the population proportion for each of the following.

- a.  $n = 200$  and  $\hat{p} = .01$
- b.  $n = 160$  and  $\hat{p} = .9875$

Explain why these confidence intervals reveal a problem when the conditions for using the normal approximation do not hold.

## Self-Review Test

1. Complete the following sentences using the terms *population parameter* and *sample statistic*.
  - a. Estimation means assigning values to a \_\_\_\_\_ based on the value of a \_\_\_\_\_.
  - b. An estimator is a \_\_\_\_\_ used to estimate a \_\_\_\_\_.
  - c. The value of a \_\_\_\_\_ is called the point estimate of the corresponding \_\_\_\_\_.
2. A 95% confidence interval for  $\mu$  can be interpreted to mean that if we take 100 samples of the same size and construct 100 such confidence intervals for  $\mu$ , then
  - a. 95 of them will not include  $\mu$
  - b. 95 will include  $\mu$
  - c. 95 will include  $\bar{x}$
3. The confidence level is denoted by
  - a.  $(1 - \alpha)100\%$
  - b.  $100\alpha\%$
  - c.  $\alpha$
4. The margin of error of the estimate for  $\mu$  is
  - a.  $z\sigma_{\bar{x}}$  (or  $ts_{\bar{x}}$ )
  - b.  $\sigma/\sqrt{n}$  (or  $s/\sqrt{n}$ )
  - c.  $\sigma_{\bar{x}}$  (or  $s_{\bar{x}}$ )
5. Which of the following assumptions is not required to use the *t* distribution to make a confidence interval for  $\mu$ ?
  - a. Either the population from which the sample is taken is (approximately) normally distributed or  $n \geq 30$ .
  - b. The population standard deviation,  $\sigma$ , is not known.
  - c. The sample size is at least 10.
6. The parameter(s) of the *t* distribution is (are)
  - a.  $n$
  - b. degrees of freedom
  - c.  $\mu$  and degrees of freedom
7. A sample of 36 vacation homes built during the past 2 years in a coastal resort region gave a mean construction cost of \$159,000 with a population standard deviation of \$27,000.
  - a. What is the point estimate of the corresponding population mean?
  - b. Make a 99% confidence interval for the mean construction cost for all vacation homes built in this region during the past 2 years. What is the margin of error here?
8. A sample of 25 malpractice lawsuits filed against doctors showed that the mean compensation awarded to the plaintiffs was \$610,425 with a standard deviation of \$94,820. Find a 95% confidence interval for the mean compensation awarded to plaintiffs of all such lawsuits. Assume that the compensations awarded to plaintiffs of all such lawsuits are normally distributed.
  - a. What is the value of the point estimate of the corresponding population proportion?
  - b. Construct a 99% confidence interval for the proportion of all American adults who will say *domestic issues* in response to the aforementioned question.
9. In a *Time Magazine/Aspen* poll of American adults conducted by the strategic research firm Penn Schoen Berland, these adults were asked, "In your opinion, what is more important for the U.S. to focus on in the next decade?" Eighty-three percent of the adults polled said *domestic issues* (*Time*, July 11, 2011). Suppose that this poll was based on a random sample of 1000 American adults.
  - a. What is the value of the point estimate of the corresponding population proportion?
  - b. Construct a 99% confidence interval for the proportion of all American adults who will say *domestic issues* in response to the aforementioned question.
10. A company that makes toaster ovens has done extensive testing on the accuracy of its temperature-setting mechanism. For a previous toaster model of this company, the standard deviation of the temperatures when the mechanism is set for 350°F is 5.78°. Assume that this is the population standard deviation for a new toaster model that uses the same temperature mechanism. How large a sample must be taken so that the estimate of the mean temperature when the mechanism is set for 350°F is within 1.25° of the population mean temperature? Use a 95% confidence level.
11. A college registrar has received numerous complaints about the online registration procedure at her college, alleging that the system is slow, confusing, and error prone. She wants to estimate the proportion of all students at this college who are dissatisfied with the online registration procedure. What is the most conservative estimate of the sample size that would limit the margin of error to be within .05 of the population proportion for a 90% confidence interval?
12. Refer to Problem 11. Assume that a preliminary study has shown that 70% of the students surveyed at this college are dissatisfied with the current online registration system. How large a sample should be taken in this case so that the margin of error is within .05 of the population proportion for a 90% confidence interval?
13. Dr. Garcia estimated the mean stress score before a statistics test for a random sample of 25 students. She found the mean and standard deviation for this sample to be 7.1 (on a scale of 1 to 10) and 1.2,

respectively. She used a 97% confidence level. However, she thinks that the confidence interval is too wide. How can she reduce the width of the confidence interval? Describe all possible alternatives. Which alternative do you think is best and why?

**\*14.** You want to estimate the mean number of hours that students at your college work per week. Briefly explain how you will conduct this study using a small sample. Take a sample of 12 students from your college who hold a job. Collect data on the number of hours that these students spent working last week. Then estimate the population mean. Choose your own confidence level. What assumptions will you make to estimate this population mean?

**\*15.** You want to estimate the proportion of people who are happy with their current jobs. Briefly explain how you will conduct this study. Take a sample of 35 persons and collect data on whether or not they are happy with their current jobs. Then estimate the population proportion. Choose your own confidence level.

## Mini-Projects

### MINI-PROJECT 8-1

A study conducted by the Oregon Employment Agency and Bureau of Labor Statistics included information on the average annual salaries for a variety of jobs identified by high school students as being careers of interest. The following table contains this information from that study.

Occupation	Average Salary (\$)	Educational Requirement
Accountant	63,018	Bachelor's degree
Programmers	70,644	Bachelor's degree
Applications software designer	88,643	Bachelor's degree
Systems software designer	104,005	Bachelor's degree
Fashion designer	66,544	Bachelor's degree
Firefighter	52,248	Postsecondary training
Forensic scientist	57,887	Bachelor's degree
Photographer	39,534	On-the-job training
Physician or surgeon	180,390	Professional degree
Police officer	59,088	On-the-job training
Preschool teacher	25,017	Associate degree
Elementary school teacher	52,549	Bachelor's degree
High school teacher	52,919	Bachelor's degree
Postsecondary school teacher	77,289	Master's degree
Veterinarian	80,788	Professional degree

*Source:* Oregon Employment Department and Bureau of Labor Statistics.

The study did not mention sample sizes for the average salary calculations, nor were any standard deviations provided. For the purpose of this mini-project, we will assume that the average salaries were calculated using random samples of 50 workers from each occupation and that the distribution of salaries is approximately normal for each occupation. Furthermore, we will assume that the sample standard deviation for each of the listed occupations is equal to 10% of the average salary. For example, the sample standard deviation of the accountants' salaries will be \$6301.80, the sample standard deviation of the programmers' salaries will be \$7064.40, and so on.

Calculate a 95% confidence interval for the population mean that corresponds to each of the sample average salaries listed in the table. Compare the intervals for the various occupations. Based on these intervals, rank these occupations from highest to lowest salary. Identify the occupations that have overlapping confidence intervals. What do you conclude about the average salaries of the occupations that have overlapping confidence intervals? Explain your reasoning.

### ■ MINI-PROJECT 8-2

Consider the data set on the heights of NFL players as given in Data Set III on the Web site for this text.

- Take a random sample of 15 players, and find a 95% confidence interval for  $\mu$ . Assume that the heights of these players are normally distributed.
- Repeat part a for samples of size 31 and 60, respectively.
- Compare the widths of your three confidence intervals.
- Now calculate the mean,  $\mu$ , of the heights of all players. Do all of your confidence intervals contain this  $\mu$ ? If not, which ones do not contain  $\mu$ ?

### ■ MINI-PROJECT 8-3

Here is a project that can involve a social activity and also show you the importance of making sure that the underlying requirements are met prior to calculating a confidence interval. Invite some of your friends over and buy a big bag of Milk Chocolate M&Ms. Take at least 40 random samples of 10 M&Ms each from the bag. Note that taking many random samples will reduce the risk of obtaining some extremely odd results. Before eating the candy, calculate the proportion of brown candies for each sample. Then, using each sample proportion, compute a 95% confidence interval for the proportion of brown candies in all M&Ms. According to the company, the population proportion is .13, that is, 13% of all M&Ms are brown. Determine what percentage of the confidence intervals contains the population proportion .13. Is this percentage close to 95%? What happens if you increase your sample size to 20, and then to 50? If you want, you can use technology to simulate those random samples, which makes the process much faster. Besides, the candy will probably be eaten by the time you get ready to take larger samples.

### ■ MINI-PROJECT 8-4

In recent years, merchants and advertisers have reacted to the increase in the purchasing power, whether direct or indirect, of tweens, teens, and young adults. In Harris Interactive's Youth Trends 2010 survey, a random sample of Americans of age 8 to 24 years old were asked whether they would purchase or impact the purchase of tickets to entertainment or sporting events over the next month ([http://www.harrisinteractive.com/vault/HI\\_TrendsTudes\\_2010\\_v09\\_i02.pdf](http://www.harrisinteractive.com/vault/HI_TrendsTudes_2010_v09_i02.pdf)). The following table shows the results for the three age groups 8 to 12, 13 to 17, and 18 to 24 year:

Age Group	Percentage Answering Yes
8–12	40
13–17	43
18–24	45

As shown in the table, 45% of the 18-to 24-year-olds said that they will purchase or impact the purchase of tickets to entertainment or sporting events over the next month.

- Using the results given in the table, calculate a 95% confidence interval for the proportion of all Americans of age 18 to 24 years who said that they will purchase or impact the purchase of tickets to entertainment or sporting events over the next month. Assume that the survey included 434 Americans of age 18 to 24 years.
- Take a random sample of 50 college students within the specified age group (18 to 24 years) and ask them the same question. Using your results, calculate a 95% confidence interval for the proportion of all Americans of age 18 to 24 years who will say that they will purchase or impact the purchase of tickets to entertainment or sporting events over the next month. If the sample size is not large enough to use the normal approximation, increase the sample size to 75.
- Compare the confidence intervals calculated in parts a and b. Are the results consistent between the two surveys?
- Is there any reason to believe that your results are not representative of the population of interest, which is all Americans of age 18 to 24 years? Explain why.

## DECIDE FOR YOURSELF

### DECIDING ABOUT WHETHER YOU CAN USE THE $t$ DISTRIBUTION

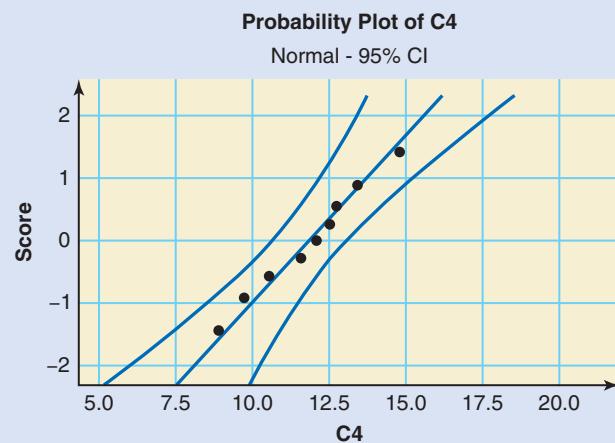
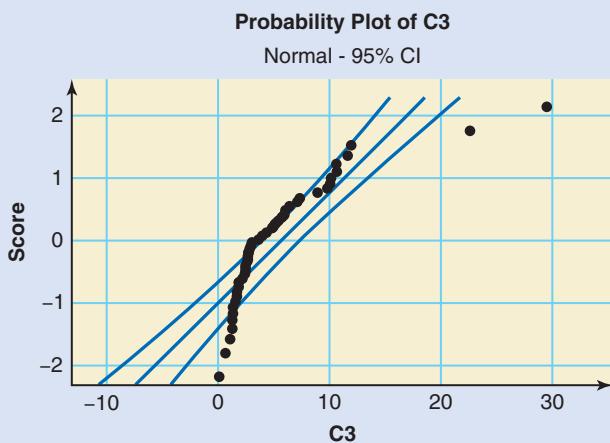
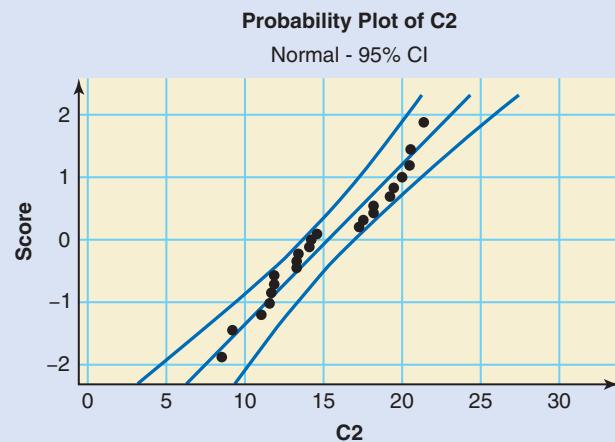
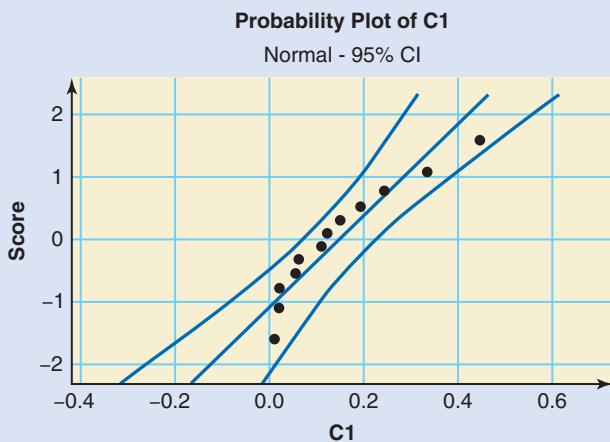
As mentioned in the beginning of Section 8.3, an underlying condition for being able to use the  $t$  distribution to estimate a population mean is that the population from which the sample is selected is normally distributed. Although we are unable to determine with absolute certainty whether or not a data set comes from a normally distributed population, there are methods that allow us to determine whether the normality assumption is reasonable. Although the majority of such methods involve performing a test of hypothesis, which we have not discussed yet, we can use a normal quantile plot, which was discussed in Chapter 6.

Using the  $t$  distribution to make a confidence interval for a population mean is one example of using the  *$t$  distribution procedures* for performing statistical inference. The  $t$  distribution procedures are known to be *robust*, which means that you can still use these procedures when one or more of the underlying assumptions have been violated, as long as certain conditions hold true. In the case of the  $t$  distribution procedures, as the sample size gets larger, the procedures

become insensitive to larger violations of the normality assumption. Two basic rules of thumb are as follows:

- (a) If your data set contains outliers, especially extreme outliers, there are methods other than the  $t$  distribution procedures that should be used. These methods are classified as nonparametric methods, and some of them are discussed in Chapter 15.
- (b) If your sample size is very small, that is,  $n < 10$ , the data set needs to be very close to being normally distributed, which means that the normal quantile plot of the data needs to be very close to being linear. As  $n$  gets larger, the  $t$  distribution procedures can be used even when the data are skewed. As  $n$  becomes 30 or larger, the  $t$  distribution procedures can be used in most cases that do not include outliers.

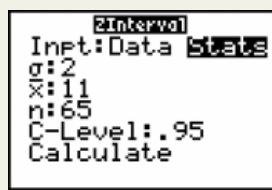
Here are the normal quantile plots for four different data sets. Based on each plot, would it be appropriate to use the  $t$  distribution procedures? Explain why or why not.



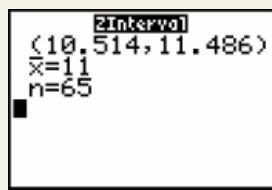
# TECHNOLOGY INSTRUCTION

## Confidence Intervals for Population Means and Proportions

### TI-84



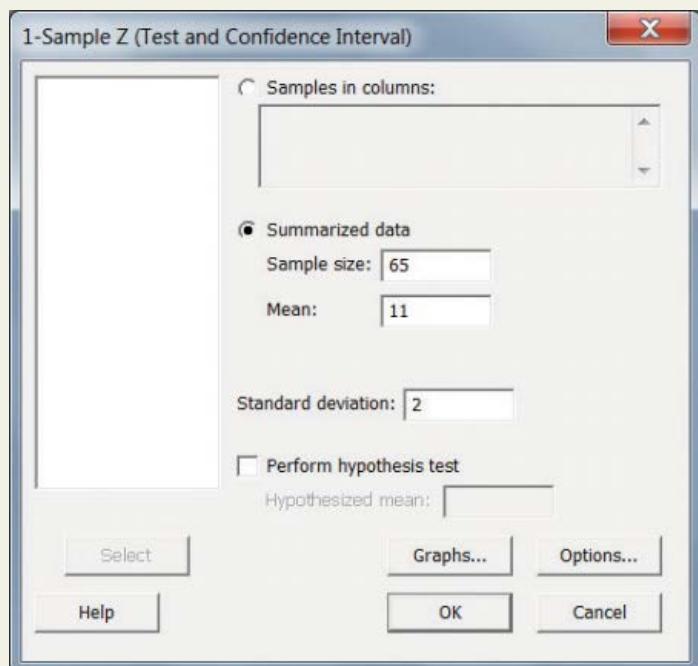
Screen 8.1



Screen 8.2

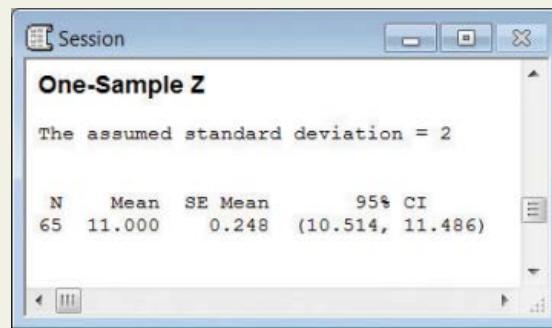
- To find a confidence interval for a population mean  $\mu$  given the population standard deviation  $\sigma$ , select **STAT >TESTS >ZInterval**. If you have the data stored in a list, select **Data** and enter the name of the list. If you have the summary statistics, choose **Stats** and enter the sample mean and size. Enter your value for  $\sigma$  and the confidence level as a decimal as **C-Level**. Select **Calculate**. (See Screens 8.1 and 8.2.)
- To find a confidence interval for a population mean  $\mu$  without knowing the population standard deviation  $\sigma$ , select **STAT >TESTS >TInterval**. If you have the data stored in a list, select **Data** and enter the name of the list. If you have the summary statistics, choose **Stats** and enter the sample mean, standard deviation, and size. Enter your confidence level as a decimal as **C-Level**. Select **Calculate**.
- To find a confidence interval for a population proportion  $p$ , select **STAT >TESTS >1-PropZInt**. Enter the number of successes as  $x$  and the sample size as  $n$ . Enter the confidence level as a decimal as **C-Level**. Select **Calculate**.

### Minitab



Screen 8.3

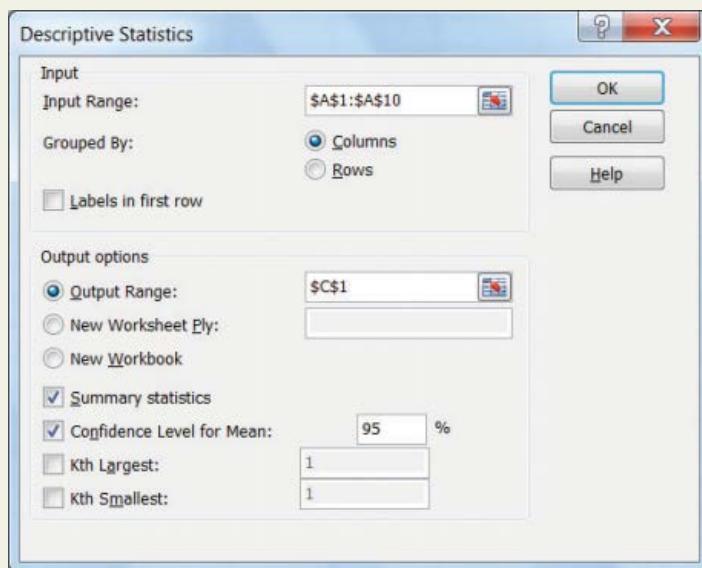
- To find a confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known, select **Stat >Basic Statistics >1-Sample Z**. If you have data on a variable entered in a column of a Minitab spreadsheet, enter the name of that column in the **Samples in columns:** box. If you know the summary statistics, click next to **Summarized data** and enter the values of the **Sample size** and **Mean** in their respective boxes. In both cases, enter the value of the population standard deviation in the **Standard deviation** box. Click the **Options** button and enter the **Confidence level**. Now click **OK** in both windows. The confidence interval will appear in the **Session** window. (See Screens 8.3 and 8.4.)
- To find a confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known, select **Stat >Basic Statistics >1-Sample t**. If you have data on a variable entered in a column of a Minitab spreadsheet, enter the name of that column in the **Samples in columns:** box. If you know the summary statistics, click next to **Summarized data** and enter the values of the **Sample size**, **Mean**, and **Sample standard deviation** in their respective boxes. Click the **Options** button and enter the **Confidence level**. Now click **OK** in both windows. The confidence interval will appear in the **Session** window.
- To find a confidence interval for a population proportion  $p$ , select **Stat >Basic Statistics > 1-Proportion**. If you have sample data (consisting of two values for success and failure) entered in a column, select **Samples in columns** and type your column name in the box. If, instead, you have the number of successes and the number of trials, select **Summarized data** and enter them. Click the **Options** button and enter the **Confidence level**. Click **OK** in both boxes. The confidence interval will appear in the **Session** window.



Screen 8.4

**Excel**

- To calculate the margin of error for a confidence interval for a population mean when the population standard deviation is unknown and the individual data values are available, first use the instructions to obtain the summary statistics (mean and standard deviation) using the Analysis ToolPak presented in the Technology Instruction section of Chapter 3. Then use the following additional step. After filling out all of the relevant information in the **Descriptive Statistics** dialog box, check the box **Confidence Level for Mean** and enter the confidence level as a percentage. Click **OK**. (See Screens 8.5 and 8.6.)



Screen 8.5

C	D
Column1	
Mean	76.7
Standard Error	2.503553
Median	79
Mode	84
Standard Deviation	7.91693
Sample Variance	62.67778
Kurtosis	0.91293
Skewness	-1.25255
Range	24
Minimum	60
Maximum	84
Sum	767
Count	10
Confidence Level(95.0%)	5.66343

Screen 8.6

- To find the margin of error for a confidence interval for a population mean when the population standard deviation  $\sigma$  is known and the sample size  $n$  and the confidence level  $1 - \alpha$  are provided, type  $=\text{CONFIDENCE.NORM}(\alpha, \sigma, n)$ . (See Screens 8.7 and 8.8.) If the population standard deviation  $\sigma$  is unknown, but the sample standard deviation  $s$ , the sample size  $n$ , and the confidence level  $1 - \alpha$  are provided, type  $=\text{CONFIDENCE.T}(\alpha, s, n)$ . (For Excel 2007 and earlier versions, the CONFIDENCE.T function is not available, and the CONFIDENCE.NORM function is CONFIDENCE.)

	A	B	C	D
1	Mean	11		
2	Standard Deviation	2		
3	Size	65		
4	Alpha	0.05		
5				
6	Margin of error	=CONFIDENCE.NORM(0.05,2,65)		
7		CONFIDENCE.NORM(alpha, standard_dev, size)		

Screen 8.7

	A	B
1	Mean	11
2	Standard Deviation	2
3	Size	65
4	Alpha	0.05
5		
6	Margin of error	0.486207225

Screen 8.8

## TECHNOLOGY ASSIGNMENTS

**TA8.1** The following data give the annual incomes (in thousands of dollars) before taxes for a sample of 36 randomly selected families from a city:

21.6	33.0	25.6	37.9	50.0	148.1
50.1	21.5	70.0	72.8	58.2	85.4
91.2	57.0	72.2	45.0	95.0	27.8
92.8	79.4	45.3	76.0	48.6	69.3
40.6	69.0	75.5	57.5	49.7	75.1
96.3	44.5	84.0	43.0	61.7	126.0

Construct a 99% confidence interval for  $\mu$  assuming that the population standard deviation is \$23.75 thousand.

**TA8.2** The following data give the checking account balances (in dollars) on a certain day for a randomly selected sample of 30 households:

500	100	650	1917	2200	500	180	3000	1500	1300
319	1500	1102	405	124	1000	134	2000	150	800
200	750	300	2300	40	1200	500	900	20	160

Construct a 97% confidence interval for  $\mu$  assuming that the population standard deviation is unknown.

**TA8.3** Refer to Data Set I that accompanies this text on the prices of various products in different cities across the country. Using the data on the cost of going to the dentist's office, make a 98% confidence interval for the population mean  $\mu$ .

**TA8.4** Refer to the Beach to Beacon 10K Road Race data set (Data Set IV) for all participants that accompanies this text. Take a sample of 100 observations from this data set.

a. Using the sample data, make a 95% confidence interval for the mean time taken to complete this race by all participants.

b. Now calculate the mean time taken to run this race by all participants. Does the confidence interval made in part a include this population mean?

**TA8.5** Repeat Technology Assignment TA8.4 for a sample of 25 observations. Assume that the distribution of times taken to run this race by all participants is approximately normal.

**TA8.6** The following data give the prices (in thousands of dollars) of 16 recently sold houses in an area.

341	163	327	204	197	203	313	279
456	228	383	289	533	399	271	381

Construct a 99% confidence interval for the mean price of all houses in this area. Assume that the distribution of prices of all houses in the given area is normal.

**TA8.7** A researcher wanted to estimate the mean contributions made to charitable causes by major companies. A random sample of 18 companies produced the following data on contributions (in millions of dollars) made by them.

1.8	.6	1.2	.3	2.6	1.9	3.4	2.6	.2
2.4	1.4	2.5	3.1	.9	1.2	2.0	.8	1.1

Make a 98% confidence interval for the mean contributions made to charitable causes by all major companies. Assume that the contributions made to charitable causes by all major companies have a normal distribution.

**TA8.8** A mail-order company promises its customers that their orders will be processed and mailed within 72 hours after an order is placed. The quality control department at the company checks from time to time to see if this promise is kept. Recently the quality control department took a sample of 200 orders and found that 176 of them were processed and mailed within 72 hours of the placement of the orders. Make a 98% confidence interval for the corresponding population proportion.

**TA8.9** One of the major problems faced by department stores is a high percentage of returns. The manager of a department store wanted to estimate the percentage of all sales that result in returns. A sample of 500 sales showed that 95 of them had products returned within the time allowed for returns. Make a 99% confidence interval for the corresponding population proportion.

**TA8.10** One of the major problems faced by auto insurance companies is the filing of fraudulent claims. An insurance company carefully investigated 1000 auto claims filed with it and found 108 of them to be fraudulent. Make a 96% confidence interval for the corresponding population proportion.

**TA8.11** Create the normal quantile plot for each of the data sets in TA8.1, 8.2, 8.3, 8.6, and 8.7. Assess whether it is reasonable to use the *t* distribution procedure to make a confidence interval for the population mean in each of these five problems. Explain your reasoning. (Note: See the Decide For Yourself section in this chapter for assistance in making the assessment.)



© Mash Audio Visuals Pvt. Ltd. Agency/Stockphoto

## Hypothesis Tests About the Mean and Proportion

### 9.1 Hypothesis Tests: An Introduction

### 9.2 Hypothesis Tests About $\mu : \sigma$ Known

#### Case Study 9–1 Average Student Debt for the Class of 2010

### 9.3 Hypothesis Tests About $\mu : \sigma$ Not Known

### 9.4 Hypothesis Tests About a Population Proportion: Large Samples

#### Case Study 9–2 Is Raising Taxes on the Rich Fair?

Will you graduate from college with debt? If yes, how much money do you think you will owe? Do you know that students who graduated from college in 2010 with loans had an average debt of \$25,250? The average debt of the class of 2010 varied a great deal from state to state, with the highest average debt for students who graduated from colleges in New Hampshire at \$31,048 and the lowest average for students who graduated from colleges in Utah at \$15,509. (See Case Study 9–1.)

This chapter introduces the second topic in inferential statistics: tests of hypotheses. In a test of hypothesis, we test a certain given theory or belief about a population parameter. We may want to find out, using some sample information, whether or not a given claim (or statement) about a population parameter is true. This chapter discusses how to make such tests of hypotheses about the population mean,  $\mu$ , and the population proportion,  $p$ .

As an example, a soft-drink company may claim that, on average, its cans contain 12 ounces of soda. A government agency may want to test whether or not such cans do contain, on average, 12 ounces of soda. As another example, according to a poll conducted by *The New York Times* and CBS News in 2012, 75% of Americans said that Supreme Court justices' decisions are sometimes influenced by their personal or political views. A researcher wants to check if this percentage is still true. In the first of these two examples we are to test a hypothesis about the population mean,  $\mu$ , and in the second example we are to test a hypothesis about the population proportion,  $p$ .

## 9.1 Hypothesis Tests: An Introduction

Why do we need to perform a test of hypothesis? Reconsider the example about soft-drink cans. Suppose we take a sample of 100 cans of the soft drink under investigation. We then find out that the mean amount of soda in these 100 cans is 11.89 ounces. Based on this result, can we state that, on average, all such cans contain less than 12 ounces of soda and that the company is lying to the public? Not until we perform a test of hypothesis can we make such an accusation. The reason is that the mean,  $\bar{x} = 11.89$  ounces, is obtained from a sample. The difference between 12 ounces (the required average amount for the population) and 11.89 ounces (the observed average amount for the sample) may have occurred only because of the sampling error (assuming that no nonsampling errors have been committed). Another sample of 100 cans may give us a mean of 12.04 ounces. Therefore, we perform a test of hypothesis to find out how large the difference between 12 ounces and 11.89 ounces is and whether or not this difference has occurred as a result of chance alone. Now, if 11.89 ounces is the mean for all cans and not for just 100 cans, then we do not need to make a test of hypothesis. Instead, we can immediately state that the mean amount of soda in all such cans is less than 12 ounces. We perform a test of hypothesis only when we are making a decision about a population parameter based on the value of a sample statistic.

### 9.1.1 Two Hypotheses

Consider as a nonstatistical example a person who has been indicted for committing a crime and is being tried in a court. Based on the available evidence, the judge or jury will make one of two possible decisions:

1. The person is not guilty.
2. The person is guilty.

At the outset of the trial, the person is presumed not guilty. The prosecutor's efforts are to prove that the person has committed the crime and, hence, is guilty.

In statistics, *the person is not guilty* is called the **null hypothesis** and *the person is guilty* is called the **alternative hypothesis**. The null hypothesis is denoted by  $H_0$ , and the alternative hypothesis is denoted by  $H_1$ . In the beginning of the trial it is assumed that the person is not guilty. The null hypothesis is usually the hypothesis that is assumed to be true to begin with. The two hypotheses for the court case are written as follows (notice the colon after  $H_0$  and  $H_1$ ):

Null hypothesis:  $H_0$ : The person is not guilty

Alternative hypothesis:  $H_1$ : The person is guilty

In a statistics example, the null hypothesis states that a given claim (or statement) about a population parameter is true. Reconsider the example of the soft-drink company's claim that, on average, its cans contain 12 ounces of soda. In reality, this claim may or may not be true. However, we will initially assume that the company's claim is true (that is, the company is not guilty of cheating and lying). To test the claim of the soft-drink company, the null hypothesis will be that the company's claim is true. Let  $\mu$  be the mean amount of soda in all cans. The company's claim will be true if  $\mu = 12$  ounces. Thus, the null hypothesis will be written as

$$H_0: \mu = 12 \text{ ounces} \quad (\text{The company's claim is true})$$

In this example, the null hypothesis can also be written as  $\mu \geq 12$  ounces because the claim of the company will still be true if the cans contain, on average, more than 12 ounces of soda. The company will be accused of cheating the public only if the cans contain, on average, less than 12 ounces of soda. However, it will not affect the test whether we use an  $=$  or a  $\geq$  sign in the null hypothesis as long as the alternative hypothesis has a  $<$  sign. Remember that in the null hypothesis (and in the alternative hypothesis also) we use a population parameter (such as  $\mu$  or  $p$ ) and not a sample statistic (such as  $\bar{x}$  or  $\hat{p}$ ).

**Definition**

**Null Hypothesis** A *null hypothesis* is a claim (or statement) about a population parameter that is assumed to be true until it is declared false.

The alternative hypothesis in our statistics example will be that the company's claim is false and its soft-drink cans contain, on average, less than 12 ounces of soda—that is,  $\mu < 12$  ounces. The alternative hypothesis will be written as

$$H_1: \mu < 12 \text{ ounces} \quad (\text{The company's claim is false})$$

**Definition**

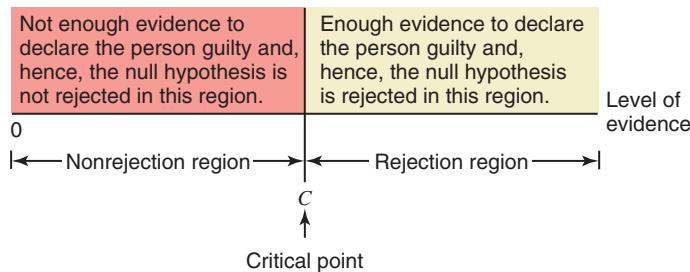
**Alternative Hypothesis** An *alternative hypothesis* is a claim about a population parameter that will be declared true if the null hypothesis is declared to be false.

Let us return to the example of the court trial. The trial begins with the assumption that the null hypothesis is true—that is, the person is not guilty. The prosecutor assembles all the possible evidence and presents it in the court to prove that the null hypothesis is false and the alternative hypothesis is true (that is, the person is guilty). In the case of our statistics example, the information obtained from a sample will be used as evidence to decide whether or not the claim of the company is true. In the court case, the decision made by the judge (or jury) depends on the amount of evidence presented by the prosecutor. At the end of the trial, the judge (or jury) will consider whether or not the evidence presented by the prosecutor is sufficient to declare the person guilty. The amount of evidence that will be considered to be sufficient to declare the person guilty depends on the discretion of the judge (or jury).

### 9.1.2 Rejection and Nonrejection Regions

In Figure 9.1, which represents the court case, the point marked 0 indicates that there is no evidence against the person being tried. The farther we move toward the right on the horizontal axis, the more convincing the evidence is that the person has committed the crime. We have arbitrarily marked a point  $C$  on the horizontal axis. Let us assume that a judge (or jury) considers any amount of evidence from point  $C$  to the right of it to be sufficient and any amount of evidence to the left of  $C$  to be insufficient to declare the person guilty. Point  $C$  is called the **critical value** or **critical point** in statistics. If the amount of evidence presented by the prosecutor falls in the area to the left of point  $C$ , the verdict will reflect that there is not enough evidence to declare the person guilty. Consequently, the accused person will be declared *not guilty*. In statistics, this decision is stated as *do not reject  $H_0$*  or failing to reject  $H_0$ . It is equivalent to saying that there is not enough evidence to declare the null hypothesis false. The area to the left of point  $C$  is called the *nonrejection region*; that is, this is the region where the null hypothesis is not rejected. However, if the amount of evidence falls at point  $C$

**Figure 9.1** Nonrejection and rejection regions for the court case.



or to the right of point  $C$ , the verdict will be that there is sufficient evidence to declare the person guilty. In statistics, this decision is stated as *reject  $H_0$*  or *the null hypothesis is false*. Rejecting  $H_0$  is equivalent to saying that *the alternative hypothesis is true*. The area to the right of point  $C$  (including point  $C$ ) is called the *rejection region*; that is, this is the region where the null hypothesis is rejected.

### 9.1.3 Two Types of Errors

We all know that a court's verdict is not always correct. If a person is declared guilty at the end of a trial, there are two possibilities.

1. The person has *not* committed the crime but is declared guilty (because of what may be false evidence).
2. The person *has* committed the crime and is rightfully declared guilty.

In the first case, the court has made an error by punishing an innocent person. In statistics, this kind of error is called a **Type I** or an  $\alpha$  (*alpha*) **error**. In the second case, because the guilty person has been punished, the court has made the correct decision. The second row in the shaded portion of Table 9.1 shows these two cases. The two columns of Table 9.1, corresponding to *the person is not guilty* and *the person is guilty*, give the two actual situations. Which one of these is true is known only to the person being tried. The two rows in this table, corresponding to *the person is not guilty* and *the person is guilty*, show the two possible court decisions.

**Table 9.1** Four Possible Outcomes for a Court Case

		Actual Situation	
		The Person Is Not Guilty	The Person Is Guilty
Court's decision	The person is not guilty	Correct decision	Type II or $\beta$ error
	The person is guilty	Type I or $\alpha$ error	Correct decision

In our statistics example, a Type I error will occur when  $H_0$  is actually true (that is, the cans do contain, on average, 12 ounces of soda), but it just happens that we draw a sample with a mean that is much less than 12 ounces and we wrongfully reject the null hypothesis,  $H_0$ . The value of  $\alpha$ , called the **significance level** of the test, represents the probability of making a Type I error. In other words,  $\alpha$  is the probability of rejecting the null hypothesis,  $H_0$ , when in fact it is true.

#### Definition

**Type I Error** A *Type I error* occurs when a true null hypothesis is rejected. The value of  $\alpha$  represents the probability of committing this type of error; that is,

$$\alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true})$$

The value of  $\alpha$  represents the *significance level* of the test.

The size of the rejection region in a statistics problem of a test of hypothesis depends on the value assigned to  $\alpha$ . In one approach to the test of hypothesis, we assign a value to  $\alpha$  before making the test. Although any value can be assigned to  $\alpha$ , commonly used values of  $\alpha$  are .01, .025, .05, and .10. Usually the value assigned to  $\alpha$  does not exceed .10 (or 10%).

Now, suppose that in the court trial case the person is declared not guilty at the end of the trial. Such a verdict does not indicate that the person has indeed *not* committed the crime. It is

possible that the person is guilty but there is not enough evidence to prove the guilt. Consequently, in this situation there are again two possibilities.

1. The person has *not* committed the crime and is declared not guilty.
2. The person *has* committed the crime but, *because of the lack of enough evidence*, is declared not guilty.

In the first case, the court's decision is correct. In the second case, however, the court has committed an error by setting a guilty person free. In statistics, this type of error is called a **Type II** or a  $\beta$  (the Greek letter *beta*) **error**. These two cases are shown in the first row of the shaded portion of Table 9.1.

In our statistics example, a Type II error will occur when the null hypothesis,  $H_0$ , is actually false (that is, the soda contained in all cans, on average, is less than 12 ounces), but it happens by chance that we draw a sample with a mean that is close to or greater than 12 ounces and we wrongfully conclude *do not reject*  $H_0$ . The value of  $\beta$  represents the probability of making a Type II error. It represents the probability that  $H_0$  is not rejected when actually  $H_0$  is false. The value of  $1 - \beta$  is called the **power of the test**. It represents the probability of not making a Type II error.

### Definition

**Type II Error** A *Type II error* occurs when a false null hypothesis is not rejected. The value of  $\beta$  represents the probability of committing a Type II error; that is,

$$\beta = P(H_0 \text{ is not rejected} \mid H_0 \text{ is false})$$

The value of  $1 - \beta$  is called the *power of the test*. It represents the probability of not making a Type II error.

The two types of errors that occur in tests of hypotheses depend on each other. We cannot lower the values of  $\alpha$  and  $\beta$  simultaneously for a test of hypothesis for a fixed sample size. Lowering the value of  $\alpha$  will raise the value of  $\beta$ , and lowering the value of  $\beta$  will raise the value of  $\alpha$ . However, we can decrease both  $\alpha$  and  $\beta$  simultaneously by increasing the sample size. The explanation of how  $\alpha$  and  $\beta$  are related and the computation of  $\beta$  are not within the scope of this text.

Table 9.2, which is similar to Table 9.1, is written for the statistics problem of a test of hypothesis. In Table 9.2 *the person is not guilty* is replaced by  $H_0$  is true, *the person is guilty* by  $H_0$  is false, and the *court's decision* by *decision*.

**Table 9.2 Four Possible Outcomes for a Test of Hypothesis**

		Actual Situation	
		$H_0$ Is True	$H_0$ Is False
Decision	Do not reject $H_0$	Correct decision	Type II or $\beta$ error
	Reject $H_0$	Type I or $\alpha$ error	Correct decision

### 9.1.4 Tails of a Test

The statistical hypothesis-testing procedure is similar to the trial of a person in court but with two major differences. The first major difference is that in a statistical test of hypothesis, the partition of the total region into rejection and nonrejection regions is not arbitrary. Instead, it depends on the value assigned to  $\alpha$  (Type I error). As mentioned earlier,  $\alpha$  is also called the significance level of the test.

The second major difference relates to the rejection region. In the court case, the rejection region is at and to the right side of the critical point, as shown in Figure 9.1. However, in statistics, the rejection region for a hypothesis-testing problem can be on both sides, with the nonrejection region in the middle, or it can be on the left side or right side of the nonrejection region. These possibilities are explained in the next three parts of this section. A test with two rejection regions is called a **two-tailed test**, and a test with one rejection region is called a **one-tailed test**. The one-tailed test is called a **left-tailed test** if the rejection region is in the left tail of the distribution curve, and it is called a **right-tailed test** if the rejection region is in the right tail of the distribution curve.

### Definition

**Tails of the Test** A *two-tailed test* has rejection regions in both tails, a *left-tailed test* has the rejection region in the left tail, and a *right-tailed test* has the rejection region in the right tail of the distribution curve.

### A Two-Tailed Test

According to the U.S. Bureau of Labor Statistics, people in the United States who had a bachelor's degree and were employed earned an average of \$1038 a week in 2010. Suppose an economist wants to check whether this mean has changed since 2010. The key word here is *changed*. The mean weekly earning of employed Americans with a bachelor's degree has changed if it has either increased or decreased since 2010. This is an example of a two-tailed test. Let  $\mu$  be the mean weekly earning of employed Americans with a bachelor's degree. The two possible decisions are as follows:

1. The mean weekly earning of employed Americans with a bachelor's degree has not changed since 2010, that is, currently  $\mu = \$1038$ .
2. The mean weekly earning of employed Americans with a bachelor's degree has changed since 2010, that is, currently  $\mu \neq \$1038$ .

We will write the null and alternative hypotheses for this test as follows:

$H_0: \mu = \$1038$  (The mean weekly earning of employed Americans with a bachelor's degree has not changed)

$H_1: \mu \neq \$1038$  (The mean weekly earning of employed Americans with a bachelor's degree has changed)

Whether a test is two-tailed or one-tailed is determined by the sign in the alternative hypothesis. If the alternative hypothesis has a *not equal to* ( $\neq$ ) sign, as in this example, it is a two-tailed test. As shown in Figure 9.2, a two-tailed test has two rejection regions, one in each tail of the distribution curve. Figure 9.2 shows the sampling distribution of  $\bar{x}$ , assuming it has a normal distribution. Assuming  $H_0$  is true,  $\bar{x}$  has a normal distribution with its mean equal to

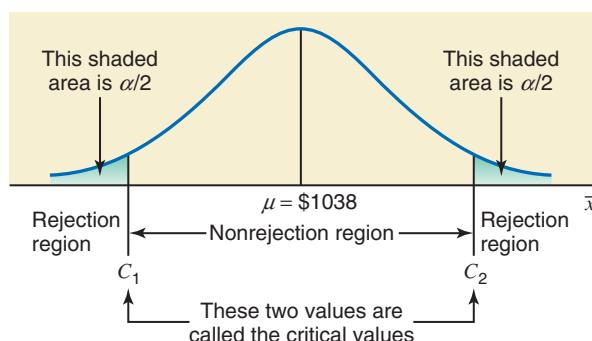


Figure 9.2 A two-tailed test.

\$1038 (the value of  $\mu$  in  $H_0$ ). In Figure 9.2, the area of each of the two rejection regions is  $\alpha/2$  and the total area of both rejection regions is  $\alpha$  (the significance level). As shown in this figure, a two-tailed test of hypothesis has two critical values that separate the two rejection regions from the nonrejection region. We will reject  $H_0$  if the value of  $\bar{x}$  obtained from the sample falls in either of the two rejection regions. We will not reject  $H_0$  if the value of  $\bar{x}$  lies in the nonrejection region. By rejecting  $H_0$ , we are saying that the difference between the value of  $\mu$  stated in  $H_0$  and the value of  $\bar{x}$  obtained from the sample is too large to have occurred because of the sampling error alone. Consequently, this difference appears to be real. By not rejecting  $H_0$ , we are saying that the difference between the value of  $\mu$  stated in  $H_0$  and the value of  $\bar{x}$  obtained from the sample is small and it may have occurred because of the sampling error alone.

### A Left-Tailed Test

Reconsider the example of the mean amount of soda in all soft-drink cans produced by a company. The company claims that these cans, on average, contain 12 ounces of soda. However, if these cans contain less than the claimed amount of soda, then the company can be accused of underfilling the cans. Suppose a consumer agency wants to test whether the mean amount of soda per can is less than 12 ounces. Note that the key phrase this time is *less than*, which indicates a left-tailed test. Let  $\mu$  be the mean amount of soda in all cans. The two possible decisions are as follows:

1. The mean amount of soda in all cans is equal to 12 ounces, that is,  $\mu = 12$  ounces.
2. The mean amount of soda in all cans is less than 12 ounces, that is,  $\mu < 12$  ounces.

The null and alternative hypotheses for this test are written as

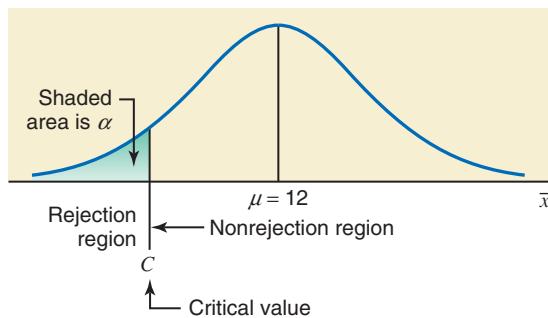
$$H_0: \mu = 12 \text{ ounces} \quad (\text{The mean is equal to 12 ounces})$$

$$H_1: \mu < 12 \text{ ounces} \quad (\text{The mean is less than 12 ounces})$$

In this case, we can also write the null hypothesis as  $H_0: \mu \geq 12$ . This will not affect the result of the test as long as the sign in  $H_1$  is *less than* ( $<$ ).

When the alternative hypothesis has a *less than* ( $<$ ) sign, as in this case, the test is always left-tailed. In a left-tailed test, the rejection region is in the left tail of the distribution curve, as shown in Figure 9.3, and the area of this rejection region is equal to  $\alpha$  (the significance level). We can observe from this figure that there is only one critical value in a left-tailed test.

**Figure 9.3** A left-tailed test.



Assuming  $H_0$  is true, the sampling distribution of  $\bar{x}$  has a mean equal to 12 ounces (the value of  $\mu$  in  $H_0$ ). We will reject  $H_0$  if the value of  $\bar{x}$  obtained from the sample falls in the rejection region; we will not reject  $H_0$  otherwise.

### A Right-Tailed Test

To illustrate the third case, according to [www.city-data.com](http://www.city-data.com), the average price of homes in West Orange, New Jersey, was \$459,204 in 2009. Suppose a real estate researcher wants to check whether the current mean price of homes in this town is higher than \$459,204. The key phrase

in this case is *higher than*, which indicates a right-tailed test. Let  $\mu$  be the current mean price of homes in this town. The two possible decisions are as follows:

1. The current mean price of homes in this town is not higher than \$459,204, that is, currently  $\mu = \$459,204$ .
2. The current mean price of homes in this town is higher than \$459,204, that is, currently  $\mu > \$459,204$ .

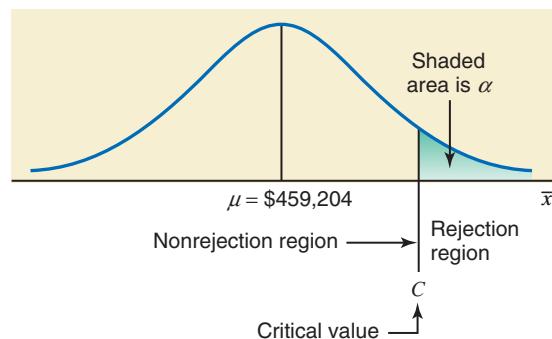
We write the null and alternative hypotheses for this test as follows:

$$H_0: \mu = \$459,204 \text{ (The current mean price of homes in this town is not higher than } \$459,204\text{)}$$

$$H_1: \mu > \$459,204 \text{ (The current mean price of homes in this area is higher than } \$459,204\text{)}$$

Note that here we can also write the null hypothesis as  $H_0: \mu \leq \$459,204$ , which states that the current mean price of homes in this area is either equal to or less than \$459,204. Again, the result of the test will not be affected by whether we use an *equal to* ( $=$ ) or a *less than or equal to* ( $\leq$ ) sign in  $H_0$  as long as the alternative hypothesis has a *greater than* ( $>$ ) sign.

When the alternative hypothesis has a *greater than* ( $>$ ) sign, the test is always right-tailed. As shown in Figure 9.4, in a right-tailed test, the rejection region is in the right tail of the distribution curve. The area of this rejection region is equal to  $\alpha$ , the significance level. Like a left-tailed test, a right-tailed test has only one critical value.



**Figure 9.4** A right-tailed test.

Again, assuming  $H_0$  is true, the sampling distribution of  $\bar{x}$  has a mean equal to \$459,204 (the value of  $\mu$  in  $H_0$ ). We will reject  $H_0$  if the value of  $\bar{x}$  obtained from the sample falls in the rejection region. Otherwise, we will not reject  $H_0$ .

Table 9.3 summarizes the foregoing discussion about the relationship between the signs in  $H_0$  and  $H_1$  and the tails of a test.

**Table 9.3** Signs in  $H_0$  and  $H_1$  and Tails of a Test

	Two-Tailed Test	Left-Tailed Test	Right-Tailed Test
Sign in the null hypothesis $H_0$	$=$	$=$ or $\geq$	$=$ or $\leq$
Sign in the alternative hypothesis $H_1$	$\neq$	$<$	$>$
Rejection region	In both tails	In the left tail	In the right tail

Note that the null hypothesis always has an *equal to* ( $=$ ) or a *greater than or equal to* ( $\geq$ ) or a *less than or equal to* ( $\leq$ ) sign, and the alternative hypothesis always has a *not equal to* ( $\neq$ ) or a *less than* ( $<$ ) or a *greater than* ( $>$ ) sign.

In this text we will use the following two procedures to make tests of hypothesis.

1. **The *p*-value approach.** Under this procedure, we calculate what is called the *p*-value for the observed value of the sample statistic. If we have a predetermined significance level, then we compare the *p*-value with this significance level and make a decision. Note that here *p* stands for probability.
2. **The critical-value approach.** In this approach, we find the critical value(s) from a table (such as the normal distribution table or the *t* distribution table) and find the value of the test statistic for the observed value of the sample statistic. Then we compare these two values and make a decision.

**Remember ►** Remember, the procedures to be learned in this chapter assume that the sample taken is a simple random sample. Also, remember that the critical point is included in the rejection region.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 9.1 Briefly explain the meaning of each of the following terms.
    - a. Null hypothesis
    - b. Alternative hypothesis
    - c. Critical point(s)
    - d. Significance level
    - e. Nonrejection region
    - f. Rejection region
    - g. Tails of a test
    - h. Two types of errors
  - 9.2 What are the four possible outcomes for a test of hypothesis? Show these outcomes by writing a table. Briefly describe the Type I and Type II errors.
  - 9.3 Explain how the tails of a test depend on the sign in the alternative hypothesis. Describe the signs in the null and alternative hypotheses for a two-tailed, a left-tailed, and a right-tailed test, respectively.
  - 9.4 Explain which of the following is a two-tailed test, a left-tailed test, or a right-tailed test.
    - a.  $H_0: \mu = 45$ ,  $H_1: \mu > 45$
    - b.  $H_0: \mu = 23$ ,  $H_1: \mu \neq 23$
    - c.  $H_0: \mu \geq 75$ ,  $H_1: \mu < 75$
- Show the rejection and nonrejection regions for each of these cases by drawing a sampling distribution curve for the sample mean, assuming that it is normally distributed.
- 9.5 Explain which of the following is a two-tailed test, a left-tailed test, or a right-tailed test.
    - a.  $H_0: \mu = 12$ ,  $H_1: \mu < 12$
    - b.  $H_0: \mu \leq 85$ ,  $H_1: \mu > 85$
    - c.  $H_0: \mu = 33$ ,  $H_1: \mu \neq 33$
- Show the rejection and nonrejection regions for each of these cases by drawing a sampling distribution curve for the sample mean, assuming that it is normally distributed.
- 9.6 Which of the two hypotheses (null and alternative) is initially assumed to be true in a test of hypothesis?
  - 9.7 Consider  $H_0: \mu = 20$  versus  $H_1: \mu < 20$ .
    - a. What type of error would you make if the null hypothesis is actually false and you fail to reject it?
    - b. What type of error would you make if the null hypothesis is actually true and you reject it?
  - 9.8 Consider  $H_0: \mu = 55$  versus  $H_1: \mu \neq 55$ .
    - a. What type of error would you make if the null hypothesis is actually false and you fail to reject it?
    - b. What type of error would you make if the null hypothesis is actually true and you reject it?

### APPLICATIONS

- 9.9 Write the null and alternative hypotheses for each of the following examples. Determine if each is a case of a two-tailed, a left-tailed, or a right-tailed test.
  - a. To test if the mean number of hours spent working per week by college students who hold jobs is different from 20 hours
  - b. To test whether or not a bank's ATM is out of service for an average of more than 10 hours per month
  - c. To test if the mean length of experience of airport security guards is different from 3 years
  - d. To test if the mean credit card debt of college seniors is less than \$1000
  - e. To test if the mean time a customer has to wait on the phone to speak to a representative of a mail-order company about unsatisfactory service is more than 12 minutes
- 9.10 Write the null and alternative hypotheses for each of the following examples. Determine if each is a case of a two-tailed, a left-tailed, or a right-tailed test.
  - a. To test if the mean amount of time spent per week watching sports on television by all adult men is different from 9.5 hours
  - b. To test if the mean amount of money spent by all customers at a supermarket is less than \$105

- c. To test whether the mean starting salary of college graduates is higher than \$47,000 per year
- d. To test if the mean waiting time at the drive-through window at a fast food restaurant during rush hour differs from 10 minutes
- e. To test if the mean hours spent per week on house chores by all housewives is less than 30

## 9.2 Hypothesis Tests About $\mu$ : $\sigma$ Known

This section explains how to perform a test of hypothesis for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known. As in Section 8.2 of Chapter 8, here also there are three possible cases as follows.

**Case I.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is normally distributed,

then we use the normal distribution to perform a test of hypothesis about  $\mu$  because from Section 7.3.1 of Chapter 7 the sampling distribution of  $\bar{x}$  is normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , assuming that  $n/N \leq .05$ .

**Case II.** If the following two conditions are fulfilled:

1. The population standard deviation  $\sigma$  is known
2. The sample size is large (i.e.,  $n \geq 30$ ),

then, again, we use the normal distribution to perform a test of hypothesis about  $\mu$  because from Section 7.3.2 of Chapter 7, due to the central limit theorem, the sampling distribution of  $\bar{x}$  is (approximately) normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , assuming that  $n/N \leq .05$ .

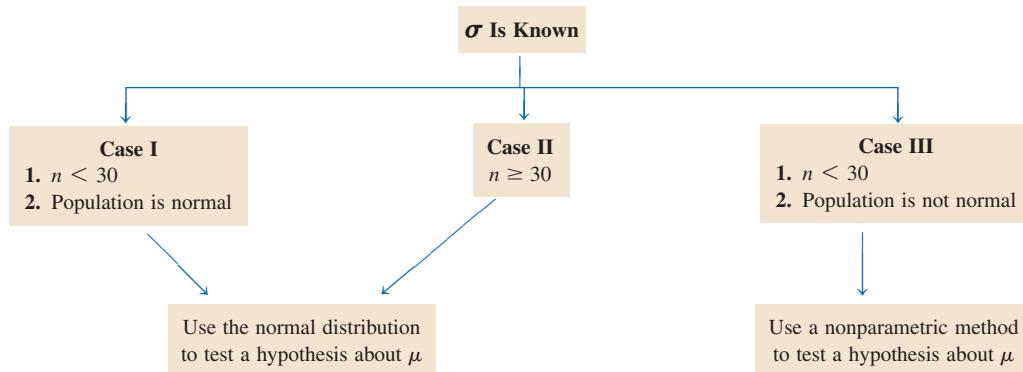
**Case III.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown),

then we use a nonparametric method (explained in Chapter 15) to perform a test of hypothesis about  $\mu$ .

This section will cover the first two cases. The procedure for performing a test of hypothesis about  $\mu$  is the same in both these cases. Note that in Case I, the population does not have to be exactly normally distributed. As long as it is close to the normal distribution without any outliers, we can use the normal distribution procedure. In Case II, although 30 is considered a large sample, if the population distribution is very different from the normal distribution, then 30 may not be a large enough sample size for the sampling distribution of  $\bar{x}$  to be normal and, hence, to use the normal distribution.

The following chart summarizes the above three cases.



Below we explain two procedures, the *p*-value approach and the critical-value approach, to test hypotheses about  $\mu$  under Cases I and II. We will use the normal distribution to perform such tests.

Note that the two approaches—the *p*-value approach and the critical-value approach—are not mutually exclusive. We do not need to use one or the other. We can use both at the same time.

### 9.2.1 The *p*-Value Approach

In this procedure, we find a probability value such that a given null hypothesis is rejected for any  $\alpha$  (significance level) greater than this value and it is not rejected for any  $\alpha$  less than this value. The **probability-value approach**, more commonly called the *p*-value approach, gives such a value. In this approach, we calculate the ***p*-value** for the test, which is defined as the smallest level of significance at which the given null hypothesis is rejected. Using this *p*-value, we state the decision. If we have a predetermined value of  $\alpha$ , then we compare the value of *p* with  $\alpha$  and make a decision.

#### Definition

***p*-Value** Assuming that the null hypothesis is true, the *p*-value can be defined as the probability that a sample statistic (such as the sample mean) is at least as far away from the hypothesized value in the direction of the alternative hypothesis as the one obtained from the sample data under consideration. Note that the *p*-value is the smallest significance level at which the null hypothesis is rejected.

Using the *p*-value approach, we reject the null hypothesis if

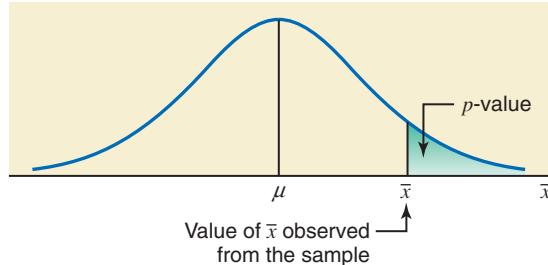
$$p\text{-value} \leq \alpha \quad \text{or} \quad \alpha \geq p\text{-value}$$

and we do not reject the null hypothesis if

$$p\text{-value} > \alpha \quad \text{or} \quad \alpha < p\text{-value}$$

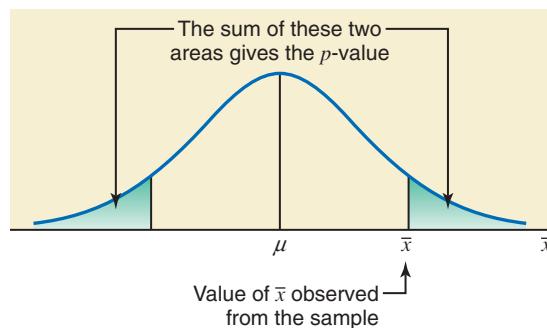
For a one-tailed test, the *p*-value is given by the area in the tail of the sampling distribution curve beyond the observed value of the sample statistic. Figure 9.5 shows the *p*-value for a right-tailed test about  $\mu$ . For a left-tailed test, the *p*-value will be the area in the lower tail of the sampling distribution curve to the left of the observed value of  $\bar{x}$ .

**Figure 9.5** The *p*-value for a right-tailed test.



For a two-tailed test, the *p*-value is twice the area in the tail of the sampling distribution curve beyond the observed value of the sample statistic. Figure 9.6 shows the *p*-value for a two-tailed test. Each of the areas in the two tails gives one-half the *p*-value.

**Figure 9.6** The *p*-value for a two-tailed test.



To find the area under the normal distribution curve beyond the sample mean  $\bar{x}$ , we first find the  $z$  value for  $\bar{x}$  using the following formula.

**Calculating the z Value for  $\bar{x}$**  When using the normal distribution, the value of  $z$  for  $\bar{x}$  for a test of hypothesis about  $\mu$  is computed as follows:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The value of  $z$  calculated for  $\bar{x}$  using this formula is also called the **observed value of  $z$** .

Then we find the area under the tail of the normal distribution curve beyond this value of  $z$ . This area gives the  $p$ -value or one-half the  $p$ -value, depending on whether it is a one-tailed test or a two-tailed test.

A test of hypothesis procedure that uses the  $p$ -value approach involves the following four steps.

### Steps to Perform a Test of Hypothesis Using the $p$ -Value Approach

1. State the null and alternative hypotheses.
2. Select the distribution to use.
3. Calculate the  $p$ -value.
4. Make a decision.

Examples 9–1 and 9–2 illustrate the calculation and use of the  $p$ -value to test a hypothesis using the normal distribution.

## ■ EXAMPLE 9–1

At Canon Food Corporation, it used to take an average of 90 minutes for new workers to learn a food processing job. Recently the company installed a new food processing machine. The supervisor at the company wants to find if the mean time taken by new workers to learn the food processing procedure on this new machine is different from 90 minutes. A sample of 20 workers showed that it took, on average, 85 minutes for them to learn the food processing procedure on the new machine. It is known that the learning times for all new workers are normally distributed with a population standard deviation of 7 minutes. Find the  $p$ -value for the test that the mean learning time for the food processing procedure on the new machine is different from 90 minutes. What will your conclusion be if  $\alpha = .01$ ?

Performing a hypothesis test using the  $p$ -value approach for a two-tailed test with the normal distribution.

**Solution** Let  $\mu$  be the mean time (in minutes) taken to learn the food processing procedure on the new machine by all workers, and let  $\bar{x}$  be the corresponding sample mean. From the given information,

$$n = 20, \quad \bar{x} = 85 \text{ minutes}, \quad \sigma = 7 \text{ minutes}, \quad \text{and} \quad \alpha = .01$$

To calculate the  $p$ -value and perform the test, we apply the following four steps.

**Step 1.** *State the null and alternative hypotheses.*

$$H_0: \mu = 90 \text{ minutes}$$

$$H_1: \mu \neq 90 \text{ minutes}$$

Note that the null hypothesis states that the mean time for learning the food processing procedure on the new machine is 90 minutes, and the alternative hypothesis states that this time is different from 90 minutes.

**Step 2.** Select the distribution to use.

Here, the population standard deviation  $\sigma$  is known, the sample size is small ( $n < 30$ ), but the population distribution is normal. Hence, the sampling distribution of  $\bar{x}$  is normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Consequently, we will use the normal distribution to find the  $p$ -value and make the test.

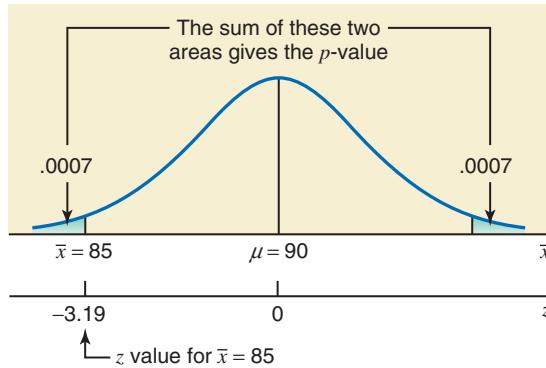
**Step 3.** Calculate the  $p$ -value.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The  $p$ -value is equal to twice the area in the tail of the sampling distribution curve of  $\bar{x}$  to the left of  $\bar{x} = 85$ , as shown in Figure 9.7. To find this area, we first find the  $z$  value for  $\bar{x} = 85$  as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{20}} = 1.56524758 \text{ minutes}$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{85 - 90}{1.56524758} = -3.19$$

**Figure 9.7** The  $p$ -value for a two-tailed test.



The area to the left of  $\bar{x} = 85$  is equal to the area under the standard normal curve to the left of  $z = -3.19$ . From the normal distribution table, the area to the left of  $z = -3.19$  is .0007. Consequently, the  $p$ -value is

$$p\text{-value} = 2(.0007) = .0014$$

**Step 4.** Make a decision.

Thus, based on the  $p$ -value of .0014, we can state that for any  $\alpha$  (significance level) greater than or equal to .0014, we will reject the null hypothesis stated in Step 1, and for any  $\alpha$  less than .0014, we will not reject the null hypothesis.

Because  $\alpha = .01$  is greater than the  $p$ -value of .0014, we reject the null hypothesis at this significance level. Therefore, we conclude that the mean time for learning the food processing procedure on the new machine is different from 90 minutes. ■

## ■ EXAMPLE 9-2

Performing a hypothesis test using the  $p$ -value approach for a one-tailed test with the normal distribution.

The management of Priority Health Club claims that its members lose an average of 10 pounds or more within the first month after joining the club. A consumer agency that wanted to check this claim took a random sample of 36 members of this health club and found that they lost an average of 9.2 pounds within the first month of membership. The population standard deviation is known to be 2.4 pounds. Find the  $p$ -value for this test. What will your decision be if  $\alpha = .01$ ? What if  $\alpha = .05$ ?

**Solution** Let  $\mu$  be the mean weight lost during the first month of membership by all members of this health club, and let  $\bar{x}$  be the corresponding mean for the sample. From the given information,

$$n = 36, \quad \bar{x} = 9.2 \text{ pounds}, \quad \text{and} \quad \sigma = 2.4 \text{ pounds}$$

The claim of the club is that its members lose, on average, 10 pounds or more within the first month of membership. To perform the test using the  $p$ -value approach, we apply the following four steps.

**Step 1.** State the null and alternative hypotheses.

$$H_0: \mu \geq 10 \quad (\text{The mean weight lost is 10 pounds or more.})$$

$$H_1: \mu < 10 \quad (\text{The mean weight lost is less than 10 pounds.})$$

**Step 2.** Select the distribution to use.

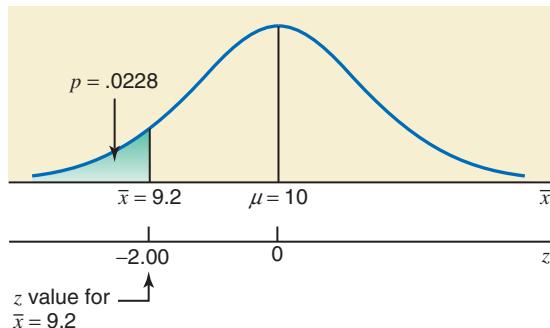
Here, the population standard deviation  $\sigma$  is known, and the sample size is large ( $n \geq 30$ ). Hence, the sampling distribution of  $\bar{x}$  is normal (due to the Central Limit Theorem) with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Consequently, we will use the normal distribution to find the  $p$ -value and perform the test.

**Step 3.** Calculate the  $p$ -value.

The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed. The  $p$ -value is given by the area to the left of  $\bar{x} = 9.2$  under the sampling distribution curve of  $\bar{x}$ , as shown in Figure 9.8. To find this area, we first find the  $z$  value for  $\bar{x} = 9.2$  as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.4}{\sqrt{36}} = .40$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{9.2 - 10}{.40} = -2.00$$



**Figure 9.8** The  $p$ -value for a left-tailed test.

The area to the left of  $\bar{x} = 9.2$  under the sampling distribution of  $\bar{x}$  is equal to the area under the standard normal curve to the left of  $z = -2.00$ . From the normal distribution table, the area to the left of  $z = -2.00$  is .0228. Consequently,

$$p\text{-value} = .0228$$

**Step 4.** Make a decision.

Thus, based on the  $p$ -value of .0228, we can state that for any  $\alpha$  (significance level) greater than or equal to .0228 we will reject the null hypothesis stated in Step 1, and for any  $\alpha$  less than .0228 we will not reject the null hypothesis.

Since  $\alpha = .01$  is less than the  $p$ -value of .0228, we do not reject the null hypothesis at this significance level. Consequently, we conclude that there is not significant evidence to reject the claim that the mean weight lost within the first month of membership by the members of this club is 10 pounds or more.

Now, because  $\alpha = .05$  is greater than the  $p$ -value of .0228, we reject the null hypothesis at this significance level. In this case we conclude that the mean weight lost within the first month of membership by the members of this club is less than 10 pounds. ■



© Christopher Futcher/iStockphoto

## 9.2.2 The Critical-Value Approach

In this procedure, we have a predetermined value of the significance level  $\alpha$ . The value of  $\alpha$  gives the total area of the rejection region(s). First we find the critical value(s) of  $z$  from the

normal distribution table for the given significance level. Then we find the value of the test statistic  $z$  for the observed value of the sample statistic  $\bar{x}$ . Finally we compare these two values and make a decision. Remember, if the test is one-tailed, there is only one critical value of  $z$ , and it is obtained by using the value of  $\alpha$  which gives the area in the left or right tail of the normal distribution curve depending on whether the test is left-tailed or right-tailed, respectively. However, if the test is two-tailed, there are two critical values of  $z$  and they are obtained by using  $\alpha/2$  area in each tail of the normal distribution curve. The value of the test statistic is obtained as follows.

**Test Statistic** In tests of hypotheses about  $\mu$  using the normal distribution, the random variable

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

is called the *test statistic*. The test statistic can be defined as a rule or criterion that is used to make the decision on whether or not to reject the null hypothesis.

A test of hypothesis procedure that uses the critical-value approach involves the following five steps.

#### Steps to Perform a Test of Hypothesis with the Critical-Value Approach

1. State the null and alternative hypotheses.
2. Select the distribution to use.
3. Determine the rejection and nonrejection regions.
4. Calculate the value of the test statistic.
5. Make a decision.

Examples 9–3 and 9–4 illustrate the use of these five steps to perform tests of hypotheses about the population mean  $\mu$ . Example 9–3 is concerned with a two-tailed test, and Example 9–4 describes a one-tailed test.

### ■ EXAMPLE 9–3

Conducting a two-tailed test of hypothesis about  $\mu$ :  $\sigma$  known and  $n \geq 30$ .

The TIV Telephone Company provides long-distance telephone service in an area. According to the company's records, the average length of all long-distance calls placed through this company in 2011 was 12.44 minutes. The company's management wanted to check if the mean length of the current long-distance calls is different from 12.44 minutes. A sample of 150 such calls placed through this company produced a mean length of 13.71 minutes. The standard deviation of all such calls is 2.65 minutes. Using a 2% significance level, can you conclude that the mean length of all current long-distance calls is different from 12.44 minutes?

**Solution** Let  $\mu$  be the mean length of all current long-distance calls placed through this company and  $\bar{x}$  be the corresponding mean for the sample. From the given information,

$$n = 150, \quad \bar{x} = 13.71 \text{ minutes}, \quad \text{and} \quad \sigma = 2.65 \text{ minutes}$$

We are to test whether or not the mean length of all current long-distance calls is different from 12.44 minutes. The significance level  $\alpha$  is .02; that is, the probability of rejecting the null hypothesis when it actually is true should not exceed .02. This is the probability of making a Type I error. We perform the test of hypothesis using the five steps as follows.

**Step 1.** State the null and alternative hypotheses.

Notice that we are testing to find whether or not the mean length of all current long-distance calls is different from 12.44 minutes. We write the null and alternative hypotheses as follows.

$H_0: \mu = 12.44$  (The mean length of all current long-distance calls is 12.44 minutes.)

$H_1: \mu \neq 12.44$  (The mean length of all current long-distance calls is different from 12.44 minutes.)

**Step 2.** Select the distribution to use.

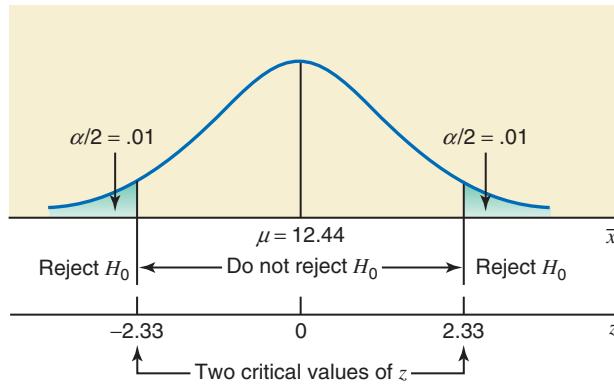
Here, the population standard deviation  $\sigma$  is known, and the sample size is large ( $n \geq 30$ ). Hence, the sampling distribution of  $\bar{x}$  is (approximately) normal (due to the Central Limit Theorem) with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Consequently, we will use the normal distribution to perform the test of this example.

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is .02. The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed with two rejection regions, one in each tail of the normal distribution curve of  $\bar{x}$ . Because the total area of both rejection regions is .02 (the significance level), the area of the rejection region in each tail is .01; that is,

$$\text{Area in each tail} = \alpha/2 = .02/2 = .01$$

These areas are shown in Figure 9.9. Two critical points in this figure separate the two rejection regions from the nonrejection region. Next, we find the  $z$  values for the two critical points using the area of the rejection region. To find the  $z$  values for these critical points, we look for .0100 and .9900 areas in the normal distribution table. From Table IV, the  $z$  values of the two critical points, as shown in Figure 9.9, are approximately  $-2.33$  and  $2.33$ .



**Figure 9.9** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

The decision to reject or not to reject the null hypothesis will depend on whether the evidence from the sample falls in the rejection or the nonrejection region. If the value of  $\bar{x}$  falls in either of the two rejection regions, we reject  $H_0$ . Otherwise, we do not reject  $H_0$ . The value of  $\bar{x}$  obtained from the sample is called the *observed value of  $\bar{x}$* . To locate the position of  $\bar{x} = 13.71$  on the sampling distribution curve of  $\bar{x}$  in Figure 9.9, we first calculate the  $z$  value for  $\bar{x} = 13.71$ . This is called the *value of the test statistic*. Then, we compare the value of the test statistic with the two critical values of  $z$ ,  $-2.33$  and  $2.33$ , shown in Figure 9.9. If the value of the test statistic is between  $-2.33$  and  $2.33$ , we do not reject  $H_0$ . If the value of the test statistic is either greater than or equal to  $2.33$  or less than or equal to  $-2.33$ , we reject  $H_0$ .

**Calculating the Value of the Test Statistic** When using the normal distribution, *the value of the test statistic  $z$*  for  $\bar{x}$  for a test of hypothesis about  $\mu$  is computed as follows:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

where

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This value of  $z$  for  $\bar{x}$  is also called the **observed value of  $z$** .

The value of  $\bar{x}$  from the sample is 13.71. We calculate the  $z$  value as follows:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} = \frac{2.65}{\sqrt{150}} = .21637159 \\ z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{13.71 - 12.44}{.21637159} = 5.87\end{aligned}$$

From  $H_0$

The value of  $\mu$  in the calculation of the  $z$  value is substituted from the null hypothesis. The value of  $z = 5.87$  calculated for  $\bar{x}$  is called the *computed value of the test statistic  $z$* . This is the value of  $z$  that corresponds to the value of  $\bar{x}$  observed from the sample. It is also called the *observed value of  $z$* .

#### Step 5. Make a decision.

In the final step we make a decision based on the location of the value of the test statistic  $z$  computed for  $\bar{x}$  in Step 4. This value of  $z = 5.87$  is greater than the critical value of  $z = 2.33$ , and it falls in the rejection region in the right tail in Figure 9.9. Hence, we reject  $H_0$  and conclude that based on the sample information, it appears that the current mean length of all such calls is not equal to 12.44 minutes.

By rejecting the null hypothesis, we are stating that the difference between the sample mean,  $\bar{x} = 13.71$  minutes, and the hypothesized value of the population mean,  $\mu = 12.44$  minutes, is too large and may not have occurred because of chance or sampling error alone. This difference seems to be real and, hence, the mean length of all such calls is currently different from 12.44 minutes. Note that the rejection of the null hypothesis does not necessarily indicate that the mean length of all such calls is currently definitely different from 12.44 minutes. It simply indicates that there is strong evidence (from the sample) that the current mean length of such calls is not equal to 12.44 minutes. There is a possibility that the current mean length of all such calls is equal to 12.44 minutes, but by the luck of the draw we selected a sample with a mean that is too far from the hypothesized mean of 12.44 minutes. If so, we have wrongfully rejected the null hypothesis  $H_0$ . This is a Type I error and its probability is .02 in this example. ■

We can use the  $p$ -value approach to perform the test of hypothesis in Example 9–3. In this example, the test is two-tailed. The  $p$ -value is equal to twice the area under the sampling distribution of  $\bar{x}$  to the right of  $\bar{x} = 13.71$ . As calculated in Step 4 above, the  $z$  value for  $\bar{x} = 13.71$  is 5.87. From the normal distribution table, the area to the right of  $z = 5.87$  is (approximately) zero. Hence, the  $p$ -value is (approximately) zero. (If you use technology, you will obtain the  $p$ -value of .000000002.) As we know from earlier discussions, we will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than or equal to the  $p$ -value. Consequently, in this example, we will reject the null hypothesis for any  $\alpha > 0$ . Since  $\alpha = .02$  here, which is greater than zero, we reject the null hypothesis.

### ■ EXAMPLE 9–4

Conducting a left-tailed test of hypothesis about  $\mu$ :  $\sigma$  known,  $n < 30$ , and population normal.

The mayor of a large city claims that the average net worth of families living in this city is at least \$300,000. A random sample of 25 families selected from this city produced a mean net worth of \$288,000. Assume that the net worths of all families in this city have a normal distribution with the population standard deviation of \$80,000. Using a 2.5% significance level, can you conclude that the mayor's claim is false?

**Solution** Let  $\mu$  be the mean net worth of families living in this city and  $\bar{x}$  be the corresponding mean for the sample. From the given information,

$$n = 25, \quad \bar{x} = \$288,000, \quad \text{and} \quad \sigma = \$80,000$$

The significance level is  $\alpha = .025$ .

**Step 1.** State the null and alternative hypotheses.

We are to test whether or not the mayor's claim is false. The mayor's claim is that the average net worth of families living in this city is at least \$300,000. Hence, the null and alternative hypotheses are as follows:

$H_0: \mu \geq \$300,000$  (The mayor's claim is true. The mean net worth is at least \$300,000.)

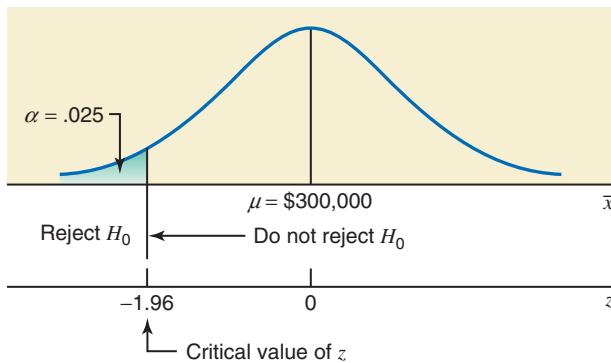
$H_1: \mu < \$300,000$  (The mayor's claim is false. The mean net worth is less than \$300,000.)

**Step 2.** Select the distribution to use.

Here, the population standard deviation  $\sigma$  is known, the sample size is small ( $n < 30$ ), but the population distribution is normal. Hence, the sampling distribution of  $\bar{x}$  is normal with its mean equal to  $\mu$  and the standard deviation equal to  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Consequently, we will use the normal distribution to perform the test.

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is  $.025$ . The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed with the rejection region in the left tail of the sampling distribution curve of  $\bar{x}$ . The critical value of  $z$ , obtained from the normal table for  $.0250$  area in the left tail, is  $-1.96$ , as shown in Figure 9.10.



**Figure 9.10** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $z$  for  $\bar{x} = \$288,000$  is calculated as follows:

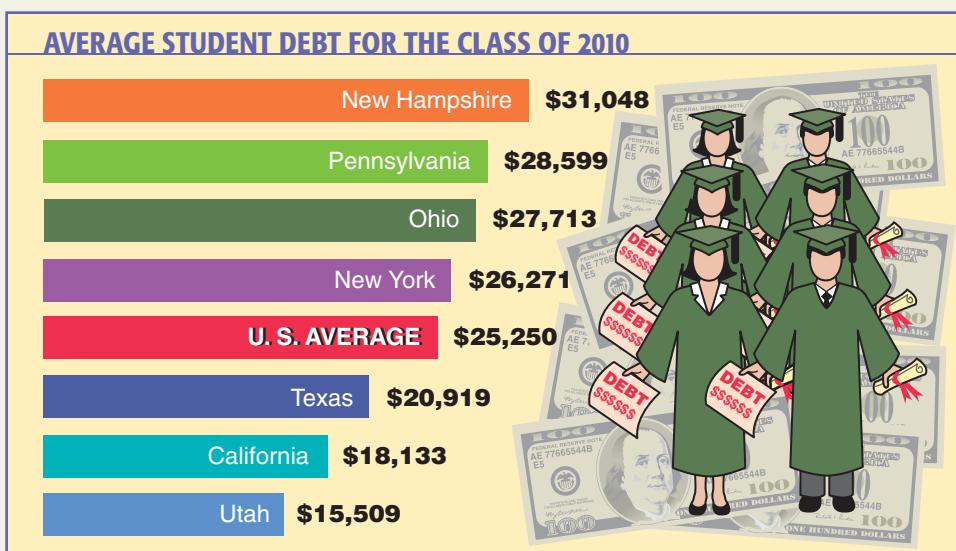
$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} = \frac{80,000}{\sqrt{25}} = \$16,000 \\ z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{288,000 - 300,000}{16,000} = -.75 \end{aligned}$$

From  $H_0$

**Step 5.** Make a decision.

The value of the test statistic  $z = -.75$  is greater than the critical value of  $z = -1.96$ , and it falls in the nonrejection region. As a result, we fail to reject  $H_0$ . Therefore, we can state that based on the sample information, it appears that the mean net worth of families in this city is not less than \$300,000. Note that we are not concluding that the mean net worth is definitely not less than \$300,000. By not rejecting the null hypothesis, we are saying that the information obtained from the sample is not strong enough to reject the null hypothesis and to conclude that the mayor's claim is false. ■

## AVERAGE STUDENT DEBT FOR THE CLASS OF 2010



Data source: The Project on Student Debt. The Institute for College Access & Success.

According to estimates by the Project on Student Debt study, two-thirds of college students who graduated in 2010 had loans to pay (<http://projectonstudentdebt.org/files/pub/classof2010.pdf>). The accompanying chart lists the U.S. average and the average of such debts for a few selected states for students in the class of 2010 who graduated with loans. Remember that these averages are based on a survey of students who graduated in 2010 with debt. For example, U.S. college graduates in 2010 owed an average of \$25,250. Students who graduated with loans from colleges in New Hampshire in 2010 had the highest average debt of \$31,048, and those who graduated with loans from colleges in Utah had the lowest average debt of \$15,509. Note that these averages are based on sample surveys. Suppose we want to find out if the average debt for students in Ohio was higher than the U.S. average of \$25,250. Suppose that the average debt for Ohio college students in the class of 2010 is based on a random sample of 900 students who graduated with loans. Assume that the standard deviation of the loans for all Ohio students in the class of 2010 was \$4800 and the significance level is 1%. The test is right-tailed because we are testing the hypothesis that the average debt for the class of 2010 in Ohio (which was \$27,713) was higher than \$25,250. The null and alternative hypotheses are

$$\begin{aligned} H_0: \mu &= \$25,250 \\ H_1: \mu &> \$25,250 \end{aligned}$$

Here,  $n = 900$ ,  $\bar{x} = \$27,713$ ,  $\sigma = \$4800$ , and  $\alpha = .01$ . The population standard deviation is known, and the sample is large. Hence, we can use the normal distribution to perform this test. Using the normal distribution to perform the test, we find that the critical value of  $z$  is 2.33 for .01 area in the right tail of the normal curve. We find the observed value of  $z$  as follows.

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} = \frac{4800}{\sqrt{900}} = \$160 \\ z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{27,713 - 25,250}{160} = 15.39 \end{aligned}$$

The value of the test statistic  $z = 15.39$  for  $\bar{x}$  is larger than the critical value of  $z = 2.33$ , and it falls in the rejection region. Consequently, we reject  $H_0$  and conclude that the average debt of students who graduated in 2010 from colleges in New Hampshire is higher than \$25,250, which is the average for the United States.

To use the  $p$ -value approach, we find the area under the normal curve to the right of  $z = 15.39$  from the normal distribution table. This area is .0000. Therefore, the  $p$ -value is .0000. Since  $\alpha = .01$  is larger than the  $p$ -value = .0000, we reject the null hypothesis.

We can use the  $p$ -value approach to perform the test of hypothesis in Example 9-4. In this example, the test is left-tailed. The  $p$ -value is given by the area under the sampling distribution of  $\bar{x}$  to the left of  $\bar{x} = \$288,000$ . As calculated in Step 4 above, the  $z$  value for  $\bar{x} = \$288,000$  is  $-.75$ . From the normal distribution table, the area to the left of  $z = -.75$  is .2266. Hence, the  $p$ -value is

**Data Source:**  
<http://projectonstudentdebt.org/files/pub/classof2010.pdf>.

.2266. We will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than or equal to the  $p$ -value. Consequently, we will reject the null hypothesis in this example for any  $\alpha \geq .2266$ . Since in this example  $\alpha = .025$ , which is less than .2266, we fail to reject the null hypothesis.

In studies published in various journals, authors usually use the terms *significantly different* and *not significantly different* when deriving conclusions based on hypothesis tests. These terms are short versions of the terms *statistically significantly different* and *statistically not significantly different*. The expression *significantly different* means that the difference between the observed value of the sample mean  $\bar{x}$  and the hypothesized value of the population mean  $\mu$  is so large that it probably did not occur because of the sampling error alone. Consequently, the null hypothesis is rejected. In other words, the difference between  $\bar{x}$  and  $\mu$  is statistically significant. Thus, the statement *significantly different* is equivalent to saying that the *null hypothesis is rejected*. In Example 9–3, we can state as a conclusion that the observed value of  $\bar{x} = 13.71$  minutes is significantly different from the hypothesized value of  $\mu = 12.44$  minutes. That is, the mean length of all current long-distance calls is different from 12.44 minutes.

On the other hand, the statement *not significantly different* means that the difference between the observed value of the sample mean  $\bar{x}$  and the hypothesized value of the population mean  $\mu$  is so small that it may have occurred just because of chance. Consequently, the null hypothesis is not rejected. Thus, the expression *not significantly different* is equivalent to saying that we *fail to reject the null hypothesis*. In Example 9–4, we can state as a conclusion that the observed value of  $\bar{x} = \$288,000$  is not significantly less than the hypothesized value of  $\mu = \$300,000$ . In other words, the current mean net worth of households in this city is not less than \$300,000.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 9.11** What are the five steps of a test of hypothesis using the critical value approach? Explain briefly.
- 9.12** What does the level of significance represent in a test of hypothesis? Explain.
- 9.13** By rejecting the null hypothesis in a test of hypothesis example, are you stating that the alternative hypothesis is true?
- 9.14** What is the difference between the critical value of  $z$  and the observed value of  $z$ ?
- 9.15** Briefly explain the procedure used to calculate the  $p$ -value for a two-tailed and for a one-tailed test, respectively.
- 9.16** Find the  $p$ -value for each of the following hypothesis tests.
  - a.  $H_0: \mu = 23$ ,  $H_1: \mu \neq 23$ ,  $n = 50$ ,  $\bar{x} = 21.25$ ,  $\sigma = 5$
  - b.  $H_0: \mu = 15$ ,  $H_1: \mu < 15$ ,  $n = 80$ ,  $\bar{x} = 13.25$ ,  $\sigma = 5.5$
  - c.  $H_0: \mu = 38$ ,  $H_1: \mu > 38$ ,  $n = 35$ ,  $\bar{x} = 40.25$ ,  $\sigma = 7.2$
- 9.17** Find the  $p$ -value for each of the following hypothesis tests.
  - a.  $H_0: \mu = 46$ ,  $H_1: \mu \neq 46$ ,  $n = 40$ ,  $\bar{x} = 49.60$ ,  $\sigma = 9.7$
  - b.  $H_0: \mu = 26$ ,  $H_1: \mu < 26$ ,  $n = 33$ ,  $\bar{x} = 24.30$ ,  $\sigma = 4.3$
  - c.  $H_0: \mu = 18$ ,  $H_1: \mu > 18$ ,  $n = 55$ ,  $\bar{x} = 20.50$ ,  $\sigma = 7.8$
- 9.18** Consider  $H_0: \mu = 29$  versus  $H_1: \mu \neq 29$ . A random sample of 25 observations taken from this population produced a sample mean of 25.3. The population is normally distributed with  $\sigma = 8$ .
  - a. Calculate the  $p$ -value.
  - b. Considering the  $p$ -value of part a, would you reject the null hypothesis if the test were made at a significance level of .05?
  - c. Considering the  $p$ -value of part a, would you reject the null hypothesis if the test were made at a significance level of .01?
- 9.19** Consider  $H_0: \mu = 72$  versus  $H_1: \mu > 72$ . A random sample of 16 observations taken from this population produced a sample mean of 75.2. The population is normally distributed with  $\sigma = 6$ .
  - a. Calculate the  $p$ -value.
  - b. Considering the  $p$ -value of part a, would you reject the null hypothesis if the test were made at a significance level of .01?
  - c. Considering the  $p$ -value of part a, would you reject the null hypothesis if the test were made at a significance level of .025?

**9.20** For each of the following examples of tests of hypotheses about  $\mu$ , show the rejection and nonrejection regions on the sampling distribution of the sample mean assuming that it is normal.

- a. A two-tailed test with  $\alpha = .05$  and  $n = 40$
- b. A left-tailed test with  $\alpha = .01$  and  $n = 20$
- c. A right-tailed test with  $\alpha = .02$  and  $n = 55$

**9.21** For each of the following examples of tests of hypotheses about  $\mu$ , show the rejection and nonrejection regions on the sampling distribution of the sample mean assuming it is normal.

- a. A two-tailed test with  $\alpha = .01$  and  $n = 100$
- b. A left-tailed test with  $\alpha = .005$  and  $n = 27$
- c. A right-tailed test with  $\alpha = .025$  and  $n = 36$

**9.22** Consider the following null and alternative hypotheses:

$$H_0: \mu = 25 \text{ versus } H_1: \mu \neq 25$$

Suppose you perform this test at  $\alpha = .05$  and reject the null hypothesis. Would you state that the difference between the hypothesized value of the population mean and the observed value of the sample mean is “statistically significant” or would you state that this difference is “statistically not significant”? Explain.

**9.23** Consider the following null and alternative hypotheses:

$$H_0: \mu = 60 \text{ versus } H_1: \mu > 60$$

Suppose you perform this test at  $\alpha = .01$  and fail to reject the null hypothesis. Would you state that the difference between the hypothesized value of the population mean and the observed value of the sample mean is “statistically significant” or would you state that this difference is “statistically not significant”? Explain.

**9.24** For each of the following significance levels, what is the probability of making a Type I error?

- a.  $\alpha = .025$
- b.  $\alpha = .05$
- c.  $\alpha = .01$

**9.25** For each of the following significance levels, what is the probability of making a Type I error?

- a.  $\alpha = .10$
- b.  $\alpha = .02$
- c.  $\alpha = .005$

**9.26** A random sample of 80 observations produced a sample mean of 86.50. Find the critical and observed values of  $z$  for each of the following tests of hypothesis using  $\alpha = .10$ . The population standard deviation is known to be 7.20.

- a.  $H_0: \mu = 91$  versus  $H_1: \mu \neq 91$
- b.  $H_0: \mu = 91$  versus  $H_1: \mu < 91$

**9.27** A random sample of 18 observations produced a sample mean of 9.24. Find the critical and observed values of  $z$  for each of the following tests of hypothesis using  $\alpha = .05$ . The population standard deviation is known to be 5.40 and the population distribution is normal.

- a.  $H_0: \mu = 8.5$  versus  $H_1: \mu \neq 8.5$
- b.  $H_0: \mu = 8.5$  versus  $H_1: \mu > 8.5$

**9.28** Consider the null hypothesis  $H_0: \mu = 625$ . Suppose that a random sample of 29 observations is taken from a normally distributed population with  $\sigma = 32$ . Using a significance level of .01, show the rejection and nonrejection regions on the sampling distribution curve of the sample mean and find the critical value(s) of  $z$  when the alternative hypothesis is as follows.

- a.  $H_1: \mu \neq 625$
- b.  $H_1: \mu > 625$
- c.  $H_1: \mu < 625$

**9.29** Consider the null hypothesis  $H_0: \mu = 5$ . A random sample of 140 observations is taken from a population with  $\sigma = 17$ . Using  $\alpha = .05$ , show the rejection and nonrejection regions on the sampling distribution curve of the sample mean and find the critical value(s) of  $z$  for the following.

- a. a right-tailed test
- b. a left-tailed test
- c. a two-tailed test

**9.30** Consider  $H_0: \mu = 100$  versus  $H_1: \mu \neq 100$ .

- a. A random sample of 64 observations produced a sample mean of 98. Using  $\alpha = .01$ , would you reject the null hypothesis? The population standard deviation is known to be 12.
- b. Another random sample of 64 observations taken from the same population produced a sample mean of 104. Using  $\alpha = .01$ , would you reject the null hypothesis? The population standard deviation is known to be 12.

Comment on the results of parts a and b.

**9.31** Consider  $H_0: \mu = 45$  versus  $H_1: \mu < 45$ .

- A random sample of 25 observations produced a sample mean of 41.8. Using  $\alpha = .025$ , would you reject the null hypothesis? The population is known to be normally distributed with  $\sigma = 6$ .
- Another random sample of 25 observations taken from the same population produced a sample mean of 43.8. Using  $\alpha = .025$ , would you reject the null hypothesis? The population is known to be normally distributed with  $\sigma = 6$ .

Comment on the results of parts a and b.

**9.32** Make the following tests of hypotheses.

- $H_0: \mu = 25, H_1: \mu \neq 25, n = 81, \bar{x} = 28.5, \sigma = 3, \alpha = .01$
- $H_0: \mu = 12, H_1: \mu < 12, n = 45, \bar{x} = 11.25, \sigma = 4.5, \alpha = .05$
- $H_0: \mu = 40, H_1: \mu > 40, n = 100, \bar{x} = 47, \sigma = 7, \alpha = .10$

**9.33** Make the following tests of hypotheses.

- $H_0: \mu = 80, H_1: \mu \neq 80, n = 33, \bar{x} = 76.5, \sigma = 15, \alpha = .10$
- $H_0: \mu = 32, H_1: \mu < 32, n = 75, \bar{x} = 26.5, \sigma = 7.4, \alpha = .01$
- $H_0: \mu = 55, H_1: \mu > 55, n = 40, \bar{x} = 60.5, \sigma = 4, \alpha = .05$

## ■ APPLICATIONS

**9.34** A consumer advocacy group suspects that a local supermarket's 10-ounce packages of cheddar cheese actually weigh less than 10 ounces. The group took a random sample of 20 such packages and found that the mean weight for the sample was 9.955 ounces. The population follows a normal distribution with the population standard deviation of .15 ounce.

- Find the  $p$ -value for the test of hypothesis with the alternative hypothesis that the mean weight of all such packages is less than 10 ounces. Will you reject the null hypothesis at  $\alpha = .01$ ?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .01$ .

**9.35** The manufacturer of a certain brand of auto batteries claims that the mean life of these batteries is 45 months. A consumer protection agency that wants to check this claim took a random sample of 24 such batteries and found that the mean life for this sample is 43.05 months. The lives of all such batteries have a normal distribution with the population standard deviation of 4.5 months.

- Find the  $p$ -value for the test of hypothesis with the alternative hypothesis that the mean life of these batteries is less than 45 months. Will you reject the null hypothesis at  $\alpha = .025$ ?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .025$ .

**9.36** A study claims that all adults spend an average of 14 hours or more on chores during a weekend. A researcher wanted to check if this claim is true. A random sample of 200 adults taken by this researcher showed that these adults spend an average of 14.65 hours on chores during a weekend. The population standard deviation is known to be 3.0 hours.

- Find the  $p$ -value for the hypothesis test with the alternative hypothesis that all adults spend more than 14 hours on chores during a weekend. Will you reject the null hypothesis at  $\alpha = .01$ ?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .01$ .

**9.37** According to the U.S. Bureau of Labor Statistics, all workers in America who had a bachelor's degree and were employed earned an average of \$1038 a week in 2010. A recent sample of 400 American workers who have a bachelor's degree showed that they earn an average of \$1060 per week. Suppose that the population standard deviation of such earnings is \$160.

- Find the  $p$ -value for the test of hypothesis with the alternative hypothesis that the current mean weekly earning of American workers who have a bachelor's degree is higher than \$1038. Will you reject the null hypothesis at  $\alpha = .025$ ?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .025$ .

**9.38** According to the U.S. Postal Service, the average weight of mail received by Americans in 2011 through the Postal Service was 57.2 pounds (*The New York Times*, December 4, 2011). One hundred randomly selected Americans were asked to keep all their mail for last year. It was found that they received an average of 55.3 pounds of mail last year. Suppose that the population standard deviation is 8.4 pounds.

- Find the  $p$ -value for the test of hypothesis with the alternative hypothesis that the average weight of mail received by all Americans last year was less than 57.2 pounds. Will you reject the null hypothesis at  $\alpha = .01$ ? Explain. What if  $\alpha = .025$ ?
- Test the hypothesis of part a using the critical-value approach. Will you reject the null hypothesis at  $\alpha = .01$ ? What if  $\alpha = .025$ ?

**9.39** A telephone company claims that the mean duration of all long-distance phone calls made by its residential customers is 10 minutes. A random sample of 100 long-distance calls made by its residential customers taken from the records of this company showed that the mean duration of calls for this sample is 9.20 minutes. The population standard deviation is known to be 3.80 minutes.

- Find the  $p$ -value for the test that the mean duration of all long-distance calls made by residential customers of this company is different from 10 minutes. If  $\alpha = .02$ , based on this  $p$ -value, would you reject the null hypothesis? Explain. What if  $\alpha = .05$ ?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .02$ . Does your conclusion change if  $\alpha = .05$ ?

**9.40** Lazarus Steel Corporation produces iron rods that are supposed to be 36 inches long. The machine that makes these rods does not produce each rod exactly 36 inches long. The lengths of the rods are normally distributed, and they vary slightly. It is known that when the machine is working properly, the mean length of the rods is 36 inches. The standard deviation of the lengths of all rods produced on this machine is always equal to .035 inch. The quality control department at the company takes a sample of 20 such rods every week, calculates the mean length of these rods, and tests the null hypothesis,  $\mu = 36$  inches, against the alternative hypothesis,  $\mu \neq 36$  inches. If the null hypothesis is rejected, the machine is stopped and adjusted. A recent sample of 20 rods produced a mean length of 36.015 inches.

- Calculate the  $p$ -value for this test of hypothesis. Based on this  $p$ -value, will the quality control inspector decide to stop the machine and adjust it if he chooses the maximum probability of a Type I error to be .02? What if the maximum probability of a Type I error is .10?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .02$ . Does the machine need to be adjusted? What if  $\alpha = .10$ ?

**9.41** At Farmer's Dairy, a machine is set to fill 32-ounce milk cartons. However, this machine does not put exactly 32 ounces of milk into each carton; the amount varies slightly from carton to carton but has a normal distribution. It is known that when the machine is working properly, the mean net weight of these cartons is 32 ounces. The standard deviation of the milk in all such cartons is always equal to .15 ounce. The quality control inspector at this company takes a sample of 25 such cartons every week, calculates the mean net weight of these cartons, and tests the null hypothesis,  $\mu = 32$  ounces, against the alternative hypothesis,  $\mu \neq 32$  ounces. If the null hypothesis is rejected, the machine is stopped and adjusted. A recent sample of 25 such cartons produced a mean net weight of 31.93 ounces.

- Calculate the  $p$ -value for this test of hypothesis. Based on this  $p$ -value, will the quality control inspector decide to stop the machine and readjust it if she chooses the maximum probability of a Type I error to be .01? What if the maximum probability of a Type I error is .05?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .01$ . Does the machine need to be adjusted? What if  $\alpha = .05$ ?

**9.42** According to Moebs Services Inc., an individual checking account at major U.S. banks costs these banks between \$350 and \$450 per year (*Time*, November 21, 2011). Suppose that the average cost of individual checking accounts at major U.S. banks was \$400 for the year 2011. A bank consultant wants to determine whether the current mean cost of such checking accounts at major U.S. banks is more than \$400 a year. A recent random sample of 150 such checking accounts taken from major U.S. banks produced a mean annual cost to them of \$410. Assume that the standard deviation of annual costs to major banks of all such checking accounts is \$60.

- Find the  $p$ -value for the test of hypothesis. Based on this  $p$ -value, would you reject the null hypothesis if the maximum probability of Type I error is to be .05? What if the maximum probability of Type I error is to be .01?
- Test the hypothesis of part a using the critical-value approach and  $\alpha = .05$ . Would you reject the null hypothesis? What if  $\alpha = .01$ ? What if  $\alpha = 0$ ?

**9.43** Records in a three-county area show that in the last few years, Girl Scouts sold an average of 47.93 boxes of cookies per year per girl scout, with a population standard deviation of 8.45 boxes per year. Fifty randomly selected Girl Scouts from the region sold an average of 46.54 boxes this year. Scout leaders are concerned that the demand for Girl Scout cookies may have decreased.

- Test at a 10% significance level whether the average number of boxes of cookies sold by all Girl Scouts in the three-county area is lower than the historical average of 47.93.
- What will your decision be in part a if the probability of a Type I error is zero? Explain.

**9.44** A journalist claims that all adults in her city spend an average of 30 hours or more per month on general reading, such as newspapers, magazines, novels, and so forth. A recent sample of 25 adults from this city showed that they spend an average of 27 hours per month on general reading. The population of such times is known to be normally distributed with the population standard deviation of 7 hours.

- a. Using a 2.5% significance level, would you conclude that the mean time spent per month on such reading by all adults in this city is less than 30 hours? Use both procedures—the  $p$ -value approach and the critical value approach.
- b. Make the test of part a using a 1% significance level. Is your decision different from that of part a? Comment on the results of parts a and b.

**9.45** A study claims that all homeowners in a town spend an average of 8 hours or more on house cleaning and gardening during a weekend. A researcher wanted to check if this claim is true. A random sample of 20 homeowners taken by this researcher showed that they spend an average of 7.68 hours on such chores during a weekend. The population of such times for all homeowners in this town is normally distributed with the population standard deviation of 2.1 hours.

- a. Using a 1% significance level, can you conclude that the claim that all homeowners spend an average of 8 hours or more on such chores during a weekend is false? Use both approaches.
- b. Make the test of part a using a 2.5% significance level. Is your decision different from the one in part a? Comment on the results of parts a and b.

**9.46** A company claims that the mean net weight of the contents of its All Taste cereal boxes is at least 18 ounces. Suppose you want to test whether or not the claim of the company is true. Explain briefly how you would conduct this test using a large sample. Assume that  $\sigma = .25$  ounce.

## 9.3 Hypothesis Tests About $\mu$ : $\sigma$ Not Known

This section explains how to perform a test of hypothesis about the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known. Here, again, there are three possible cases as follows.

**Case I.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is not known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is normally distributed,

then we use the  $t$  distribution to perform a test of hypothesis about  $\mu$ .

**Case II.** If the following two conditions are fulfilled:

1. The population standard deviation  $\sigma$  is not known
2. The sample size is large (i.e.,  $n \geq 30$ ),

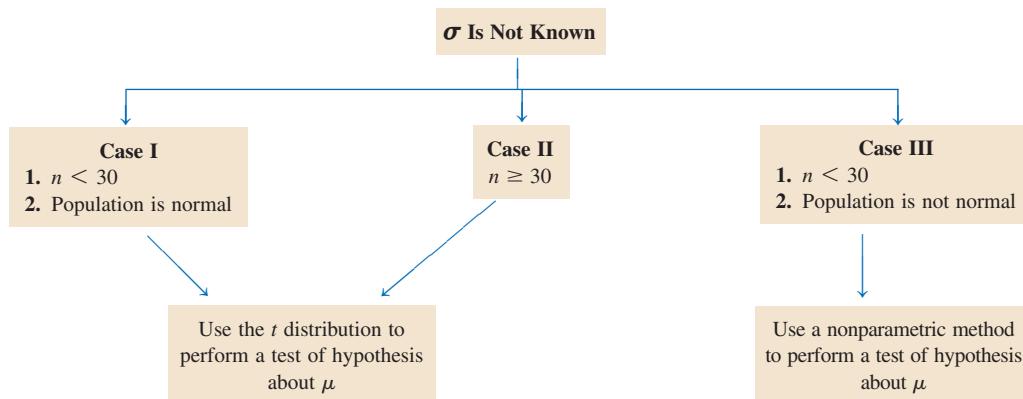
then again we use the  $t$  distribution to perform a test of hypothesis about  $\mu$ .

**Case III.** If the following three conditions are fulfilled:

1. The population standard deviation  $\sigma$  is not known
2. The sample size is small (i.e.,  $n < 30$ )
3. The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown),

then we use a nonparametric method to perform a test of hypothesis about  $\mu$ .

The following chart summarizes the above three cases.



Below we discuss Cases I and II and learn how to use the  $t$  distribution to perform a test of hypothesis about  $\mu$  when  $\sigma$  is not known. When the conditions mentioned for Case I or Case II are satisfied, the random variable

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

has a  $t$  distribution. Here, the  $t$  is called the **test statistic** to perform a test of hypothesis about a population mean  $\mu$ .

**Test Statistic** The value of the *test statistic*  $t$  for the sample mean  $\bar{x}$  is computed as

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The value of  $t$  calculated for  $\bar{x}$  by using this formula is also called the **observed value** of  $t$ .

In Section 9.2, we discussed two procedures, the  $p$ -value approach and the critical-value approach, to test hypotheses about  $\mu$  when  $\sigma$  is known. In this section also we will use these two procedures to test hypotheses about  $\mu$  when  $\sigma$  is not known. The steps used in these procedures are the same as in Section 9.2. The only difference is that we will be using the  $t$  distribution in place of the normal distribution.

### 9.3.1 The $p$ -Value Approach

To use the  $p$ -value approach to perform a test of hypothesis about  $\mu$  using the  $t$  distribution, we will use the same four steps that we used in such a procedure in Section 9.2.1. Although the  $p$ -value can be obtained by using a technology very easily, we can use Table V of Appendix C to find a **range for the  $p$ -value** when technology is not available. Note that when using the  $t$  distribution and Table V, we cannot find the exact  $p$ -value but only a range within which it falls.

Examples 9–5 and 9–6 illustrate the  $p$ -value procedure to test a hypothesis about  $\mu$  using the  $t$  distribution.

### ■ EXAMPLE 9–5

Finding a  $p$ -value and making a decision for a two-tailed test of hypothesis about  $\mu$ :  $\sigma$  not known,  $n < 30$ , and population normal.

A psychologist claims that the mean age at which children start walking is 12.5 months. Carol wanted to check if this claim is true. She took a random sample of 18 children and found that the mean age at which these children started walking was 12.9 months with a standard deviation of .80 month. It is known that the ages at which all children start walking are approximately normally distributed. Find the  $p$ -value for the test that the mean age at which all children start walking is different from 12.5 months. What will your conclusion be if the significance level is 1%?

**Solution** Let  $\mu$  be the mean age at which all children start walking, and let  $\bar{x}$  be the corresponding mean for the sample. From the given information,

$$n = 18, \quad \bar{x} = 12.9 \text{ months}, \quad \text{and} \quad s = .80 \text{ month}$$

The claim of the psychologist is that the mean age at which children start walking is 12.5 months. To calculate the  $p$ -value and to make the decision, we apply the following four steps.

**Step 1. State the null and alternative hypotheses.**

We are to test if the mean age at which all children start walking is different from 12.5 months. Hence, the null and alternative hypotheses are

$$H_0: \mu = 12.5 \quad (\text{The mean walking age is 12.5 months.})$$

$$H_1: \mu \neq 12.5 \quad (\text{The mean walking age is different from 12.5 months.})$$

**Step 2.** Select the distribution to use.

In this example, we do not know the population standard deviation  $\sigma$ , the sample size is small ( $n < 30$ ), and the population is normally distributed. Hence, it is Case I mentioned in the beginning of this section. Consequently, we will use the  $t$  distribution to find the  $p$ -value for this test.

**Step 3.** Calculate the  $p$ -value.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. To find the  $p$ -value, first we find the degrees of freedom and the  $t$  value for  $\bar{x} = 12.9$  months. Then, the  $p$ -value is equal to twice the area in the tail of the  $t$  distribution curve beyond this  $t$  value for  $\bar{x} = 12.9$  months. This  $p$ -value is shown in Figure 9.11. We find this  $p$ -value as follows:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{.80}{\sqrt{18}} = .18856181$$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{12.9 - 12.5}{.18856181} = 2.121$$

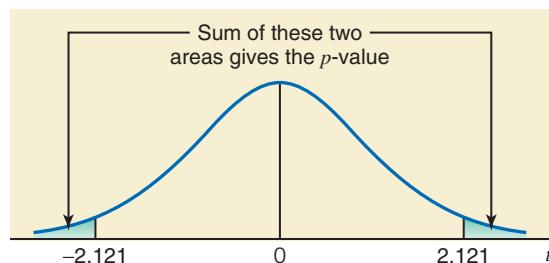
↓ From  $H_0$

and  $df = n - 1 = 18 - 1 = 17$

Now we can find the range for the  $p$ -value. To do so, we go to Table V of Appendix C (the  $t$  distribution table) and find the row of  $df = 17$ . In this row, we find the two values of  $t$  that cover  $t = 2.121$ . From Table V, for  $df = 17$ , these two values of  $t$  are 2.110 and 2.567. The test statistic  $t = 2.121$  falls between these two values. Now look in the top row of this table to find the areas in the tail of the  $t$  distribution curve that correspond to 2.110 and 2.567. These two areas are .025 and .01, respectively. In other words, the area in the upper tail of the  $t$  distribution curve for  $df = 17$  and  $t = 2.110$  is .025, and the area in the upper tail of the  $t$  distribution curve for  $df = 17$  and  $t = 2.567$  is .01. Because it is a two-tailed test, the  $p$ -value for  $t = 2.121$  is between  $2(.025) = .05$  and  $2(.01) = .02$ , which can be written as

$$.02 < p\text{-value} < .05$$

Note that by using Table V of Appendix C, we cannot find the exact  $p$ -value but only a range for it. If we have access to technology, we can find the exact  $p$ -value by using technology. If we use technology for this example, we will obtain a  $p$ -value of .049.



**Figure 9.11** The required  $p$ -value.

**Step 4.** Make a decision.

Thus, we can state that for any  $\alpha$  greater than or equal to .05 (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For any  $\alpha$  less than or equal to .02 (the lower limit of the  $p$ -value range), we will not reject the null hypothesis. However, if  $\alpha$  is between .02 and .05, we cannot make a decision. Note that if we use technology, then the  $p$ -value we will obtain for this example is .049, and we can make a decision for any value of  $\alpha$ . For our example,  $\alpha = .01$ , which is less than the lower limit of the  $p$ -value range of .02. As a result, we fail to reject  $H_0$  and conclude that the mean age at which all children start walking is not significantly different from 12.5 months. As a result, we can state that the difference between the hypothesized population mean and the sample mean is so small that it may have occurred because of sampling error. ■

Finding a *p*-value and making a decision for a left-tailed test of hypothesis about  $\mu$ :  $\sigma$  not known and  $n \geq 30$ .

## ■ EXAMPLE 9–6

Grand Auto Corporation produces auto batteries. The company claims that its top-of-the-line Never Die batteries are good, on average, for at least 65 months. A consumer protection agency tested 45 such batteries to check this claim. It found that the mean life of these 45 batteries is 63.4 months, and the standard deviation is 3 months. Find the *p*-value for the test that the mean life of all such batteries is less than 65 months. What will your conclusion be if the significance level is 2.5%?

**Solution** Let  $\mu$  be the mean life of all such auto batteries, and let  $\bar{x}$  be the corresponding mean for the sample. From the given information,

$$n = 45, \quad \bar{x} = 63.4 \text{ months}, \quad \text{and} \quad s = 3 \text{ months}$$

The claim of the company is that the mean life of these batteries is at least 65 months. To calculate the *p*-value and to make the decision, we apply the following four steps.

**Step 1. State the null and alternative hypotheses.**

We are to test if the mean life of these batteries is at least 65 months. Hence, the null and alternative hypotheses are

$$\begin{aligned} H_0: \mu &\geq 65 && (\text{The mean life of batteries is at least 65 months.}) \\ H_1: \mu &< 65 && (\text{The mean life of batteries is less than 65 months.}) \end{aligned}$$

**Step 2. Select the distribution to use.**

In this example, we do not know the population standard deviation  $\sigma$ , and the sample size is large ( $n \geq 30$ ). Hence, it is Case II mentioned in the beginning of this section. Consequently, we will use the *t* distribution to find the *p*-value for this test.

**Step 3. Calculate the *p*-value.**

The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed. To find the *p*-value, first we find the degrees of freedom and the *t* value for  $\bar{x} = 63.4$  months. Then, the *p*-value is given by the area in the tail of the *t* distribution curve beyond this *t* value for  $\bar{x} = 63.4$  months. This *p*-value is shown in Figure 9.12. We find this *p*-value as follows:

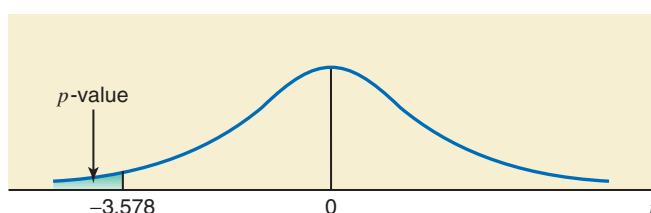
$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{45}} = .44721360$$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{63.4 - 65}{.44721360} = -3.578$$

and

$$df = n - 1 = 45 - 1 = 44$$

Now we can find the range for the *p*-value. To do so, we go to Table V of Appendix C (the *t* distribution table) and find the row of  $df = 44$ . In this row, we find the two values of *t* that cover  $t = 3.578$ . Note that we use the positive value of the test statistic *t*, although our test statistic has a negative value. From Table V, for  $df = 44$ , the largest value of *t* is 3.286, for which the area in the tail of the *t* distribution is .001. This means that the area to the left of  $t = -3.286$  is .001. Because  $-3.578$  is smaller than  $-3.286$ , the area to the left of  $t = -3.578$



**Figure 9.12** The required *p*-value.

is smaller than .001. Therefore, the  $p$ -value for  $t = -3.578$  is less than .001, which can be written as

$$p\text{-value} < .001$$

Thus, here the  $p$ -value has only the upper limit of .001. In other words, the  $p$ -value for this example is less than .001. If we use technology for this example, we will obtain a  $p$ -value of .00043.

#### Step 4. Make a decision.

Thus, we can state that for any  $\alpha$  greater than or equal to .001 (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For our example  $\alpha = .025$ , which is greater than the upper limit of the  $p$ -value of .001. As a result, we reject  $H_0$  and conclude that the mean life of such batteries is less than 65 months. Therefore, we can state that the difference between the hypothesized population mean of 65 months and the sample mean of 63.4 is too large to be attributed to sampling error alone. ■

### 9.3.2 The Critical-Value Approach

In this procedure, as mentioned in Section 9.2.2, we have a predetermined value of the significance level  $\alpha$ . The value of  $\alpha$  gives the total area of the rejection region(s). First we find the critical value(s) of  $t$  from the  $t$  distribution table in Appendix C for the given degrees of freedom and the significance level. Then we find the value of the test statistic  $t$  for the observed value of the sample statistic  $\bar{x}$ . Finally we compare these two values and make a decision. Remember, if the test is one-tailed, there is only one critical value of  $t$ , and it is obtained by using the value of  $\alpha$ , which gives the area in the left or right tail of the  $t$  distribution curve, depending on whether the test is left-tailed or right-tailed, respectively. However, if the test is two-tailed, there are two critical values of  $t$ , and they are obtained by using  $\alpha/2$  area in each tail of the  $t$  distribution curve. The value of the test statistic  $t$  is obtained as mentioned earlier in this section.

Examples 9–7 and 9–8 describe the procedure to test a hypothesis about  $\mu$  using the critical-value approach and the  $t$  distribution.

### ■ EXAMPLE 9–7

Refer to Example 9–5. A psychologist claims that the mean age at which children start walking is 12.5 months. Carol wanted to check if this claim is true. She took a random sample of 18 children and found that the mean age at which these children started walking was 12.9 months with a standard deviation of .80 month. Using a 1% significance level, can you conclude that the mean age at which all children start walking is different from 12.5 months? Assume that the ages at which all children start walking have an approximate normal distribution.

Conducting a two-tailed test of hypothesis about  $\mu$ :  $\sigma$  unknown,  $n < 30$ , and population normal.

**Solution** Let  $\mu$  be the mean age at which all children start walking, and let  $\bar{x}$  be the corresponding mean for the sample. Then, from the given information,

$$n = 18, \quad \bar{x} = 12.9 \text{ months}, \quad s = .80 \text{ month}, \quad \text{and} \quad \alpha = .01$$

#### Step 1. State the null and alternative hypotheses.

We are to test if the mean age at which all children start walking is different from 12.5 months. The null and alternative hypotheses are

$$H_0: \mu = 12.5 \quad (\text{The mean walking age is 12.5 months.})$$

$$H_1: \mu \neq 12.5 \quad (\text{The mean walking age is different from 12.5 months.})$$

#### Step 2. Select the distribution to use.

In this example, the population standard deviation  $\sigma$  is not known, the sample size is small ( $n < 30$ ), and the population is normally distributed. Hence, it is Case I mentioned in the beginning of Section 9.3. Consequently, we will use the  $t$  distribution to perform the test in this example.



Cohen Ostrow/Digital Vision/Getty Images

**Step 3.** Determine the rejection and nonrejection regions.

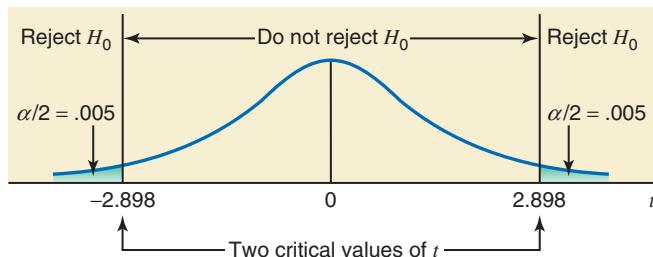
The significance level is .01. The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed and the rejection region lies in both tails. The area of the rejection region in each tail of the  $t$  distribution curve is

$$\text{Area in each tail} = \alpha/2 = .01/2 = .005$$

$$df = n - 1 = 18 - 1 = 17$$

From the  $t$  distribution table, the critical values of  $t$  for 17 degrees of freedom and .005 area in each tail of the  $t$  distribution curve are  $-2.898$  and  $2.898$ . These values are shown in Figure 9.13.

**Figure 9.13** The critical values of  $t$ .

**Step 4.** Calculate the value of the test statistic.

We calculate the value of the test statistic  $t$  for  $\bar{x} = 12.9$  as follows:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{.80}{\sqrt{18}} = .18856181$$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{12.9 - 12.5}{.18856181} = 2.121$$

From  $H_0$

**Step 5.** Make a decision.

The value of the test statistic  $t = 2.121$  falls between the two critical points,  $-2.898$  and  $2.898$ , which is the nonrejection region. Consequently, we fail to reject  $H_0$ . As a result, we can state that the difference between the hypothesized population mean and the sample mean is so small that it may have occurred because of sampling error. The mean age at which children start walking is not significantly different from 12.5 months. ■

### ■ EXAMPLE 9–8

Conducting a right-tailed test of hypothesis about  $\mu$ :  $\sigma$  unknown and  $n \geq 30$ .



PhotoDisc, Inc./Getty Images

The management at Massachusetts Savings Bank is always concerned about the quality of service provided to its customers. With the old computer system, a teller at this bank could serve, on average, 22 customers per hour. The management noticed that with this service rate, the waiting time for customers was too long. Recently the management of the bank installed a new computer system, expecting that it would increase the service rate and consequently make the customers happier by reducing the waiting time. To check if the new computer system is more efficient than the old system, the management of the bank took a random sample of 70 hours and found that during these hours the mean number of customers served by tellers was 27 per hour with a standard deviation of 2.5. Testing at a 1% significance level, would you conclude that the new computer system is more efficient than the old computer system?

**Solution** Let  $\mu$  be the mean number of customers served per hour by a teller using the new system, and let  $\bar{x}$  be the corresponding mean for the sample. Then, from the given information,

$$n = 70 \text{ hours}, \quad \bar{x} = 27 \text{ customers}, \quad s = 2.5 \text{ customers}, \quad \text{and } \alpha = .01$$

**Step 1.** State the null and alternative hypotheses.

We are to test whether or not the new computer system is more efficient than the old system. The new computer system will be more efficient than the old system if the mean number of customers served per hour by using the new computer system is significantly more than 22; otherwise, it will not be more efficient. The null and alternative hypotheses are

$$H_0: \mu = 22 \quad (\text{The new computer system is not more efficient.})$$

$$H_1: \mu > 22 \quad (\text{The new computer system is more efficient.})$$

**Step 2.** Select the distribution to use.

In this example, the population standard deviation  $\sigma$  is not known and the sample size is large ( $n \geq 30$ ). Hence, it is Case II mentioned in the beginning of Section 9.3. Consequently, we will use the  $t$  distribution to perform the test for this example.

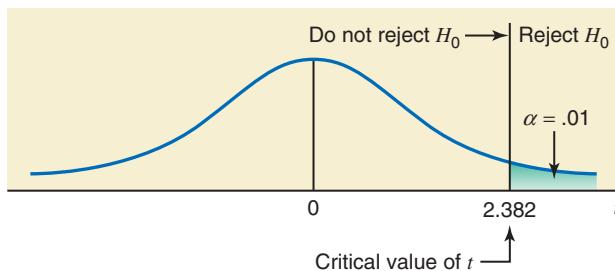
**Step 3.** Determine the rejection and nonrejection regions.

The significance level is .01. The  $>$  sign in the alternative hypothesis indicates that the test is right-tailed and the rejection region lies in the right tail of the  $t$  distribution curve.

$$\text{Area in the right tail} = \alpha = .01$$

$$df = n - 1 = 70 - 1 = 69$$

From the  $t$  distribution table, the critical value of  $t$  for 69 degrees of freedom and .01 area in the right tail is 2.382. This value is shown in Figure 9.14.



**Figure 9.14** The critical value of  $t$ .

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $\bar{x} = 27$  is calculated as follows:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2.5}{\sqrt{70}} = .29880715$$

From  $H_0$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{27 - 22}{.29880715} = 16.733$$

**Step 5.** Make a decision.

The value of the test statistic  $t = 16.733$  is greater than the critical value of  $t = 2.382$ , and it falls in the rejection region. Consequently, we reject  $H_0$ . As a result, we conclude that the value of the sample mean is too large compared to the hypothesized value of the population mean, and the difference between the two may not be attributed to chance alone. The mean number of customers served per hour using the new computer system is more than 22. The new computer system is more efficient than the old computer system. ■

**Note: What If the Sample Size Is Large and the Number of df Is Not In the t Distribution Table?**

In the above section when  $\sigma$  is not known, we used the  $t$  distribution to perform tests of hypothesis about  $\mu$  in Cases I and II. Note that in Case II, the sample size is large. If we have access to technology, it does not matter how large the sample size is, we can always use the

*t* distribution. However, if we are using the *t* distribution table (Table V of Appendix C), this may pose a problem. Usually such a table only goes up to a certain number of degrees of freedom. For example, Table V in Appendix C only goes up to 75 degrees of freedom. Thus, if the sample size is larger than 76 (with *df* more than 75) here, we cannot use Table V to find the critical value(s) of *t* to make a decision in this section. In such a situation when *n* is large and is not included in the *t* distribution table, there are two options:

1. Use the *t* value from the last row (the row of  $\infty$ ) in Table V of Appendix C.
2. Use the normal distribution as an approximation to the *t* distribution.

To use the normal distribution as an approximation to the *t* distribution to make a test of hypothesis about  $\mu$ , the procedure is exactly like the one in Section 9.2, except that now we will replace  $\sigma$  by  $s$ , and  $\sigma_{\bar{x}}$  by  $s_{\bar{x}}$ .

Note that the *t* values obtained from the last row of the *t* distribution table are the same as will be obtained from the normal distribution table for the same areas in the upper tail or lower tail of the distribution. Again, note that here we can use the normal distribution as a convenience and as an approximation, but if we can, we should use the *t* distribution by using technology. Exercise 9.71 at the end of this section presents such a situation.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**9.47** Briefly explain the conditions that must hold true to use the *t* distribution to make a test of hypothesis about the population mean.

**9.48** For each of the following examples of tests of hypothesis about  $\mu$ , show the rejection and nonrejection regions on the *t* distribution curve.

- a. A two-tailed test with  $\alpha = .02$  and  $n = 20$
- b. A left-tailed test with  $\alpha = .01$  and  $n = 16$
- c. A right-tailed test with  $\alpha = .05$  and  $n = 18$

**9.49** For each of the following examples of tests of hypothesis about  $\mu$ , show the rejection and nonrejection regions on the *t* distribution curve.

- a. A two-tailed test with  $\alpha = .01$  and  $n = 15$
- b. A left-tailed test with  $\alpha = .005$  and  $n = 25$
- c. A right-tailed test with  $\alpha = .025$  and  $n = 22$

**9.50** A random sample of 14 observations taken from a population that is normally distributed produced a sample mean of 212.37 and a standard deviation of 16.35. Find the critical and observed values of *t* and the ranges for the *p*-value for each of the following tests of hypotheses, using  $\alpha = .10$ .

- a.  $H_0: \mu = 205$  versus  $H_1: \mu \neq 205$
- b.  $H_0: \mu = 205$  versus  $H_1: \mu > 205$

**9.51** A random sample of 8 observations taken from a population that is normally distributed produced a sample mean of 44.98 and a standard deviation of 6.77. Find the critical and observed values of *t* and the ranges for the *p*-value for each of the following tests of hypotheses, using  $\alpha = .05$ .

- a.  $H_0: \mu = 50$  versus  $H_1: \mu \neq 50$
- b.  $H_0: \mu = 50$  versus  $H_1: \mu < 50$

**9.52** Consider the null hypothesis  $H_0: \mu = 100$ . Suppose that a random sample of 35 observations is taken from this population to perform this test. Using a significance level of .01, show the rejection and nonrejection regions and find the critical value(s) of *t* when the alternative hypothesis is as follows.

- a.  $H_1: \mu \neq 100$
- b.  $H_1: \mu > 100$
- c.  $H_1: \mu < 100$

**9.53** Consider the null hypothesis  $H_0: \mu = 12.80$ . A random sample of 58 observations is taken from this population to perform this test. Using  $\alpha = .05$ , show the rejection and nonrejection regions on the sampling distribution curve of the sample mean and find the critical value(s) of *t* for the following.

- a. a right-tailed test
- b. a left-tailed test
- c. a two-tailed test

**9.54** Consider  $H_0: \mu = 80$  versus  $H_1: \mu \neq 80$  for a population that is normally distributed.

- a. A random sample of 25 observations taken from this population produced a sample mean of 77 and a standard deviation of 8. Using  $\alpha = .01$ , would you reject the null hypothesis?

- b.** Another random sample of 25 observations taken from the same population produced a sample mean of 86 and a standard deviation of 6. Using  $\alpha = .01$ , would you reject the null hypothesis?

Comment on the results of parts a and b.

- 9.55** Consider  $H_0: \mu = 40$  versus  $H_1: \mu > 40$ .

- A random sample of 64 observations taken from this population produced a sample mean of 43 and a standard deviation of 5. Using  $\alpha = .025$ , would you reject the null hypothesis?
- Another random sample of 64 observations taken from the same population produced a sample mean of 41 and a standard deviation of 7. Using  $\alpha = .025$ , would you reject the null hypothesis?

Comment on the results of parts a and b.

- 9.56** Perform the following tests of hypothesis.

- $H_0: \mu = 285, H_1: \mu < 285, n = 55, \bar{x} = 267.80, s = 42.90, \alpha = .05$
- $H_0: \mu = 10.70, H_1: \mu \neq 10.70, n = 47, \bar{x} = 12.025, s = 4.90, \alpha = .01$
- $H_0: \mu = 147,500, H_1: \mu > 147,500, n = 41, \bar{x} = 149,812, s = 22,972, \alpha = .10$

- 9.57** Perform the following tests of hypotheses for data coming from a normal distribution.

- $H_0: \mu = 94.80, H_1: \mu < 94.80, n = 12, \bar{x} = 92.87, s = 5.34, \alpha = .10$
- $H_0: \mu = 18.70, H_1: \mu \neq 18.70, n = 25, \bar{x} = 20.05, s = 2.99, \alpha = .05$
- $H_0: \mu = 59, H_1: \mu > 59, n = 7, \bar{x} = 59.42, s = .418, \alpha = .01$

## ■ APPLICATIONS

- 9.58** The police that patrol a heavily traveled highway claim that the average driver exceeds the 65 miles per hour speed limit by more than 10 miles per hour. Seventy-two randomly selected cars were clocked by airplane radar. The average speed was 77.40 miles per hour, and the standard deviation of the speeds was 5.90 miles per hour. Find the range for the  $p$ -value for this test. What will your conclusion be using this  $p$ -value range and  $\alpha = .02$ ?

- 9.59** According to an estimate, the average age at first marriage for women in the United States was 26.1 years in 2010 (*Time*, March 21, 2011). A recent sample of 60 women from New Jersey who got married for the first time this year showed that their average age at first marriage was 27.2 years with a standard deviation of 3.5 years. Using a 2.5% significance level and the critical-value approach, can you conclude that the average age for women in New Jersey who got married for the first time this year is higher than 26.1 years? Find the range for the  $p$ -value for this test. What will your conclusion be using this  $p$ -value range and  $\alpha = .025$ ?

- 9.60** The president of a university claims that the mean time spent partying by all students at this university is not more than 7 hours per week. A random sample of 40 students taken from this university showed that they spent an average of 9.50 hours partying the previous week with a standard deviation of 2.3 hours. Test at a 2.5% significance level whether the president's claim is true. Explain your conclusion in words.

- 9.61** The mean balance of all checking accounts at a bank on December 31, 2011, was \$850. A random sample of 55 checking accounts taken recently from this bank gave a mean balance of \$780 with a standard deviation of \$230. Using a 1% significance level, can you conclude that the mean balance of such accounts has decreased during this period? Explain your conclusion in words. What if  $\alpha = .025$ ?

- 9.62** A soft-drink manufacturer claims that its 12-ounce cans do not contain, on average, more than 30 calories. A random sample of 64 cans of this soft drink, which were checked for calories, contained a mean of 32 calories with a standard deviation of 3 calories. Does the sample information support the alternative hypothesis that the manufacturer's claim is false? Use a significance level of 5%. Find the range for the  $p$ -value for this test. What will your conclusion be using this  $p$ -value and  $\alpha = .05$ ?

- 9.63** According to an estimate, the average price of homes in Martha's Vineyard, Massachusetts, was \$650,000 in 2011 (*USA TODAY*, August 11, 2011). A recent random sample of 70 homes from Martha's Vineyard showed that their average price is \$674,000 with a standard deviation of \$94,500. Using a 2% significance level, can you conclude that the current average price of homes in Martha's Vineyard is different from \$650,000? Use both the  $p$ -value and critical-value approaches.

- 9.64** A paint manufacturing company claims that the mean drying time for its paints is not longer than 45 minutes. A random sample of 20 gallons of paints selected from the production line of this company showed that the mean drying time for this sample is 49.50 minutes with a standard deviation of 3 minutes. Assume that the drying times for these paints have a normal distribution.

- Using a 1% significance level, would you conclude that the company's claim is true?
- What is the Type I error in this exercise? Explain in words. What is the probability of making such an error?

**9.65** The manager of a restaurant in a large city claims that waiters working in all restaurants in his city earn an average of \$150 or more in tips per week. A random sample of 25 waiters selected from restaurants of this city yielded a mean of \$139 in tips per week with a standard deviation of \$28. Assume that the weekly tips for all waiters in this city have a normal distribution.

- Using a 1% significance level, can you conclude that the manager's claim is true? Use both approaches.
- What is the Type I error in this exercise? Explain. What is the probability of making such an error?

**9.66** A business school claims that students who complete a 3-month typing course can type, on average, at least 1200 words an hour. A random sample of 25 students who completed this course typed, on average, 1125 words an hour with a standard deviation of 85 words. Assume that the typing speeds for all students who complete this course have an approximate normal distribution.

- Suppose the probability of making a Type I error is selected to be zero. Can you conclude that the claim of the business school is true? Answer without performing the five steps of a test of hypothesis.
- Using a 5% significance level, can you conclude that the claim of the business school is true? Use both approaches.

**9.67** According to an estimate, 2 years ago the average age of all CEOs of medium-sized companies in the United States was 58 years. Jennifer wants to check if this is still true. She took a random sample of 70 such CEOs and found their mean age to be 55 years with a standard deviation of 6 years.

- Suppose that the probability of making a Type I error is selected to be zero. Can you conclude that the current mean age of all CEOs of medium-sized companies in the Untied States is different from 58 years?
- Using a 1% significance level, can you conclude that the current mean age of all CEOs of medium-sized companies in the United States is different from 58 years? Use both approaches.

**9.68** A past study claimed that adults in America spent an average of 18 hours a week on leisure activities. A researcher wanted to test this claim. She took a sample of 12 adults and asked them about the time they spend per week on leisure activities. Their responses (in hours) are as follows.

13.6    14.0    24.5    24.6    22.9    37.7    14.6    14.5    21.5    21.0    17.8    21.4

Assume that the times spent on leisure activities by all American adults are normally distributed. Using a 10% significance level, can you conclude that the average amount of time spent by American adults on leisure activities has changed? (*Hint:* First calculate the sample mean and the sample standard deviation for these data using the formulas learned in Sections 3.1.1 and 3.2.2 of Chapter 3. Then make the test of hypothesis about  $\mu$ .)

**9.69** The past records of a supermarket show that its customers spend an average of \$95 per visit at this store. Recently the management of the store initiated a promotional campaign according to which each customer receives points based on the total money spent at the store, and these points can be used to buy products at the store. The management expects that as a result of this campaign, the customers should be encouraged to spend more money at the store. To check whether this is true, the manager of the store took a sample of 14 customers who visited the store. The following data give the money (in dollars) spent by these customers at this supermarket during their visits.

109.15	136.01	107.02	116.15	101.53	109.29	110.79
94.83	100.91	97.94	104.30	83.54	67.59	120.44

Assume that the money spent by all customers at this supermarket has a normal distribution. Using a 5% significance level, can you conclude that the mean amount of money spent by all customers at this supermarket after the campaign was started is more than \$95? (*Hint:* First calculate the sample mean and the sample standard deviation for these data using the formulas learned in Sections 3.1.1 and 3.2.2 of Chapter 3. Then make the test of hypothesis about  $\mu$ .)

**9.70** According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of \$4129 for family health insurance coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose a recent random sample of 25 workers with employer-provided health insurance selected from a city paid an average premium of \$4517 for family health insurance coverage with a standard deviation of \$580. Assume that such premiums paid by all such workers in this city are normally distributed. Does the sample information support the alternative hypothesis that the average premium for such coverage paid by all such workers in this city is different from \$4129? Use a 5% significance level. Use both the  $p$ -value approach and the critical-value approach.

**9.71** According to an estimate, the average total parent and student debt for new college graduates was \$34,400 in 2010–11 (*Time*, October 31, 2011). A random sample of 500 of this year's graduates showed that their average such debt is \$38,460 with a standard deviation of \$5600. Do the data provide significant evidence at a 1% significance level to conclude that the current average total parent and student debt for new graduates is higher than \$34,400? Use both the  $p$ -value approach and the critical-value approach.

\***9.72** The manager of a service station claims that the mean amount spent on gas by its customers is \$15.90 per visit. You want to test if the mean amount spent on gas at this station is different from \$15.90 per visit. Briefly explain how you would conduct this test when  $\sigma$  is not known.

\***9.73** A tool manufacturing company claims that its top-of-the-line machine that is used to manufacture bolts produces an average of 88 or more bolts per hour. A company that is interested in buying this machine wants to check this claim. Suppose you are asked to conduct this test. Briefly explain how you would do so when  $\sigma$  is not known.

## 9.4 Hypothesis Tests About a Population Proportion: Large Samples

Often we want to conduct a test of hypothesis about a population proportion. For example, according to a 2011 *Time/Money Magazine* survey, 70% of Americans of age 18 years and older said that they had cut back on vacation and entertainment due to bad economic conditions (*Time*, October 10, 2011). An economist may want to check whether this percentage has changed since 2011. As another example, a mail-order company claims that 90% of all orders it receives are shipped within 72 hours. The company's management may want to determine from time to time whether or not this claim is true.

This section presents the procedure to perform tests of hypotheses about the population proportion,  $p$ , for large samples. The procedures to make such tests are similar in many respects to the ones for the population mean,  $\mu$ . Again, the test can be two-tailed or one-tailed. We know from Chapter 7 that when the sample size is large, the sample proportion,  $\hat{p}$ , is approximately normally distributed with its mean equal to  $p$  and standard deviation equal to  $\sqrt{pq/n}$ . Hence, we use the normal distribution to perform a test of hypothesis about the population proportion,  $p$ , for a large sample. As was mentioned in Chapters 7 and 8, in the case of a proportion, the sample size is considered to be large when  $np$  and  $nq$  are both greater than 5.

**Test Statistic** The value of the *test statistic*  $z$  for the sample proportion,  $\hat{p}$ , is computed as

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \quad \text{where} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

The value of  $p$  that is used in this formula is the one from the null hypothesis. The value of  $q$  is equal to  $1 - p$ .

The value of  $z$  calculated for  $\hat{p}$  using the above formula is also called the **observed value of  $z$** .

In Section 9.2, we discussed two procedures, the *p*-value approach and the critical-value approach, to test hypotheses about  $\mu$ . Here too we will use these two procedures to test hypotheses about  $p$ . The steps used in these procedures are the same as in Section 9.2. The only difference is that we will be making tests of hypotheses about  $p$  rather than about  $\mu$ .

### 9.4.1 The *p*-Value Approach

To use the *p*-value approach to perform a test of hypothesis about  $p$ , we will use the same four steps that we used in such a procedure in Section 9.2. Although the *p*-value for a test of hypothesis about  $p$  can be obtained very easily by using technology, we can use Table IV of Appendix C to find this *p*-value when technology is not available.

Examples 9–9 and 9–10 illustrate the *p*-value procedure to test a hypothesis about  $p$  for a large sample.

#### ■ EXAMPLE 9–9

In a 2011 National Institute on Alcohol Abuse and Alcoholism survey, 33% of American adults said that they had never consumed alcohol (*USA TODAY*, November 17, 2011). Suppose that this result is true for the 2011 population of American adults. In a recent random sample of 2300 adult Americans, 35% said that they had never consumed alcohol. Find the *p*-value to test the hypothesis that the current percentage of American adults who have never consumed alcohol is different from 33%. What is your conclusion if the significance level is 5%?

*Finding a p-value and making a decision for a two-tailed test of hypothesis about  $p$ : large sample.*

**Solution** Let  $p$  be the current proportion of all American adults who have never consumed alcohol and  $\hat{p}$  be the corresponding sample proportion. Then, from the given information,

$$n = 2300, \quad \hat{p} = .35, \quad \text{and} \quad \alpha = .05$$

In 2011, 33% of American adults said that they had never consumed alcohol. Hence,

$$p = .33 \quad \text{and} \quad q = 1 - p = 1 - .33 = .67$$

To calculate the  $p$ -value and to make a decision, we apply the following four steps.

**Step 1. State the null and alternative hypotheses.**

The current percentage of American adults who have never consumed alcohol will not be different from 33% if  $p = .33$ , and the current percentage will be different from 33% if  $p \neq .33$ . The null and alternative hypotheses are as follows:

$$H_0: p = .33 \quad (\text{The current percentage is not different from 33%})$$

$$H_1: p \neq .33 \quad (\text{The current percentage is different from 33%})$$

**Step 2. Select the distribution to use.**

To check whether the sample is large, we calculate the values of  $np$  and  $nq$ .

$$np = 2300(.33) = 759 \quad \text{and} \quad nq = 2300(.67) = 1541$$

Since  $np$  and  $nq$  are both greater than 5, we can conclude that the sample size is large. Consequently, we will use the normal distribution to find the  $p$ -value for this test.

**Step 3 Calculate the  $p$ -value.**

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The  $p$ -value is equal to twice the area in the tail of the normal distribution curve to the right of  $z$  for  $\hat{p} = .35$ . This  $p$ -value is shown in Figure 9.15. To find this  $p$ -value, first we find the test statistic  $z$  for  $\hat{p} = .35$  as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.33)(.67)}{2300}} = .00980461$$

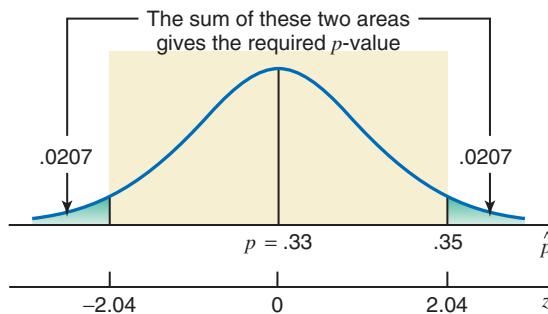
$\downarrow$  From  $H_0$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.35 - .33}{.00980461} = 2.04$$

Now we find the area to the right of  $z = 2.04$  from the normal distribution table. This area is  $1 - .9793 = .0207$ . Consequently, the  $p$ -value is

$$p\text{-value} = 2(.0207) = .0414$$

**Figure 9.15** The required  $p$ -value.



**Step 4. Make a decision.**

Thus, we can state that for any  $\alpha$  greater than or equal to .0414 we will reject the null hypothesis, and for any  $\alpha$  less than .0414 we will not reject the null hypothesis. In our example

$\alpha = .05$ , which is greater than the  $p$ -value of .0414. As a result, we reject  $H_0$  and conclude that the current percentage of American adults who have never consumed alcohol is significantly different from .33. Consequently, we can state that the difference between the hypothesized population proportion of .33 and the sample proportion of .35 is too large to be attributed to sampling error alone when  $\alpha = .05$ . ■

## ■ EXAMPLE 9–10

When working properly, a machine that is used to make chips for calculators does not produce more than 4% defective chips. Whenever the machine produces more than 4% defective chips, it needs an adjustment. To check if the machine is working properly, the quality control department at the company often takes samples of chips and inspects them to determine if they are good or defective. One such random sample of 200 chips taken recently from the production line contained 12 defective chips. Find the  $p$ -value to test the hypothesis whether or not the machine needs an adjustment. What would your conclusion be if the significance level is 2.5%?

*Finding a p-value and making a decision for a right-tailed test of hypothesis about  $p$ : large sample.*

**Solution** Let  $p$  be the proportion of defective chips in all chips produced by this machine, and let  $\hat{p}$  be the corresponding sample proportion. Then, from the given information,

$$n = 200, \quad \hat{p} = 12/200 = .06, \quad \text{and} \quad \alpha = .025$$

When the machine is working properly, it does not produce more than 4% defective chips. Hence, assuming that the machine is working properly, we obtain

$$p = .04 \quad \text{and} \quad q = 1 - p = 1 - .04 = .96$$

To calculate the  $p$ -value and to make a decision, we apply the following four steps.

**Step 1.** *State the null and alternative hypotheses.*

The machine will not need an adjustment if the percentage of defective chips is 4% or less, and it will need an adjustment if this percentage is greater than 4%. Hence, the null and alternative hypotheses are as follows:

$$H_0: p \leq .04 \quad (\text{The machine does not need an adjustment.})$$

$$H_1: p > .04 \quad (\text{The machine needs an adjustment.})$$

**Step 2.** *Select the distribution to use.*

To check if the sample is large, we calculate the values of  $np$  and  $nq$ :

$$np = 200(.04) = 8 \quad \text{and} \quad nq = 200(.96) = 192$$

Since  $np$  and  $nq$  are both greater than 5, we can conclude that the sample size is large. Consequently, we will use the normal distribution to find the  $p$ -value for this test.

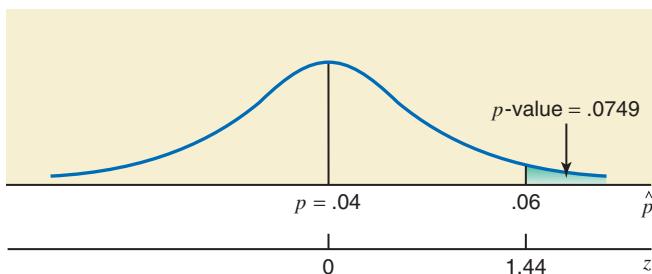
**Step 3.** *Calculate the p-value.*

The  $>$  sign in the alternative hypothesis indicates that the test is right-tailed. The  $p$ -value is given by the area in the upper tail of the normal distribution curve to the right of  $z$  for  $\hat{p} = .06$ . This  $p$ -value is shown in Figure 9.16. To find this  $p$ -value, first we find the test statistic  $z$  for  $\hat{p} = .06$  as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.04)(.96)}{200}} = .01385641$$

$\downarrow$  From  $H_0$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.06 - .04}{.01385641} = 1.44$$

**Figure 9.16** The required  $p$ -value.

Now we find the area to the right of  $z = 1.44$  from the normal distribution table. This area is  $1 - .9251 = .0749$ . Consequently, the  $p$ -value is

$$p\text{-value} = .0749$$

**Step 4. Make a decision.**

Thus, we can state that for any  $\alpha$  greater than or equal to  $.0749$  we will reject the null hypothesis, and for any  $\alpha$  less than  $.0749$  we will not reject the null hypothesis. For our example,  $\alpha = .025$ , which is less than the  $p$ -value of  $.0749$ . As a result, we fail to reject  $H_0$  and conclude that the machine does not need an adjustment. ■

### 9.4.2 The Critical-Value Approach

In this procedure, as mentioned in Section 9.2.2, we have a predetermined value of the significance level  $\alpha$ . The value of  $\alpha$  gives the total area of the rejection region(s). First we find the critical value(s) of  $z$  from the normal distribution table for the given significance level. Then we find the value of the test statistic  $z$  for the observed value of the sample statistic  $\hat{p}$ . Finally we compare these two values and make a decision. Remember, if the test is one-tailed, there is only one critical value of  $z$ , and it is obtained by using the value of  $\alpha$ , which gives the area in the left or right tail of the normal distribution curve, depending on whether the test is left-tailed or right-tailed, respectively. However, if the test is two-tailed, there are two critical values of  $z$ , and they are obtained by using  $\alpha/2$  area in each tail of the normal distribution curve. The value of the test statistic  $z$  is obtained as mentioned earlier in this section.

Examples 9–11 and 9–12 describe the procedure to test a hypothesis about  $p$  using the critical-value approach and the normal distribution.

#### ■ EXAMPLE 9–11

Making a two-tailed test of hypothesis about  $p$  using the critical-value approach: large sample

Refer to Example 9–9. In a 2011 National Institute on Alcohol Abuse and Alcoholism survey, 33% of American adults said that they had never consumed alcohol (*USA TODAY*, November 17, 2011). Suppose this result is true for the 2011 population of American adults. In a recent random sample of 2300 adult Americans, 35% said that they have never consumed alcohol. Using a 5% significance level, can you conclude that the current percentage of American adults who have never consumed alcohol is different from 33%?

**Solution** Let  $p$  be the current proportion of all American adults who have never consumed alcohol and  $\hat{p}$  be the corresponding sample proportion. Then, from the given information,

$$n = 2300, \quad \hat{p} = .35, \quad \text{and} \quad \alpha = .05$$

In 2011, 33% of American adults said that they had never consumed alcohol. Hence,

$$p = .33 \quad \text{and} \quad q = 1 - p = 1 - .33 = .67$$

To use the critical-value approach to perform a test of hypothesis, we apply the following five steps.

**Step 1.** State the null and alternative hypotheses.

The current percentage of American adults who have never consumed alcohol will not be different from 33% if  $p = .33$ , and the current percentage will be different from 33% if  $p \neq .33$ . The null and alternative hypotheses are as follows:

$$H_0: p = .33 \quad (\text{The current percentage is not different from } 33\%.)$$

$$H_1: p \neq .33 \quad (\text{The current percentage is different from } 33\%.)$$

**Step 2.** Select the distribution to use.

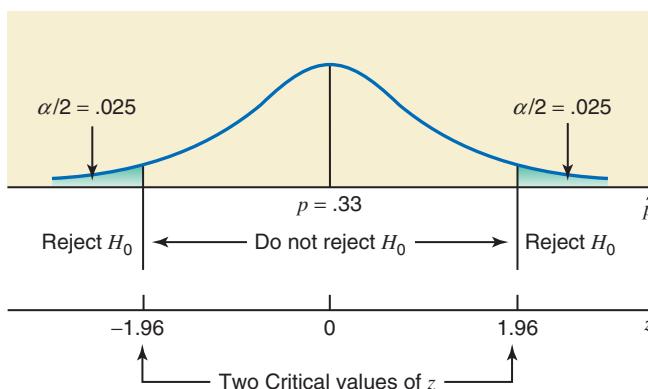
To check if the sample is large, we calculate the values of  $np$  and  $nq$ .

$$np = 2300(.33) = 759 \quad \text{and} \quad nq = 2300(.67) = 1541$$

Since  $np$  and  $nq$  are both greater than 5, we can conclude that the sample size is large. Consequently, we will use the normal distribution to make the test.

**Step 3.** Determine the rejection and nonrejection regions.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The significance level is .05. Therefore, the total area of the two rejection regions is .05, and the rejection region in each tail of the sampling distribution of  $\hat{p}$  is  $\alpha/2 = .05/2 = .025$ . The critical values of  $z$ , obtained from the standard normal distribution table, are  $-1.96$  and  $1.96$ , as shown in Figure 9.17.



**Figure 9.17** The critical values of  $z$ .

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $z$  for  $\hat{p} = .35$  is calculated as follows.

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.33)(.67)}{2300}} = .00980461$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.35 - .33}{.00980461} = 2.04$$

**Step 5.** Make a decision.

The value of the test statistic  $z = 2.04$  for  $\hat{p}$  falls in the rejection region. As a result, we reject  $H_0$  and conclude that the current percentage of American adults who have never consumed alcohol is significantly different from .33. Consequently, we can state that the difference between the hypothesized population proportion of .33 and the sample proportion of .35 is too large to be attributed to sampling error alone when  $\alpha = .05$ . ■

*Conducting a left-tailed test of hypothesis about  $p$  using the critical-value approach: large sample.*

### ■ EXAMPLE 9-12

Direct Mailing Company sells computers and computer parts by mail. The company claims that at least 90% of all orders are mailed within 72 hours after they are received. The quality control department at the company often takes samples to check if this claim is valid. A recently taken sample of 150 orders showed that 129 of them were mailed within 72 hours. Do you think the company's claim is true? Use a 2.5% significance level.

**Solution** Let  $p$  be the proportion of all orders that are mailed by the company within 72 hours, and let  $\hat{p}$  be the corresponding sample proportion. Then, from the given information,

$$n = 150, \quad \hat{p} = 129/150 = .86, \quad \text{and} \quad \alpha = .025$$

The company claims that at least 90% of all orders are mailed within 72 hours. Assuming that this claim is true, the values of  $p$  and  $q$  are

$$p = .90 \quad \text{and} \quad q = 1 - p = 1 - .90 = .10$$

**Step 1.** State the null and alternative hypotheses.

The null and alternative hypotheses are

$$H_0: p \geq .90 \quad (\text{The company's claim is true.})$$

$$H_1: p < .90 \quad (\text{The company's claim is false.})$$

**Step 2.** Select the distribution to use.

We first check whether  $np$  and  $nq$  are both greater than 5:

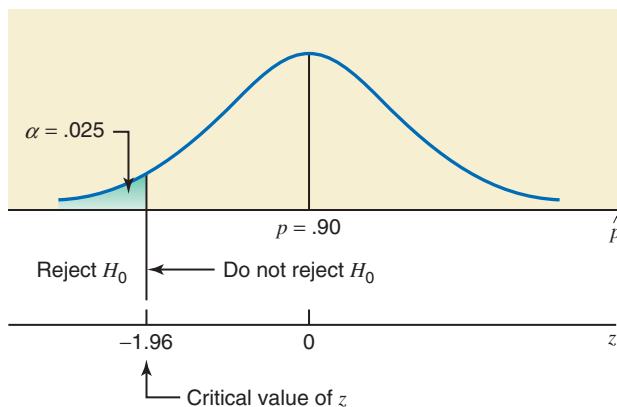
$$np = 150(.90) = 135 > 5 \quad \text{and} \quad nq = 150(.10) = 15 > 5$$

Consequently, the sample size is large. Therefore, we use the normal distribution to make the hypothesis test about  $p$ .

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is .025. The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed, and the rejection region lies in the left tail of the sampling distribution of  $\hat{p}$  with its area equal to .025. As shown in Figure 9.18, the critical value of  $z$ , obtained from the normal distribution table for .0250 area in the left tail, is  $-1.96$ .

**Figure 9.18** Critical value of  $z$ .



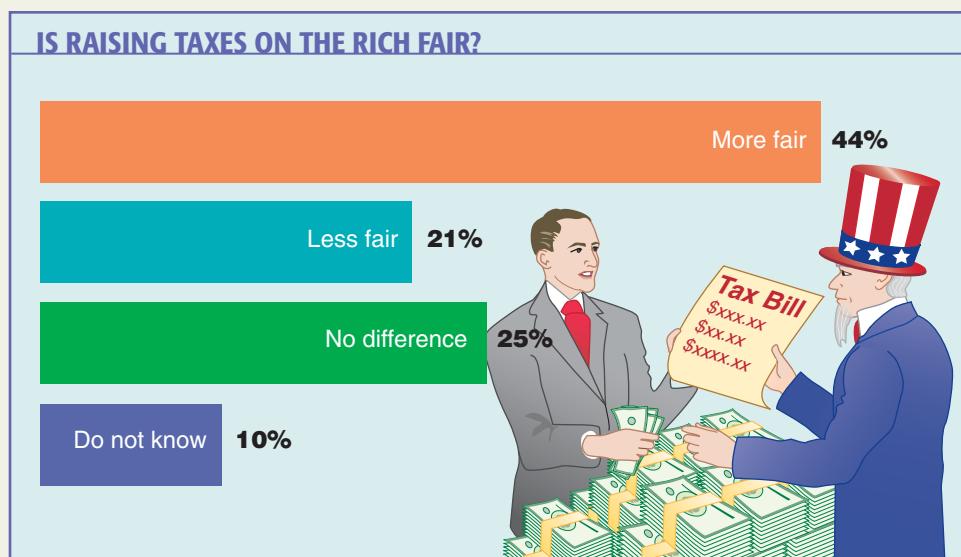
**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $z$  for  $\hat{p} = .86$  is calculated as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.90)(.10)}{150}} = .02449490$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.86 - .90}{.02449490} = -1.63$$

From  $H_0$



Data source: Pew Research Center for the People & the Press national survey of American adults conducted July 12-15, 2012.

In a national poll of American adults conducted July 12–15, 2012, by the Pew Research Center for the People & the Press, adults were asked whether raising taxes on the rich (defined as those earning more than \$250,000 per year) would make the tax system more fair. The responses of the adults polled are shown in the accompanying graph. Of these adults, 44% said that raising taxes on the rich would make the tax system more fair, 21% said that it would make the system less fair, 25% thought it would make no difference, and 10% did not know.

Suppose that we want to check whether the current percentage of American adults who will say that raising taxes on the rich would make the tax system more fair is different from 44%. Suppose we take a sample of 1600 American adults and ask them the same question, and 48% of them say that raising taxes on the rich will make the system more fair. Let us choose a significance level of 1%. The test is two-tailed. The null and alternative hypotheses are

$$H_0: p = .44$$

$$H_1: p \neq .44$$

Here,  $n = 1600$ ,  $\hat{p} = .48$ ,  $\alpha = .01$ , and  $\alpha/2 = .005$ . The sample is large. (The reader should check that  $np$  and  $nq$  are both greater than 5.) Using the normal distribution for the test, we find that the critical values of  $z$  for .0050 and .9950 areas to the left are  $-2.58$  and  $2.58$ , respectively. We find the observed value of  $z$  as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.44)(.56)}{1600}} = .01240967$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.48 - .44}{.01240967} = 3.22$$

The value of the test statistic  $z = 3.22$  for  $\hat{p}$  is larger than the upper critical value of  $z = 2.58$ , and it falls in the rejection region. Consequently, we reject  $H_0$  and conclude that the current percentage of American adults who hold the opinion that raising taxes on the rich would make the tax system more fair is significantly different from 44%.

We can use the  $p$ -value approach too. From the normal distribution table, the area under the normal curve to the right of  $z = 3.22$  is .0006. Therefore, the  $p$ -value is  $2(.0006) = .0012$ . Since  $\alpha = .01$  is larger than .0012, we reject the null hypothesis.

*Data Source:* <http://www.peoplepress.org/2012/07/16/raising-taxes-on-rich-seen-as-good-for-economy-fairness/>.

**Step 5. Make a decision.**

The value of the test statistic  $z = -1.63$  is greater than the critical value of  $z = -1.96$ , and it falls in the nonrejection region. Therefore, we fail to reject  $H_0$ . We can state that the difference between the sample proportion and the hypothesized value of the population proportion is small, and this difference may have occurred owing to chance alone. Therefore, the proportion of all orders that are mailed within 72 hours is at least 90%, and the company's claim seems to be true. ■

**EXERCISES****CONCEPTS AND PROCEDURES**

**9.74** Explain when a sample is large enough to use the normal distribution to make a test of hypothesis about the population proportion.

**9.75** In each of the following cases, do you think the sample size is large enough to use the normal distribution to make a test of hypothesis about the population proportion? Explain why or why not.

- a.  $n = 40$  and  $p = .11$
- b.  $n = 100$  and  $p = .73$
- c.  $n = 80$  and  $p = .05$
- d.  $n = 50$  and  $p = .14$

**9.76** In each of the following cases, do you think the sample size is large enough to use the normal distribution to make a test of hypothesis about the population proportion? Explain why or why not.

- a.  $n = 30$  and  $p = .65$
- b.  $n = 70$  and  $p = .05$
- c.  $n = 60$  and  $p = .06$
- d.  $n = 900$  and  $p = .17$

**9.77** For each of the following examples of tests of hypothesis about the population proportion, show the rejection and nonrejection regions on the graph of the sampling distribution of the sample proportion.

- a. A two-tailed test with  $\alpha = .10$
- b. A left-tailed test with  $\alpha = .01$
- c. A right-tailed test with  $\alpha = .05$

**9.78** For each of the following examples of tests of hypothesis about the population proportion, show the rejection and nonrejection regions on the graph of the sampling distribution of the sample proportion.

- a. A two-tailed test with  $\alpha = .05$
- b. A left-tailed test with  $\alpha = .02$
- c. A right-tailed test with  $\alpha = .025$

**9.79** A random sample of 500 observations produced a sample proportion equal to .38. Find the critical and observed values of  $z$  for each of the following tests of hypotheses using  $\alpha = .05$ .

- a.  $H_0: p = .30$  versus  $H_1: p > .30$
- b.  $H_0: p = .30$  versus  $H_1: p \neq .30$

**9.80** A random sample of 200 observations produced a sample proportion equal to .60. Find the critical and observed values of  $z$  for each of the following tests of hypotheses using  $\alpha = .01$ .

- a.  $H_0: p = .63$  versus  $H_1: p < .63$
- b.  $H_0: p = .63$  versus  $H_1: p \neq .63$

**9.81** Consider the null hypothesis  $H_0: p = .65$ . Suppose a random sample of 1000 observations is taken to perform this test about the population proportion. Using  $\alpha = .05$ , show the rejection and nonrejection regions and find the critical value(s) of  $z$  for a

- a. left-tailed test
- b. two-tailed test
- c. right-tailed test

**9.82** Consider the null hypothesis  $H_0: p = .25$ . Suppose a random sample of 400 observations is taken to perform this test about the population proportion. Using  $\alpha = .01$ , show the rejection and nonrejection regions and find the critical value(s) of  $z$  for a

- a. left-tailed test
- b. two-tailed test
- c. right-tailed test

**9.83** Consider  $H_0: p = .70$  versus  $H_1: p \neq .70$ .

- A random sample of 600 observations produced a sample proportion equal to .68. Using  $\alpha = .01$ , would you reject the null hypothesis?
- Another random sample of 600 observations taken from the same population produced a sample proportion equal to .76. Using  $\alpha = .01$ , would you reject the null hypothesis?

Comment on the results of parts a and b.

**9.84** Consider  $H_0: p = .45$  versus  $H_1: p < .45$ .

- A random sample of 400 observations produced a sample proportion equal to .42. Using  $\alpha = .025$ , would you reject the null hypothesis?
- Another random sample of 400 observations taken from the same population produced a sample proportion of .39. Using  $\alpha = .025$ , would you reject the null hypothesis?

Comment on the results of parts a and b.

**9.85** Make the following hypothesis tests about  $p$ .

- $H_0: p = .45$ ,  $H_1: p \neq .45$ ,  $n = 100$ ,  $\hat{p} = .49$ ,  $\alpha = .10$
- $H_0: p = .72$ ,  $H_1: p < .72$ ,  $n = 700$ ,  $\hat{p} = .64$ ,  $\alpha = .05$
- $H_0: p = .30$ ,  $H_1: p > .30$ ,  $n = 200$ ,  $\hat{p} = .33$ ,  $\alpha = .01$

**9.86** Make the following hypothesis tests about  $p$ .

- $H_0: p = .57$ ,  $H_1: p \neq .57$ ,  $n = 800$ ,  $\hat{p} = .50$ ,  $\alpha = .05$
- $H_0: p = .26$ ,  $H_1: p < .26$ ,  $n = 400$ ,  $\hat{p} = .23$ ,  $\alpha = .01$
- $H_0: p = .84$ ,  $H_1: p > .84$ ,  $n = 250$ ,  $\hat{p} = .85$ ,  $\alpha = .025$

## ■ APPLICATIONS

**9.87** According to the U.S. Census Bureau, 11% of children in the United States lived with at least one grandparent in 2009 (*USA TODAY*, June 30, 2011). Suppose that in a recent sample of 1600 children, 224 were found to be living with at least one grandparent. At a 5% significance level, can you conclude that the proportion of all children in the United States who currently live with at least one grandparent is higher than .11? Use both the  $p$ -value and the critical-value approaches.

**9.88** According to a book published in 2011, 45% of the undergraduate students in the United States show almost no gain in learning in their first 2 years of college (Richard Arum *et al.*, *Academically Adrift*, University of Chicago Press, Chicago, 2011). A recent sample of 1500 undergraduate students showed that this percentage is 38%. Can you reject the null hypothesis at a 1% significance level in favor of the alternative that the percentage of undergraduate students in the United States who show almost no gain in learning in their first 2 years of college is currently lower than 45%? Use both the  $p$ -value and the critical-value approaches.

**9.89** According to a *New York Times/CBS* News poll conducted during June 24–28, 2011, 55% of the American adults polled said that owning a home is a *very important part* of the American Dream (*The New York Times*, June 30, 2011). Suppose this result was true for the population of all American adults in 2011. In a recent poll of 1800 American adults, 61% said that owning a home is a *very important part* of the American Dream. Perform a hypothesis test to determine whether it is reasonable to conclude that the percentage of all American adults who currently hold this opinion is higher than 55%. Use a 2% significance level, and use both the  $p$ -value and the critical-value approaches.

**9.90** Beginning in the second half of 2011, there were widespread protests in many American cities that were primarily against Wall Street corruption and the increasing gap between the rich and the poor in America. According to a *Time Magazine/ABT SRBI* poll conducted by telephone during October 9–10, 2011, 86% of adults who were familiar with those protests agreed that Wall Street and lobbyists have too much influence in Washington (*The New York Times*, October 22, 2011). Assume that 86% of all American adults in 2011 believed that Wall Street and lobbyists have too much influence in Washington. A recent random sample of 2000 American adults showed that 1780 of them believe that Wall Street and lobbyists have too much influence in Washington. Using a 5% significance level, perform a test of hypothesis to determine whether the current percentage of American adults who believe that Wall Street and lobbyists have too much influence in Washington is higher than 86%. Use both the  $p$ -value and the critical-value approaches.

**9.91** According to a Pew Research Center nationwide telephone survey of American adults conducted by phone between March 15 and April 24, 2011, 75% of adults said that college education has become too expensive for most people and they cannot afford it (*Time*, May 30, 2011). Suppose that this result is true for the 2011 population of American adults. In a recent poll of 1600 American adults, 1160 said that college education has become too expensive for most people and they cannot afford it. Using a 1% significance level, perform a test of hypothesis to determine whether the current percentage of American adults

who will say that college education has become too expensive for most people and they cannot afford it is lower than 75%. Use both the  $p$ -value and the critical-value approaches.

**9.92** According to a Pew Research Center nationwide telephone survey conducted between March 15 and April 24, 2011, 55% of college graduates said that college education prepared them for a job (*Time*, May 30, 2011). Suppose this result was true of all college graduates at that time. In a recent sample of 2100 college graduates, 60% said that college education prepared them for a job. Is there significant evidence at a 1% significance level to conclude that the current percentage of all college graduates who will say that college education prepared them for a job is different from 55%? Use both the  $p$ -value and the critical-value approaches.

**9.93** A food company is planning to market a new type of frozen yogurt. However, before marketing this yogurt, the company wants to find what percentage of the people like it. The company's management has decided that it will market this yogurt only if at least 35% of the people like it. The company's research department selected a random sample of 400 persons and asked them to taste this yogurt. Of these 400 persons, 112 said they liked it.

- a. Testing at a 2.5% significance level, can you conclude that the company should market this yogurt?
- b. What will your decision be in part a if the probability of making a Type I error is zero? Explain.
- c. Make the test of part a using the  $p$ -value approach and  $\alpha = .025$ .

**9.94** A mail-order company claims that at least 60% of all orders are mailed within 48 hours. From time to time the quality control department at the company checks if this promise is fulfilled. Recently the quality control department at this company took a sample of 400 orders and found that 208 of them were mailed within 48 hours of the placement of the orders.

- a. Testing at a 1% significance level, can you conclude that the company's claim is true?
- b. What will your decision be in part a if the probability of making a Type I error is zero? Explain.
- c. Make the test of part a using the  $p$ -value approach and  $\alpha = .01$ .

**9.95** Brooklyn Corporation manufactures DVDs. The machine that is used to make these DVDs is known to produce not more than 5% defective DVDs. The quality control inspector selects a sample of 200 DVDs each week and inspects them for being good or defective. Using the sample proportion, the quality control inspector tests the null hypothesis  $p \leq .05$  against the alternative hypothesis  $p > .05$ , where  $p$  is the proportion of DVDs that are defective. She always uses a 2.5% significance level. If the null hypothesis is rejected, the production process is stopped to make any necessary adjustments. A recent sample of 200 DVDs contained 17 defective DVDs.

- a. Using a 2.5% significance level, would you conclude that the production process should be stopped to make necessary adjustments?
- b. Perform the test of part a using a 1% significance level. Is your decision different from the one in part a?

Comment on the results of parts a and b.

**9.96** Shulman Steel Corporation makes bearings that are supplied to other companies. One of the machines makes bearings that are supposed to have a diameter of 4 inches. The bearings that have a diameter of either more or less than 4 inches are considered defective and are discarded. When working properly, the machine does not produce more than 7% of bearings that are defective. The quality control inspector selects a sample of 200 bearings each week and inspects them for the size of their diameters. Using the sample proportion, the quality control inspector tests the null hypothesis  $p \leq .07$  against the alternative hypothesis  $p > .07$ , where  $p$  is the proportion of bearings that are defective. He always uses a 2% significance level. If the null hypothesis is rejected, the machine is stopped to make any necessary adjustments. One sample of 200 bearings taken recently contained 22 defective bearings.

- a. Using a 2% significance level, will you conclude that the machine should be stopped to make necessary adjustments?
- b. Perform the test of part a using a 1% significance level. Is your decision different from the one in part a?

Comment on the results of parts a and b.

**\*9.97** Two years ago, 75% of the customers of a bank said that they were satisfied with the services provided by the bank. The manager of the bank wants to know if this percentage of satisfied customers has changed since then. She assigns this responsibility to you. Briefly explain how you would conduct such a test.

**\*9.98** A study claims that 65% of students at all colleges and universities hold off-campus (part-time or full-time) jobs. You want to check if the percentage of students at your school who hold off-campus jobs is different from 65%. Briefly explain how you would conduct such a test. Collect data from 40 students at your school on whether or not they hold off-campus jobs. Then, calculate the proportion of students in this sample who hold off-campus jobs. Using this information, test the hypothesis. Select your own significance level.

## USES AND MISUSES...

### FOLLOW THE RECIPE

Hypothesis testing is one of the most powerful and dangerous tools of statistics. It allows us to make statements about a population and attach a degree of uncertainty to these statements. Pick up a newspaper and flip through it; rare will be the day when the paper does not contain a story featuring a statistical result, often reported with a significance level. Given that the subjects of these reports—public health, the environment, and so on—are important to our lives, it is critical that we perform the statistical calculations and interpretations properly. The first step, one that you should look for when reading statistical results, is proper formulation/specification.

*Formulation or specification*, simply put, is the list of steps you perform when constructing a hypothesis test. In this chapter, these steps are: stating the null and alternative hypotheses; selecting the appropriate distribution; and determining the rejection and nonrejection regions. Once these steps are performed, all you need to do is to calculate the *p*-value or the test statistic to complete the hypothesis test. It is important to beware of traps in the specification.

Though it might seem obvious, stating the hypothesis properly can be difficult. For hypotheses around a population mean, the null and alternative hypotheses are mathematical statements that do not overlap and also provide no holes. Suppose that a confectioner states that the average mass of his chocolate bars is 100 grams. The null hypothesis is that the mass of the bars is 100 grams, and the alternative hypothesis is that the mass of the bars is not 100 grams. When you take a sample of chocolate bars and measure their masses, all possibilities for the sample mean will fall within one of your decision regions. The problem is a little more difficult for hypotheses based on proportions. Make sure that you only have two categories. For example, if you are trying to determine the percentage of the population that has blonde hair, your groups are “blonde” and “not blonde.” You need to decide how to categorize bald people before you conduct this experiment: Do not include bald people in the survey.

Finally, beware of numerical precision. When your sample is large and you assume that it has a normal distribution, the rejection region for a two-tailed test using the normal distribution with a significance level of 5% will be values of the sample mean that are farther than 1.96 standard deviations from the assumed mean. When you perform your calculations, the sample mean may fall on the border of your decision region. Remember that there is measurement error and sample error that you cannot account for. In this case, it is probably best to adjust your significance level so that the sample mean falls squarely in a decision region.

### THE POWER OF NEGATIVE THINKING

In the beginning of this chapter, you learned about Type I and Type II errors. If this is your first statistics class, you might think that this is your first exposure to the concepts of Type I and Type II errors, but

that is not the case. As a matter of fact, if you can recall having a medical test, you must have had an interaction with a variety of concepts related to hypothesis testing, including Type I and Type II errors.

In a typical medical test, the assumption (our null hypothesis) is that you do not have the condition for which you are being tested. If the assumption is true, the doctor knows what should happen in the test. If the test results are different from the *normal* range, the doctor has data that would allow the assumption to be rejected. Whenever the null hypothesis is rejected (that is, the test results demonstrate that the person has the condition for which he or she was tested), the medical result is called a *positive* test result. If the doctor fails to reject the null hypothesis (if the test results do not demonstrate that you have the condition), the medical result is called a *negative* test result.

As with other types of hypothesis tests, medical tests are not perfect. Sometimes people are (falsely) diagnosed as having a condition or illness when they actually do not have it. In medical terminology, this Type I error is referred to as a *false positive*. Similarly, the result of a test may (wrongly) indicate that a person does not have a condition or illness when actually he or she has it. This Type II error is called a *false negative* in medical terminology.

Companies that develop medical tests perform intensive research and clinical tests to reduce the risk of making both these types of errors. Specifically, data are collected on the *sensitivity* and the *specificity* of medical tests. In the context of an illness, the sensitivity of a test is the proportion of all people with the illness who are identified by the test as actually having it. For example, suppose 100 students on a college campus have been identified by throat culture as having strep throat. All these 100 students are tested for strep throat using another type of test. Suppose 97 of the 100 tests come back positive with the second test. Then the sensitivity of the (second) test is .97 (or 97%), and the probability of a false negative (Type II error) is one minus the sensitivity, which is .03 here.

The specificity of a test refers to how well a test identifies that a healthy person does not have a given disease. Using the strep throat reference again, suppose that 400 students have been identified by throat culture as not having strep throat. All these 400 students are given a new strep test, and 394 of them are shown to have a negative result (that is, they are identified as not having strep throat). Then the specificity of the test is  $394/400 = .985$  (or 98.5%), and the probability of a false positive (Type I error) is one minus the specificity, which is .015 here.

Of course, low probabilities for both types of errors are very important in medical testing. A false positive can result in an individual obtaining unnecessary, often expensive, and sometime debilitating treatment, whereas a false negative can allow a disease to progress to an advanced stage when early detection could have helped to save a person’s life.

## Glossary

**$\alpha$**  The significance level of a test of hypothesis that denotes the probability of rejecting a null hypothesis when it actually is true. (The probability of committing a Type I error.)

**Alternative hypothesis** A claim about a population parameter that will be true if the null hypothesis is false.

**$\beta$**  The probability of not rejecting a null hypothesis when it actually is false. (The probability of committing a Type II error.)

**Critical value or critical point** One or two values that divide the whole region under the sampling distribution of a sample statistic into rejection and nonrejection regions.

**Left-tailed test** A test in which the rejection region lies in the left tail of the distribution curve.

**Null hypothesis** A claim about a population parameter that is assumed to be true until proven otherwise.

**Observed value of  $z$  or  $t$**  The value of  $z$  or  $t$  calculated for a sample statistic such as the sample mean or the sample proportion.

**One-tailed test** A test in which there is only one rejection region, either in the left tail or in the right tail of the distribution curve.

**p-value** The smallest significance level at which a null hypothesis can be rejected.

**Right-tailed test** A test in which the rejection region lies in the right tail of the distribution curve.

**Significance level** The value of  $\alpha$  that gives the probability of committing a Type I error.

**Test statistic** The value of  $z$  or  $t$  calculated for a sample statistic such as the sample mean or the sample proportion.

**Two-tailed test** A test in which there are two rejection regions, one in each tail of the distribution curve.

**Type I error** An error that occurs when a true null hypothesis is rejected.

**Type II error** An error that occurs when a false null hypothesis is not rejected.

## Supplementary Exercises



**9.99** Consider the following null and alternative hypotheses:

$$H_0: \mu = 120 \quad \text{versus} \quad H_1: \mu > 120$$

A random sample of 81 observations taken from this population produced a sample mean of 123.5. The population standard deviation is known to be 15.

- If this test is made at a 2.5% significance level, would you reject the null hypothesis? Use the critical-value approach.
- What is the probability of making a Type I error in part a?
- Calculate the  $p$ -value for the test. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .01$ ? What if  $\alpha = .05$ ?

**9.100** Consider the following null and alternative hypotheses:

$$H_0: \mu = 40 \quad \text{versus} \quad H_1: \mu \neq 40$$

A random sample of 64 observations taken from this population produced a sample mean of 38.4. The population standard deviation is known to be 6.

- If this test is made at a 2% significance level, would you reject the null hypothesis? Use the critical-value approach.
- What is the probability of making a Type I error in part a?
- Calculate the  $p$ -value for the test. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .01$ ? What if  $\alpha = .05$ ?

**9.101** Consider the following null and alternative hypotheses:

$$H_0: p = .82 \quad \text{versus} \quad H_1: p \neq .82$$

A random sample of 600 observations taken from this population produced a sample proportion of .86.

- If this test is made at a 2% significance level, would you reject the null hypothesis? Use the critical-value approach.
- What is the probability of making a Type I error in part a?
- Calculate the  $p$ -value for the test. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .025$ ? What if  $\alpha = .01$ ?

**9.102** Consider the following null and alternative hypotheses:

$$H_0: p = .44 \quad \text{versus} \quad H_1: p < .44$$

A random sample of 450 observations taken from this population produced a sample proportion of .39.

- a. If this test is made at a 2% significance level, would you reject the null hypothesis? Use the critical-value approach.
- b. What is the probability of making a Type I error in part a?
- c. Calculate the  $p$ -value for the test. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .01$ ? What if  $\alpha = .025$ ?

**9.103** According to the American Time Use Survey, Americans watched television each weekday for an average of 151 minutes in 2011 (*Time*, July 11, 2011). Suppose that this result is true for the 2011 population of all American adults. A recent sample of 120 American adults showed that they watch television each weekday for an average of 162 minutes. Assume that the population standard deviation for times spent watching television each weekday by American adults is 30 minutes.

- a. Find the  $p$ -value for the test of hypothesis with the alternative hypothesis that the current average time spent watching television each weekday by American adults is higher than 151 minutes. What is your conclusion at  $\alpha = .01$ ?
- b. Test the hypothesis of part a using the critical-value approach and  $\alpha = .01$ .

**9.104** The mean consumption of water per household in a city was 1245 cubic feet per month. Due to a water shortage because of a drought, the city council campaigned for water use conservation by households. A few months after the campaign was started, the mean consumption of water for a sample of 100 households was found to be 1175 cubic feet per month. The population standard deviation is given to be 250 cubic feet.

- a. Find the  $p$ -value for the hypothesis test that the mean consumption of water per household has decreased due to the campaign by the city council. Would you reject the null hypothesis at  $\alpha = .025$ ?
- b. Make the test of part a using the critical-value approach and  $\alpha = .025$ .

**9.105** A highway construction zone has a posted speed limit of 40 miles per hour. Workers working at the site claim that the mean speed of vehicles passing through this construction zone is at least 50 miles per hour. A random sample of 36 vehicles passing through this zone produced a mean speed of 48 miles per hour. The population standard deviation is known to be 4 miles per hour.

- a. Do you think the sample information is consistent with the workers' claim? Use  $\alpha = .025$ .
- b. What is the Type I error in this case? Explain. What is the probability of making this error?
- c. Will your conclusion of part a change if the probability of making a Type I error is zero?
- d. Find the  $p$ -value for the test of part a. What is your decision if  $\alpha = .025$ ?

**9.106** According to an estimate, the average age at first marriage for men in the United States was 28.2 years in 2010 (*Time*, March 21, 2011). A recent sample of 200 men from Ohio who got married for the first time this year showed that their average age at first marriage was 27.1 years. Assume that the population standard deviation for the distribution of ages at first marriage of all men from Ohio who got married for the first time this year is 5.8 years.

- a. Using the critical-value approach, can you conclude that the average age at first marriage for all men from Ohio who got married for the first time this year is lower than 28.2 years? Use  $\alpha = .01$ .
- b. What is the Type I error in part a? Explain. What is the probability of making this error in part a?
- c. Will your conclusion of part a change if the probability of making a Type I error is zero?
- d. Calculate the  $p$ -value for the test of part a. What is your conclusion if  $\alpha = .01$ ?

**9.107** A real estate agent claims that the mean living area of all single-family homes in his county is at most 2400 square feet. A random sample of 50 such homes selected from this county produced the mean living area of 2540 square feet and a standard deviation of 472 square feet.

- a. Using  $\alpha = .05$ , can you conclude that the real estate agent's claim is true?
- b. What will your conclusion be if  $\alpha = .01$ ?

Comment on the results of parts a and b.

**9.108** According to Moebs Services Inc., the cost of an individual checking account at U.S. community banks to these banks was between \$175 and \$200 in 2011 (*Time*, November 21, 2011). Suppose that the average annual cost of individual checking accounts at U.S. community banks to these banks was \$190 in 2011. A recent sample of 40 individual checking accounts selected from U.S. community banks showed that they cost these banks an average of \$211 per year with a standard deviation of \$35.

- a. Using  $\alpha = .025$ , can you conclude that the current average annual cost of individual checking accounts at U.S. community banks to these banks is higher than \$190? Use the critical-value approach.
- b. Find the range of the  $p$ -value for the test of part a. What is your conclusion with  $\alpha = .025$ ?

**9.109** Customers often complain about long waiting times at restaurants before the food is served. A restaurant claims that it serves food to its customers, on average, within 15 minutes after the order is placed. A local newspaper journalist wanted to check if the restaurant's claim is true. A sample of 36 customers showed that the mean time taken to serve food to them was 15.75 minutes with a standard deviation of 2.4 minutes. Using the sample mean, the journalist says that the restaurant's claim is false. Do you think the journalist's conclusion is fair to the restaurant? Use a 1% significance level to answer this question.

**9.110** The customers at a bank complained about long lines and the time they had to spend waiting for service. It is known that the customers at this bank had to wait 8 minutes, on average, before being served. The management made some changes to reduce the waiting time for its customers. A sample of 60 customers taken after these changes were made produced a mean waiting time of 7.5 minutes with a standard deviation of 2.1 minutes. Using this sample mean, the bank manager displayed a huge banner inside the bank mentioning that the mean waiting time for customers has been reduced by new changes. Do you think the bank manager's claim is justifiable? Use a 2.5% significance level to answer this question. Use both approaches.

**9.111** The administrative office of a hospital claims that the mean waiting time for patients to get treatment in its emergency ward is 25 minutes. A random sample of 16 patients who received treatment in the emergency ward of this hospital produced a mean waiting time of 27.5 minutes with a standard deviation of 4.8 minutes. Using a 1% significance level, test whether the mean waiting time at the emergency ward is different from 25 minutes. Assume that the waiting times for all patients at this emergency ward have a normal distribution.

**9.112** An earlier study claimed that U.S. adults spent an average of 114 minutes per day with their family. A recently taken sample of 25 adults from a city showed that they spend an average of 109 minutes per day with their family. The sample standard deviation is 11 minutes. Assume that the times spent by adults with their family have an approximate normal distribution.

- Using a 1% significance level, test whether the mean time spent currently by all adults with their families in this city is different from 114 minutes a day.
- Suppose the probability of making a Type I error is zero. Can you make a decision for the test of part a without going through the five steps of hypothesis testing? If yes, what is your decision? Explain.

**9.113** A computer company that recently introduced a new software product claims that the mean time taken to learn how to use this software is not more than 2 hours for people who are somewhat familiar with computers. A random sample of 12 such persons was selected. The following data give the times taken (in hours) by these persons to learn how to use this software.

1.75	2.25	2.40	1.90	1.50	2.75
2.15	2.25	1.80	2.20	3.25	2.60

Test at a 1% significance level whether the company's claim is true. Assume that the times taken by all persons who are somewhat familiar with computers to learn how to use this software are approximately normally distributed.

**9.114** A company claims that its 8-ounce low-fat yogurt cups contain, on average, at most 150 calories per cup. A consumer agency wanted to check whether or not this claim is true. A random sample of 10 such cups produced the following data on calories.

147	159	153	146	144	161	163	153	143	158
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Test using a 2.5% significance level whether the company's claim is true. Assume that the numbers of calories for such cups of yogurt produced by this company have an approximate normal distribution.

**9.115** According to the U.S. Census Bureau, 69% of children under the age of 18 years in the United States lived with two parents in 2009. Suppose that in a recent sample of 2000 children, 1298 were living with two parents.

- Using the critical value approach and  $\alpha = .05$ , test whether the current percentage of all children under the age of 18 years in the United States who live with two parents is different from 69%.
- How do you explain the Type I error in part a? What is the probability of making this error in part a?
- Calculate the  $p$ -value for the test of part a. What is your conclusion if  $\alpha = .05$ ?

**9.116** In a *Time Magazine/Aspen* poll of American adults conducted by the strategic research firm, Penn Schoen Berland, these adults were asked, "In your opinion, what is more important for the U.S. to focus on in the next decade?" Eighty-three percent of the adults polled said *domestic issues* (*Time*, July 11, 2011). Assume that this percentage is true for the 2011 population of American adults. In a recent random sample of 1400 adults, 1078 held this opinion.

- Using the critical-value approach and  $\alpha = .01$ , test whether the current percentage of American adults who hold the above opinion is less than 83%.

- b. How do you explain the Type I error in part a? What is the probability of making this error in part a?
- c. Calculate the  $p$ -value for the test of part a. What is your conclusion if  $\alpha = .01$ ?

**9.117** More and more people are abandoning national brand products and buying store brand products to save money. The president of a company that produces national brand coffee claims that 40% of the people prefer to buy national brand coffee. A random sample of 700 people who buy coffee showed that 259 of them buy national brand coffee. Using  $\alpha = .01$ , can you conclude that the percentage of people who buy national brand coffee is different from 40%? Use both approaches to make the test.

**9.118** In a poll conducted by *The New York Times* and CBS News, 44% of Americans approve of the job that the Supreme Court is doing (*The New York Times*, June 8, 2012). Assume that this percentage was true for the population of Americans at the time of this poll was conducted. A recent poll of 1300 Americans showed that 39% of them approve of the job that the Supreme Court is doing. At a 2% significance level, can you conclude that the current proportion of Americans who approve of the job that the Supreme Court is doing is different from .44?

**9.119** Mong Corporation makes auto batteries. The company claims that 80% of its LL70 batteries are good for 70 months or longer. A consumer agency wanted to check if this claim is true. The agency took a random sample of 40 such batteries and found that 75% of them were good for 70 months or longer.

- a. Using a 1% significance level, can you conclude that the company's claim is false?
- b. What will your decision be in part a if the probability of making a Type I error is zero?

Explain.

**9.120** Dartmouth Distribution Warehouse makes deliveries of a large number of products to its customers. To keep its customers happy and satisfied, the company's policy is to deliver on time at least 90% of all the orders it receives from its customers. The quality control inspector at the company quite often takes samples of orders delivered and checks to see whether this policy is maintained. A recent sample of 90 orders taken by this inspector showed that 75 of them were delivered on time.

- a. Using a 2% significance level, can you conclude that the company's policy is maintained?
- b. What will your decision be in part a if the probability of making a Type I error is zero? Explain.

## Advanced Exercises

**9.121** Professor Hansen believes that some people have the ability to predict in advance the outcome of a spin of a roulette wheel. He takes 100 student volunteers to a casino. The roulette wheel has 38 numbers, each of which is equally likely to occur. Of these 38 numbers, 18 are red, 18 are black, and 2 are green. Each student is to place a series of five bets, choosing either a red or a black number before each spin of the wheel. Thus, a student who bets on red has an 18/38 chance of winning that bet. The same is true of betting on black.

- a. Assuming random guessing, what is the probability that a particular student will win all five of his or her bets?
- b. Suppose for each student we formulate the hypothesis test  
 $H_0$ : The student is guessing  
 $H_1$ : The student has some predictive ability  
 Suppose we reject  $H_0$  only if the student wins all five bets. What is the significance level?
- c. Suppose that 2 of the 100 students win all five of their bets. Professor Hansen says, "For these two students we can reject  $H_0$  and conclude that we have found two students with some ability to predict." What do you make of Professor Hansen's conclusion?

**9.122** Acme Bicycle Company makes derailleurs for mountain bikes. Usually no more than 4% of these parts are defective, but occasionally the machines that make them get out of adjustment and the rate of defectives exceeds 4%. To guard against this, the chief quality control inspector takes a random sample of 130 derailleurs each week and checks each one for defects. If too many of these parts are defective, the machines are shut down and adjusted. To decide how many parts must be defective to shut down the machines, the company's statistician has set up the hypothesis test

$$H_0: p \leq .04 \quad \text{versus} \quad H_1: p > .04$$

where  $p$  is the proportion of defectives among all derailleurs being made currently. Rejection of  $H_0$  would call for shutting down the machines. For the inspector's convenience, the statistician would like the rejection region to have the form, "Reject  $H_0$  if the number of defective parts is  $C$  or more." Find the value of  $C$  that will make the significance level (approximately) .05.

**9.123** Alpha Airline claims that only 15% of its flights arrive more than 10 minutes late. Let  $p$  be the proportion of all of Alpha's flights that arrive more than 10 minutes late. Consider the hypothesis test

$$H_0: p \leq .15 \quad \text{versus} \quad H_1: p > .15$$

Suppose we take a random sample of 50 flights by Alpha Airline and agree to reject  $H_0$  if 9 or more of them arrive late. Find the significance level for this test.

**9.124** The standard therapy that is used to treat a disorder cures 60% of all patients in an average of 140 visits. A health care provider considers supporting a new therapy regime for the disorder if it is effective in reducing the number of visits while retaining the cure rate of the standard therapy. A study of 200 patients with the disorder who were treated by the new therapy regime reveals that 108 of them were cured in an average of 132 visits with a standard deviation of 38 visits. What decision should be made using a .01 level of significance?

**9.125** The package of Sylvania CFL 65-watt replacement bulbs that use only 16 watts claims that these bulbs have an average life of 8000 hours. Assume that the standard deviation of lives of these light bulbs is 400 hours. A skeptical consumer does not think that these light bulbs last as long as the manufacturer claims, and she decides to test 52 randomly selected light bulbs. She has set up the decision rule that if the average life of these 52 light bulbs is less than or equal to 7890 hours, then she will reject company's claim and conclude that the company has printed too high an average life on the packages, and she will write them a letter to that effect. Approximately what significance level is she using? If she decides instead on the decision rule that if the average life of these 52 light bulbs is less than or equal to 7857 hours she will reject the null hypothesis that company's claim is true, then approximately what significance level is she using? Interpret the values you get.

**9.126** Thirty percent of all people who are inoculated with the current vaccine that is used to prevent a disease contract the disease within a year. The developer of a new vaccine that is intended to prevent this disease wishes to test for significant evidence that the new vaccine is more effective.

- a. Determine the appropriate null and alternative hypotheses.
- b. The developer decides to study 100 randomly selected people by inoculating them with the new vaccine. If 84 or more of them do not contract the disease within a year, the developer will conclude that the new vaccine is superior to the old one. What significance level is the developer using for the test?
- c. Suppose 20 people inoculated with the new vaccine are studied and the new vaccine is concluded to be better than the old one if fewer than 3 people contract the disease within a year. What is the significance level of the test?

**9.127** Since 1984, all automobiles have been manufactured with a middle tail-light. You have been hired to answer the following question: Is the middle tail-light effective in reducing the number of rear-end collisions? You have available to you any information you could possibly want about all rear-end collisions involving cars built before 1984. How would you conduct an experiment to answer the question? In your answer, include things like (a) the precise meaning of the unknown parameter you are testing; (b)  $H_0$  and  $H_1$ ; (c) a detailed explanation of what sample data you would collect to draw a conclusion; and (d) any assumptions you would make, particularly about the characteristics of cars built before 1984 versus those built since 1984.

**9.128** Before a championship football game, the referee is given a special commemorative coin to toss to decide which team will kick the ball first. Two minutes before game time, he receives an anonymous tip that the captain of one of the teams may have substituted a biased coin that has a 70% chance of showing heads each time it is tossed. The referee has time to toss the coin 10 times to test it. He decides that if it shows 8 or more heads in 10 tosses, he will reject this coin and replace it with another coin. Let  $p$  be the probability that this coin shows heads when it is tossed once.

- a. Formulate the relevant null and alternative hypotheses (in terms of  $p$ ) for the referee's test.
- b. Using the referee's decision rule, find  $\alpha$  for this test.

**9.129** In Las Vegas, Nevada, and Atlantic City, New Jersey, tests are performed often on the various gambling devices used in casinos. For example, dice are often tested to determine if they are balanced. Suppose you are assigned the task of testing a die, using a two-tailed test to make sure that the probability of a 2-spot is 1/6. Using the 5% significance level, determine how many 2-spots you would have to obtain to reject the null hypothesis when your sample size is

- a. 120
- b. 1200
- c. 12,000

Calculate the value of  $\hat{p}$  for each of these three cases. What can you say about the relationship between (1) the difference between  $\hat{p}$  and 1/6 that is necessary to reject the null hypothesis and (2) the sample size as it gets larger?

**9.130** A statistician performs the test  $H_0: \mu = 15$  versus  $H_1: \mu \neq 15$  and finds the  $p$ -value to be .4546.

- a. The statistician performing the test does not tell you the value of the sample mean and the value of the test statistic. Despite this, you have enough information to determine the pair of  $p$ -values associated with the following alternative hypotheses.

- i.  $H_1: \mu < 15$       ii.  $H_1: \mu > 15$

Note that you will need more information to determine which  $p$ -value goes with which alternative. Determine the pair of  $p$ -values. Here the value of the sample mean is the same in both cases.

- b. Suppose the statistician tells you that the value of the test statistic is negative. Match the  $p$ -values with the alternative hypotheses.

Note that the result for one of the two alternatives implies that the sample mean is not on the same side of  $\mu = 15$  as the rejection region. Although we have not discussed this scenario in the book, it is important to recognize that there are many real-world scenarios in which this type of situation does occur. For example, suppose the EPA is to test whether or not a company is exceeding a specific pollution level. If the average discharge level obtained from the sample falls below the threshold (mentioned in the null hypothesis), then there would be no need to perform the hypothesis test.

**9.131** You read an article that states “50 hypothesis tests of  $H_0: \mu = 35$  versus  $H_1: \mu \neq 35$  were performed using  $\alpha = .05$  on 50 different samples taken from the same population with a mean of 35. Of these, 47 tests failed to reject the null hypothesis.” Explain why this type of result is not surprising.

## Self-Review Test

1. A test of hypothesis is always about
  - a. a population parameter
  - b. a sample statistic
  - c. a test statistic
2. A Type I error is committed when
  - a. a null hypothesis is not rejected when it is actually false
  - b. a null hypothesis is rejected when it is actually true
  - c. an alternative hypothesis is rejected when it is actually true
3. A Type II error is committed when
  - a. a null hypothesis is not rejected when it is actually false
  - b. a null hypothesis is rejected when it is actually true
  - c. an alternative hypothesis is rejected when it is actually true
4. A critical value is the value
  - a. calculated from sample data
  - b. determined from a table (e.g., the normal distribution table or other such tables)
  - c. neither a nor b
5. The computed value of a test statistic is the value
  - a. calculated for a sample statistic
  - b. determined from a table (e.g., the normal distribution table or other such tables)
  - c. neither a nor b
6. The observed value of a test statistic is the value
  - a. calculated for a sample statistic
  - b. determined from a table (e.g., the normal distribution table or other such tables)
  - c. neither a nor b
7. The significance level, denoted by  $\alpha$ , is
  - a. the probability of committing a Type I error
  - b. the probability of committing a Type II error
  - c. neither a nor b
8. The value of  $\beta$  gives the
  - a. probability of committing a Type I error
  - b. probability of committing a Type II error
  - c. power of the test

9. The value of  $1 - \beta$  gives the
  - a. probability of committing a Type I error
  - b. probability of committing a Type II error
  - c. power of the test
10. A two-tailed test is a test with
  - a. two rejection regions
  - b. two nonrejection regions
  - c. two test statistics
11. A one-tailed test
  - a. has one rejection region
  - b. has one nonrejection region
  - c. both a and b
12. The smallest level of significance at which a null hypothesis is rejected is called
  - a.  $\alpha$
  - b.  $p$ -value
  - c.  $\beta$
13. The sign in the alternative hypothesis in a two-tailed test is always
  - a.  $<$
  - b.  $>$
  - c.  $\neq$
14. The sign in the alternative hypothesis in a left-tailed test is always
  - a.  $<$
  - b.  $>$
  - c.  $\neq$
15. The sign in the alternative hypothesis in a right-tailed test is always
  - a.  $<$
  - b.  $>$
  - c.  $\neq$
16. According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of \$921 for single (one person) health insurance coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose that a recent random sample of 100 workers with employer-provided health insurance selected from a large city paid an average premium of \$946 for single health insurance coverage. Assume that such premiums paid by all such workers in this city have a standard deviation of \$110.
  - a. Using the critical-value approach and a 1% significance level, can you conclude that the current average such premium paid by all such workers in this city is different from \$921?
  - b. Using the critical-value approach and a 2.5% significance level, can you conclude that the current average such premium paid by all such workers in this city is higher than \$921?
  - c. What is the Type I error in parts a and b? What is the probability of making this error in each of parts a and b?
  - d. Calculate the  $p$ -value for the test of part a. What is your conclusion if  $\alpha = .01$ ?
  - e. Calculate the  $p$ -value for the test of part b. What is your conclusion if  $\alpha = .025$ ?
17. A minor league baseball executive has become concerned about the slow pace of games played in her league, fearing that it will lower attendance. She meets with the league's managers and umpires and discusses guidelines for speeding up the games. Before the meeting, the mean duration of nine-inning games was 3 hours, 5 minutes (i.e., 185 minutes). A random sample of 36 nine-inning games after the meeting showed a mean of 179 minutes with a standard deviation of 12 minutes.
  - a. Testing at a 1% significance level, can you conclude that the mean duration of nine-inning games has decreased after the meeting?
  - b. What is the Type I error in part a? What is the probability of making this error?
  - c. What will your decision be in part a if the probability of making a Type I error is zero? Explain.
  - d. Find the range for the  $p$ -value for the test of part a. What is your decision based on this  $p$ -value?
18. An editor of a New York publishing company claims that the mean time taken to write a textbook is at least 31 months. A sample of 16 textbook authors found that the mean time taken by them to write a textbook was 25 months with a standard deviation of 7.2 months.
  - a. Using a 2.5% significance level, would you conclude that the editor's claim is true? Assume that the time taken to write a textbook is normally distributed for all textbook authors.
  - b. What is the Type I error in part a? What is the probability of making this error?
  - c. What will your decision be in part a if the probability of making a Type I error is .001?
19. A financial advisor claims that less than 50% of adults in the United States have a will. A random sample of 1000 adults showed that 450 of them have a will.
  - a. At a 5% significance level, can you conclude that the percentage of people who have a will is less than 50%?
  - b. What is the Type I error in part a? What is the probability of making this error?
  - c. What would your decision be in part a if the probability of making a Type I error were zero? Explain.
  - d. Find the  $p$ -value for the test of hypothesis mentioned in part a. Using this  $p$ -value, will you reject the null hypothesis if  $\alpha = .05$ ? What if  $\alpha = .01$ ?

## Mini-Projects

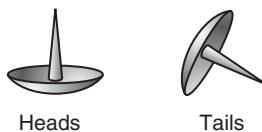
### MINI-PROJECT 9-1

Refer to the NFL data (Data Set III) on the Web site for this text and to the information on this data set in Appendix B. According to the information contained in this data set, the mean height of players who were on the rosters of NFL teams on October 31 of the 2011 NFL season was 73.99 inches.

- a. Create the appropriate graph(s) of the heights of all NFL players to determine whether the population of heights is approximately normally distributed. What is your conclusion? (See Appendix 6.1 in Chapter 6.)
- b. Take a random sample of 15 players from this NFL data file. Test  $H_0: \mu = 73.99$  inches against  $H_1: \mu \neq 73.99$  inches, using  $\alpha = .05$ .
- c. Repeat part b for samples of 31 and 45 players, respectively.
- d. Did any of the three tests in parts b and c lead to the conclusion that the mean height of all NFL players is different from 73.99 inches? If you were to repeat this process a large number of times using the same sample size, what percentage of samples would be expected to lead to this conclusion?

### MINI-PROJECT 9-2

A thumbtack that is tossed on a desk can land in one of the two ways shown in the following illustration:



Brad and Dan cannot agree on the likelihood of obtaining a head or a tail. Brad argues that obtaining a tail is more likely than obtaining a head because of the shape of the tack. If the tack had no point at all, it would resemble a coin that has the same probability of coming up heads or tails when tossed. But the longer the point, the less likely it is that the tack will stand up on its head when tossed. Dan believes that as the tack lands tails, the point causes the tack to jump around and come to rest in the heads position. Brad and Dan need you to settle their dispute. Do you think the tack is equally likely to land heads or tails? To investigate this question, find an ordinary thumbtack and toss it a large number of times (say, 100 times).

- a. What is the meaning, in words, of the unknown parameter in this problem?
- b. Set up the null and alternative hypotheses and compute the  $p$ -value based on your results from tossing the tack.
- c. How would you answer the original question now? If you decide the tack is not fair, do you side with Brad or Dan?
- d. What would you estimate the value of the parameter in part a to be? Find a 90% confidence interval for this parameter.
- e. After doing this experiment, do you think 100 tosses are enough to infer the nature of your tack? Using your result as a preliminary estimate, determine how many tosses would be necessary to be 95% certain of having 4% accuracy; that is, the margin of error of estimate is 4%. Have you observed enough tosses?

### MINI-PROJECT 9-3

Collect pennies in the amount of \$5. Do not obtain rolls of pennies from a bank because many such rolls will consist solely of new pennies. Treat these 500 pennies as your population. Determine the ages, in years, of all these pennies. Calculate the mean and standard deviation of these ages and denote them by  $\mu$  and  $\sigma$ , respectively.

- a. Take a random sample of 10 pennies from these 500. Find the average age of these 10 pennies, which is the value of  $\bar{x}$ . Perform a test with the null hypothesis that  $\mu$  is equal to the value obtained for all 500 pennies and the alternative hypothesis that  $\mu$  is not equal to this value. Use a significance level of .10.
- b. Suppose you repeat the procedure of part a nine more times. How many times would you expect to reject the null hypothesis? Now actually repeat the procedure of part a nine more times, making

sure that you put the 10 pennies selected each time back in the population and that you mix all pennies well before taking a sample. How many times did you reject the null hypothesis? Note that you can enter the ages of these 500 pennies in a technology and then use that technology to take samples and make tests of hypothesis.

- Repeat parts a and b for a sample size of 25. Did you reject the null hypothesis more often with a sample size of 10 or a sample size of 25?

### MINI-PROJECT 9-4

In the article “*Flipping Out—Think a Coin Toss Has a 50–50 Chance? Think Again*” ([www.thebigmoney.com/articles/hey-wait-minute/2009/07/28/flipping-out?page=0&g=1](http://www.thebigmoney.com/articles/hey-wait-minute/2009/07/28/flipping-out?page=0&g=1)), author David E. Adler discusses how a team of Stanford researchers concluded that when you flip the coin, whichever side of a coin is facing up when you place it on your thumb to flip it is more likely to be the side that faces up after the coin is flipped. To test this concept, you are going to flip a coin 100 times. To simplify the record keeping, you should perform all 100 flips with head facing up when you place the coin on your thumb to flip it or perform all 100 flips with tail facing up when you place the coin on your thumb to flip it. Do your best to use the same amount of force each time you flip the coin.

- If you had head facing up when you placed the coin on your thumb, calculate the sample proportion of flips in which head occurred. (Or you can perform this experiment with tail facing up.) Perform a test with the null hypothesis that the side that started up will land up 50% of the time versus the alternative hypothesis that the side that started up will land up more than 50% of the time. Use a significance level of 5%.
- The Stanford research group concluded that the side that started up will land up 51% of the time. Use the data from your 100 flips to test the null hypothesis that the side that started up will land up 51% of the time versus the alternative hypothesis that the side that started up will not land 51% of the time. Use a significance level of 5%.
- Suppose that you were really bored one day and decided to repeat this experiment four times, using more flips each time, as shown in the following table.

Number of flips	Number of flips in which the side that started up also landed up
500	255
1000	510
5000	2550
10000	5100

For each of these four cases, calculate the test statistic and  $p$ -value for the hypothesis test described in part a. Based on your results, what can you conclude about the number of repetitions needed to distinguish between a result that occurs 50% of the time and the one that occurs 51% of the time?

## DECIDE FOR YOURSELF STATISTICAL AND PRACTICAL SIGNIFICANCE

The hypothesis-testing procedure helps us to make a conclusion regarding a claim or statement, and often this claim or statement is about the value of a parameter or the relationship between two or more parameters. When we reject the null hypothesis, we conclude that the result is statistically significant at the given significance level of  $\alpha$ . So, what exactly does the term “statistically significant” mean? Using the single-sample analogy, statistically significant implies that the value of a point estimator (such as a sample mean or sample proportion) of a parameter is far enough (in terms of the standard deviation or standard error) from the hypothesized value of the parameter so that it falls in the most extreme  $\alpha \times 100\%$  of the area under the sampling distribution curve.

Now the logical follow-up question is: “What does *statistically significant* imply with regard to my specific application?” Unlike the first question, which has a specific answer, the answer to this ques-

tion is: “It depends.” In any hypothesis test, one must consider the practical significance of the result. For example, suppose a new gasoline additive has been invented and the company that produces it claims that it increases average gas mileage. A fleet of cars of a specific model, based on EPA numbers, obtains an average of 448 miles per tank full of gas without this additive. A random sample of 25 such cars is selected. Each car is driven on a tank full of gas with this additive added to the gas. The sample mean for these 25 cars is found to be 453 miles per tank full of gas, with a sample standard deviation of 22 miles. To understand the difference between the statistical significance and practical significance, find answers to the following questions.

1. Perform the appropriate hypothesis test using the  $t$  distribution to determine if the average mileage per tank full of gas increases with

the additive. Use a 5% significance level. Is this increase statistically significant? Assume that the population is normally distributed.

**2.** Now suppose we use a sample of 100 cars instead of 25 cars, but the values of the means and the standard deviation remain the same. Perform the above hypothesis test again and see if your answer changes with this larger sample size.

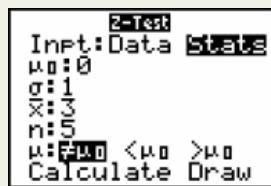
**3.** Regardless of the sample size, discuss whether the result (453 miles versus 448 miles) is *practically significant*, that is, whether or

not the increase is meaningful to the everyday driver. Suppose it is recommended that the additive should be used every 3000 miles. Assuming that the price of gas is \$3.44 (national average price per gallon on November 1, 2011, per [www.fuelgauge-report.aaa.com](http://www.fuelgauge-report.aaa.com)) per gallon and the gas tank holds 16 gallons of gas, calculate the savings in gas expenditure per mile. Then multiply this number by 3000 to obtain the savings per application of the additive. Assuming that the additive is not free, is it worth using it?

## TECHNOLOGY INSTRUCTION

### Hypothesis Testing

#### TI-84



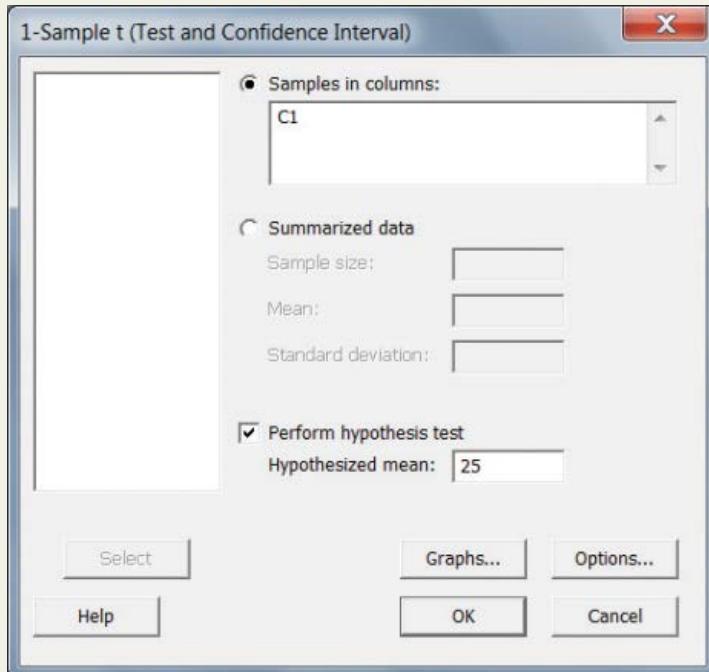
Screen 9.1

- To test a hypothesis about a population mean  $\mu$  given the population standard deviation  $\sigma$ , select **STAT >TESTS >ZTest**. If you have the data stored in a list, select **Data**, and enter the name of the list. If you have the summary statistics, choose **Stats**, and enter the sample mean and size. Enter  $\mu_0$ , the constant value for the population mean from your null hypothesis. Enter your value for  $\sigma$ , and select which alternative hypothesis you are using. Select **Calculate**. (See Screen 9.1.)
- To test a hypothesis about a population mean  $\mu$  without knowing the population standard deviation  $\sigma$ , select **STAT >TESTS >TTest**. If you have the data stored in a list, select **Data**, and enter the name of the list. If you have the summary statistics, choose **Stats**, and enter the sample mean, standard deviation, and size. Enter  $\mu_0$ , the constant value for the population mean from your null hypothesis. Select which alternative hypothesis you are using. Select **Calculate**.
- To test a hypothesis about a population proportion  $p$ , select **STAT >TESTS >1-PropZTest**. Enter the constant value for  $p$  from the null hypothesis as  $p_0$ . Enter the number of successes as  $x$  and the sample size as  $n$ . Select the alternative hypothesis you are using. Select **Calculate**.

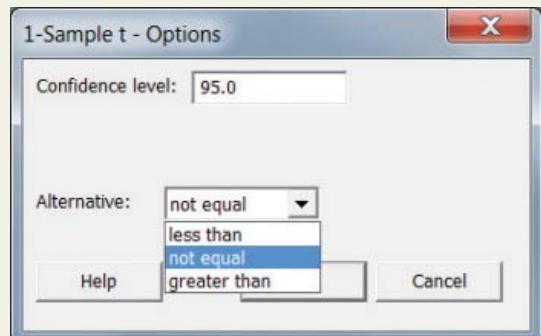
#### Minitab

- To perform a hypothesis test for the population mean  $\mu$  when the population standard deviation  $\sigma$  is given, select **Stat >Basic Statistics >1-Sample Z**. If you have your data entered in a column, enter the name of that column in the **Samples in columns:** box. Instead, if you know the summary statistics, click next to **Summarized data** and enter the values of the **Sample size** and **Mean** in their respective boxes. In both cases, enter the value of the population standard deviation in the **Standard deviation** box. Enter the value of  $\mu$  from the null hypothesis in the **Test mean:** box. Click on the **Options** button and select the appropriate alternative hypothesis from the **Alternative** box. Click **OK** in both windows. The output will appear in the **Session** window, which will give the  $p$ -value for the test. Based on this  $p$ -value, you can make a decision.
- To perform a hypothesis test for the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known, select **Stat >Basic Statistics >1-Sample t**. If you have your data entered in a column, enter the name of that column in the **Samples in columns:** box. Instead, if you know the summary statistics, click next to **Summarized data** and enter the values of the **Sample size**, **Sample standard deviation**, and **Mean** in their respective boxes. Check the **Perform Hypothesis Test** box and enter the value of  $\mu$  from the null

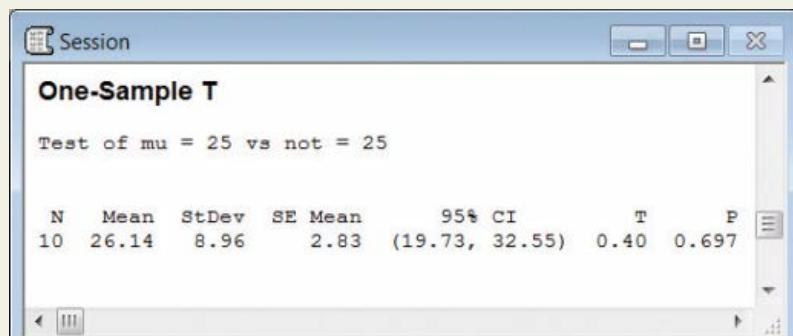
hypothesis in the **Test mean:** box. (See **Screen 9.2**.) Click on the **Options** button, and select the appropriate alternative hypothesis from the **Alternative** box. (See **Screen 9.3**.) Click **OK** in both windows. The output will appear in the **Session** window, which will give the *p*-value for the test. (See **Screen 9.4**.) Based on this *p*-value, you can make a decision.



Screen 9.2



Screen 9.3



Screen 9.4

3. To perform a hypothesis test for the population proportion *p*, select **Stat >Basic Statistics >1 Proportion**. If you have sample data (consisting of values for successes and failures) entered in a column, enter the name of that column in the **Samples in columns:** box. Instead, if you know the number of trials and number of successes, click next to **Summarized data**, and enter the required values in the **Number of trials:** and **Number of events:** boxes, respectively. Click on the **Options** button, and enter the value of the proportion from the null hypothesis in the **Test proportion:** box. Select the appropriate alternative hypothesis from the **Alternative** box, and check the box next to **Use test and interval based on normal distribution**. Click **OK** in both windows. The output will appear in the **Session** window, which will give the *p*-value for the test. Based on this *p*-value, you can make a decision.

**Excel**

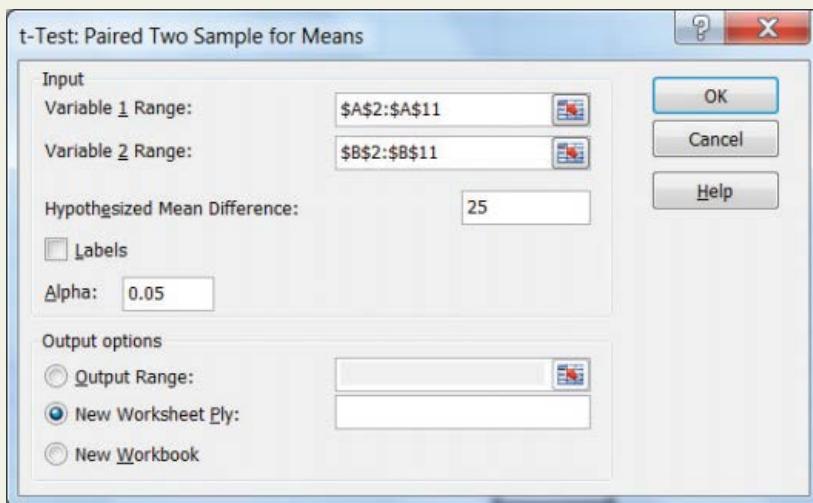
	A	B
1	data	
2	14.3	0
3	25.2	0
4	22.5	0
5	38.3	0
6	16.9	0
7	26.7	0
8	19.5	0
9	23.1	0
10	41	0
11	33.9	0

Screen 9.5

The Data Analysis ToolPak does not contain a preprogrammed function for a test about a population mean in which the population standard deviation is known. The Excel function **ZTEST** works easily only in specific situations, and it requires substantial adjustment in a number of situations, so it will not be discussed here.

The Data Analysis ToolPak also does not contain a preprogrammed function for a test about a population mean in which the population standard deviation is unknown. However, the function used for the paired *t*-test, which is covered in Chapter 10, can be manipulated relatively easily in order to produce results for a one-sample *t*-test. (Note: The Excel function **TTEST** has issues similar to the **ZTEST** function.)

1. Create a second column of data that is the same length as the data that you wish to analyze. All of the entries in the second column of data should be zero. (See Screen 9.5.)
2. Click the **Data** tab, then click the **Data Analysis** button within the **Analysis** group. From the **Data Analysis** window that will appear, select **t-test: Paired Two Sample for Means**.
3. Enter the location of the data you wish to analyze in the **Variable 1 Range** box. Enter the location of the column of zeroes in the **Variable 2 Range** box. Enter the value for  $\mu$  in the null hypothesis in the **Hypothesized Mean Difference** box. Enter the significance level, as a decimal, in the **Alpha** box. Choose how you wish the output to appear. (See Screen 9.6.) Click **OK**.



Screen 9.6

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		Variable 1	Variable 2
4	Mean	26.14	0
5	Variance	80.24933	0
6	Observations	10	10
7	Pearson Correlation	#DIV/0!	
8	Hypothesized Mean Difference	25	
9	df	9	
10	t Stat	0.402424	
11	P(T<=t) one-tail	0.348381	
12	t Critical one-tail	1.833113	
13	P(T<=t) two-tail	0.696762	
14	t Critical two-tail	2.262157	

Screen 9.7

4. The two lines in the output that you will need to determine the *p*-value are the lines labeled **t Stat** and **P(T<=t) two-tail**. (See Screen 9.7.) If the alternative hypothesis is two-tailed, the value in the **P(T<=t) two-tail** box is the *p*-value for the test. If the alternative hypothesis is one-tailed, use the following set of rules:

- a. If the hypothesis test is left-tailed and the value of **t Stat** is negative OR the hypothesis test is right-tailed and the value of **t Stat** is positive, the *p*-value of the test is equal to one-half the value in the **P(T<=t) two-tail** box.
- b. If the hypothesis test is left-tailed and the value of **t Stat** is positive OR the hypothesis test is right-tailed and the value of **t Stat** is negative, the *p*-value of the test is equal to 1 minus one-half the value in the **P(T<=t) two-tail** box.

## TECHNOLOGY ASSIGNMENTS

**TA9.1** According to Freddie Mac (<http://www.freddiemac.com/pmms/pmms30.htm>), the average interest rate on a 30-year fixed-rate mortgage in October 2011 was 4.07%. The following data represent the interest rates on 50 randomly selected 30-year fixed-rate mortgages approved during the week of November 7–11, 2011:

4.19	4.43	4.43	4.29	4.16	4.42	3.76	4.07	3.78	3.97
4.15	4.04	3.86	3.96	4.03	4.45	4.40	4.07	4.30	3.98
3.96	3.68	4.64	3.95	3.95	3.83	4.26	4.28	4.30	3.87
3.89	4.10	4.33	3.84	3.94	3.78	4.11	4.19	4.26	3.91
3.96	4.19	4.10	3.73	4.04	4.69	3.88	4.34	3.93	4.09

- a. Perform a graphical analysis to determine whether the assumption that the distribution of all 30-year fixed-rate mortgage rates is normal is a reasonable assumption.
- b. Test at a 5% significance level whether the average interest rate on all 30-year fixed-rate mortgages granted during the week of November 7–11, 2011, was different from 4.07%.

**TA9.2** Some colleges are known for their excellent cafeteria food, so much so that the term “Freshman 15” has been coined to refer to the amount of weight that students gain during their freshman year at college. The following data represent the amount of weight gained by 40 randomly selected students from a college during their freshman year. Note that a negative value implies that a student lost weight.

21.1	17.7	25.1	9.8	25.9	5.3	0.3	23.4	22.4	7.6
25.5	15.9	24.2	27.5	-3.0	8.7	13.6	11.2	13.5	7.8
17.8	5.9	-2.4	2.9	0.7	-1.0	25.7	18.0	28.7	3.2
2.2	26.7	24.5	10.5	25.5	-3.2	-0.5	8.0	5.7	-4.6

Although this college is happy about the reputation of its food service, it is concerned about the health issues of substantial weight gains. As a result, it distributed nutrition pamphlets to students in an attempt to reduce the amount of weight gain. Perform a hypothesis test at a 10% significance level to determine whether the average weight gain by all freshmen at this college during the first year is less than 15 pounds.

**TA9.3** General Logs Banana Bombs cereal is sold in 10.40-ounce packages. Because the cereal is sold by weight, the number of pieces of Banana Bombs varies from box to box. The following values represent the number of pieces in 19 boxes of Banana Bombs.

686	695	690	681	683	705	724	701	689	698
715	703	711	676	686	695	697	707	693	

Perform a hypothesis test to determine whether the average number of pieces in all 10.40-ounce boxes of Banana Bombs is different from 700. Assume that the distribution of the number of pieces in a 10.40-ounce box is approximately normal. Use  $\alpha = .05$ .

**TA9.4** According to a basketball coach, the mean height of all male college basketball players is 74 inches. A random sample of 25 such players produced the following data on their heights.

68	76	74	83	77	76	69	67	71	74	79	85	69
78	75	78	68	72	83	79	82	76	69	70	81	

Test at a 2% significance level whether the mean height of all male college basketball players is different from 74 inches. Assume that the heights of all male college basketball players are (approximately) normally distributed.

**TA9.5** A past study claimed that adults in America spent an average of 18 hours a week on leisure activities. A researcher took a sample of 10 adults from a town and asked them about the time they spend per week on leisure activities. Their responses (in hours) follow.

14	25	22	38	16	26	19	23	41	33
----	----	----	----	----	----	----	----	----	----

Assume that the times spent on leisure activities by all adults are normally distributed and the population standard deviation is 3 hours. Using a 5% significance level, can you conclude that the claim of the earlier study is true?

**TA9.6** Take a random sample of 150 runners from Data Set IV, which contains the 2011 Beach to Beacon 10K road race results. Use these data to perform 10 hypothesis tests on the variable that identifies whether a runner is from Maine, where  $p$  is the proportion of all runners who are from Maine. Specifically, perform a test for each of the sets of hypotheses

$$H_0: p = .70 \quad \text{versus} \quad H_1: p \neq .70$$

through

$$H_0: p = .79 \quad \text{versus} \quad H_1: p \neq .79$$

incrementing the hypothesized value by .01. Which null hypotheses are rejected at the 10% significance level? What conclusions can you make about the value of  $p$  based on the results of these tests?

**TA9.7** A mail-order company claims that at least 60% of all orders it receives are mailed within 48 hours. From time to time the quality control department at the company checks if this promise is kept. Recently, the quality control department at this company took a sample of 400 orders and found that 224 of them were mailed within 48 hours of the placement of the orders. Test at a 1% significance level whether or not the company's claim is true.

# CHAPTER 10



© christopherandt/Stockphoto

## Estimation and Hypothesis Testing: Two Populations

### 10.1 Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Known

### 10.2 Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Unknown but Equal

#### Case Study 10-1 One-Way Commute Times for Six Cities

### 10.3 Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Unknown and Unequal

### 10.4 Inferences About the Difference Between Two Population Means for Paired Samples

### 10.5 Inferences About the Difference Between Two Population Proportions for Large and Independent Samples

#### Case Study 10-2 Do You Worry About Your Weight?

Do you commute to work in a car alone or on a motorbike? How long does your typical one-way commute take? Do you know how much time adults, who drive a car alone or ride a motorbike as their main mode of transportation to work or school, spend commuting? A 2011 IBM Commuter Pain Survey of adults aged 18 to 65 years gathered data on typical commuting times from 20 cities around the world. For example, the average commute times for samples of such adults were found to be 30.3 minutes for New York City, 30.6 minutes for Chicago, and 29.5 minutes for Los Angeles, while the average for all 20 cities was 33 minutes. (See Case Study 10-1.)

Chapters 8 and 9 discussed the estimation and hypothesis-testing procedures for  $\mu$  and  $p$  involving a single population. This chapter extends the discussion of estimation and hypothesis-testing procedures to the difference between two population means and the difference between two population proportions. For example, we may want to make a confidence interval for the difference between the mean prices of houses in California and in New York, or we may want to test the hypothesis that the mean price of houses in California is different from that in New York. As another example, we may want to make a confidence interval for the difference between the proportions of all male and female adults who abstain from drinking, or we may want to test the hypothesis that the proportion of all adult men who abstain from drinking is different from the proportion of all adult women who abstain from drinking. Constructing confidence intervals and testing hypotheses about population parameters are referred to as *making inferences*.

## 10.1 Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Known

Let  $\mu_1$  be the mean of the first population and  $\mu_2$  be the mean of the second population. Suppose we want to make a confidence interval and test a hypothesis about the difference between these two population means, that is,  $\mu_1 - \mu_2$ . Let  $\bar{x}_1$  be the mean of a sample taken from the first population and  $\bar{x}_2$  be the mean of a sample taken from the second population. Then,  $\bar{x}_1 - \bar{x}_2$  is the sample statistic that is used to make an interval estimate and to test a hypothesis about  $\mu_1 - \mu_2$ . This section discusses how to make confidence intervals and test hypotheses about  $\mu_1 - \mu_2$  when certain conditions (to be explained later in this section) are satisfied. First we explain the concepts of independent and dependent samples.

### 10.1.1 Independent Versus Dependent Samples

Two samples are **independent** if they are drawn from two different populations and the elements of one sample have no relationship to the elements of the second sample. If the elements of the two samples are somehow related, then the samples are said to be **dependent**. Thus, in two independent samples, the selection of one sample has no effect on the selection of the second sample.

#### Definition

**Independent Versus Dependent Samples** Two samples drawn from two populations are *independent* if the selection of one sample from one population does not affect the selection of the second sample from the second population. Otherwise, the samples are *dependent*.

Examples 10–1 and 10–2 illustrate independent and dependent samples, respectively.

#### ■ EXAMPLE 10–1

Suppose we want to estimate the difference between the mean salaries of all male and all female executives. To do so, we draw two samples, one from the population of male executives and another from the population of female executives. These two samples are *independent* because they are drawn from two different populations, and the samples have no effect on each other. ■

Illustrating two independent samples.

#### ■ EXAMPLE 10–2

Suppose we want to estimate the difference between the mean weights of all participants before and after a weight loss program. To accomplish this, suppose we take a sample of 40 participants and measure their weights before and after the completion of this program. Note that these two samples include the same 40 participants. This is an example of two *dependent* samples. ■

Illustrating two dependent samples.

This section and Sections 10.2, 10.3, and 10.5 discuss how to make confidence intervals and test hypotheses about the difference between two population parameters when samples are independent. Section 10.4 discusses how to make confidence intervals and test hypotheses about the difference between two population means when samples are dependent.

### 10.1.2 Mean, Standard Deviation, and Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

Suppose we select two (independent) samples from two different populations that are referred to as population 1 and population 2. Let

- $\mu_1$  = the mean of population 1
- $\mu_2$  = the mean of population 2
- $\sigma_1$  = the standard deviation of population 1
- $\sigma_2$  = the standard deviation of population 2
- $n_1$  = the size of the sample drawn from population 1
- $n_2$  = the size of the sample drawn from population 2
- $\bar{x}_1$  = the mean of the sample drawn from population 1
- $\bar{x}_2$  = the mean of the sample drawn from population 2

Then, as we discussed in Chapters 8 and 9, if

1. The standard deviation  $\sigma_1$  of population 1 is known
2. At least one of the following two conditions is fulfilled:
  - i. The sample is large (i.e.,  $n_1 \geq 30$ )
  - ii. If the sample size is small, then the population from which the sample is drawn is normally distributed

then the sampling distribution of  $\bar{x}_1$  is normal with its mean equal to  $\mu_1$  and the standard deviation equal to  $\sigma_1/\sqrt{n_1}$ , assuming that  $n_1/N_1 \leq .05$ .

Similarly, if

1. The standard deviation  $\sigma_2$  of population 2 is known
2. At least one of the following two conditions is fulfilled:
  - i. The sample is large (i.e.,  $n_2 \geq 30$ )
  - ii. If the sample size is small, then the population from which the sample is drawn is normally distributed

then the sampling distribution of  $\bar{x}_2$  is normal with its mean equal to  $\mu_2$  and the standard deviation equal to  $\sigma_2/\sqrt{n_2}$ , assuming that  $n_2/N_2 \leq .05$ .

Using these results, we can make the following statements about the mean, the standard deviation, and the shape of the sampling distribution of  $\bar{x}_1 - \bar{x}_2$ .

If the following conditions are satisfied,

1. The two samples are independent
2. The standard deviations  $\sigma_1$  and  $\sigma_2$  of the two populations are known
3. At least one of the following two conditions is fulfilled:
  - i. Both samples are large (i.e.,  $n_1 \geq 30$  and  $n_2 \geq 30$ )
  - ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed

then the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  is (approximately) normally distributed with its mean and standard deviation,<sup>1</sup> respectively,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

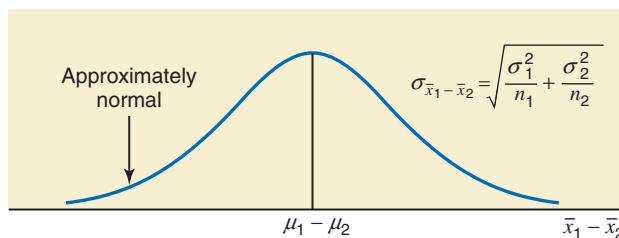
$$\text{and } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In these cases, we can use the normal distribution to make a confidence interval and test a hypothesis about  $\mu_1 - \mu_2$ . Figure 10.1 shows the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  when the above conditions are fulfilled.

<sup>1</sup>The formula for the standard deviation of  $\bar{x}_1 - \bar{x}_2$  can also be written as

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}$$

where  $\sigma_{\bar{x}_1} = \sigma_1/\sqrt{n_1}$  and  $\sigma_{\bar{x}_2} = \sigma_2/\sqrt{n_2}$ .

**Figure 10.1** The sampling distribution of  $\bar{x}_1 - \bar{x}_2$ .

**Sampling Distribution, Mean, and Standard Deviation of  $\bar{x}_1 - \bar{x}_2$**  When the conditions listed on the previous page are satisfied, the *sampling distribution* of  $\bar{x}_1 - \bar{x}_2$  is (approximately) normal with its *mean* and *standard deviation* as, respectively,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note that to apply the procedures learned in this chapter, the samples selected must be simple random samples.

### 10.1.3 Interval Estimation of $\mu_1 - \mu_2$

By constructing a confidence interval for  $\mu_1 - \mu_2$ , we find the difference between the means of two populations. For example, we may want to find the difference between the mean heights of male and female adults. The difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ , is the point estimator of the difference between the two population means,  $\mu_1 - \mu_2$ . When the conditions mentioned earlier in this section hold true, we use the normal distribution to make a confidence interval for the difference between the two population means. The following formula gives the interval estimation for  $\mu_1 - \mu_2$ .

**Confidence Interval for  $\mu_1 - \mu_2$**  When using the normal distribution, the  $(1 - \alpha)100\%$  confidence interval for  $\mu_1 - \mu_2$  is

$$(\bar{x}_1 - \bar{x}_2) \pm z\sigma_{\bar{x}_1 - \bar{x}_2}$$

The value of  $z$  is obtained from the normal distribution table for the given confidence level. The value of  $\sigma_{\bar{x}_1 - \bar{x}_2}$  is calculated as explained earlier. Here,  $\bar{x}_1 - \bar{x}_2$  is the point estimator of  $\mu_1 - \mu_2$ .

Note that in the real world,  $\sigma_1$  and  $\sigma_2$  are never known. Consequently we will never use the procedures of this section, but we are discussing these procedures in this book for the information of the readers.

Example 10–3 illustrates the procedure to construct a confidence interval for  $\mu_1 - \mu_2$  using the normal distribution.

### ■ EXAMPLE 10–3

According to the Kaiser Family Foundation surveys in 2011 and 2010, the average annual premium for employer-sponsored health insurance for family coverage was \$15,073 in 2011 and \$13,770 in 2010 (*USA TODAY*, September 29, 2011). Suppose that these averages are based on random samples of 250 and 200 employees who had such employer-sponsored health insurance plans for 2011 and 2010, respectively. Further assume that the population standard deviations for 2011 and 2010 were \$2160 and \$1990, respectively. Let  $\mu_1$  and  $\mu_2$  be the population means for such annual premiums for the years 2011 and 2010, respectively.

*Constructing a confidence interval for  $\mu_1 - \mu_2$ :  $\sigma_1$  and  $\sigma_2$  known, and samples are large.*

- (a) What is the point estimate of  $\mu_1 - \mu_2$ ?
- (b) Construct a 97% confidence interval for  $\mu_1 - \mu_2$ .

**Solution** Let us refer to the employees who had employer-sponsored health insurance for family coverage in 2011 as population 1 and those for 2010 as population 2. Then the respective samples are samples 1 and 2. Let  $\bar{x}_1$  and  $\bar{x}_2$  be the means of the two samples, respectively. From the given information,

$$\text{For 2011: } n_1 = 250, \quad \bar{x}_1 = \$15,073, \quad \sigma_1 = \$2160$$

$$\text{For 2010: } n_2 = 200, \quad \bar{x}_2 = \$13,770, \quad \sigma_2 = \$1990$$

- (a) The point estimate of  $\mu_1 - \mu_2$  is given by the value of  $\bar{x}_1 - \bar{x}_2$ . Thus,

$$\text{Point estimate of } \mu_1 - \mu_2 = \$15,073 - \$13,770 = \$1303$$

- (b) The confidence level is  $1 - \alpha = .97$ . From the normal distribution table, the values of  $z$  for .0150 and .9850 areas to the left are  $-2.17$  and  $2.17$ , respectively. Hence, we will use  $z = 2.17$  in the confidence interval formula. First we calculate the standard deviation of  $\bar{x}_1 - \bar{x}_2$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2}$ , as follows:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2160)^2}{250} + \frac{(1990)^2}{200}} = \$196.1196064$$

Next, substituting all the values in the confidence interval formula, we obtain a 97% confidence interval for  $\mu_1 - \mu_2$  as

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm z\sigma_{\bar{x}_1 - \bar{x}_2} &= (\$15,073 - \$13,770) \pm 2.17(196.1196064) \\ &= 1303 \pm 425.58 = \$877.42 \text{ to } \$1728.58 \end{aligned}$$

Thus, with 97% confidence we can state that the difference between the average annual premiums for employer-sponsored health insurance for family coverage in 2011 and 2010 is between \$877.42 and \$1728.58. The value  $z\sigma_{\bar{x}_1 - \bar{x}_2} = \$425.58$  is called the margin of error for this estimate. ■

Note that in Example 10–3 both sample sizes were large and the population standard deviations were known. If the standard deviations of the two populations are known, at least one of the sample sizes is small, and both populations are normally distributed, we use the normal distribution to make a confidence interval for  $\mu_1 - \mu_2$ . The procedure in this case is exactly the same as in Example 10–3.

#### 10.1.4 Hypothesis Testing About $\mu_1 - \mu_2$

It is often necessary to compare the means of two populations. For example, we may want to know if the mean price of houses in Chicago is the same as that in Los Angeles. Similarly, we may be interested in knowing if, on average, American children spend fewer hours in school than Japanese children do. In both these cases, we will perform a test of hypothesis about  $\mu_1 - \mu_2$ . The alternative hypothesis in a test of hypothesis may be that the means of the two populations are different, or that the mean of the first population is greater than the mean of the second population, or that the mean of the first population is less than the mean of the second population. These three situations are described next.

1. Testing an alternative hypothesis that the means of two populations are different is equivalent to  $\mu_1 \neq \mu_2$ , which is the same as  $\mu_1 - \mu_2 \neq 0$ .
2. Testing an alternative hypothesis that the mean of the first population is greater than the mean of the second population is equivalent to  $\mu_1 > \mu_2$ , which is the same as  $\mu_1 - \mu_2 > 0$ .
3. Testing an alternative hypothesis that the mean of the first population is less than the mean of the second population is equivalent to  $\mu_1 < \mu_2$ , which is the same as  $\mu_1 - \mu_2 < 0$ .

The procedure that is followed to perform a test of hypothesis about the difference between two population means is similar to the one that is used to test hypotheses about single-population parameters in Chapter 9. The procedure involves the same five steps for the critical-value approach that were used in Chapter 9 to test hypotheses about  $\mu$  and  $p$ . Here, again, if the following

conditions are satisfied, we will use the normal distribution to make a test of hypothesis about  $\mu_1 - \mu_2$ .

1. The two samples are independent.
2. The standard deviations  $\sigma_1$  and  $\sigma_2$  of the two populations are known.
3. At least one of the following two conditions is fulfilled:
  - i. Both samples are large (i.e.,  $n_1 \geq 30$  and  $n_2 \geq 30$ )
  - ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed

**Test Statistic  $z$  for  $\bar{x}_1 - \bar{x}_2$**  When using the normal distribution, the value of the *test statistic  $z$*  for  $\bar{x}_1 - \bar{x}_2$  is computed as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

The value of  $\mu_1 - \mu_2$  is substituted from  $H_0$ . The value of  $\sigma_{\bar{x}_1 - \bar{x}_2}$  is calculated as earlier in this section.

Example 10–4 shows how to make a test of hypothesis about  $\mu_1 - \mu_2$ .

## ■ EXAMPLE 10–4

Refer to Example 10–3 about the average annual premiums for employer-sponsored health insurance for family coverage in 2011 and 2010. Test at a 1% significance level whether the population means for the two years are different.

**Solution** From the information given in Example 10–3,

$$\text{For 2011: } n_1 = 250, \quad \bar{x}_1 = \$15,073, \quad \sigma_1 = \$2160$$

$$\text{For 2010: } n_2 = 200, \quad \bar{x}_2 = \$13,770, \quad \sigma_2 = \$1990$$

Let  $\mu_1$  and  $\mu_2$  be the population means for such annual premiums for the years 2011 and 2010, respectively. Let  $\bar{x}_1$  and  $\bar{x}_2$  be the corresponding sample means.

**Step 1. State the null and alternative hypotheses.**

We are to test whether the two population means are different. The two possibilities are:

- i. The mean annual premiums for the years 2011 and 2010 are not different. In other words,  $\mu_1 = \mu_2$ , which can be written as  $\mu_1 - \mu_2 = 0$ .
- ii. The mean annual premiums for the years 2011 and 2010 are different. That is,  $\mu_1 \neq \mu_2$ , which can be written as  $\mu_1 - \mu_2 \neq 0$ .

Considering these two possibilities, the null and alternative hypotheses are, respectively,

$$H_0: \mu_1 - \mu_2 = 0 \quad (\text{The two population means are not different.})$$

$$H_1: \mu_1 - \mu_2 \neq 0 \quad (\text{The two population means are different.})$$

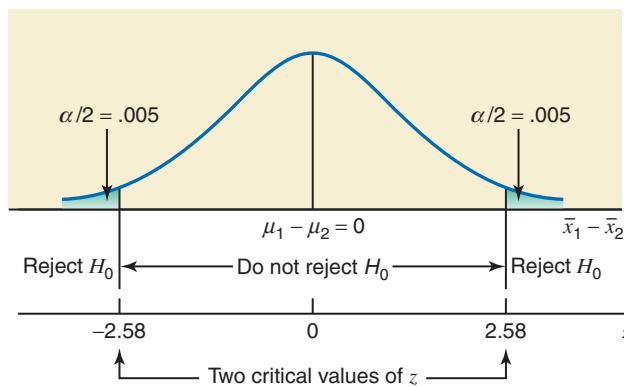
**Step 2. Select the distribution to use.**

Here, the population standard deviations,  $\sigma_1$  and  $\sigma_2$ , are known, and both samples are large ( $n_1 \geq 30$  and  $n_2 \geq 30$ ). Therefore, the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  is approximately normal, and we use the normal distribution to perform the hypothesis test.

**Step 3. Determine the rejection and nonrejection regions.**

The significance level is given to be .01. The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The area in each tail of the normal distribution curve is  $\alpha/2 = .01/2 = .005$ . The critical values of  $z$  for .0050 and .9950 areas to the left are (approximately) –2.58 and 2.58 from Table IV of Appendix C. These values are shown in Figure 10.2.

Making a two-tailed test of hypothesis about  $\mu_1 - \mu_2$ :  $\sigma_1$  and  $\sigma_2$  are known, and samples are large.

**Figure 10.2** Rejection and nonrejection regions.**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $z$  for  $\bar{x}_1 - \bar{x}_2$  is computed as follows:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2160)^2}{250} + \frac{(1990)^2}{200}} = \$196.1196064$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\$15,073 - \$13,770) - 0}{196.1196064} = 6.64$$

From  $H_0$

**Step 5.** Make a decision.

Because the value of the test statistic  $z = 6.64$  falls in the rejection region, we reject the null hypothesis  $H_0$ . Therefore, we conclude that the average annual premiums for employer-sponsored health insurance for family coverage were different for 2011 and 2010.

**Using the *p*-Value to Make a Decision**

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2. Then in Step 3 we calculate the value of the test statistic  $z$  (as done in Step 4) and find the *p*-value for this  $z$  from the normal distribution table. In Step 4, the  $z$  value for  $\bar{x}_1 - \bar{x}_2$  was calculated to be 6.64. In this example, the test is two-tailed. The *p*-value is equal to twice the area under the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  to the right of  $z = 6.64$ . From the normal distribution table (Table IV in Appendix C), the area to the right of  $z = 6.64$  is (approximately) zero. Therefore, the *p*-value is zero. As we know from Chapter 9, we will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than or equal to the *p*-value. Consequently, in this example, we will reject the null hypothesis for (almost) any  $\alpha > 0$ . Since  $\alpha = .01$  in this example, which is greater than zero, we reject the null hypothesis. ■

**EXERCISES****CONCEPTS AND PROCEDURES**

**10.1** Briefly explain the meaning of independent and dependent samples. Give one example of each.

**10.2** Describe the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  for two independent samples when  $\sigma_1$  and  $\sigma_2$  are known and either both sample sizes are large or both populations are normally distributed. What are the mean and standard deviation of this sampling distribution?

**10.3** The following information is obtained from two independent samples selected from two normally distributed populations.

$$n_1 = 18 \quad \bar{x}_1 = 7.82 \quad \sigma_1 = 2.35$$

$$n_2 = 15 \quad \bar{x}_2 = 5.99 \quad \sigma_2 = 3.17$$

a. What is the point estimate of  $\mu_1 - \mu_2$ ?

b. Construct a 99% confidence interval for  $\mu_1 - \mu_2$ . Find the margin of error for this estimate.

**10.4** The following information is obtained from two independent samples selected from two populations.

$$\begin{array}{lll} n_1 = 650 & \bar{x}_1 = 1.05 & \sigma_1 = 5.22 \\ n_2 = 675 & \bar{x}_2 = 1.54 & \sigma_2 = 6.80 \end{array}$$

- a. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 95% confidence interval for  $\mu_1 - \mu_2$ . Find the margin of error for this estimate.

**10.5** Refer to the information given in Exercise 10.3. Test at a 5% significance level if the two population means are different.

**10.6** Refer to the information given in Exercise 10.4. Test at a 1% significance level if the two population means are different.

**10.7** Refer to the information given in Exercise 10.4. Test at a 5% significance level if  $\mu_1$  is less than  $\mu_2$ .

**10.8** Refer to the information given in Exercise 10.3. Test at a 1% significance level if  $\mu_1$  is greater than  $\mu_2$ .

## ■ APPLICATIONS

**10.9** In parts of the eastern United States, whitetail deer are a major nuisance to farmers and homeowners, frequently damaging crops, gardens, and landscaping. A consumer organization arranges a test of two of the leading deer repellents A and B on the market. Fifty-six unfenced gardens in areas having high concentrations of deer are used for the test. Twenty-nine gardens are chosen at random to receive repellent A, and the other 27 receive repellent B. For each of the 56 gardens, the time elapsed between application of the repellent and the appearance in the garden of the first deer is recorded. For repellent A, the mean time is 101 hours. For repellent B, the mean time is 92 hours. Assume that the two populations of elapsed times have normal distributions with population standard deviations of 15 and 10 hours, respectively.

- a. Let  $\mu_1$  and  $\mu_2$  be the population means of elapsed times for the two repellents, respectively. Find the point estimate of  $\mu_1 - \mu_2$ .
- b. Find a 97% confidence interval for  $\mu_1 - \mu_2$ .
- c. Test at a 2% significance level whether the mean elapsed times for repellents A and B are different. Use both approaches, the critical-value and  $p$ -value, to perform this test.

**10.10** The U.S. Department of Labor collects data on unemployment insurance payments made to unemployed people in different states. Suppose that during 2011 a random sample of 1000 unemployed people in Florida received an average weekly unemployment benefit of \$219.65, while a random sample of 900 unemployed people in Mississippi received an average weekly unemployment benefit of \$191.47. Assume that the population standard deviations of 2011 weekly unemployment benefits paid to all unemployed workers in Florida and Mississippi were \$35.15 and \$28.22, respectively. (Note: A 2011 study by DailyFinance.com (<http://www.dailyfinance.com/2011/05/12/unemployment-benefits-best-worst-states/>) rated Mississippi and Florida as the two worst states for unemployment benefits.)

- a. Let  $\mu_1$  and  $\mu_2$  be the means of weekly unemployment benefits paid to all unemployed workers during 2011 in Florida and Mississippi, respectively. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 96% confidence interval for  $\mu_1 - \mu_2$ .
- c. Using a 2% significance level, can you conclude that the means of all weekly unemployment benefits paid to all unemployed workers during 2011 in Florida and Mississippi are different? Use both the  $p$ -value and the critical-value approaches to make this test.

**10.11** A local college cafeteria has a self-service soft ice cream machine. The cafeteria provides bowls that can hold up to 16 ounces of ice cream. The food service manager is interested in comparing the average amount of ice cream dispensed by male students to the average amount dispensed by female students. A measurement device was placed on the ice cream machine to determine the amounts dispensed. Random samples of 85 male and 78 female students who got ice cream were selected. The sample averages were 7.23 and 6.49 ounces for the male and female students, respectively. Assume that the population standard deviations are 1.22 and 1.17 ounces, respectively.

- a. Let  $\mu_1$  and  $\mu_2$  be the population means of ice cream amounts dispensed by all male and all female students at this college, respectively. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 95% confidence interval for  $\mu_1 - \mu_2$ .
- c. Using a 1% significance level, can you conclude that the average amount of ice cream dispensed by all male college students is larger than the average amount dispensed by all female college students? Use both approaches to make this test.

**10.12** Employees of a large corporation are concerned about the declining quality of medical services provided by their group health insurance. A random sample of 100 office visits by employees of this corporation to primary care physicians during 2004 found that the doctors spent an average of 19 minutes with

each patient. This year a random sample of 108 such visits showed that doctors spent an average of 15.5 minutes with each patient. Assume that the standard deviations for the two populations are 2.7 and 2.1 minutes, respectively.

- Construct a 95% confidence interval for the difference between the two population means for these two years.
- Using a 2.5% level of significance, can you conclude that the mean time spent by doctors with each patient is lower for this year than for 2004?
- What would your decision be in part b if the probability of making a Type I error were zero? Explain.

**10.13** A car magazine is comparing the total repair costs incurred during the first three years on two sports cars, the T-999 and the XPY. Random samples of 45 T-999s and 51 XPYs are taken. All 96 cars are 3 years old and have similar mileages. The mean of repair costs for the 45 T-999 cars is \$3300 for the first 3 years. For the 51 XPY cars, this mean is \$3850. Assume that the standard deviations for the two populations are \$800 and \$1000, respectively.

- Construct a 99% confidence interval for the difference between the two population means.
- Using a 1% significance level, can you conclude that such mean repair costs are different for these two types of cars?
- What would your decision be in part b if the probability of making a Type I error were zero? Explain.

**10.14** The management at New Century Bank claims that the mean waiting time for all customers at its branches is less than that at the Public Bank, which is its main competitor. A business consulting firm took a sample of 200 customers from the New Century Bank and found that they waited an average of 4.5 minutes before being served. Another sample of 300 customers taken from the Public Bank showed that these customers waited an average of 4.75 minutes before being served. Assume that the standard deviations for the two populations are 1.2 and 1.5 minutes, respectively.

- Make a 97% confidence interval for the difference between the two population means.
- Test at a 2.5% significance level whether the claim of the management of the New Century Bank is true.
- Calculate the  $p$ -value for the test of part b. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .01$ ? What if  $\alpha = .05$ ?

**10.15** Maine Mountain Dairy claims that its 8-ounce low-fat yogurt cups contain, on average, fewer calories than the 8-ounce low-fat yogurt cups produced by a competitor. A consumer agency wanted to check this claim. A sample of 27 such yogurt cups produced by this company showed that they contained an average of 141 calories per cup. A sample of 25 such yogurt cups produced by its competitor showed that they contained an average of 144 calories per cup. Assume that the two populations are normally distributed with population standard deviations of 5.5 and 6.4 calories, respectively.

- Make a 98% confidence interval for the difference between the mean number of calories in the 8-ounce low-fat yogurt cups produced by the two companies.
- Test at a 1% significance level whether Maine Mountain Dairy's claim is true.
- Calculate the  $p$ -value for the test of part b. Based on this  $p$ -value, would you reject the null hypothesis if  $\alpha = .05$ ? What if  $\alpha = .025$ ?

## 10.2

### Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Unknown but Equal

This section discusses making a confidence interval and testing a hypothesis about the difference between the means of two populations,  $\mu_1 - \mu_2$ , assuming that the standard deviations,  $\sigma_1$  and  $\sigma_2$ , of these populations are not known but are assumed to be equal. There are some other conditions, explained below, that must be fulfilled to use the procedures discussed in this section.

If the following conditions are satisfied,

- The two samples are independent
- The standard deviations  $\sigma_1$  and  $\sigma_2$  of the two populations are unknown, but they can be assumed to be equal, that is,  $\sigma_1 = \sigma_2$

3. At least one of the following two conditions is fulfilled:
- Both samples are large (i.e.,  $n_1 \geq 30$  and  $n_2 \geq 30$ )
  - If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed

then we use the  $t$  distribution to make a confidence interval and test a hypothesis about the difference between the means of two populations,  $\mu_1 - \mu_2$ .

When the standard deviations of the two populations are equal, we can use  $\sigma$  for both  $\sigma_1$  and  $\sigma_2$ . Because  $\sigma$  is unknown, we replace it by its point estimator  $s_p$ , which is called the **pooled sample standard deviation** (hence, the subscript  $p$ ). The value of  $s_p$  is computed by using the information from the two samples as follows.

**Pooled Standard Deviation for Two Samples** The *pooled standard deviation for two samples* is computed as

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where  $n_1$  and  $n_2$  are the sizes of the two samples and  $s_1^2$  and  $s_2^2$  are the variances of the two samples, respectively. Here  $s_p$  is an estimator of  $\sigma$ .

In this formula,  $n_1 - 1$  are the degrees of freedom for sample 1,  $n_2 - 1$  are the degrees of freedom for sample 2, and  $n_1 + n_2 - 2$  are the *degrees of freedom for the two samples taken together*. Note that  $s_p$  is an estimator of the standard deviation,  $\sigma$ , of each of the two populations.

When  $s_p$  is used as an estimator of  $\sigma$ , the standard deviation  $\sigma_{\bar{x}_1 - \bar{x}_2}$  of  $\bar{x}_1 - \bar{x}_2$  is estimated by  $s_{\bar{x}_1 - \bar{x}_2}$ . The value of  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated by using the following formula.

**Estimator of the Standard Deviation of  $\bar{x}_1 - \bar{x}_2$**  The *estimator of the standard deviation of  $\bar{x}_1 - \bar{x}_2$*  is

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Now we are ready to discuss the procedures that are used to make confidence intervals and test hypotheses about  $\mu_1 - \mu_2$  for independent samples selected from two populations with unknown but equal standard deviations.

### 10.2.1 Interval Estimation of $\mu_1 - \mu_2$

As was mentioned earlier in this chapter, the difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ , is the point estimator of the difference between the two population means,  $\mu_1 - \mu_2$ . The following formula gives the confidence interval for  $\mu_1 - \mu_2$  when the  $t$  distribution is used and the conditions mentioned earlier in this section are fulfilled.

**Confidence Interval for  $\mu_1 - \mu_2$**  The  $(1 - \alpha)100\%$  confidence interval for  $\mu_1 - \mu_2$  is

$$(\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2}$$

where the value of  $t$  is obtained from the  $t$  distribution table for the given confidence level and  $n_1 + n_2 - 2$  degrees of freedom, and  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated as explained earlier.

Example 10–5 describes the procedure to make a confidence interval for  $\mu_1 - \mu_2$  using the  $t$  distribution.

### ■ EXAMPLE 10–5

*Constructing a confidence interval for  $\mu_1 - \mu_2$ : two independent samples, unknown but equal  $\sigma_1$  and  $\sigma_2$ .*



© ansonsaw/iStockphoto

A consumer agency wanted to estimate the difference in the mean amounts of caffeine in two brands of coffee. The agency took a sample of 15 one-pound jars of Brand I coffee that showed the mean amount of caffeine in these jars to be 80 milligrams per jar with a standard deviation of 5 milligrams. Another sample of 12 one-pound jars of Brand II coffee gave a mean amount of caffeine equal to 77 milligrams per jar with a standard deviation of 6 milligrams. Construct a 95% confidence interval for the difference between the mean amounts of caffeine in one-pound jars of these two brands of coffee. Assume that the two populations are normally distributed and that the standard deviations of the two populations are equal.

**Solution** Let  $\mu_1$  and  $\mu_2$  be the mean amounts of caffeine per jar in all 1-pound jars of Brands I and II, respectively, and let  $\bar{x}_1$  and  $\bar{x}_2$  be the means of the two respective samples. From the given information,

$$\begin{array}{lll} \text{Brand I coffee: } & n_1 = 15 & \bar{x}_1 = 80 \text{ milligrams} \\ & s_1 = 5 \text{ milligrams} \\ \text{Brand II coffee: } & n_2 = 12 & \bar{x}_2 = 77 \text{ milligrams} \\ & s_2 = 6 \text{ milligrams} \end{array}$$

The confidence level is  $1 - \alpha = .95$ .

Here,  $\sigma_1$  and  $\sigma_2$  are unknown but assumed to be equal, the samples are independent (taken from two different populations), and the sample sizes are small but the two populations are normally distributed. Hence, we will use the  $t$  distribution to make the confidence interval for  $\mu_1 - \mu_2$  as all conditions mentioned in the beginning of this section are satisfied.

First we calculate the standard deviation of  $\bar{x}_1 - \bar{x}_2$  as follows. Note that since it is assumed that  $\sigma_1$  and  $\sigma_2$  are equal, we will use  $s_p$  to calculate  $s_{\bar{x}_1 - \bar{x}_2}$ .

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(15 - 1)(5)^2 + (12 - 1)(6)^2}{15 + 12 - 2}} = 5.46260011$$

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (5.46260011) \sqrt{\frac{1}{15} + \frac{1}{12}} = 2.11565593$$

Next, to find the  $t$  value from the  $t$  distribution table, we need to know the area in each tail of the  $t$  distribution curve and the degrees of freedom.

$$\text{Area in each tail} = \alpha/2 = (1 - .95)/2 = .025$$

$$\text{Degrees of freedom} = n_1 + n_2 - 2 = 15 + 12 - 2 = 25$$

The  $t$  value for  $df = 25$  and .025 area in the right tail of the  $t$  distribution curve is 2.060. The 95% confidence interval for  $\mu_1 - \mu_2$  is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2} &= (80 - 77) \pm 2.060(2.11565593) \\ &= 3 \pm 4.36 = \mathbf{-1.36 \text{ to } 7.36 \text{ milligrams}} \end{aligned}$$

Thus, with 95% confidence we can state that based on these two sample results, the difference in the mean amounts of caffeine in 1-pound jars of these two brands of coffee lies between  $-1.36$  and  $7.36$  milligrams. Because the lower limit of the interval is negative, it is possible that the mean amount of caffeine is greater in the second brand than in the first brand of coffee.

Note that the value of  $\bar{x}_1 - \bar{x}_2$ , which is  $80 - 77 = 3$ , gives the point estimate of  $\mu_1 - \mu_2$ . The value of  $ts_{\bar{x}_1 - \bar{x}_2}$ , which is 4.36, is the margin of error. ■

## 10.2.2 Hypothesis Testing About $\mu_1 - \mu_2$

When the conditions mentioned in the beginning of Section 10.2 are satisfied, the  $t$  distribution is applied to make a hypothesis test about the difference between two population means. The test statistic in this case is  $t$ , which is calculated as follows.

**Test Statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$**  The value of the *test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$*  is computed as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

The value of  $\mu_1 - \mu_2$  in this formula is substituted from the null hypothesis, and  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated as explained earlier in Section 10.2.1.

Examples 10–6 and 10–7 illustrate how a test of hypothesis about the difference between two population means for independent samples that are selected from two populations with equal standard deviations is conducted using the  $t$  distribution.

### ■ EXAMPLE 10–6

A sample of 14 cans of Brand I diet soda gave the mean number of calories of 23 per can with a standard deviation of 3 calories. Another sample of 16 cans of Brand II diet soda gave the mean number of calories of 25 per can with a standard deviation of 4 calories. At a 1% significance level, can you conclude that the mean numbers of calories per can are different for these two brands of diet soda? Assume that the calories per can of diet soda are normally distributed for each of the two brands and that the standard deviations for the two populations are equal.

*Making a two-tailed test of hypothesis about  $\mu_1 - \mu_2$ : two independent samples, and unknown but equal  $\sigma_1$  and  $\sigma_2$*

**Solution** Let  $\mu_1$  and  $\mu_2$  be the mean numbers of calories per can for diet soda of Brand I and Brand II, respectively, and let  $\bar{x}_1$  and  $\bar{x}_2$  be the means of the respective samples. From the given information,

$$\text{Brand I diet soda: } n_1 = 14 \quad \bar{x}_1 = 23 \quad s_1 = 3$$

$$\text{Brand II diet soda: } n_2 = 16 \quad \bar{x}_2 = 25 \quad s_2 = 4$$

The significance level is  $\alpha = .01$ .

**Step 1.** *State the null and alternative hypotheses.*

We are to test for the difference in the mean numbers of calories per can for the two brands. The null and alternative hypotheses are, respectively,

$$H_0: \mu_1 - \mu_2 = 0 \quad (\text{The mean numbers of calories are not different})$$

$$H_1: \mu_1 - \mu_2 \neq 0 \quad (\text{The mean numbers of calories are different})$$

**Step 2.** *Select the distribution to use.*

Here, the two samples are independent,  $\sigma_1$  and  $\sigma_2$  are unknown but equal, and the sample sizes are small but both populations are normally distributed. Hence, all conditions mentioned in the beginning of Section 10.2 are fulfilled. Consequently, we will use the  $t$  distribution.

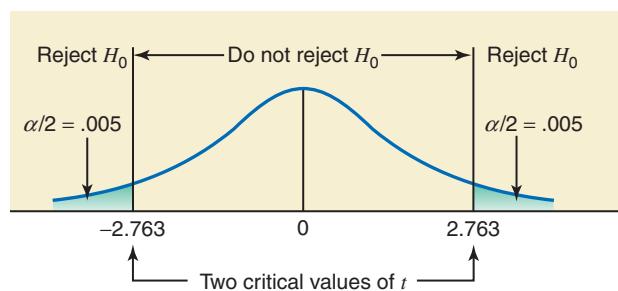
**Step 3.** *Determine the rejection and nonrejection regions.*

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The significance level is .01. Hence,

$$\text{Area in each tail} = \alpha/2 = .01/2 = .005$$

$$\text{Degrees of freedom} = n_1 + n_2 - 2 = 14 + 16 - 2 = 28$$

The critical values of  $t$  for  $df = 28$  and .005 area in each tail of the  $t$  distribution curve are  $-2.763$  and  $2.763$ , as shown in Figure 10.3.

**Figure 10.3** Rejection and nonrejection regions.**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$  is computed as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(14 - 1)(3)^2 + (16 - 1)(4)^2}{14 + 16 - 2}} = 3.57071421$$

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3.57071421) \sqrt{\frac{1}{14} + \frac{1}{16}} = 1.30674760$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(23 - 25) - 0}{1.30674760} = -1.531$$

From  $H_0$

**Step 5.** Make a decision.

Because the value of the test statistic  $t = -1.531$  for  $\bar{x}_1 - \bar{x}_2$  falls in the nonrejection region, we fail to reject the null hypothesis. Consequently we conclude that there is no difference in the mean numbers of calories per can for the two brands of diet soda. The difference in  $\bar{x}_1$  and  $\bar{x}_2$  observed for the two samples may have occurred due to sampling error only.

### Using the *p*-Value to Make a Decision

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2 of this example. Then in Step 3, we calculate the value of the test statistic  $t$  (as done in Step 4 above) and then find the *p*-value for this  $t$  from the *t* distribution table (Table V of Appendix C) or by using technology. In Step 4 above, the *t*-value for  $\bar{x}_1 - \bar{x}_2$  was calculated to be  $-1.531$ . In this example, the test is two-tailed. Therefore, the *p*-value is equal to twice the area under the *t* distribution curve to the left of  $t = -1.531$ . If we have access to technology, we can use it to find the exact *p*-value, which will be  $.137$ . If we use the *t* distribution table, we can only find the range for the *p*-value. From Table V of Appendix C, for  $df = 28$ , the two values that include  $1.531$  are  $1.313$  and  $1.701$ . (Note that we use the positive value of  $t$ , although our  $t$  is negative.) Thus, the test statistic  $t = -1.531$  falls between  $-1.313$  and  $-1.701$ . The areas in the *t* distribution table that correspond to  $1.313$  and  $1.701$  are  $.10$  and  $.05$ , respectively. Because it is a two-tailed test, the *p*-value for  $t = -1.531$  is between  $2(.10) = .20$  and  $2(.05) = .10$ , which can be written as

$$.10 < p\text{-value} < .20$$

As we know from Chapter 9, we will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than or equal to the *p*-value. Consequently, in this example, we will reject the null hypothesis for any  $\alpha \geq .20$  using the above range and not reject it for  $\alpha < .10$ . If we use technology, we will reject the null hypothesis for  $\alpha \geq .137$ . Since  $\alpha = .01$  in this example, which is smaller than both  $.10$  and  $.137$ , we fail to reject the null hypothesis. ■

### EXAMPLE 10-7

A sample of 40 children from New York State showed that the mean time they spend watching television is 28.50 hours per week with a standard deviation of 4 hours. Another sample of 35 children from California showed that the mean time spent by them watching television is 23.25 hours per week with a standard deviation of 5 hours. Using a 2.5% significance level, can you conclude that the mean time spent watching television by children in New York State is greater than that for children in California? Assume that the standard deviations for the two populations are equal.

Making a right-tailed test of hypothesis about  $\mu_1 - \mu_2$ : two independent samples,  $\sigma_1$  and  $\sigma_2$  unknown but equal, and both samples are large.

**Solution** Let the children from New York State be referred to as population 1 and those from California as population 2. Let  $\mu_1$  and  $\mu_2$  be the mean time spent watching television by children in populations 1 and 2, respectively, and let  $\bar{x}_1$  and  $\bar{x}_2$  be the mean time spent watching television by children in the respective samples. From the given information,

$$\begin{array}{lll} \text{New York: } & n_1 = 40 & \bar{x}_1 = 28.50 \text{ hours} \\ & & s_1 = 4 \text{ hours} \\ \text{California: } & n_2 = 35 & \bar{x}_2 = 23.25 \text{ hours} \\ & & s_2 = 5 \text{ hours} \end{array}$$

The significance level is  $\alpha = .025$ .



© Dori O'Connell/iStockphoto

**Step 1.** *State the null and alternative hypotheses.*

The two possible decisions are:

1. The mean time spent watching television by children in New York State is not greater than that for children in California. This can be written as  $\mu_1 = \mu_2$  or  $\mu_1 - \mu_2 = 0$ .
2. The mean time spent watching television by children in New York State is greater than that for children in California. This can be written as  $\mu_1 > \mu_2$  or  $\mu_1 - \mu_2 > 0$ .

Hence, the null and alternative hypotheses are, respectively,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Note that the null hypothesis can also be written as  $\mu_1 - \mu_2 \leq 0$ .

**Step 2.** *Select the distribution to use.*

Here, the two samples are independent (taken from two different populations),  $\sigma_1$  and  $\sigma_2$  are unknown but assumed to be equal, and both samples are large. Hence, all conditions mentioned in the beginning of Section 10.2 are fulfilled. Consequently, we use the  $t$  distribution to make the test.

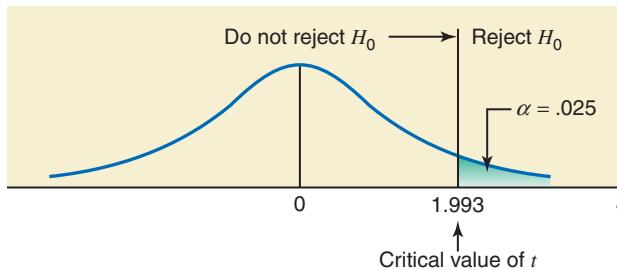
**Step 3.** *Determine the rejection and nonrejection regions.*

The  $>$  sign in the alternative hypothesis indicates that the test is right-tailed. The significance level is  $.025$ .

Area in the right tail of the  $t$  distribution =  $\alpha = .025$

Degrees of freedom =  $n_1 + n_2 - 2 = 40 + 35 - 2 = 73$

From the  $t$  distribution table, the critical value of  $t$  for  $df = 73$  and  $.025$  area in the right tail of the  $t$  distribution is 1.993. This value is shown in Figure 10.4.



**Figure 10.4** Rejection and nonrejection regions.

**Step 4.** *Calculate the value of the test statistic.*

The value of the test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$  is computed as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(40 - 1)(4)^2 + (35 - 1)(5)^2}{40 + 35 - 2}} = 4.49352655$$

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (4.49352655) \sqrt{\frac{1}{40} + \frac{1}{35}} = 1.04004930$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(28.50 - 23.25) - 0}{1.04004930} \stackrel{\downarrow}{=} 5.048 \quad \text{From } H_0$$

## ONE-WAY COMMUTE TIMES FOR SIX CITIES



Data source: 2011 IBM Commuter Pain Survey of adults age 18 to 65 years old.

The accompanying chart shows the average one-way commute times for six cities. These commute times are based on the IBM Commuter Pain Survey conducted in 2011 of adults age 18 to 65 years old selected from 20 cities around the world who drive a car alone or a motorbike as the main mode of transportation to work or school. We show the commuting times for only six of these 20 cities in this chart. We can pick any two cities from these six to make a confidence interval for the difference in their commute times and to test a hypothesis about whether the mean commute time for one selected city is lower than the mean commute time in another city. Suppose we pick Toronto and Chicago as the two cities. The sample sizes, as given in the chart, are 294 and 288 for Toronto and Chicago, respectively, and the mean commute times are 29.8 and 30.6 minutes, respectively. To make such a confidence interval and to test this hypothesis, we also need to know the standard deviations of these commute times. Suppose that the sample standard deviations of the one-way commute times in these two cities are 11 and 12 minutes, respectively. Also assume that although the population standard deviations are not known, they are (approximately) equal.

Let  $\mu_1$  and  $\mu_2$  be the mean one-way commute times for all adults age 18 to 65 years old in Toronto and Chicago, respectively, who drive a car alone or a motorbike as the main mode of transportation to work or school. Let  $\bar{x}_1$  and  $\bar{x}_2$  be the corresponding sample means. Then, from the given information:

$$\text{For Toronto: } n_1 = 294 \quad \bar{x}_1 = 29.8 \text{ minutes} \quad s_1 = 11 \text{ minutes}$$

$$\text{For Chicago: } n_2 = 288 \quad \bar{x}_2 = 30.6 \text{ minutes} \quad s_2 = 12 \text{ minutes}$$

Below we make a confidence interval for  $\mu_1 - \mu_2$  and test a hypothesis about  $\mu_1 - \mu_2$  for this example.

### 1. Confidence interval for $\mu_1 - \mu_2$

Suppose we want to make a 98% confidence interval for  $\mu_1 - \mu_2$ . The area in each tail of the  $t$  distribution and the degrees of freedom are

$$\text{Area in each tail} = \alpha/2 = (1 - .98)/2 = .01$$

$$\text{Degrees of freedom} = n_1 + n_2 - 2 = 294 + 288 - 2 = 580$$

### Step 5. Make a decision.

Because the value of the test statistic  $t = 5.048$  for  $\bar{x}_1 - \bar{x}_2$  falls in the rejection region (see Figure 10.4), we reject the null hypothesis  $H_0$ . Hence, we conclude that children in New York State spend more time, on average, watching TV than children in California.

Because  $df = 580$  is not in the  $t$  distribution table, we will use the last row of Table V to obtain the  $t$  value for .01 area in the right tail. This  $t$  value is 2.326.

We calculate the standard deviation of  $\bar{x}_1 - \bar{x}_2$  as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(294 - 1)(11)^2 + (288 - 1)(12)^2}{294 + 288 - 2}} = 11.50569574$$

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (11.50569574) \sqrt{\frac{1}{294} + \frac{1}{288}} = .95390356$$

Hence, the 98% confidence interval for  $\mu_1 - \mu_2$  is

$$(\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2} = (29.8 - 30.6) \pm 2.326 (.95390356) = -.80 \pm 2.22$$

**= -3.02 to 1.42 minutes**

Thus, the 98% confidence interval for  $\mu_1 - \mu_2$  is -3.02 to 1.42 minutes. As the first number of this interval is negative and the second number is positive, we can state that such mean commute time could be lower for Toronto by (at most) 3.02 minutes or it could be higher for Toronto by (at most) 1.42 minutes for a 98% confidence level.

## 2. Test of hypothesis about $\mu_1 - \mu_2$

Suppose we want to test, at a 1% significance level, if the mean commuting time mentioned above for all such commuters in Toronto is lower than the one for Chicago. In other words, we are to test if  $\mu_1$  is less than  $\mu_2$ . The null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 < \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 < 0$$

Note that the test is left-tailed. Because the population standard deviations are not known, we will use the  $t$  distribution. The area in the left tail of the  $t$  distribution and the degrees of freedom are

$$\text{Area in the left tail} = \alpha = .01$$

$$\text{Degrees of freedom} = n_1 + n_2 - 2 = 294 + 288 - 2 = 580$$

Because  $df = 580$  is not in the  $t$  distribution table, we will use the last row of Table V to obtain the  $t$  value for .01 area in the left tail. This  $t$  value is -2.326.

As calculated earlier, the standard deviation of  $\bar{x}_1 - \bar{x}_2$  is

$$s_{\bar{x}_1 - \bar{x}_2} = .95390356$$

The value of the test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$  is computed as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(29.8 - 30.6) - 0}{.95390356} = -.839$$

From  $H_0$

Because the value of the test statistic  $t = -.839$  is larger than the critical value of  $t = -2.326$ , it falls in the nonrejection region. Consequently, we fail to reject the null hypothesis and conclude that the mean commuting time in Toronto is not lower than the one in Chicago.

We can also use the  $p$ -value approach to make this decision. In this example, the test is left-tailed. As calculated above, the  $t$  value for  $\bar{x}_1 - \bar{x}_2$  is -.839. From the last row of the  $t$  distribution table, -.839 is greater than -1.282. Therefore, the  $p$ -value is higher than .10. (Actually, if you use technology, you will obtain a  $p$ -value of .201.) Since  $\alpha = .01$  in this example, it is smaller than this  $p$ -value, and we fail to reject the null hypothesis and conclude that the mean commuting time in Toronto is not lower than the one in Chicago.

*Note:* We are thankful to IBM for providing us the data from IBM Commuter Pain Survey 2011.

## Using the $p$ -Value to Make a Decision

To use the  $p$ -value approach to make the above decision, we keep Steps 1 and 2 of this example. Then in Step 3, we calculate the value of the test statistic  $t$  (as done in Step 4 above) and then find the  $p$ -value for this  $t$  from the  $t$  distribution table (Table V of Appendix C) or by using technology. In Step 4 above, the  $t$ -value for  $\bar{x}_1 - \bar{x}_2$  was calculated to be 5.048. In this

example, the test is right-tailed. Therefore, the  $p$ -value is equal to the area under the  $t$  distribution curve to the right of  $t = 5.048$ . If we have access to technology, we can use it to find the exact  $p$ -value, which will be .000. If we use the  $t$  distribution table, for  $df = 73$ , the value of the test statistic  $t = 5.048$  is larger than 3.206. Therefore, the  $p$ -value for  $t = 5.048$  is less than .001, which can be written as

$$p\text{-value} < .001$$

Since we will reject the null hypothesis for any  $\alpha$  (significance level) greater than or equal to the  $p$ -value, here we reject the null hypothesis because  $\alpha = .025$  is greater than both the  $p$ -values, .001 obtained above from the table and .000 obtained by using technology. Note that obtaining the  $p$ -value = .000 from technology does not mean that the  $p$ -value is zero. It means that when it is rounded to three digits after the decimal, it is .000. ■

### Note: What If the Sample Sizes Are Large and the Number of $df$ Are Not in the $t$ Distribution Table?

In this section, we used the  $t$  distribution to make confidence intervals and perform tests of hypothesis about  $\mu_1 - \mu_2$ . When both sample sizes are large, it does not matter how large the sample sizes are if we are using technology. However, if we are using the  $t$  distribution table (Table V of Appendix C), this may pose a problem if samples are too large. Table V in Appendix C goes up to only 75 degrees of freedom. Thus, if the degrees of freedom are larger than 75, we cannot use Table V to find the critical value(s) of  $t$ . As mentioned in Chapters 8 and 9, in such a situation, there are two options:

1. Use the  $t$  value from the last row (the row of  $\infty$ ) in Table V.
2. Use the normal distribution as an approximation to the  $t$  distribution.

A few of the exercises at the end of this section present such situations.

## EXERCISES

### CONCEPTS AND PROCEDURES

**10.16** Explain what conditions must hold true to use the  $t$  distribution to make a confidence interval and to test a hypothesis about  $\mu_1 - \mu_2$  for two independent samples selected from two populations with unknown but equal standard deviations.

**10.17** The following information was obtained from two independent samples selected from two normally distributed populations with unknown but equal standard deviations.

$$n_1 = 21 \quad \bar{x}_1 = 13.97 \quad s_1 = 3.78$$

$$n_2 = 20 \quad \bar{x}_2 = 15.55 \quad s_2 = 3.26$$

- a. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 95% confidence interval for  $\mu_1 - \mu_2$ .

**10.18** The following information was obtained from two independent samples selected from two populations with unknown but equal standard deviations.

$$n_1 = 55 \quad \bar{x}_1 = 90.40 \quad s_1 = 11.60$$

$$n_2 = 50 \quad \bar{x}_2 = 86.30 \quad s_2 = 10.25$$

- a. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 99% confidence interval for  $\mu_1 - \mu_2$ .

**10.19** Refer to the information given in Exercise 10.17. Test at a 5% significance level if the two population means are different.

**10.20** Refer to the information given in Exercise 10.18. Test at a 1% significance level if the two population means are different.

**10.21** Refer to the information given in Exercise 10.17. Test at a 1% significance level if  $\mu_1$  is less than  $\mu_2$ .

**10.22** Refer to the information given in Exercise 10.18. Test at a 5% significance level if  $\mu_1$  is greater than  $\mu_2$ .

- 10.23** The following information was obtained from two independent samples selected from two normally distributed populations with unknown but equal standard deviations.

Sample 1: 47.7 46.9 51.9 34.1 65.8 61.5 50.2 40.8 53.1 46.1 47.9 45.7 49.0  
 Sample 2: 50.0 47.4 32.7 48.8 54.0 46.3 42.5 40.8 39.0 68.2 48.5 41.8

- Let  $\mu_1$  be the mean of population 1 and  $\mu_2$  be the mean of population 2. What is the point estimate of  $\mu_1 - \mu_2$ ?
- Construct a 98% confidence interval for  $\mu_1 - \mu_2$ .
- Test at a 1% significance level if  $\mu_1$  is greater than  $\mu_2$ .

- 10.24** The following information was obtained from two independent samples selected from two normally distributed populations with unknown but equal standard deviations.

Sample 1: 2.18 2.23 1.96 2.24 2.72 1.87 2.68 2.15 2.49 2.05  
 Sample 2: 1.82 1.26 2.00 1.89 1.73 2.03 1.43 2.05 1.54 2.50 1.99 2.13

- Let  $\mu_1$  be the mean of population 1 and  $\mu_2$  be the mean of population 2. What is the point estimate of  $\mu_1 - \mu_2$ ?
- Construct a 99% confidence interval for  $\mu_1 - \mu_2$ .
- Test at a 2.5% significance level if  $\mu_1$  is lower than  $\mu_2$ .

## ■ APPLICATIONS

- 10.25** The standard recommendation for automobile oil changes is once every 3000 miles. A local mechanic is interested in determining whether people who drive more expensive cars are more likely to follow the recommendation. Independent random samples of 45 customers who drive luxury cars and 40 customers who drive compact lower-price cars were selected. The average distance driven between oil changes was 3187 miles for the luxury car owners and 3214 miles for the compact lower-price cars. The sample standard deviations were 42.40 and 50.70 miles for the luxury and compact groups, respectively. Assume that the population distributions of the distances between oil changes have the same standard deviation for the two populations.

- Construct a 95% confidence interval for the difference in the mean distances between oil changes for all luxury cars and all compact lower-price cars.
- Using a 1% significance level, can you conclude that the mean distance between oil changes is less for all luxury cars than that for all compact lower-price cars?

- 10.26** A town that recently started a single-stream recycling program provided 60-gallon recycling bins to 25 randomly selected households and 75-gallon recycling bins to 22 randomly selected households. The total volume of recycling over a 10-week period was measured for each of the households. The average total volumes were 382 and 415 gallons for the households with the 60- and 75-gallon bins, respectively. The sample standard deviations were 52.5 and 43.8 gallons, respectively. Assume that the 10-week total volumes of recycling are approximately normally distributed for both groups and that the population standard deviations are equal.

- Construct a 98% confidence interval for the difference in the mean volumes of 10-week recycling for the households with the 60- and 75-gallon bins.
- Using a 2% significance level, can you conclude that the average 10-week recycling volume of all households having 60-gallon containers is different from the average volume of all households that have 75-gallon containers?

- 10.27** An insurance company wants to know if the average speed at which men drive cars is greater than that of women drivers. The company took a random sample of 27 cars driven by men on a highway and found the mean speed to be 72 miles per hour with a standard deviation of 2.2 miles per hour. Another sample of 18 cars driven by women on the same highway gave a mean speed of 68 miles per hour with a standard deviation of 2.5 miles per hour. Assume that the speeds at which all men and all women drive cars on this highway are both normally distributed with the same population standard deviation.

- Construct a 98% confidence interval for the difference between the mean speeds of cars driven by all men and all women on this highway.
- Test at a 1% significance level whether the mean speed of cars driven by all men drivers on this highway is greater than that of cars driven by all women drivers.

- 10.28** A high school counselor wanted to know if tenth-graders at her high school tend to have more free time than the twelfth-graders. She took random samples of 25 tenth-graders and 23 twelfth-graders. Each student was asked to record the amount of free time he or she had in a typical week. The mean for the tenth-graders was found to be 29 hours of free time per week with a standard deviation of 7.0 hours. For the twelfth-graders, the mean was 22 hours of free time per week with a standard deviation of 6.2 hours. Assume that the two populations are normally distributed with equal but unknown population standard deviations.

- Make a 90% confidence interval for the difference between the corresponding population means.
- Test at a 5% significance level whether the two population means are different.

**10.29** A company claims that its medicine, Brand A, provides faster relief from pain than another company's medicine, Brand B. A researcher tested both brands of medicine on two groups of randomly selected patients. The results of the test are given in the following table. The mean and standard deviation of relief times are in minutes.

Brand	Sample Size	Mean of Relief Times	Standard Deviation of Relief Times
A	25	44	11
B	22	49	9

Assume that the two populations are normally distributed with unknown but equal standard deviations.

- a. Construct a 99% confidence interval for the difference between the mean relief times for the two brands of medicine.
- b. Test at a 1% significance level whether the mean relief time for Brand A is less than that for Brand B.

**10.30** A consumer organization tested two paper shredders, the Piranha and the Crocodile, designed for home use. Each of 10 randomly selected volunteers shredded 100 sheets of paper with the Piranha, and then another sample of 10 randomly selected volunteers each shredded 100 sheets with the Crocodile. The Piranha took an average of 203 seconds to shred 100 sheets with a standard deviation of 6 seconds. The Crocodile took an average of 187 seconds to shred 100 sheets with a standard deviation of 5 seconds. Assume that the shredding times for both machines are normally distributed with equal but unknown standard deviations.

- a. Construct a 99% confidence interval for the difference between the two population means.
- b. Using a 1% significance level, can you conclude that the mean time taken by the Piranha to shred 100 sheets is higher than that for the Crocodile?
- c. What would your decision be in part b if the probability of making a Type I error were zero? Explain.

**10.31** Quadro Corporation has two supermarket stores in a city. The company's quality control department wanted to check if the customers are equally satisfied with the service provided at these two stores. A sample of 380 customers selected from Supermarket I produced a mean satisfaction index of 7.6 (on a scale of 1 to 10, 1 being the lowest and 10 being the highest) with a standard deviation of .75. Another sample of 370 customers selected from Supermarket II produced a mean satisfaction index of 8.1 with a standard deviation of .59. Assume that the customer satisfaction index for each supermarket has unknown but same population standard deviation.

- a. Construct a 98% confidence interval for the difference between the mean satisfaction indexes for all customers for the two supermarkets.
- b. Test at a 1% significance level whether the mean satisfaction indexes for all customers for the two supermarkets are different.

**10.32** According to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011 ([money.cnn.com/2011/08/19/pf/credit\\_card\\_issuance/index.htm](http://money.cnn.com/2011/08/19/pf/credit_card_issuance/index.htm)). Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average credit limits of \$2635 and \$2887, respectively. Suppose that the sample standard deviations for these two samples were \$365 and \$412, respectively, and the assumption that the population standard deviations are equal for the two populations is reasonable.

- a. Let  $\mu_1$  and  $\mu_2$  be the average credit limits on all credit cards issued in January 2011 and in May 2011, respectively. What is the point estimate of  $\mu_1 - \mu_2$ ?
- b. Construct a 98% confidence interval for  $\mu_1 - \mu_2$ .
- c. Using a 1% significance level, can you conclude that the average credit limit for all new credit cards issued in January 2011 was lower than the corresponding average for all credit cards issued in May 2011? Use both the  $p$ -value and the critical-value approaches to make this test.

## 10.3

### Inferences About the Difference Between Two Population Means for Independent Samples: $\sigma_1$ and $\sigma_2$ Unknown and Unequal

Section 10.2 explained how to make inferences about the difference between two population means using the  $t$  distribution when the standard deviations of the two populations are unknown but equal and certain other assumptions hold true. Now, what if all other assumptions

of Section 10.2 hold true, but the population standard deviations are not only unknown but also unequal? In this case, the procedures used to make confidence intervals and to test hypotheses about  $\mu_1 - \mu_2$  remain similar to the ones we learned in Sections 10.2.1 and 10.2.2, except for two differences. When the population standard deviations are unknown and not equal, the degrees of freedom are no longer given by  $n_1 + n_2 - 2$ , and the standard deviation of  $\bar{x}_1 - \bar{x}_2$  is not calculated using the pooled standard deviation  $s_p$ .

### Degrees of Freedom

If

1. The two samples are independent
2. The standard deviations  $\sigma_1$  and  $\sigma_2$  of the two populations are unknown and unequal, that is,  $\sigma_1 \neq \sigma_2$
3. At least one of the following two conditions is fulfilled:
  - i. Both samples are large (i.e.,  $n_1 \geq 30$  and  $n_2 \geq 30$ )
  - ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed

then the  $t$  distribution is used to make inferences about  $\mu_1 - \mu_2$ , and the *degrees of freedom* for the  $t$  distribution are given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

The number given by this formula is always rounded down for  $df$ .

Because the standard deviations of the two populations are not known, we use  $s_{\bar{x}_1 - \bar{x}_2}$  as a point estimator of  $\sigma_{\bar{x}_1 - \bar{x}_2}$ . The following formula is used to calculate the standard deviation  $s_{\bar{x}_1 - \bar{x}_2}$  of  $\bar{x}_1 - \bar{x}_2$ .

**Estimate of the Standard Deviation of  $\bar{x}_1 - \bar{x}_2$**  The value of  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated as

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

### 10.3.1 Interval Estimation of $\mu_1 - \mu_2$

Again, the difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ , is the point estimator of the difference between the two population means,  $\mu_1 - \mu_2$ . The following formula gives the confidence interval for  $\mu_1 - \mu_2$  when the  $t$  distribution is used and the conditions mentioned earlier in this section are satisfied.

**Confidence Interval for  $\mu_1 - \mu_2$**  The  $(1 - \alpha)100\%$  confidence interval for  $\mu_1 - \mu_2$  is

$$(\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2}$$

where the value of  $t$  is obtained from the  $t$  distribution table for a given confidence level and the degrees of freedom are given by the formula mentioned earlier, and  $s_{\bar{x}_1 - \bar{x}_2}$  is also calculated as explained earlier.

Example 10–8 describes how to construct a confidence interval for  $\mu_1 - \mu_2$  when the standard deviations of the two populations are unknown and unequal.

*Constructing a confidence interval for  $\mu_1 - \mu_2$ : two independent samples,  $\sigma_1$  and  $\sigma_2$  unknown and unequal.*

## ■ EXAMPLE 10-8

According to Example 10-5 of Section 10.2.1, a sample of 15 one-pound jars of coffee of Brand I showed that the mean amount of caffeine in these jars is 80 milligrams per jar with a standard deviation of 5 milligrams. Another sample of 12 one-pound coffee jars of Brand II gave a mean amount of caffeine equal to 77 milligrams per jar with a standard deviation of 6 milligrams. Construct a 95% confidence interval for the difference between the mean amounts of caffeine in one-pound coffee jars of these two brands. Assume that the two populations are normally distributed and that the standard deviations of the two populations are not equal.

**Solution** Let  $\mu_1$  and  $\mu_2$  be the mean amounts of caffeine per jar in all 1-pound jars of Brands I and II, respectively, and let  $\bar{x}_1$  and  $\bar{x}_2$  be the means of the two respective samples.

From the given information,

$$\begin{array}{lll} \text{Brand I coffee: } & n_1 = 15 & \bar{x}_1 = 80 \text{ milligrams} \\ & & s_1 = 5 \text{ milligrams} \\ \text{Brand II coffee: } & n_2 = 12 & \bar{x}_2 = 77 \text{ milligrams} \\ & & s_2 = 6 \text{ milligrams} \end{array}$$

The confidence level is  $1 - \alpha = .95$ .

First, we calculate the standard deviation of  $\bar{x}_1 - \bar{x}_2$  as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(5)^2}{15} + \frac{(6)^2}{12}} = 2.16024690$$

Next, to find the  $t$  value from the  $t$  distribution table, we need to know the area in each tail of the  $t$  distribution curve and the degrees of freedom.

$$\text{Area in each tail} = \alpha/2 = (1 - .95)/2 = .025$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{(5)^2}{15} + \frac{(6)^2}{12}\right)^2}{\frac{\left(\frac{(5)^2}{15}\right)^2}{15 - 1} + \frac{\left(\frac{(6)^2}{12}\right)^2}{12 - 1}} = 21.42 \approx 21$$

Note that the degrees of freedom are always rounded down as in this calculation. From the  $t$  distribution table, the  $t$  value for  $df = 21$  and .025 area in the right tail of the  $t$  distribution curve is 2.080. The 95% confidence interval for  $\mu_1 - \mu_2$  is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2} &= (80 - 77) \pm 2.080(2.16024690) \\ &= 3 \pm 4.49 = \mathbf{-1.49 \text{ to } 7.49} \end{aligned}$$

Thus, with 95% confidence we can state that based on these two sample results, the difference in the mean amounts of caffeine in 1-pound jars of these two brands of coffee is between  $-1.49$  and  $7.49$  milligrams. ■

Comparing this confidence interval with the one obtained in Example 10-5, we observe that the two confidence intervals are very close. From this we can conclude that even if the standard deviations of the two populations are not equal and we use the procedure of Section 10.2.1 to make a confidence interval for  $\mu_1 - \mu_2$ , the margin of error will be small as long as the difference between the two population standard deviations is not too large.

### 10.3.2 Hypothesis Testing About $\mu_1 - \mu_2$

When the standard deviations of the two populations are unknown and unequal along with the other conditions of Section 10.2 holding true, we use the  $t$  distribution to make a test of hypothesis about  $\mu_1 - \mu_2$ . This procedure differs from the one in Section 10.2.2 only in the calculation of degrees of freedom for the  $t$  distribution and the standard deviation of  $\bar{x}_1 - \bar{x}_2$ . The  $df$  and the standard deviation of  $\bar{x}_1 - \bar{x}_2$  in this case are given by the formulas used in Section 10.3.1.

**Test Statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$**  The value of the *test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$*  is computed as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

The value of  $\mu_1 - \mu_2$  in this formula is substituted from the null hypothesis, and  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated as explained earlier.

Example 10–9 illustrates the procedure used to conduct a test of hypothesis about  $\mu_1 - \mu_2$  when the standard deviations of the two populations are unknown and unequal.

## ■ EXAMPLE 10–9

According to Example 10–6 of Section 10.2.2, a sample of 14 cans of Brand I diet soda gave the mean number of calories per can of 23 with a standard deviation of 3 calories. Another sample of 16 cans of Brand II diet soda gave the mean number of calories of 25 per can with a standard deviation of 4 calories. Test at a 1% significance level whether the mean numbers of calories per can of diet soda are different for these two brands. Assume that the calories per can of diet soda are normally distributed for each of these two brands and that the standard deviations for the two populations are not equal.

Making a two-tailed test of hypothesis about  $\mu_1 - \mu_2$ : two independent samples, and unknown and unequal  $\sigma_1$  and  $\sigma_2$

**Solution** Let  $\mu_1$  and  $\mu_2$  be the mean numbers of calories for all cans of diet soda of Brand I and Brand II, respectively, and let  $\bar{x}_1$  and  $\bar{x}_2$  be the means of the respective samples. From the given information,

$$\begin{array}{lll} \text{Brand I diet soda:} & n_1 = 14 & \bar{x}_1 = 23 & s_1 = 3 \\ \text{Brand II diet soda:} & n_2 = 16 & \bar{x}_2 = 25 & s_2 = 4 \end{array}$$

The significance level is  $\alpha = .01$ .

### Step 1. State the null and alternative hypotheses.

We are to test for the difference in the mean numbers of calories per can for the two brands. The null and alternative hypotheses are, respectively,

$$\begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & \text{(The mean numbers of calories are not different.)} \\ H_1: \mu_1 - \mu_2 \neq 0 & \text{(The mean numbers of calories are different.)} \end{array}$$

### Step 2. Select the distribution to use.

Here, the two samples are independent,  $\sigma_1$  and  $\sigma_2$  are unknown and unequal, the sample sizes are small, but both populations are normally distributed. Hence, all conditions mentioned in the beginning of Section 10.3 are fulfilled. Consequently, we use the  $t$  distribution to make the test.

### Step 3. Determine the rejection and nonrejection regions.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The significance level is .01. Hence,

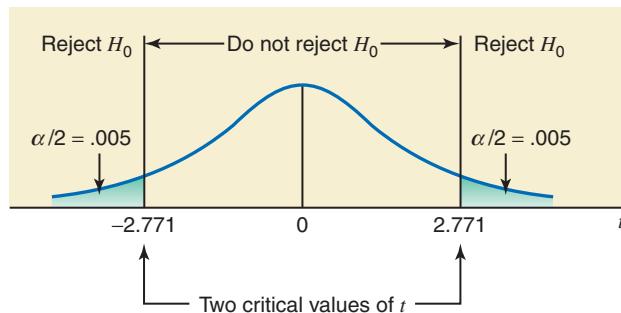
$$\text{Area in each tail} = \alpha/2 = .01/2 = .005$$

The degrees of freedom are calculated as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{(3)^2}{14} + \frac{(4)^2}{16}\right)^2}{\frac{\left(\frac{(3)^2}{14}\right)^2}{14 - 1} + \frac{\left(\frac{(4)^2}{16}\right)^2}{16 - 1}} = 27.41 \approx 27$$

From the  $t$  distribution table, the critical values of  $t$  for  $df = 27$  and  $.005$  area in each tail of the  $t$  distribution curve are  $-2.771$  and  $2.771$ . These values are shown in Figure 10.5.

**Figure 10.5** Rejection and nonrejection regions.



**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $\bar{x}_1 - \bar{x}_2$  is computed as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(3)^2}{14} + \frac{(4)^2}{16}} = 1.28173989$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(23 - 25) - 0}{1.28173989} = -1.560$$

From  $H_0$

**Step 5.** Make a decision.

Because the value of the test statistic  $t = -1.560$  for  $\bar{x}_1 - \bar{x}_2$  falls in the nonrejection region, we fail to reject the null hypothesis. Hence, there is no difference in the mean numbers of calories per can for the two brands of diet soda. The difference in  $\bar{x}_1$  and  $\bar{x}_2$  observed for the two samples may have occurred due to sampling error only.

### Using the $p$ -Value to Make a Decision

We can use the  $p$ -value approach to make the above decision. To do so, we keep Steps 1 and 2 of this example. Then in Step 3 we calculate the value of the test statistic  $t$  (as done in Step 4 above) and then find the  $p$ -value for this  $t$  from the  $t$  distribution table (Table V of Appendix C) or by using technology. In Step 4 above, the  $t$ -value for  $\bar{x}_1 - \bar{x}_2$  was calculated to be  $-1.560$ . In this example, the test is two-tailed. Therefore, the  $p$ -value is equal to twice the area under the  $t$  distribution curve to the left of  $t = -1.560$ . If we have access to technology, we can use it to find the exact  $p$ -value, which will be  $.130$ . If we use the  $t$  distribution table, we can only find the range for the  $p$ -value. From Table V of Appendix C, for  $df = 27$ , the two values that include  $1.560$  are  $1.314$  and  $1.703$ . (Note that we use the positive value of  $t$ , although our  $t$  is negative.) Thus, the test statistic  $t = -1.560$  falls between  $-1.314$  and  $-1.703$ . The areas in the  $t$  distribution table that correspond to  $1.314$  and  $1.703$  are  $.10$  and  $.05$ , respectively. Because it is a two-tailed test, the  $p$ -value for  $t = -1.560$  is between  $2(.10) = .20$  and  $2(.05) = .10$ , which can be written as

$$.10 < p\text{-value} < .20$$

Since we will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than the  $p$ -value, we will reject the null hypothesis in this example for any  $\alpha \geq .20$  using the above range and not reject for  $\alpha \leq .10$ . If we use technology, we will reject the null hypothesis for  $\alpha \geq .130$ . Since  $\alpha = .01$  in this example, which is smaller than both  $.10$  and  $.130$ , we fail to reject the null hypothesis. ■

**Remember ▶** The degrees of freedom for the procedures to make a confidence interval and to test a hypothesis about  $\mu_1 - \mu_2$  learned in Sections 10.3.1 and 10.3.2 are always rounded down.

## EXERCISES

### CONCEPTS AND PROCEDURES

**10.33** Assuming that the two populations are normally distributed with unequal and unknown population standard deviations, construct a 95% confidence interval for  $\mu_1 - \mu_2$  for the following.

$$n_1 = 14 \quad \bar{x}_1 = 109.43 \quad s_1 = 2.26$$

$$n_2 = 15 \quad \bar{x}_2 = 113.88 \quad s_2 = 5.84$$

**10.34** Assuming that the two populations have unequal and unknown population standard deviations, construct a 99% confidence interval for  $\mu_1 - \mu_2$  for the following.

$$n_1 = 48 \quad \bar{x}_1 = .863 \quad s_1 = .176$$

$$n_2 = 46 \quad \bar{x}_2 = .796 \quad s_2 = .068$$

**10.35** Refer to Exercise 10.33. Test at a 5% significance level if the two population means are different.

**10.36** Refer to Exercise 10.34. Test at a 1% significance level if the two population means are different.

**10.37** Refer to Exercise 10.33. Test at a 1% significance level if  $\mu_1$  is less than  $\mu_2$ .

**10.38** Refer to Exercise 10.34. Test at a 2.5% significance level if  $\mu_1$  is greater than  $\mu_2$ .

### APPLICATIONS

**10.39** According to the information given in Exercise 10.25, a sample of 45 customers who drive luxury cars showed that their average distance driven between oil changes was 3187 miles with a sample standard deviation of 42.40 miles. Another sample of 40 customers who drive compact lower-price cars resulted in an average distance of 3214 miles with a standard deviation of 50.70 miles. Suppose that the standard deviations for the two populations are not equal.

- Construct a 95% confidence interval for the difference in the mean distance between oil changes for all luxury cars and all compact lower-price cars.
- Using a 1% significance level, can you conclude that the mean distance between oil changes is lower for all luxury cars than for all compact lower-price cars?
- Suppose that the sample standard deviations were 28.9 and 61.4 miles, respectively. Redo parts a and b. Discuss any changes in the results.

**10.40** As mentioned in Exercise 10.26, a town that recently started a single-stream recycling program provided 60-gallon recycling bins to 25 randomly selected households and 75-gallon recycling bins to 22 randomly selected households. The average total volumes of recycling over a 10-week period were 382 and 415 gallons for the two groups, respectively, with standard deviations of 52.5 and 43.8 gallons, respectively. Suppose that the standard deviations for the two populations are not equal.

- Construct a 98% confidence interval for the difference in the mean volumes of 10-week recycling for the households with the 60- and 75-gallon bins.
- Using a 2% significance level, can you conclude that the average 10-week recycling volume of all households having 60-gallon containers is different from the average 10-week recycling volume of all households that have 75-gallon containers?
- Suppose that the sample standard deviations were 59.3 and 33.8 gallons, respectively. Redo parts a and b. Discuss any changes in the results.

**10.41** According to Exercise 10.27, an insurance company wants to know if the average speed at which men drive cars is higher than that of women drivers. The company took a random sample of 27 cars driven by men on a highway and found the mean speed to be 72 miles per hour with a standard deviation of 2.2 miles per hour. Another sample of 18 cars driven by women on the same highway gave a mean speed of 68 miles per hour with a standard deviation of 2.5 miles per hour. Assume that the speeds at which all men and all women drive cars on this highway are both normally distributed with unequal population standard deviations.

- Construct a 98% confidence interval for the difference between the mean speeds of cars driven by all men and all women on this highway.
- Test at a 1% significance level whether the mean speed of cars driven by all men drivers on this highway is higher than that of cars driven by all women drivers.
- Suppose that the sample standard deviations were 1.9 and 3.4 miles per hour, respectively. Redo parts a and b. Discuss any changes in the results.

**10.42** Refer to Exercise 10.28. Now assume that the two populations are normally distributed with unequal and unknown population standard deviations.

- Make a 90% confidence interval for the difference between the corresponding population means.
- Test at a 5% significance level whether the two population means are different.
- Suppose that the sample standard deviations were 9.5 and 5.1 hours, respectively. Redo parts a and b. Discuss any changes in the results.

**10.43** As mentioned in Exercise 10.29, a company claims that its medicine, Brand A, provides faster relief from pain than another company's medicine, Brand B. A researcher tested both brands of medicine on two groups of randomly selected patients. The results of the test are given in the following table. The mean and standard deviation of relief times are in minutes.

Brand	Sample Size	Mean of Relief Times	Standard Deviation of Relief Times
A	25	44	11
B	22	49	9

Assume that the two populations are normally distributed with unknown and unequal standard deviations.

- Construct a 99% confidence interval for the difference between the mean relief times for the two brands of medicine.
- Test at a 1% significance level whether the mean relief time for Brand A is less than that for Brand B.
- Suppose that the sample standard deviations were 13.3 and 7.2 minutes, respectively. Redo parts a and b. Discuss any changes in the results.

**10.44** Refer to Exercise 10.30. Now assume that the shredding times for both paper shredders are normally distributed with unequal and unknown standard deviations.

- Construct a 99% confidence interval for the difference between the two population means.
- Using a 1% significance level, can you conclude that the mean time taken by the Piranha to shred 100 sheets is higher than that for the Crocodile?
- Suppose that the sample standard deviations were 7.40 and 4.60 seconds, respectively. Redo parts a and b. Discuss any changes in the results.
- What would your decision be in part b if the probability of making a Type I error were zero? Explain.

**10.45** As mentioned in Exercise 10.31, Quadro Corporation has two supermarkets in a city. The company's quality control department wanted to check if the customers are equally satisfied with the service provided at these two stores. A sample of 380 customers selected from Supermarket I produced a mean satisfaction index of 7.6 (on a scale of 1 to 10, 1 being the lowest and 10 being the highest) with a standard deviation of .75. Another sample of 370 customers selected from Supermarket II produced a mean satisfaction index of 8.1 with a standard deviation of .59. Assume that the customer satisfaction index for each supermarket has an unknown and different population standard deviation.

- Construct a 98% confidence interval for the difference between the mean satisfaction indexes for all customers for the two supermarkets.
- Test at a 1% significance level whether the mean satisfaction indexes for all customers for the two supermarkets are different.
- Suppose that the sample standard deviations were .88 and .39, respectively. Redo parts a and b. Discuss any changes in the results.

**10.46** Refer to Exercise 10.32. As mentioned in that exercise, according to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 new credit cards issued in January 2011 and 500 new credit cards issued in May 2011 had average credit limits of \$2635 and \$2887, respectively. Suppose that the sample standard deviations for these two samples were \$365 and \$412, respectively. Now assume that the population standard deviations for the two populations are unknown and not equal.

- Let  $\mu_1$  and  $\mu_2$  be the average credit limits on all credit cards issued in January 2011 and in May 2011, respectively. What is the point estimate of  $\mu_1 - \mu_2$ ?
- Construct a 98% confidence interval for  $\mu_1 - \mu_2$ .
- Using a 1% significance level, can you conclude that the average credit limit for all new credit cards issued in January 2011 was lower than the corresponding average for all credit cards issued in May 2011? Use both the  $p$ -value and the critical-value approaches to make this test.

## 10.4 Inferences About the Difference Between Two Population Means for Paired Samples

Sections 10.1, 10.2, and 10.3 were concerned with estimation and hypothesis testing about the difference between two population means when the two samples were drawn independently from two different populations. This section describes estimation and hypothesis-testing procedures for the difference between two population means when the samples are dependent.

In a case of two dependent samples, two data values—one for each sample—are collected from the same source (or element) and, hence, these are also called **paired or matched samples**. For example, we may want to make inferences about the mean weight loss for members of a health club after they have gone through an exercise program for a certain period of time. To do so, suppose we select a sample of 15 members of this health club and record their weights before and after the program. In this example, both sets of data are collected from the same 15 persons, once before and once after the program. Thus, although there are two samples, they contain the same 15 persons. This is an example of paired (or dependent or matched) samples. The procedures to make confidence intervals and test hypotheses in the case of paired samples are different from the ones for independent samples discussed in earlier sections of this chapter.

### Definition

**Paired or Matched Samples** Two samples are said to be *paired or matched samples* when for each data value collected from one sample there is a corresponding data value collected from the second sample, and both these data values are collected from the same source.

As another example of paired samples, suppose an agronomist wants to measure the effect of a new brand of fertilizer on the yield of potatoes. To do so, he selects 10 pieces of land and divides each piece into two portions. Then he randomly assigns one of the two portions from each piece of land to grow potatoes without using fertilizer (or using some other brand of fertilizer). The second portion from each piece of land is used to grow potatoes with the new brand of fertilizer. Thus, he will have 10 pairs of data values. Then, using the procedure to be discussed in this section, he will make inferences about the difference in the mean yields of potatoes with and without the new fertilizer.

The question arises, why does the agronomist not choose 10 pieces of land on which to grow potatoes without using the new brand of fertilizer and another 10 pieces of land to grow potatoes by using the new brand of fertilizer? If he does so, the effect of the fertilizer might be confused with the effects due to soil differences at different locations. Thus, he will not be able to isolate the effect of the new brand of fertilizer on the yield of potatoes. Consequently, the results will not be reliable. By choosing 10 pieces of land and then dividing each of them into two portions, the researcher decreases the possibility that the difference in the productivities of different pieces of land affects the results.

In paired samples, the difference between the two data values for each element of the two samples is denoted by  $d$ . This value of  $d$  is called the **paired difference**. We then treat all the values of  $d$  as one sample and make inferences applying procedures similar to the ones used for one-sample cases in Chapters 8 and 9. Note that because each source (or element) gives a pair of values (one for each of the two data sets), each sample contains the same number of values. That is, both samples are of the same size. Therefore, we denote the (common) **sample size** by  $n$ , which gives the number of paired difference values denoted by  $d$ . The **degrees of freedom** for the paired samples are  $n - 1$ . Let

$\mu_d$  = the mean of the paired differences for the population

$\sigma_d$  = the standard deviation of the paired differences for the population, which is usually not known

$\bar{d}$  = the mean of the paired differences for the sample

$s_d$  = the standard deviation of the paired differences for the sample

$n$  = the number of paired difference values

**Mean and Standard Deviation of the Paired Differences for Two Samples** The values of the mean and standard deviation,  $\bar{d}$  and  $s_d$ , respectively, of paired differences for two samples are calculated as<sup>2</sup>

$$\bar{d} = \frac{\sum d}{n}$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n - 1}}$$

In paired samples, instead of using  $\bar{x}_1 - \bar{x}_2$  as the sample statistic to make inferences about  $\mu_1 - \mu_2$ , we use the sample statistic  $\bar{d}$  to make inferences about  $\mu_d$ . Actually the value of  $\bar{d}$  is always equal to  $\bar{x}_1 - \bar{x}_2$ , and the value of  $\mu_d$  is always equal to  $\mu_1 - \mu_2$ .

**Sampling Distribution, Mean, and Standard Deviation of  $\bar{d}$**  If  $\sigma_d$  is known and either the sample size is large ( $n \geq 30$ ) or the population is normally distributed, then the *sampling distribution* of  $\bar{d}$  is approximately normal with its *mean* and *standard deviation* given as, respectively,

$$\mu_{\bar{d}} = \mu_d \quad \text{and} \quad \sigma_{\bar{d}} = \frac{\sigma_d}{\sqrt{n}}$$

Thus, if the standard deviation  $\sigma_d$  of the population paired differences is known and either the sample size is large (i.e.,  $n \geq 30$ ) or the population of paired differences is normally distributed (with  $n < 30$ ), then the normal distribution can be used to make a confidence interval and to test a hypothesis about  $\mu_d$ . However, usually  $\sigma_d$  is not known. Then, if the standard deviation  $\sigma_d$  of the population paired differences is unknown and either the sample size is large (i.e.,  $n \geq 30$ ) or the population of paired differences is normally distributed (with  $n < 30$ ), then the *t* distribution is used to make a confidence interval and to test a hypothesis about  $\mu_d$ .

#### Making Inferences About $\mu_d$ If

1. The standard deviation  $\sigma_d$  of the population of paired differences is unknown
  2. At least one of the following two conditions is fulfilled:
    - i. The sample size is large (i.e.,  $n \geq 30$ )
    - ii. If the sample size is small, then the population of paired differences is normally distributed
- then the *t* distribution is used to make inferences about  $\mu_d$ . The standard deviation  $\sigma_{\bar{d}}$  of  $\bar{d}$  is estimated by  $s_{\bar{d}}$ , which is calculated as

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

Sections 10.4.1 and 10.4.2 describe the procedures that are used to make a confidence interval and to test a hypothesis about  $\mu_d$  under the above conditions. The inferences are made using the *t* distribution.

#### 10.4.1 Interval Estimation of $\mu_d$

The mean  $\bar{d}$  of paired differences for paired samples is the point estimator of  $\mu_d$ . The following formula is used to construct a confidence interval for  $\mu_d$  when the *t* distribution is used.

<sup>2</sup>The basic formula used to calculate  $s_d$  is

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}}$$

However, we will not use this formula to make calculations in this chapter.

**Confidence Interval for  $\mu_d$**  The  $(1 - \alpha)100\%$  confidence interval for  $\mu_d$  is

$$\bar{d} \pm ts_{\bar{d}}$$

where the value of  $t$  is obtained from the  $t$  distribution table for the given confidence level and  $n - 1$  degrees of freedom, and  $s_{\bar{d}}$  is calculated as explained earlier.

Example 10–10 illustrates the procedure to construct a confidence interval for  $\mu_d$ .

## ■ EXAMPLE 10–10

A researcher wanted to find the effect of a special diet on systolic blood pressure. She selected a sample of seven adults and put them on this dietary plan for 3 months. The following table gives the systolic blood pressures (in mm Hg) of these seven adults before and after the completion of this plan.

Before	210	180	195	220	231	199	224
After	193	186	186	223	220	183	233

Let  $\mu_d$  be the mean reduction in the systolic blood pressures due to this special dietary plan for the population of all adults. Construct a 95% confidence interval for  $\mu_d$ . Assume that the population of paired differences is (approximately) normally distributed.

Constructing a confidence interval for  $\mu_d$ : paired samples,  $\sigma_d$  unknown,  $n < 30$ , and population normal.

**Solution** Because the information obtained is from paired samples, we will make the confidence interval for the paired difference mean  $\mu_d$  of the population using the paired difference mean  $\bar{d}$  of the sample. Let  $d$  be the difference in the systolic blood pressure of an adult before and after this special dietary plan. Then,  $d$  is obtained by subtracting the systolic blood pressure after the plan from the systolic blood pressure before the plan. The third column of Table 10.1 lists the values of  $d$  for the seven adults. The fourth column of the table records the values of  $d^2$ , which are obtained by squaring each of the  $d$  values.

**Table 10.1**

Difference			
Before	After	$d$	$d^2$
210	193	17	289
180	186	-6	36
195	186	9	81
220	223	-3	9
231	220	11	121
199	183	16	256
224	233	-9	81
		$\Sigma d = 35$	$\Sigma d^2 = 873$

The values of  $\bar{d}$  and  $s_d$  are calculated as follows:

$$\bar{d} = \frac{\Sigma d}{n} = \frac{35}{7} = 5.00$$

$$s_d = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2/n}{n-1}} = \sqrt{\frac{873 - (35)^2/7}{7-1}} = 10.78579312$$

Hence, the standard deviation of  $\bar{d}$  is

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{10.78579312}{\sqrt{7}} = 4.07664661$$

Here,  $\sigma_d$  is not known, the sample size is small, but the population is normally distributed. Hence, we will use the  $t$  distribution to make the confidence interval. For the 95% confidence interval, the area in each tail of the  $t$  distribution curve is

$$\text{Area in each tail} = \alpha/2 = (1 - .95)/2 = .025$$

The degrees of freedom are

$$df = n - 1 = 7 - 1 = 6$$

From the  $t$  distribution table, the  $t$  value for  $df = 6$  and .025 area in the right tail of the  $t$  distribution curve is 2.447. Therefore, the 95% confidence interval for  $\mu_d$  is

$$\bar{d} \pm ts_{\bar{d}} = 5.00 \pm 2.447(4.07664661) = 5.00 \pm 9.98 = \mathbf{-4.98 \text{ to } 14.98}$$

Thus, we can state with 95% confidence that the mean difference between systolic blood pressures before and after the given dietary plan for all adult participants is between  $-4.98$  and  $14.98$  mm Hg. ■

### 10.4.2 Hypothesis Testing About $\mu_d$

A hypothesis about  $\mu_d$  is tested by using the sample statistic  $\bar{d}$ . This section illustrates the case of the  $t$  distribution only. Earlier in this section we learned what conditions should hold true to use the  $t$  distribution to test a hypothesis about  $\mu_d$ . The following formula is used to calculate the value of the test statistic  $t$  when testing a hypothesis about  $\mu_d$ .

**Test Statistic  $t$  for  $\bar{d}$**  The value of the *test statistic  $t$  for  $\bar{d}$*  is computed as follows:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

The critical value of  $t$  is found from the  $t$  distribution table for the given significance level and  $n - 1$  degrees of freedom.

Examples 10–11 and 10–12 illustrate the hypothesis-testing procedure for  $\mu_d$ .

#### ■ EXAMPLE 10–11

Conducting a left-tailed test of hypothesis about  $\mu_d$  for paired samples:  $\sigma_d$  not known, small sample but normally distributed population.

A company wanted to know if attending a course on “how to be a successful salesperson” can increase the average sales of its employees. The company sent six of its salespersons to attend this course. The following table gives the 1-week sales of these salespersons before and after they attended this course.

Before	12	18	25	9	14	16
After	18	24	24	14	19	20

Using a 1% significance level, can you conclude that the mean weekly sales for all salespersons increase as a result of attending this course? Assume that the population of paired differences has a normal distribution.

**Solution** Because the data are for paired samples, we test a hypothesis about the paired differences mean  $\mu_d$  of the population using the paired differences mean  $\bar{d}$  of the sample.

Let

$$d = (\text{Weekly sales before the course}) - (\text{Weekly sales after the course})$$

In Table 10.2, we calculate  $d$  for each of the six salespersons by subtracting the sales after the course from the sales before the course. The fourth column of the table lists the values of  $d^2$ .

**Table 10.2**

Before	After	Difference	
		$d$	$d^2$
12	18	-6	36
18	24	-6	36
25	24	1	1
9	14	-5	25
14	19	-5	25
16	20	-4	16
		$\Sigma d = -25$	$\Sigma d^2 = 139$

The values of  $\bar{d}$  and  $s_d$  are calculated as follows:

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-25}{6} = -4.17$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{139 - \frac{(-25)^2}{6}}{6-1}} = 2.63944439$$

The standard deviation of  $\bar{d}$  is

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{2.63944439}{\sqrt{6}} = 1.07754866$$

**Step 1.** State the null and alternative hypotheses.

We are to test if the mean weekly sales for all salespersons increase as a result of taking the course. Let  $\mu_1$  be the mean weekly sales for all salespersons before the course and  $\mu_2$  the mean weekly sales for all salespersons after the course. Then  $\mu_d = \mu_1 - \mu_2$ . The mean weekly sales for all salespersons will increase due to attending the course if  $\mu_1$  is less than  $\mu_2$ , which can be written as  $\mu_1 - \mu_2 < 0$  or  $\mu_d < 0$ . Consequently, the null and alternative hypotheses are, respectively,

$$H_0: \mu_d = 0 \quad (\mu_1 - \mu_2 = 0 \text{ or the mean weekly sales do not increase})$$

$$H_1: \mu_d < 0 \quad (\mu_1 - \mu_2 < 0 \text{ or the mean weekly sales do increase})$$

Note that we can also write the null hypothesis as  $\mu_d \geq 0$ .

**Step 2.** Select the distribution to use.

Here  $\sigma_d$  is unknown, the sample size is small ( $n < 30$ ), but the population of paired differences is normally distributed. Therefore, we use the  $t$  distribution to conduct the test.

**Step 3.** Determine the rejection and nonrejection regions.

The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed. The significance level is .01. Hence,

$$\text{Area in left tail} = \alpha = .01$$

$$\text{Degrees of freedom} = n - 1 = 6 - 1 = 5$$

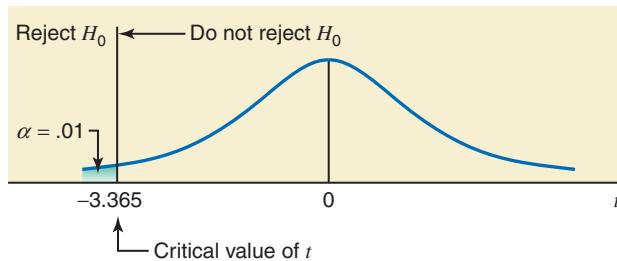
The critical value of  $t$  for  $df = 5$  and .01 area in the left tail of the  $t$  distribution curve is -3.365. This value is shown in Figure 10.6.

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $\bar{d}$  is computed as follows:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{-4.17 - 0}{1.07754866} = -3.870$$

From  $H_0$

**Figure 10.6** Rejection and nonrejection regions.**Step 5. Make a decision.**

Because the value of the test statistic  $t = -3.870$  for  $\bar{d}$  falls in the rejection region, we reject the null hypothesis. Consequently, we conclude that the mean weekly sales for all salespersons increase as a result of this course.

**Using the *p*-Value to Make a Decision**

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2 of this example. Then in Step 3, we calculate the value of the test statistic  $t$  for  $\bar{d}$  (as done in Step 4 above) and then find the *p*-value for this  $t$  from the *t* distribution table (Table V of Appendix C) or by using technology. If we have access to technology, we can use it to find the exact *p*-value, which will be .006. By using Table V, we can find the range of the *p*-value. From Table V, for  $df = 5$ , the test statistic  $t = -3.870$  falls between  $-3.365$  and  $-4.032$ . The areas in the *t* distribution table that correspond to  $-3.365$  and  $-4.032$  are .01 and .005, respectively. Because it is a left-tailed test, the *p*-value is between .01 and .005, which can be written as

$$.005 < p\text{-value} < .01$$

Since we will reject the null hypothesis for any  $\alpha$  (significance level) that is greater than or equal to the *p*-value, we will reject the null hypothesis in this example for any  $\alpha \geq .006$  using the technology and  $\alpha \geq .01$  using the above range. Since  $\alpha = .01$  in this example, which is larger than .006 obtained from technology, we reject the null hypothesis. Also, because  $\alpha$  is equal to .01, using the *p*-value range we reject the null hypothesis. ■

**EXAMPLE 10–12**

*Making a two-tailed test of hypothesis about  $\mu_d$  for paired samples:  $\sigma_d$  not known, small sample but normally distributed population.*

Refer to Example 10–10. The table that gives the blood pressures (in mm Hg) of seven adults before and after the completion of a special dietary plan is reproduced here.

Before	210	180	195	220	231	199	224
After	193	186	186	223	220	183	233

Let  $\mu_d$  be the mean of the differences between the systolic blood pressures before and after completing this special dietary plan for the population of all adults. Using a 5% significance level, can you conclude that the mean of the paired differences  $\mu_d$  is different from zero? Assume that the population of paired differences is (approximately) normally distributed.

**Solution** Table 10.3 gives  $d$  and  $d^2$  for each of the seven adults (blood pressure values in mm Hg).

The values of  $\bar{d}$  and  $s_d$  are calculated as follows:

$$\bar{d} = \frac{\sum d}{n} = \frac{35}{7} = 5.00$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{873 - \frac{(35)^2}{7}}{7-1}} = 10.78579312$$

**Table 10.3**

Before	After	Difference	
		$d$	$d^2$
210	193	17	289
180	186	-6	36
195	186	9	81
220	223	-3	9
231	220	11	121
199	183	16	256
224	233	-9	81
		$\Sigma d = 35$	$\Sigma d^2 = 873$

Hence, the standard deviation of  $\bar{d}$  is

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{10.78579312}{\sqrt{7}} = 4.07664661$$

**Step 1.** State the null and alternative hypotheses.

$H_0: \mu_d = 0$  (The mean of the paired differences is not different from zero.)

$H_1: \mu_d \neq 0$  (The mean of the paired differences is different from zero.)

**Step 2.** Select the distribution to use.

Here  $\sigma_d$  is unknown, the sample size is small, but the population of paired differences is (approximately) normal. Hence, we use the  $t$  distribution to make the test.

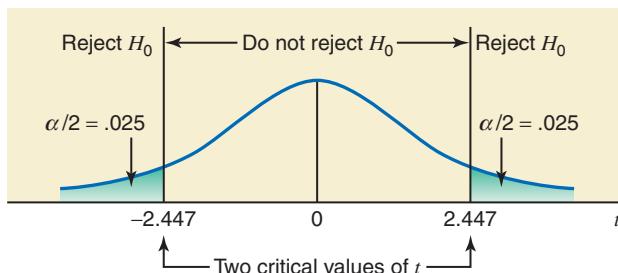
**Step 3.** Determine the rejection and nonrejection regions.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. The significance level is .05.

Area in each tail of the curve =  $\alpha/2 = .05/2 = .025$

Degrees of freedom =  $n - 1 = 7 - 1 = 6$

The two critical values of  $t$  for  $df = 6$  and .025 area in each tail of the  $t$  distribution curve are  $-2.447$  and  $2.447$ . These values are shown in Figure 10.7.



**Figure 10.7** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $\bar{d}$  is computed as follows:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{5.00 - 0}{4.07664661} \xrightarrow{\text{From } H_0} 1.226$$

**Step 5.** Make a decision.

Because the value of the test statistic  $t = 1.226$  for  $\bar{d}$  falls in the nonrejection region, we fail to reject the null hypothesis. Hence, we conclude that the mean of the population paired

differences is not different from zero. In other words, we can state that the mean of the differences between the systolic blood pressures before and after completing this special dietary plan for the population of all adults is not different from zero.

### Using the *p*-Value to Make a Decision

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2 of this example. Then in Step 3, we calculate the value of the test statistic  $t$  for  $\bar{d}$  (as done in Step 4 above) and then find the *p*-value for this  $t$  from the *t* distribution table (Table V of Appendix C) or by using technology. If we have access to technology, we can use it to find the exact *p*-value, which will be .266. By using Table V, we can find the range of the *p*-value. From Table V, for  $df = 6$ , the test statistic  $t = 1.226$  is less than 1.440. The area in the *t* distribution table that corresponds to 1.440 is .10. Because it is a two-tailed test, the *p*-value is greater than  $2(.10) = .20$ , which can be written as

$$p\text{-value} > .20$$

Since  $\alpha = .05$  in this example, which is smaller than .20 and also .266 (obtained from technology), we fail to reject the null hypothesis. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**10.47** Explain when would you use the paired-samples procedure to make confidence intervals and test hypotheses.

**10.48** Find the following confidence intervals for  $\mu_d$ , assuming that the populations of paired differences are normally distributed.

- a.  $n = 11$ ,  $\bar{d} = 25.4$ ,  $s_d = 13.5$ , confidence level = 99%
- b.  $n = 23$ ,  $\bar{d} = 13.2$ ,  $s_d = 4.8$ , confidence level = 95%
- c.  $n = 18$ ,  $\bar{d} = 34.6$ ,  $s_d = 11.7$ , confidence level = 90%

**10.49** Find the following confidence intervals for  $\mu_d$ , assuming that the populations of paired differences are normally distributed.

- a.  $n = 12$ ,  $\bar{d} = 17.5$ ,  $s_d = 6.3$ , confidence level = 99%
- b.  $n = 27$ ,  $\bar{d} = 55.9$ ,  $s_d = 14.7$ , confidence level = 95%
- c.  $n = 16$ ,  $\bar{d} = 29.3$ ,  $s_d = 8.3$ , confidence level = 90%

**10.50** Perform the following tests of hypotheses, assuming that the populations of paired differences are normally distributed.

- a.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d \neq 0$ ,  $n = 9$ ,  $\bar{d} = 6.7$ ,  $s_d = 2.5$ ,  $\alpha = .10$
- b.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d > 0$ ,  $n = 22$ ,  $\bar{d} = 14.8$ ,  $s_d = 6.4$ ,  $\alpha = .05$
- c.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d < 0$ ,  $n = 17$ ,  $\bar{d} = -9.3$ ,  $s_d = 4.8$ ,  $\alpha = .01$

**10.51** Conduct the following tests of hypotheses, assuming that the populations of paired differences are normally distributed.

- a.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d \neq 0$ ,  $n = 26$ ,  $\bar{d} = 9.6$ ,  $s_d = 3.9$ ,  $\alpha = .05$
- b.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d > 0$ ,  $n = 15$ ,  $\bar{d} = 8.8$ ,  $s_d = 4.7$ ,  $\alpha = .01$
- c.  $H_0: \mu_d = 0$ ,  $H_1: \mu_d < 0$ ,  $n = 20$ ,  $\bar{d} = -7.4$ ,  $s_d = 2.3$ ,  $\alpha = .10$

### APPLICATIONS

**10.52** A company sent seven of its employees to attend a course in building self-confidence. These employees were evaluated for their self-confidence before and after attending this course. The following table gives the scores (on a scale of 1 to 15, 1 being the lowest and 15 being the highest score) of these employees before and after they attended the course.

Before	8	5	4	9	6	9	5
After	10	8	5	11	6	7	9

- Construct a 95% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is equal to the score of an employee before attending the course minus the score of the same employee after attending the course.
- Test at a 1% significance level whether attending this course increases the mean score of employees.

Assume that the population of paired differences has a normal distribution.

- 10.53** Several retired bicycle racers are coaching a large group of young prospects. They randomly select seven of their riders to take part in a test of the effectiveness of a new dietary supplement that is supposed to increase strength and stamina. Each of the seven riders does a time trial on the same course. Then they all take the dietary supplement for 4 weeks. All other aspects of their training program remain as they were prior to the time trial. At the end of the 4 weeks, these riders do another time trial on the same course. The times (in minutes) recorded by each rider for these trials before and after the 4-week period are shown in the following table.

Before	103	97	111	95	102	96	108
After	100	95	104	101	96	91	101

- Construct a 99% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is equal to the time taken before the dietary supplement minus the time taken after the dietary supplement.
- Test at a 2.5% significance level whether taking this dietary supplement results in faster times in the time trials.

Assume that the population of paired differences is (approximately) normally distributed.

- 10.54** One type of experiment that might be performed by an exercise physiologist is as follows: Each person in a random sample is tested in a weight room to determine the heaviest weight with which he or she can perform an incline press five times with his or her dominant arm (defined as the hand that a person uses for writing). After a significant rest period, the same weight is determined for each individual's nondominant arm. The physiologist is interested in the differences in the weights pressed by each arm. The following data represent the maximum weights (in pounds) pressed by each arm for a random sample of 18 fifteen-year old girls. Assume that the differences in weights pressed by each arm for all fifteen-year old girls are approximately normally distributed.

Subject	Dominant		Nondominant		Subject	Dominant		Nondominant	
	Arm	Arm	Arm	Arm		Arm	Arm	Arm	Arm
1	59		53		10	47		38	
2	32		30		11	40		35	
3	27		24		12	36		36	
4	18		20		13	21		25	
5	42		40		14	51		48	
6	12		12		15	30		30	
7	29		24		16	32		31	
8	33		34		17	14		14	
9	22		22		18	26		27	

- Make a 99% confidence interval for the mean of the paired differences for the two populations, where a paired difference is equal to the maximum weight for the dominant arm minus the maximum weight for the nondominant arm.
- Using a 1% significance level, can you conclude that the average paired difference as defined in part a is positive?

- 10.55** The Bath Heritage Days, which take place in Bath, Maine, have been popular for, among other things, an eating contest. In 2009, the contest switched from blueberry pie to a Whoopie Pie, which consists of two large, chocolate cake-like cookies filled with a large amount of vanilla cream. Suppose the contest involves eating nine Whoopie Pies, each weighing  $1/3$  pound. The following data represent the times (in seconds) taken by each of the 13 contestants (all of whom finished all nine Whoopie Pies) to eat the first Whoopie Pie and the last (ninth) Whoopie Pie.

Contestant	1	2	3	4	5	6	7	8	9	10	11	12	13
First pie	49	59	66	49	63	70	77	59	64	69	60	58	71
Last pie	49	74	92	93	91	73	103	59	85	94	84	87	111

- Make a 95% confidence interval for the mean of the population paired differences, where a paired difference is equal to the time taken to eat the ninth pie (which is the last pie) minus the time taken to eat the first pie.
- Using a 10% significance level, can you conclude that the average time taken to eat the ninth pie (which is the last pie) is at least 15 seconds more than the average time taken to eat the first pie.

Assume that the population of paired differences is (approximately) normally distributed.

**10.56** The manufacturer of a gasoline additive claims that the use of this additive increases gasoline mileage. A random sample of six cars was selected, and these cars were driven for 1 week without the gasoline additive and then for 1 week with the gasoline additive. The following table gives the miles per gallon for these cars without and with the gasoline additive.

Without	24.6	28.3	18.9	23.7	15.4	29.5
With	26.3	31.7	18.2	25.3	18.3	30.9

- Construct a 99% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is equal to the miles per gallon without the gasoline additive minus the miles per gallon with the gasoline additive.
- Using a 2.5% significance level, can you conclude that the use of the gasoline additive increases the gasoline mileage?

Assume that the population of paired differences is (approximately) normally distributed.

**10.57** A factory that emits airborne pollutants is testing two different brands of filters for its smokestacks. The factory has two smokestacks. One brand of filter (Filter I) is placed on one smokestack, and the other brand (Filter II) is placed on the second smokestack. Random samples of air released from the smokestacks are taken at different times throughout the day. Pollutant concentrations are measured from both stacks at the same time. The following data represent the pollutant concentrations (in parts per million) for samples taken at 20 different times after passing through the filters. Assume that the differences in concentration levels at all times are approximately normally distributed.

Time	Filter I	Filter II	Time	Filter I	Filter II
1	24	26	11	11	9
2	31	30	12	8	10
3	35	33	13	14	17
4	32	28	14	17	16
5	25	23	15	19	16
6	25	28	16	19	18
7	29	24	17	25	27
8	30	33	18	20	22
9	26	22	19	23	27
10	18	18	20	32	31

- Make a 95% confidence interval for the mean of the population paired differences, where a paired difference is equal to the pollutant concentration passing through Filter I minus the pollutant concentration passing through Filter II.
- Using a 5% significance level, can you conclude that the average paired difference for concentration levels is different from zero?

## 10.5

### Inferences About the Difference Between Two Population Proportions for Large and Independent Samples

Quite often we need to construct a confidence interval and test a hypothesis about the difference between two population proportions. For instance, we may want to estimate the difference between the proportions of defective items produced on two different machines. If  $p_1$  and  $p_2$  are the proportions of defective items produced on the first and second machine, respectively, then

we are to make a confidence interval for  $p_1 - p_2$ . Alternatively, we may want to test the hypothesis that the proportion of defective items produced on Machine I is different from the proportion of defective items produced on Machine II. In this case, we are to test the null hypothesis  $p_1 - p_2 = 0$  against the alternative hypothesis  $p_1 - p_2 \neq 0$ .

This section discusses how to make a confidence interval and test a hypothesis about  $p_1 - p_2$  for two large and independent samples. The sample statistic that is used to make inferences about  $p_1 - p_2$  is  $\hat{p}_1 - \hat{p}_2$ , where  $\hat{p}_1$  and  $\hat{p}_2$  are the proportions for two large and independent samples. As discussed in Chapter 7, we determine a sample proportion by dividing the number of elements in the sample that possess a given attribute by the sample size. Thus,

$$\hat{p}_1 = x_1/n_1 \quad \text{and} \quad \hat{p}_2 = x_2/n_2$$

where  $x_1$  and  $x_2$  are the number of elements that possess a given characteristic in the two samples and  $n_1$  and  $n_2$  are the sizes of the two samples, respectively.

### 10.5.1 Mean, Standard Deviation, and Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

As discussed in Chapter 7, for a large sample, the sample proportion  $\hat{p}$  is (approximately) normally distributed with mean  $p$  and standard deviation  $\sqrt{pq/n}$ . Hence, for two large and independent samples of sizes  $n_1$  and  $n_2$ , respectively, their sample proportions  $\hat{p}_1$  and  $\hat{p}_2$  are (approximately) normally distributed with means  $p_1$  and  $p_2$  and standard deviations  $\sqrt{p_1 q_1 / n_1}$  and  $\sqrt{p_2 q_2 / n_2}$ , respectively. Using these results, we can make the following statements about the shape of the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  and its mean and standard deviation.

**Mean, Standard Deviation, and Sampling Distribution of  $\hat{p}_1 - \hat{p}_2$**  For two large and independent samples, the *sampling distribution* of  $\hat{p}_1 - \hat{p}_2$  is (approximately) normal, with its *mean* and *standard deviation* given as

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

and

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

respectively, where  $q_1 = 1 - p_1$  and  $q_2 = 1 - p_2$ .

Thus, to construct a confidence interval and to test a hypothesis about  $p_1 - p_2$  for large and independent samples, we use the normal distribution. As was indicated in Chapter 7, in the case of proportion, the sample is large if  $np$  and  $nq$  are both greater than 5. In the case of two samples, both sample sizes are large if  $n_1 p_1$ ,  $n_1 q_1$ ,  $n_2 p_2$ , and  $n_2 q_2$  are all greater than 5.

### 10.5.2 Interval Estimation of $p_1 - p_2$

The difference between two sample proportions  $\hat{p}_1 - \hat{p}_2$  is the point estimator for the difference between two population proportions  $p_1 - p_2$ . Because we do not know  $p_1$  and  $p_2$  when we are making a confidence interval for  $p_1 - p_2$ , we cannot calculate the value of  $\sigma_{\hat{p}_1 - \hat{p}_2}$ . Therefore, we use  $s_{\hat{p}_1 - \hat{p}_2}$  as the point estimator of  $\sigma_{\hat{p}_1 - \hat{p}_2}$  in the interval estimation. We construct the confidence interval for  $p_1 - p_2$  using the following formula.

**Confidence Interval for  $p_1 - p_2$**  The  $(1 - \alpha)100\%$  confidence interval for  $p_1 - p_2$  is

$$(\hat{p}_1 - \hat{p}_2) \pm z s_{\hat{p}_1 - \hat{p}_2}$$

where the value of  $z$  is read from the normal distribution table for the given confidence level, and  $s_{\hat{p}_1 - \hat{p}_2}$  is calculated as

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example 10–13 describes the procedure that is used to make a confidence interval for the difference between two population proportions for large samples.

### ■ EXAMPLE 10–13

*Constructing a confidence interval for  $p_1 - p_2$ ; large and independent samples.*



© Andrey Armyagov/iStockphoto

A researcher wanted to estimate the difference between the percentages of users of two toothpastes who will never switch to another toothpaste. In a sample of 500 users of Toothpaste A taken by this researcher, 100 said that they will never switch to another toothpaste. In another sample of 400 users of Toothpaste B taken by the same researcher, 68 said that they will never switch to another toothpaste.

- Let  $p_1$  and  $p_2$  be the proportions of all users of Toothpastes A and B, respectively, who will never switch to another toothpaste. What is the point estimate of  $p_1 - p_2$ ?
- Construct a 97% confidence interval for the difference between the proportions of all users of the two toothpastes who will never switch.

**Solution** Let  $p_1$  and  $p_2$  be the proportions of all users of Toothpastes A and B, respectively, who will never switch to another toothpaste, and let  $\hat{p}_1$  and  $\hat{p}_2$  be the respective sample proportions. Let  $x_1$  and  $x_2$  be the number of users of Toothpastes A and B, respectively, in the two samples who said that they will never switch to another toothpaste. From the given information,

$$\text{Toothpaste A: } n_1 = 500 \text{ and } x_1 = 100$$

$$\text{Toothpaste B: } n_2 = 400 \text{ and } x_2 = 68$$

The two sample proportions are calculated as follows:

$$\hat{p}_1 = x_1/n_1 = 100/500 = .20$$

$$\hat{p}_2 = x_2/n_2 = 68/400 = .17$$

Then,

$$\hat{q}_1 = 1 - .20 = .80 \text{ and } \hat{q}_2 = 1 - .17 = .83$$

- The point estimate of  $p_1 - p_2$  is as follows:

$$\text{Point estimate of } p_1 - p_2 = \hat{p}_1 - \hat{p}_2 = .20 - .17 = .03$$

- The values of  $n_1\hat{p}_1$ ,  $n_1\hat{q}_1$ ,  $n_2\hat{p}_2$ , and  $n_2\hat{q}_2$  are

$$n_1\hat{p}_1 = 500(.20) = 100 \quad n_1\hat{q}_1 = 500(.80) = 400$$

$$n_2\hat{p}_2 = 400(.17) = 68 \quad n_2\hat{q}_2 = 400(.83) = 332$$

Because each of these values is greater than 5, both sample sizes are large. Consequently we use the normal distribution to make a confidence interval for  $p_1 - p_2$ . The standard deviation of  $\hat{p}_1 - \hat{p}_2$  is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(.20)(.80)}{500} + \frac{(.17)(.83)}{400}} = .02593742$$

The  $z$  value for a 97% confidence level, obtained from the normal distribution table is 2.17. The 97% confidence interval for  $p_1 - p_2$  is

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) \pm z s_{\hat{p}_1 - \hat{p}_2} &= (.20 - .17) \pm 2.17(.02593742) \\ &= .03 \pm .056 = \mathbf{-.026 \text{ to } .086} \end{aligned}$$

Thus, with 97% confidence we can state that the difference between the two population proportions is between  $-.026$  and  $.086$ .

Note that here  $\hat{p}_1 - \hat{p}_2 = .03$  gives the point estimate of  $p_1 - p_2$  and  $z s_{\hat{p}_1 - \hat{p}_2} = .056$  is the margin of error of the estimate. ■

### 10.5.3 Hypothesis Testing About $p_1 - p_2$

In this section we learn how to test a hypothesis about  $p_1 - p_2$  for two large and independent samples. The procedure involves the same five steps we have used previously. Once again, we calculate the standard deviation of  $\hat{p}_1 - \hat{p}_2$  as

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

When a test of hypothesis about  $p_1 - p_2$  is performed, usually the null hypothesis is  $p_1 = p_2$  and the values of  $p_1$  and  $p_2$  are not known. Assuming that the null hypothesis is true and  $p_1 = p_2$ , a common value of  $p_1$  and  $p_2$ , denoted by  $\bar{p}$ , is calculated by using one of the following two formulas:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \text{or} \quad \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Which of these formulas is used depends on whether the values of  $x_1$  and  $x_2$  or the values of  $\hat{p}_1$  and  $\hat{p}_2$  are known. Note that  $x_1$  and  $x_2$  are the number of elements in each of the two samples that possess a certain characteristic. This value of  $\bar{p}$  is called the **pooled sample proportion**. Using the value of the pooled sample proportion, we compute an estimate of the standard deviation of  $\hat{p}_1 - \hat{p}_2$  as follows:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p} \bar{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where  $\bar{q} = 1 - \bar{p}$ .

**Test Statistic z for  $\hat{p}_1 - \hat{p}_2$**  The value of the *test statistic z for  $\hat{p}_1 - \hat{p}_2$*  is calculated as

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

The value of  $p_1 - p_2$  is substituted from  $H_0$ , which usually is zero.

Examples 10–14 and 10–15 illustrate the procedure to test hypotheses about the difference between two population proportions for large samples.

#### ■ EXAMPLE 10–14

Reconsider Example 10–13 about the percentages of users of two toothpastes who will never switch to another toothpaste. At a 1% significance level, can you conclude that the proportion of users of Toothpaste A who will never switch to another toothpaste is higher than the proportion of users of Toothpaste B who will never switch to another toothpaste?

Making a right-tailed test of hypothesis about  $p_1 - p_2$ : large and independent samples.

**Solution** Let  $p_1$  and  $p_2$  be the proportions of all users of Toothpastes A and B, respectively, who will never switch to another toothpaste, and let  $\hat{p}_1$  and  $\hat{p}_2$  be the corresponding sample proportions. Let  $x_1$  and  $x_2$  be the number of users of Toothpastes A and B, respectively, in the two samples who said that they will never switch to another toothpaste. From the given information,

Toothpaste A:  $n_1 = 500$  and  $x_1 = 100$

Toothpaste B:  $n_2 = 400$  and  $x_2 = 68$

The significance level is  $\alpha = .01$ . The two sample proportions are calculated as follows:

$$\hat{p}_1 = x_1/n_1 = 100/500 = .20$$

$$\hat{p}_2 = x_2/n_2 = 68/400 = .17$$

**Step 1.** State the null and alternative hypotheses.

We are to test if the proportion of users of Toothpaste A who will never switch to another toothpaste is higher than the proportion of users of Toothpaste B who will never switch to

another toothpaste. In other words, we are to test whether  $p_1$  is greater than  $p_2$ . This can be written as  $p_1 - p_2 > 0$ . Thus, the two hypotheses are

$$H_0: p_1 = p_2 \quad \text{or} \quad p_1 - p_2 = 0 \quad (p_1 \text{ is the same as } p_2)$$

$$H_1: p_1 > p_2 \quad \text{or} \quad p_1 - p_2 > 0 \quad (p_1 \text{ is greater than } p_2)$$

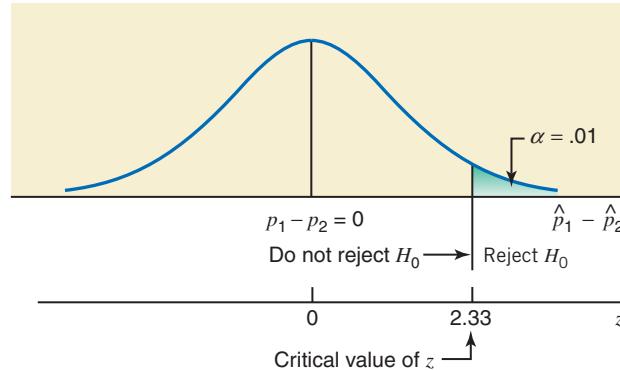
**Step 2.** Select the distribution to use.

As shown in Example 10–13,  $n_1\hat{p}_1$ ,  $n_1\hat{q}_1$ ,  $n_2\hat{p}_2$ , and  $n_2\hat{q}_2$  are all greater than 5. Consequently both samples are large, and we use the normal distribution to make the test.

**Step 3.** Determine the rejection and nonrejection regions.

The  $>$  sign in the alternative hypothesis indicates that the test is right-tailed. From the normal distribution table, for a .01 significance level, the critical value of  $z$  is 2.33 for .9900 area to the left. This is shown in Figure 10.8.

**Figure 10.8** Rejection and nonrejection regions.



**Step 4.** Calculate the value of the test statistic.

The pooled sample proportion is

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{100 + 68}{500 + 400} = .187 \quad \text{and} \quad \bar{q} = 1 - \bar{p} = 1 - .187 = .813$$

The estimate of the standard deviation of  $\hat{p}_1 - \hat{p}_2$  is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(.187)(.813)\left(\frac{1}{500} + \frac{1}{400}\right)} = .02615606$$

The value of the test statistic  $z$  for  $\hat{p}_1 - \hat{p}_2$  is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{(.20 - .17) - 0}{.02615606} = 1.15$$

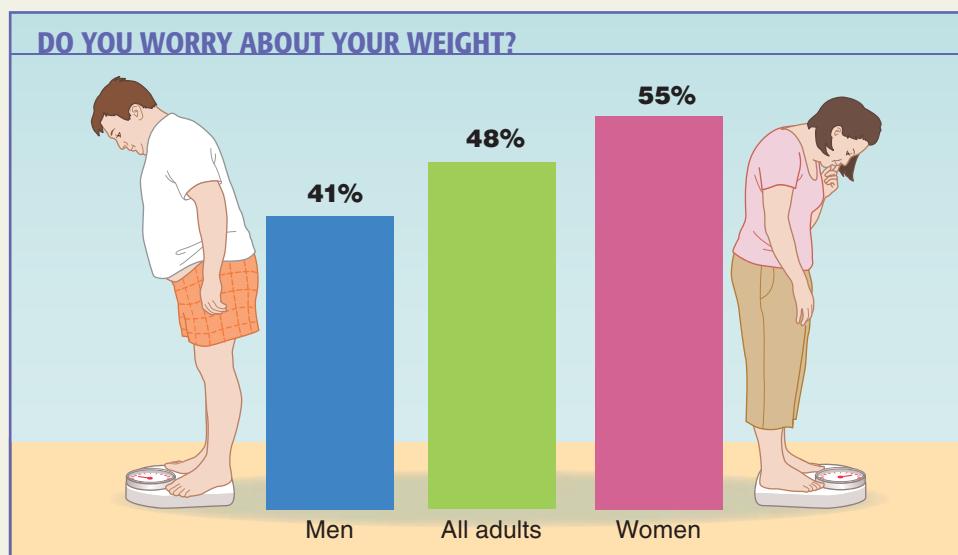
↓    ↓  
From  $H_0$

**Step 5.** Make a decision.

Because the value of the test statistic  $z = 1.15$  for  $\hat{p}_1 - \hat{p}_2$  falls in the nonrejection region, we fail to reject the null hypothesis. Therefore, we conclude that the proportion of users of Toothpaste A who will never switch to another toothpaste is not greater than the proportion of users of Toothpaste B who will never switch to another toothpaste.

### Using the *p*-Value to Make a Decision

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2 above. Then in Step 3, we calculate the value of the test statistic  $z$  (as done in Step 4 above) and find the *p*-value for this  $z$  from the normal distribution table. In Step 4 above, the  $z$ -value for  $\hat{p}_1 - \hat{p}_2$  was calculated to be 1.15. In this example, the test is right-tailed. The *p*-value is given by the area under the normal distribution curve to the right of  $z = 1.15$ . From the normal distribution table (Table IV of Appendix C), this area is  $1 - .8749 = .1251$ . Hence, the



Data source: Gallup poll of 1014 adults aged 18 and older conducted July 9-12, 2012.

The accompanying chart, reproduced here from Case Study 4–1, shows the percentage of men and women who worry about their weight at least some of the time. These results are based on a Gallup poll of 520 U.S. men and 494 U.S. women aged 18 years and older conducted July 9–12, 2012 (<http://www.gallup.com/poll/155903/Gender-Gap-Personal-Weight-Worries-Narrows.aspx>). As the numbers in the chart show, according to this survey, 41% of men and 55% of women aged 18 and older said that they worry about their weight at least some of the time. Using the information given in the chart, we can make a confidence interval and perform a test of hypothesis for the difference in the percentages for these two groups.

Let  $p_1$  and  $p_2$  be the proportions of all men and women aged 18 years and older, respectively, who worry about their weight at least some of the time. Let  $\hat{p}_1$  and  $\hat{p}_2$  be the corresponding sample proportions. Then, from the given information:

$$\begin{aligned} \text{For men: } n_1 &= 520 & \hat{p}_1 &= .41 & \hat{q}_1 &= 1 - .41 = .59 \\ \text{For women: } n_2 &= 494 & \hat{p}_2 &= .55 & \hat{q}_2 &= 1 - .55 = .45 \end{aligned}$$

Below we make a confidence interval and test a hypothesis about  $p_1 - p_2$ .

### 1. Confidence interval for $p_1 - p_2$

Suppose we want to make a 99% confidence interval for  $p_1 - p_2$ . The z value from Table IV in Appendix C for a 99% confidence level is 2.58. The standard deviation of  $\hat{p}_1 - \hat{p}_2$  is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = \sqrt{\frac{(41)(59)}{520} + \frac{(55)(45)}{494}} = .03108383$$

Hence, the 99% confidence interval for  $p_1 - p_2$  is

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) \pm z s_{\hat{p}_1 - \hat{p}_2} &= (.41 - .55) \pm 2.58(.03108383) = -.14 \pm .08 \\ &= -.22 \text{ to } -.06 \text{ or } -22\% \text{ to } -6\% \end{aligned}$$

Thus, we can say with 99% confidence that the difference in the proportions of all men and women aged 18 years and older who worry about their weight at least some of the time is in the interval  $-.22$  to  $-.06$  or  $-22\%$  to  $-6\%$ . Note that this confidence interval is very wide, as the difference between the lower and upper boundaries of the interval is 16%. By taking larger samples, we can lower this width of the confidence interval.

### 2. Test of hypothesis about $p_1 - p_2$

Suppose we want to test, at a 1% significance level, whether the proportion of all men aged 18 years and older who worry about their weight at least some of the time is lower than the proportion of all women

## DO YOU WORRY ABOUT YOUR WEIGHT?

## (CONTINUED)

aged 18 years and older who worry about their weight at least some of the time. In other words, we are to test if  $p_1$  is lower than  $p_2$ . The null and alternative hypotheses are

$$H_0: p_1 = p_2 \quad \text{or} \quad p_1 - p_2 = 0$$

$$H_1: p_1 < p_2 \quad \text{or} \quad p_1 - p_2 < 0$$

Note that the test is left-tailed. For  $\alpha = .01$ , the critical value of  $z$  from the normal distribution table for .0100 is  $-2.33$ . Thus, we will reject the null hypothesis if the observed value of  $z$  is  $-2.33$  or smaller. The pooled sample proportion is

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{520(.41) + 494(.55)}{520 + 494} = .47820513$$

and

$$\bar{q} = 1 - \bar{p} = 1 - .47820513 = .52179487$$

The estimate of the standard deviation of  $\hat{p}_1 - \hat{p}_2$  is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p} \bar{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{(.47820513)(.52179487) \left( \frac{1}{520} + \frac{1}{494} \right)} = .03138418$$

The value of the test statistic  $z$  for  $\hat{p}_1 - \hat{p}_2$  is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{(.41 - .55) - 0}{.03138418} = -4.46$$

Since the observed value of  $z = -4.46$  is smaller than the critical value of  $-2.33$ , we reject the null hypothesis. As a result we conclude that  $p_1$  is lower than  $p_2$  and that the proportion of all men aged 18 years and older who worry about their weight at least some of the time is lower than the proportion of all women aged 18 years and older who worry about their weight at least some of the time.

We can also use the  $p$ -value approach to make this decision. In this example, the test is left-tailed. As calculated above, the  $z$  value for  $\hat{p}_1 - \hat{p}_2$  is  $-4.46$ . From the normal distribution table, the area to the left of  $z = -4.46$  is (approximately)  $.0000$ . Hence, the  $p$ -value is (approximately)  $.0000$ . (By using technology, we obtain the  $p$ -value of  $.000004$ .) Since  $\alpha = .01$  in this example is greater than  $.0000$ , we reject the null hypothesis and conclude that the proportion of all men aged 18 years and older who worry about their weight at least some of the time is lower than the proportion of all women aged 18 years and older who worry about their weight at least some of the time.

$p$ -value is  $.1251$ . We reject the null hypothesis for any  $\alpha$  (significance level) greater than or equal to the  $p$ -value; in this example, we will reject the null hypothesis for any  $\alpha \geq .1251$  or  $12.51\%$ . Because  $\alpha = .01$  here, which is less than  $.1251$ , we fail to reject the null hypothesis. ■

### ■ EXAMPLE 10-15

Conducting a two-tailed test of hypothesis about  $p_1 - p_2$ : large and independent samples.

According to a 2011 survey of college freshmen by UCLA's Cooperative Institutional Research Program, 39.5% of freshmen said that they had spent 6 or more hours a week studying or doing homework as high school seniors (*USA TODAY*, January 26, 2012). This percentage was 37.3% in the 2010 survey of freshmen by the same institution. The sample sizes for these surveys are usually very large, but for this example suppose the samples included 2000 freshmen in 2010 and 2200 freshmen in 2011. Test whether the proportions of 2010 and 2011 freshmen who spent 6 or more hours a week studying or doing homework as high school seniors are different. Use a 1% significance level.

**Solution** Let  $p_1$  and  $p_2$  be the proportions of all freshmen in 2010 and 2011, respectively, who spent 6 or more hours a week studying or doing homework as high school seniors. Let  $\hat{p}_1$  and  $\hat{p}_2$  be the corresponding sample proportions. From the given information

For 2010 freshmen:	$n_1 = 2000$	$\text{and}$
	$\hat{p}_1 = .373$	
For 2011 freshmen:	$n_2 = 2200$	$\text{and}$
	$\hat{p}_2 = .395$	

The significance level is  $\alpha = .01$ .

**Step 1.** State the null and alternative hypotheses.

The null and alternative hypotheses are, respectively,

$$H_0: p_1 - p_2 = 0 \quad (\text{The two population proportions are not different.})$$

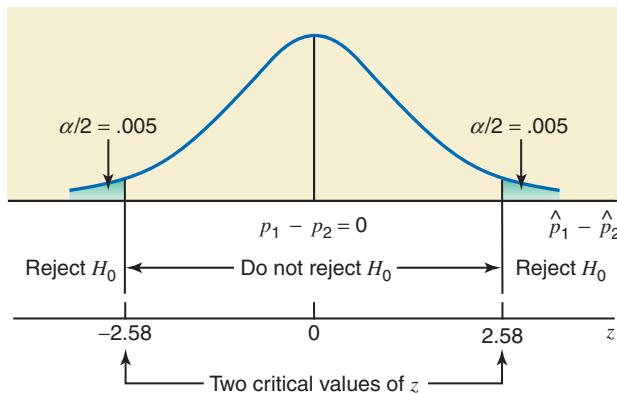
$$H_1: p_1 - p_2 \neq 0 \quad (\text{The two population proportions are different.})$$

**Step 2.** Select the distribution to use.

Because the samples are large and independent, we apply the normal distribution to make the test. (The reader should check that  $n_1\hat{p}_1$ ,  $n_1\hat{q}_1$ ,  $n_2\hat{p}_2$ , and  $n_2\hat{q}_2$  are all greater than 5.)

**Step 3.** Determine the rejection and nonrejection regions.

The  $\neq$  sign in the alternative hypothesis indicates that the test is two-tailed. For a 1% significance level, the critical values of  $z$  are  $-2.58$  and  $2.58$ . Note that to find these two critical values, we look for .0050 and .9950 areas in Table IV of Appendix C. These values are shown in Figure 10.9.



**Figure 10.9** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

The pooled sample proportion is

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{2000(.373) + 2200(.395)}{2000 + 2200} = .385$$

$$\bar{q} = 1 - \bar{p} = 1 - .385 = .615$$

The estimate of the standard deviation of  $\hat{p}_1 - \hat{p}_2$  is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(.385)(.615)\left(\frac{1}{2000} + \frac{1}{2200}\right)} = .01503371$$

The value of the test statistic  $z$  for  $\hat{p}_1 - \hat{p}_2$  is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{(.373 - .395) - 0}{.01503371} = -1.46$$

**Step 5.** Make a decision.

Because the value of the test statistic  $z = -1.46$  falls in the nonrejection region, we fail to reject the null hypothesis  $H_0$ . Therefore, we conclude that the proportions of all freshmen in 2010 and 2011 who spent 6 or more hours a week studying or doing homework as high school seniors are not different.

### Using the *p*-Value to Make a Decision

We can use the *p*-value approach to make the above decision. To do so, we keep Steps 1 and 2. Then in Step 3, we calculate the value of the test statistic  $z$  (as done in Step 4) and find the *p*-value for this  $z$  from the normal distribution table. In Step 4, the  $z$ -value for  $\hat{p}_1 - \hat{p}_2$  was

calculated to be  $-1.46$ . In this example, the test is two-tailed. The  $p$ -value is given by twice the area under the normal distribution curve to the left of  $z = -1.46$ . From the normal distribution table (Table IV of Appendix C), the area to the left of  $z = -1.46$  is  $.0721$ . Hence, the  $p$ -value is  $2(.0721) = .1442$ . As we know, we will reject the null hypothesis for any  $\alpha$  (significance level) greater than or equal to the  $p$ -value. Since  $\alpha = .01$  in this example, which is smaller than  $.1442$ , we fail to reject the null hypothesis. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**10.58** What is the shape of the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  for two large samples? What are the mean and standard deviation of this sampling distribution?

**10.59** When are the samples considered large enough for the sampling distribution of the difference between two sample proportions to be (approximately) normal?

**10.60** Construct a 99% confidence interval for  $p_1 - p_2$  for the following.

$$n_1 = 300, \hat{p}_1 = .55, n_2 = 200, \hat{p}_2 = .62$$

**10.61** Construct a 95% confidence interval for  $p_1 - p_2$  for the following.

$$n_1 = 100, \hat{p}_1 = .81, n_2 = 150, \hat{p}_2 = .77$$

**10.62** Refer to the information given in Exercise 10.60. Test at a 1% significance level if the two population proportions are different.

**10.63** Refer to the information given in Exercise 10.61. Test at a 5% significance level if  $p_1 - p_2$  is different from zero.

**10.64** Refer to the information given in Exercise 10.60. Test at a 1% significance level if  $p_1$  is less than  $p_2$ .

**10.65** Refer to the information given in Exercise 10.61. Test at a 2% significance level if  $p_1$  is greater than  $p_2$ .

**10.66** A sample of 500 observations taken from the first population gave  $x_1 = 305$ . Another sample of 600 observations taken from the second population gave  $x_2 = 348$ .

- a. Find the point estimate of  $p_1 - p_2$ .
- b. Make a 97% confidence interval for  $p_1 - p_2$ .
- c. Show the rejection and nonrejection regions on the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  for  $H_0: p_1 = p_2$  versus  $H_1: p_1 > p_2$ . Use a significance level of 2.5%.
- d. Find the value of the test statistic  $z$  for the test of part c.
- e. Will you reject the null hypothesis mentioned in part c at a significance level of 2.5%?

**10.67** A sample of 1000 observations taken from the first population gave  $x_1 = 290$ . Another sample of 1200 observations taken from the second population gave  $x_2 = 396$ .

- a. Find the point estimate of  $p_1 - p_2$ .
- b. Make a 98% confidence interval for  $p_1 - p_2$ .
- c. Show the rejection and nonrejection regions on the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  for  $H_0: p_1 = p_2$  versus  $H_1: p_1 < p_2$ . Use a significance level of 1%.
- d. Find the value of the test statistic  $z$  for the test of part c.
- e. Will you reject the null hypothesis mentioned in part c at a significance level of 1%?

### APPLICATIONS

**10.68** The global recession has led more and more people to move in with relatives, which has resulted in a large number of multigenerational households. An October 2011 Pew Research Center poll showed that 11.5% of people living in multigenerational households were living below the poverty level, and 14.6% of people living in other types of households were living below the poverty level ([www.pewsocialtrends.org/2011/10/03/fighting-poverty-in-a-bad-economy-americans-move-in-with-relatives/?src=prc-headline](http://www.pewsocialtrends.org/2011/10/03/fighting-poverty-in-a-bad-economy-americans-move-in-with-relatives/?src=prc-headline)). Suppose that these results were based on samples of 1000 people living in multigenerational households and 2000 people living in other types of households.

- a. Let  $p_1$  be the proportion of all people in multigenerational households who live below the poverty level and  $p_2$  be the proportion of all people in other types of households who live below the poverty level. Construct a 98% confidence interval for  $p_1 - p_2$ .
- b. Using a 2.5% significance level, can you conclude that  $p_1$  is less than  $p_2$ ? Use the critical-value approach.
- c. Repeat part b using the  $p$ -value approach.

**10.69** A November 2011 Gallup poll asked American adults about their views of healthcare and the healthcare system in the United States. Although feelings about the quality of healthcare were positive, the same cannot be said about the quality of the healthcare system. According to this study, 29% of Independents and 27% of Democrats rated the healthcare system as being excellent or good ([www.gallup.com/poll/150788/Americans-Maintain-Negative-View-Healthcare-Coverage.aspx](http://www.gallup.com/poll/150788/Americans-Maintain-Negative-View-Healthcare-Coverage.aspx)). Suppose that these results were based on samples of 1200 Independents and 1300 Democrats.

- Let  $p_1$  and  $p_2$  be the proportions of all Independents and all Democrats, respectively, who will rate the healthcare system as being excellent or good. Construct a 97% confidence interval for  $p_1 - p_2$ .
- Using a 1% significance level, can you conclude that  $p_1$  is different from  $p_2$ ? Use both the critical-value and the  $p$ -value approaches.

**10.70** According to Pew Research Center surveys, 79% of U.S. adults were using the Internet in January 2011 and 83% were using it in January 2012 (*USA TODAY*, January 26, 2012). Suppose that these percentages are based on random samples of 1800 U.S. adults in January 2011 and 1900 in January 2012.

- Let  $p_1$  and  $p_2$  be the proportions of all U.S. adults who were using the Internet in January 2011 and January 2012, respectively. Construct a 98% confidence interval for  $p_1 - p_2$ .
- Using a 1% significance level, can you conclude that  $p_1$  is lower than  $p_2$ ? Use both the critical-value and the  $p$ -value approaches.

**10.71** A state that requires periodic emission tests of cars operates two emission test stations, A and B, in one of its towns. Car owners have complained of lack of uniformity of procedures at the two stations, resulting in different failure rates. A sample of 400 cars at Station A showed that 53 of those failed the test; a sample of 470 cars at Station B found that 51 of those failed the test.

- What is the point estimate of the difference between the two population proportions?
- Construct a 95% confidence interval for the difference between the two population proportions.
- Testing at a 5% significance level, can you conclude that the two population proportions are different? Use both the critical-value and the  $p$ -value approaches.

**10.72** The management of a supermarket chain wanted to investigate if the percentages of men and women who prefer to buy national brand products over the store brand products are different. A sample of 600 men shoppers at the company's supermarkets showed that 246 of them prefer to buy national brand products over the store brand products. Another sample of 700 women shoppers at the company's supermarkets showed that 266 of them prefer to buy national brand products over the store brand products.

- What is the point estimate of the difference between the two population proportions?
- Construct a 98% confidence interval for the difference between the proportions of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products.
- Testing at a 1% significance level, can you conclude that the proportions of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products are different?

**10.73** The lottery commissioner's office in a state wanted to find if the percentages of men and women who play the lottery often are different. A sample of 500 men taken by the commissioner's office showed that 160 of them play the lottery often. Another sample of 300 women showed that 66 of them play the lottery often.

- What is the point estimate of the difference between the two population proportions?
- Construct a 99% confidence interval for the difference between the proportions of all men and all women who play the lottery often.
- Testing at a 1% significance level, can you conclude that the proportions of all men and all women who play the lottery often are different?

**10.74** A mail-order company has two warehouses, one on the West Coast and the second on the East Coast. The company's policy is to mail all orders placed with it within 72 hours. The company's quality control department checks quite often whether or not this policy is maintained at the two warehouses. A recently taken sample of 400 orders placed with the warehouse on the West Coast showed that 364 of them were mailed within 72 hours. Another sample of 300 orders placed with the warehouse on the East Coast showed that 279 of them were mailed within 72 hours.

- Construct a 97% confidence interval for the difference between the proportions of all orders placed at the two warehouses that are mailed within 72 hours.
- Using a 2.5% significance level, can you conclude that the proportion of all orders placed at the warehouse on the West Coast that are mailed within 72 hours is lower than the corresponding proportion for the warehouse on the East Coast?

**10.75** A company that has many department stores in the southern states wanted to find at two such stores the percentage of sales for which at least one of the items was returned. A sample of 800 sales randomly selected from Store A showed that for 280 of them at least one item was returned. Another

sample of 900 sales randomly selected from Store B showed that for 279 of them at least one item was returned.

- Construct a 98% confidence interval for the difference between the proportions of all sales at the two stores for which at least one item is returned.
- Using a 1% significance level, can you conclude that the proportions of all sales for which at least one item is returned is higher for Store A than for Store B?

## USES AND MISUSES...

## STATISTICS AND HEALTH-RELATED STUDIES

While watching the news or browsing through a newspaper, we are likely to see or read the results of some new medical study. Often, the study result is really eye-catching, such as the following headline from *USA TODAY*: "Chocolate Lowers Heart Stroke Risk" (<http://yourlife.usatoday.com/health/healthcare/studies/story/2011-08-29/Chocolate-lowers-heart-stroke-risk/50174422/1>). Certainly a headline such as this one can grab your attention, especially if you are a chocolate lover. However, it is important to find out about the type of study that was conducted and determine whether it indicated an association (i.e., it showed a potential link between chocolate consumption and stroke risk) or a causal relationship (i.e., it showed that consuming chocolate actually does lower the risk of a stroke).

Two primary types of medical studies are (1) case-control studies and (b) cohort studies. A case-control study is an observational study that essentially works *backward*. In the case of the aforementioned story, people in the study were classified into two different groups—those who had a stroke, and those who did not have a stroke. The people were interviewed and asked about their chocolate consumption habits. The study revealed that the proportion of people who consumed chocolate was higher in the nonstroke group than in the stroke group. As noted in the *USA TODAY* article, this result regarding the relationship was consistent across a number of independent case-control studies, which were combined into a single study using a process called *meta-analysis*. So, with this information, this set of case-control studies leads to a new question: Does consuming chocolate actually reduce the risk of having a stroke, or is there something else that is the cause of these results?

When a case-control study identifies a potential causal link, the next step is to perform a cohort study (i.e., a *clinical trial*). In a cohort study, individuals are selected to participate in the study, and then these individuals are divided into different groups, with each group receiving a specific *treatment*. In this case, members in one group would eat chocolate regularly, whereas the members in the other group would not. The participants would be observed over a long period of time, and the proportion of people who have a stroke in each group would be compared. If the result of this study is consistent with the result of the case-control study, more research would be conducted to determine the biological or chemical link between chocolate and stroke prevention.

One question you might ask at this point is: Why would we perform a case-control study instead of commencing with a cohort study from the start? The reason is that cohort studies, especially those involving health-related issues, are expensive and involve a great deal of time. In the chocolate-and-stroke example, the researchers will have to wait to determine whether the individuals in the study have a stroke or not. If you do not know whether there is an association between chocolate consumption and stroke risk, you would probably not want to spend the time and money to perform this study. Case-control studies, on the other hand, are relatively inexpensive and easy to perform because databases are maintained on health events, and, hence, the results of the study can be obtained by accessing databases and performing interviews. The results of a cheaper case-control study would allow us to determine whether it is worth spending the money to perform a cohort study that provides data to help make a more substantial decision.

## Glossary

**d** The difference between two matched values in two samples collected from the same source. It is called the paired difference.

**$\bar{d}$**  The mean of the paired differences for a sample.

**Independent samples** Two samples drawn from two populations such that the selection of one does not affect the selection of the other.

**Paired or matched samples** Two samples drawn in such a way that they include the same elements and two data values are ob-

tained from each element, one for each sample. Also called **dependent samples**.

**$\mu_d$**  The mean of the paired differences for the population.

**$s_d$**  The standard deviation of the paired differences for a sample.

**$\sigma_d$**  The standard deviation of the paired differences for the population.

## Supplementary Exercises

**10.76** A consulting agency was asked by a large insurance company to investigate if business majors were better salespersons than those with other majors. A sample of 20 salespersons with a business degree showed that they sold an average of 11 insurance policies per week. Another sample of 25 salespersons with a degree other than business showed that they sold an average of 9 insurance policies per week.

Assume that the two populations are normally distributed with population standard deviations of 1.80 and 1.35 policies per week, respectively.

- a. Construct a 99% confidence interval for the difference between the two population means.
- b. Using a 1% significance level, can you conclude that persons with a business degree are better salespersons than those who have a degree in another area?

**10.77** According to an estimate, the average earnings of female workers who are not union members are \$909 per week and those of female workers who are union members are \$1035 per week. Suppose that these average earnings are calculated based on random samples of 1500 female workers who are not union members and 2000 female workers who are union members. Further assume that the standard deviations for the two corresponding populations are \$70 and \$90, respectively.

- a. Construct a 95% confidence interval for the difference between the two population means.
- b. Test at a 2.5% significance level whether the mean weekly earnings of female workers who are not union members are less than those of female workers who are union members.

**10.78** An economist was interested in studying the impact of the recession on dining out, including drive-thru meals at fast food restaurants. A random sample of forty-eight families of four with discretionary incomes between \$300 and \$400 per week indicated that they reduced their spending on dining out by an average of \$31.47 per week, with a sample standard deviation of \$10.95. Another random sample of 42 families of five with discretionary incomes between \$300 and \$400 per week reduced their spending on dining out by an average \$35.28 per week, with a sample standard deviation of \$12.37. (Note that the two groups of families are differentiated by the number of family members.) Assume that the distributions of reductions in weekly dining-out spendings for the two groups have the same population standard deviation.

- a. Construct a 90% confidence interval for the difference in the mean weekly reduction in dining-out spending levels for the two populations.
- b. Using a 5% significance level, can you conclude that the average weekly spending reduction for all families of four with discretionary incomes between \$300 and \$400 per week is less than the average weekly spending reduction for all families of five with discretionary incomes between \$300 and \$400 per week?

**10.79** According to a report in *The New York Times*, in the United States, accountants and auditors earn an average of \$70,130 a year and loan officers earn \$67,960 a year (Jessica Silver-Greenberg, *The New York Times*, April 22, 2012). Suppose that these estimates are based on random samples of 1650 accountants and auditors and 1820 loan officers. Further assume that the sample standard deviations of the salaries of the two groups are \$14,400 and \$13,600, respectively, and the population standard deviations are equal for the two groups.

- a. Construct a 98% confidence interval for the difference in the mean salaries of the two groups—accountants and auditors, and loan officers.
- b. Using a 1% significance level, can you conclude that the average salary of accountants and auditors is higher than that of loan officers?

**10.80** The manager of a factory has devised a detailed plan for evacuating the building as quickly as possible in the event of a fire or other emergency. An industrial psychologist believes that workers actually leave the factory faster at closing time without following any system. The company holds fire drills periodically in which a bell sounds and workers leave the building according to the system. The evacuation time for each drill is recorded. For comparison, the psychologist also records the evacuation time when the bell sounds for closing time each day. A random sample of 36 fire drills showed a mean evacuation time of 5.1 minutes with a standard deviation of 1.1 minutes. A random sample of 37 days at closing time showed a mean evacuation time of 4.2 minutes with a standard deviation of 1.0 minute.

- a. Construct a 99% confidence interval for the difference between the two population means.
- b. Test at a 5% significance level whether the mean evacuation time is smaller at closing time than during fire drills.

Assume that the evacuation times at closing time and during fire drills have equal but unknown population standard deviations.

**10.81** Two local post offices are interested in knowing the average number of Christmas cards that are mailed out from the towns that they serve. A random sample of 80 households from Town A showed that they mailed an average of 28.55 Christmas cards with a standard deviation of 10.30. The corresponding values of the mean and standard deviation produced by a random sample of 58 households from Town B were 33.67 and 8.97 Christmas cards. Assume that the distributions of the numbers of Christmas cards mailed by all households from both these towns have the same population standard deviation.

- a. Construct a 95% confidence interval for the difference in the average numbers of Christmas cards mailed by all households in these two towns.

- b. Using a 10% significance level, can you conclude that the average number of Christmas cards mailed out by all households in Town A is different from the corresponding average for Town B?

**10.82** Refer to Exercise 10.78. Now answer the questions of parts a and b there without assuming that the standard deviations are the same for the two populations but under the following two situations.

- Using the sample standard deviations given in Exercise 10.78.
- Using sample standard deviations of \$7.17 and \$15.80 for families of four and families of five, respectively.

**10.83** Repeat Exercise 10.79 assuming that the population standard deviations are not equal for the two groups, but considering the following two situations.

- Using the sample standard deviations given in Exercise 10.79.
- Using a sample standard deviation of \$16,700 for accountants and auditors and \$7900 for loan officers.

**10.84** Repeat Exercise 10.80 without assuming that the standard deviations for the two populations are the same but considering the following two situations.

- Using the sample standard deviations given in Exercise 10.80.
- Using sample standard deviations of 1.33 and .72 for fire drills and closing time, respectively.

**10.85** Repeat Exercise 10.81 without assuming that the standard deviations for the two populations are the same but considering the following two situations.

- Using the sample standard deviations given in Exercise 10.81.
- Using sample standard deviations of 6.85 and 11.97 for Town A and Town B, respectively.

**10.86** The owner of a mosquito-infested fishing camp in Alaska wants to test the effectiveness of two rival brands of mosquito repellents, X and Y. During the first month of the season, eight people are chosen at random from those guests who agree to take part in the experiment. For each of these guests, Brand X is randomly applied to one arm and Brand Y is applied to the other arm. These guests fish for 4 hours, then the owner counts the number of bites on each arm. The table below shows the number of bites on the arm with Brand X and those on the arm with Brand Y for each guest.

Guest	A	B	C	D	E	F	G	H
Brand X	12	23	18	36	8	27	22	32
Brand Y	9	20	21	27	6	18	15	25

- Construct a 95% confidence interval for the mean  $\mu_d$  of population paired differences, where a paired difference is defined as the number of bites on the arm with Brand X minus the number of bites on the arm with Brand Y.
- Test at a 5% significance level whether the mean number of bites on the arm with Brand X and the mean number of bites on the arm with Brand Y are different for all such guests.

Assume that the population of paired differences has a normal distribution.

**10.87** A random sample of nine students was selected to test for the effectiveness of a special course designed to improve memory. The following table gives the scores in a memory test given to these students before and after this course.

Before	43	57	48	65	81	49	38	69	58
After	49	56	55	77	89	57	36	64	69

- Construct a 95% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is defined as the difference between the memory test scores of a student before and after attending this course.
- Test at a 1% significance level whether this course makes any statistically significant improvement in the memory of all students.

Assume that the population of paired differences has a normal distribution.

**10.88** In a random sample of 800 men aged 25 to 35 years, 24% said they live with one or both parents. In another sample of 850 women of the same age group, 18% said that they live with one or both parents.

- a. Construct a 95% confidence interval for the difference between the proportions of all men and all women aged 25 to 35 years who live with one or both parents.
- b. Test at a 2% significance level whether the two population proportions are different.
- c. Repeat the test of part b using the  $p$ -value approach.

**10.89** A November 2011 Pew Research Center poll asked American social media users about their use of social media (such as Facebook, Twitter, MySpace, or LinkedIn). The study is based on a national telephone survey of 2277 adult social media users conducted from April 26 to May 22, 2011 ([www.pewinternet.org/Reports/2011/Why-Americans-Use-Social-Media/Main-report.aspx](http://www.pewinternet.org/Reports/2011/Why-Americans-Use-Social-Media/Main-report.aspx)). According to this survey, 16% of 30- to 49-year-old and 18% of 50- to 64-year-old social media users cited connecting with others with common hobbies or interests as a major reason for using social networking sites. Suppose that this survey included 562 social media users in the 30 to 49 age group and 624 in the 50 to 64 age group.

- a. Let  $p_1$  and  $p_2$  be the proportions of all social media users in the age groups 30 to 49 years and 50 to 64 years, respectively, who will cite connecting with others with common hobbies or interests as a major reason for using social networking sites. Construct a 95% confidence interval for  $p_1 - p_2$ .
- b. Using a 1% significance level, can you conclude that  $p_1$  is different from  $p_2$ ? Use both the critical-value and the  $p$ -value approaches.

**10.90** A May 2011 Harris Interactive poll asked American adult women, "How often do you think women of your age, who have no special risk factors for breast cancer, should have a mammogram to check for breast cancer?" Fifty percent of women age 40 to 49 years and 56% of women age 50 years or older said annually ([www.harrisinteractive.com/NewsRoom/PressReleases/tabid/446/ctl/ReadCustomDefault/mid/1506/ArticleId/769/Default.aspx](http://www.harrisinteractive.com/NewsRoom/PressReleases/tabid/446/ctl/ReadCustomDefault/mid/1506/ArticleId/769/Default.aspx)). Suppose that these results were based on samples of 1055 women age 40 to 49 years and 1240 women age 50 years or older.

- a. Let  $p_1$  and  $p_2$  be the proportions of all women age 40 to 49 years and age 50 years or older, respectively, who will say that women of their age with no special risk factors for breast cancer should have an annual mammogram. Construct a 98% confidence interval for  $p_1 - p_2$ .
- b. Using a 1% significance level, can you conclude that  $p_1$  is less than  $p_2$ ? Use both the critical-value and the  $p$ -value approaches.

**10.91** According to a Randstad Global Work Monitor survey, 52% of men and 43% of women said that working part-time hinders their career opportunities (*USA TODAY*, October 6, 2011). Suppose that these results are based on random samples of 1350 men and 1480 women.

- a. Let  $p_1$  and  $p_2$  be the proportions of all men and all women, respectively, who will say that working part-time hinders their career opportunities. Construct a 95% confidence interval for  $p_1 - p_2$ .
- b. Using a 2% significance level, can you conclude that  $p_1$  and  $p_2$  are different? Use both the critical-value and the  $p$ -value approaches.

## Advanced Exercises

**10.92** Manufacturers of two competing automobile models, Gofer and Diplomat, each claim to have the lowest mean fuel consumption. Let  $\mu_1$  be the mean fuel consumption in miles per gallon (mpg) for the Gofer and  $\mu_2$  the mean fuel consumption in mpg for the Diplomat. The two manufacturers have agreed to a test in which several cars of each model will be driven on a 100-mile test run. Then the fuel consumption, in mpg, will be calculated for each test run. The average of the mpg for all 100-mile test runs for each model gives the corresponding mean. Assume that for each model the gas mileages for the test runs are normally distributed with  $\sigma = 2$  mpg. Note that each car is driven for one and only one 100-mile test run.

- a. How many cars (i.e., sample size) for each model are required to estimate  $\mu_1 - \mu_2$  with a 90% confidence level and with a margin of error of estimate of 1.5 mpg? Use the same number of cars (i.e., sample size) for each model.
- b. If  $\mu_1$  is actually 33 mpg and  $\mu_2$  is actually 30 mpg, what is the probability that five cars for each model would yield  $\bar{x}_1 \geq \bar{x}_2$ ?

**10.93** Maria and Ellen both specialize in throwing the javelin. Maria throws the javelin a mean distance of 200 feet with a standard deviation of 10 feet, whereas Ellen throws the javelin a mean distance of 210 feet with a standard deviation of 12 feet. Assume that the distances each of these athletes throws the javelin are normally distributed with these population means and standard deviations. If Maria and Ellen each throw the javelin once, what is the probability that Maria's throw is longer than Ellen's?

**10.94** A new type of sleeping pill is tested against an older, standard pill. Two thousand insomniacs are randomly divided into two equal groups. The first group is given the old pill, and the second group receives the new pill. The time required to fall asleep after the pill is administered is recorded for each person. The

results of the experiment are given in the following table, where  $\bar{x}$  and  $s$  represent the mean and standard deviation, respectively, for the times required to fall asleep for people in each group after the pill is taken.

	Group 1 (Old Pill)	Group 2 (New Pill)
$n$	1000	1000
$\bar{x}$	15.4 minutes	15.0 minutes
$s$	3.5 minutes	3.0 minutes

Consider the test of hypothesis  $H_0: \mu_1 - \mu_2 = 0$  versus  $H_1: \mu_1 - \mu_2 > 0$ , where  $\mu_1$  and  $\mu_2$  are the mean times required for all potential users to fall asleep using the old pill and the new pill, respectively.

- Find the  $p$ -value for this test.
- Does your answer to part a indicate that the result is statistically significant? Use  $\alpha = .025$ .
- Find a 95% confidence interval for  $\mu_1 - \mu_2$ .
- Does your answer to part c imply that this result is of great *practical significance*?

**10.95** Gamma Corporation is considering the installation of governors on cars driven by its sales staff. These devices would limit the car speeds to a preset level, which is expected to improve fuel economy. The company is planning to test several cars for fuel consumption without governors for 1 week. Then governors would be installed in the same cars, and fuel consumption will be monitored for another week. Gamma Corporation wants to estimate the mean difference in fuel consumption with a margin of error of estimate of 2 mpg with a 90% confidence level. Assume that the differences in fuel consumption are normally distributed and that previous studies suggest that an estimate of  $s_d = 3$  mpg is reasonable. How many cars should be tested? (Note that the critical value of  $t$  will depend on  $n$ , so it will be necessary to use trial and error.)

**10.96** Refer to Exercise 10.95. Suppose Gamma Corporation decides to test governors on seven cars. However, the management is afraid that the speed limit imposed by the governors will reduce the number of contacts the salespersons can make each day. Thus, both the fuel consumption and the number of contacts made are recorded for each car/salesperson for each week of the testing period, both before and after the installation of governors.

Salesperson	Number of Contacts		Fuel Consumption (mpg)	
	Before	After	Before	After
A	50	49	25	26
B	63	60	21	24
C	42	47	27	26
D	55	51	23	25
E	44	50	19	24
F	65	60	18	22
G	66	58	20	23

Suppose that as a statistical analyst with the company, you are directed to prepare a brief report that includes statistical analysis and interpretation of the data. Management will use your report to help decide whether or not to install governors on all salespersons' cars. Use 90% confidence intervals and .05 significance levels for any hypothesis tests to make suggestions. Assume that the differences in fuel consumption and the differences in the number of contacts are both normally distributed.

**10.97** Two competing airlines, Alpha and Beta, fly a route between Des Moines, Iowa, and Wichita, Kansas. Each airline claims to have a lower percentage of flights that arrive late. Let  $p_1$  be the proportion of Alpha's flights that arrive late and  $p_2$  the proportion of Beta's flights that arrive late.

- You are asked to observe a random sample of arrivals for each airline to estimate  $p_1 - p_2$  with a 90% confidence level and a margin of error of estimate of .05. How many arrivals for each airline would you have to observe? (Assume that you will observe the same number of arrivals,  $n$ , for each airline. To be sure of taking a large enough sample, use  $p_1 = p_2 = .50$  in your calculations for  $n$ .)
- Suppose that  $p_1$  is actually .30 and  $p_2$  is actually .23. What is the probability that a sample of 100 flights for each airline (200 in all) would yield  $\hat{p}_1 \geq \hat{p}_2$ ?

**10.98** Refer to Exercise 10.56, in which a random sample of six cars was selected to test a gasoline additive. The six cars were driven for 1 week without the gasoline additive and then for 1 week with the additive. The data reproduced here from that exercise show miles per gallon without and with the additive.

Without	24.6	28.3	18.9	23.7	15.4	29.5
With	26.3	31.7	18.2	25.3	18.3	30.9

Suppose that instead of the study with 6 cars, a random sample of 12 cars is selected and these cars are divided randomly into two groups of 6 cars each. The cars in the first group are driven for 1 week without the additive, and the cars in the second group are driven for 1 week with the additive. Suppose that the top row of the table lists the gas mileages for the 6 cars without the additive, and the bottom row gives the gas mileages for the cars with the additive. Assume that the distributions of the gas mileages with or without the additive are (approximately) normal with equal but unknown standard deviations.

- a. Would a paired sample test as described in Section 10.4 be appropriate in this case? Why or why not? Explain.
- b. If the paired sample test is inappropriate here, carry out a suitable test of whether the mean gas mileage is lower without the additive. Use  $\alpha = .025$ .
- c. Compare your conclusion in part b with the result of the hypothesis test in Exercise 10.56.

**10.99** Does the use of cellular telephones increase the risk of brain tumors? Suppose that a manufacturer of cell phones hires you to answer this question because of concern about public liability suits. How would you conduct an experiment to address this question? Be specific. Explain how you would observe, how many observations you would take, and how you would analyze the data once you collect them. What are your null and alternative hypotheses? Would you want to use a higher or a lower significance level for the test? Explain.

**10.100** We wish to estimate the difference between the mean scores on a standardized test of students taught by Instructors A and B. The scores of all students taught by Instructor A have a normal distribution with a standard deviation of 15, and the scores of all students taught by Instructor B have a normal distribution with a standard deviation of 10. To estimate the difference between the two means, you decide that the same number of students from each instructor's class should be observed.

- a. Assuming that the sample size is the same for each instructor's class, how large a sample should be taken from each class to estimate the difference between the mean scores of the two populations to within 5 points with 90% confidence?
- b. Suppose that samples of the size computed in part a will be selected in order to test for the difference between the two population mean scores using a .05 level of significance. How large does the difference between the two sample means have to be for you to conclude that the two population means are different?
- c. Explain why a paired-samples design would be inappropriate for comparing the scores of Instructor A versus Instructor B.

**10.101** The weekly weight losses of all dieters on Diet I have a normal distribution with a mean of 1.3 pounds and a standard deviation of .4 pound. The weekly weight losses of all dieters on Diet II have a normal distribution with a mean of 1.5 pounds and a standard deviation of .7 pound. A random sample of 25 dieters on Diet I and another sample of 36 dieters on Diet II are observed.

- a. What is the probability that the difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ , will be within  $-.15$  to  $.15$ , that is,  $-.15 < \bar{x}_1 - \bar{x}_2 < .15$ ?
- b. What is the probability that the average weight loss  $\bar{x}_1$  for dieters on Diet I will be greater than the average weight loss  $\bar{x}_2$  for dieters on Diet II?
- c. If the average weight loss of the 25 dieters using Diet I is computed to be 2.0 pounds, what is the probability that the difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ , will be within  $-.15$  to  $.15$ , that is,  $-.15 < \bar{x}_1 - \bar{x}_2 < .15$ ?
- d. Suppose you conclude that the assumption  $-.15 < \mu_1 - \mu_2 < .15$  is reasonable. What does this mean to a person who chooses one of these diets?

**10.102** Sixty-five percent of all male voters and 40% of all female voters favor a particular candidate. A sample of 100 male voters and another sample of 100 female voters will be polled. What is the probability that at least 10 more male voters than female voters will favor this candidate?

## Self-Review Test

1. To test the hypothesis that the mean blood pressure of university professors is lower than that of company executives, which of the following would you use?

- a. A left-tailed test
- b. A two-tailed test
- c. A right-tailed test

2. Briefly explain the meaning of independent and dependent samples. Give one example of each of these cases.
3. A company psychologist wanted to test if company executives have job-related stress scores higher than those of university professors. He took a sample of 40 executives and 50 professors and tested them for job-related stress. The sample of 40 executives gave a mean stress score of 7.6. The sample of 50 professors produced a mean stress score of 5.4. Assume that the standard deviations of the two populations are .8 and 1.3, respectively.
  - a. Construct a 99% confidence interval for the difference between the mean stress scores of all executives and all professors.
  - b. Test at a 2.5% significance level whether the mean stress score of all executives is higher than that of all professors.
4. A sample of 20 alcoholic fathers showed that they spend an average of 2.3 hours per week playing with their children with a standard deviation of .54 hour. A sample of 25 nonalcoholic fathers gave a mean of 4.6 hours per week with a standard deviation of .8 hour.
  - a. Construct a 95% confidence interval for the difference between the mean times spent per week playing with their children by all alcoholic and all nonalcoholic fathers.
  - b. Test at a 1% significance level whether the mean time spent per week playing with their children by all alcoholic fathers is less than that of nonalcoholic fathers.

Assume that the times spent per week playing with their children by all alcoholic and all nonalcoholic fathers both are normally distributed with equal but unknown standard deviations.

5. Repeat Problem 4 assuming that the times spent per week playing with their children by all alcoholic and all nonalcoholic fathers both are normally distributed with unequal and unknown standard deviations.
6. Lake City has two shops, Zeke's and Elmer's, that handle the majority of the town's auto body repairs. Seven cars that were damaged in collisions were taken to both shops for written estimates of the repair costs. These estimates (in dollars) are shown in the following table.

Zeke's	1058	544	1349	1296	676	998	1698
Elmer's	995	540	1175	1350	605	970	1520

- a. Construct a 99% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is equal to Zeke's estimate minus Elmer's estimate.
- b. Test at a 5% significance level whether the mean  $\mu_d$  of the population paired differences is different from zero.

Assume that the population of paired differences is (approximately) normally distributed.

7. A sample of 500 male registered voters showed that 57% of them voted in the last presidential election. Another sample of 400 female registered voters showed that 55% of them voted in the same election.

- a. Construct a 97% confidence interval for the difference between the proportions of all male and all female registered voters who voted in the last presidential election.
- b. Test at a 1% significance level whether the proportion of all male voters who voted in the last presidential election is different from that of all female voters.

## Mini-Projects

### ■ MINI-PROJECT 10-1

Suppose that a new cold-prevention drug was tested in a randomized, placebo-controlled, double-blind experiment during the month of January. One thousand healthy adults were randomly divided into two groups of 500 each—a treatment group and a control group. The treatment group was given the new drug, and the control group received a placebo. During the month, 40 people in the treatment group and 120 people in the control group caught a cold. Explain how to construct a 95% confidence interval for the difference between the relevant population proportions. Also describe an appropriate hypothesis test, using the given data, to evaluate the effectiveness of this new drug for cold prevention.

Find a similar article in a journal of medicine, psychology, or other field that lends itself to confidence intervals and hypothesis tests for differences in two means or proportions. First explain how to make the confidence intervals and hypothesis tests; then do so using the data given in the article.

## ■ MINI-PROJECT 10-2

A researcher conjectures that cities in the more populous states of the United States tend to have higher costs for doctors' visits. Using "CITY DATA" that accompany this text, select a random sample of 10 cities from the six most populous states (California, Texas, New York, Florida, Illinois and Pennsylvania). Then take a random sample of 10 cities from the remaining states in the data set. For each of the 20 cities, record the average cost of a doctor's visit. Assume that such costs are approximately normally distributed for all cities in each of the two groups of states. Further assume that the cities you selected make random samples of all cities for the two groups of states. Assume that the standard deviations for the two groups are unequal and unknown.

- Construct a 95% confidence interval for the difference in the mean costs of doctors' visits for all cities in the two groups of states.
- At a 5% level of significance, can you conclude that the average cost of a doctor's visit for all cities in the six most populous states is higher than that of a doctor's visit for all cities in the remaining states?

## ■ MINI-PROJECT 10-3

Many different kinds of analyses have been performed on the salaries of professional athletes. Perform a hypothesis test of whether or not the average salaries of players in two sports are different by taking independent random samples of 35 players each from any two sports of your choice from Major League Baseball (MLB), the National Football League (NFL), the National Basketball Association (NBA), and the National Hockey League (NHL). (Note: A good Internet reference for such data is <http://www.usatoday.com/sports/salaries/index.htm>.) After you take samples, do the following.

- For each player, calculate the weekly salary. For your information, the approximate length (in weeks) of a season is 32.5 for the MLB, 22.5 for the NFL, 28 for the NBA, and 29.5 for the NHL. This length of a season does not include the playoffs, but it does include training camp and the preseason games because each player is expected to participate in these events. Players may receive bonuses for making the playoffs, but these are not included in their base salaries. You may ignore such bonuses.
- Perform a hypothesis test to determine if the average weekly salaries are the same for the two sports that you selected. Use a significance level of 5%. Make certain to indicate whether you decide to use the pooled variance assumption or not, and justify your selection.
- Perform a hypothesis test on the same data to determine if the average annual salaries are the same for the two sports that you selected. Explain why you could get a different answer (with regard to rejecting or failing to reject the null hypothesis) when using the weekly salaries versus the annual salaries.

## ■ MINI-PROJECT 10-4

As reported in *USA TODAY* of August 27, 2009, a 3M Privacy Filters survey asked American adults what seat they prefer on a plane when they fly: window, middle, or aisle seat. Obtain random samples of 60 male and 60 female college students and ask the following question: When you fly on a plane, do you prefer to have a window seat or a nonwindow seat? Perform a hypothesis test to determine if the proportion of female college students who prefer to have a window seat when flying is different from the proportion of male college students who prefer to have a window seat when flying. Use a 5% significance level.

## ■ MINI-PROJECT 10-5

The following table gives the average expenditure expected in order to raise a child born in 2008 through the age of 17 years for families in each of three income groups. These estimates are based on a USDA Center for Nutrition Policy and Promotion study reported in *USA Today* of August 21, 2009.

Family Income	Mean Expenditure
Less than \$56,870	\$159,870
\$56,870 to \$98,470	\$221,190
More than \$98,470	\$366,660

- Suppose that the equal variance assumption is reasonable for the lower and middle income groups (the first two groups listed in the table). Using the "Less than \$56,870" group as population 1 and the "\$56,870 to \$98,470" group as population 2, determine the largest possible value of the pooled

standard deviation that would cause one to reject the null hypothesis  $H_0: \mu_1 = \mu_2$  in favor of the hypothesis  $H_1: \mu_1 < \mu_2$  at a 5% significance level when  $n_1 = n_2 = 10$ .

- b. Repeat part a for the following sample sizes
  - i.  $n_1 = n_2 = 15$
  - ii.  $n_1 = n_2 = 20$
  - iii.  $n_1 = n_2 = 30$
- c. The sample size for three groups (combined) in the study was more than 3000. Based on your results in parts a and b, do you think that it would have been possible to obtain a pooled standard deviation that would have caused you not to reject  $H_0: \mu_1 = \mu_2$  in favor of the hypothesis  $H_1: \mu_1 < \mu_2$  at the 5% significance level? Explain why.

## DECIDE FOR YOURSELF DECIDING ABOUT HOW TO DESIGN A STUDY

By now, you might feel that you have learned almost everything there is to know about statistics. In some ways, you have learned a great deal. When using the  $p$ -value approach, the rule to reject a null hypothesis whenever the  $p$ -value is less than or equal to the significance level never changes. If you know this rule, you do not have to worry about changing it. You have also learned the basic concept of a confidence interval, which will also never change. However, one of the most important lessons to learn in statistics is how to conduct a valid study. Design of experiments and sampling design are two areas of statistics that are dedicated to determining the proper way to plan a study before any data are collected. Without a proper plan, the time and money spent on the study could be a complete waste if the results are not valid.

Consider the example of gasoline additive mentioned in the Decide for Yourself section of Chapter 9. In that section, we discussed

performing a single-sample procedure. However, the same problem could be addressed by using some of the procedures learned in this chapter.

1. Describe how that analysis could be performed by selecting two independent samples of cars. Be specific about how the treatments are applied/assigned to the cars, whether there are any special considerations as to how the cars are selected, and the specific measurements that would be compared.
2. Answer question 1 assuming that we use a paired-sample procedure instead of a two independent samples procedure.
3. Discuss the strengths and weaknesses of the three procedures (including the single sample procedure discussed in Chapter 9). Which method would you prefer and why? Explain.

### TECHNOLOGY INSTRUCTION

#### TI-84

```
2-SampTTest
Inpt:Data Stats
List1:L1
List2:L2
Freq1:1
Freq2:10
μ1:<math>\mu_1</math> < <math>\mu_2</math>
μ2:<math>\mu_2</math>
Pooled:No Yes

```

Screen 10.1

```
2-SampTTest
μ1 ≠ μ2
t = -1.282837671
P = .2197778667
df = 14.425734
x̄1 = 22.36
x̄2 = 22.96
```

Screen 10.2

#### Confidence Intervals and Hypothesis Tests for Two Populations

1. To perform a hypothesis test about the difference between the means of two populations with independent samples, select **STAT >TESTS >2-SampTtest**. If the data are stored in lists, select **Data**, and enter the names of the lists. If, instead, you have summary statistics for the two samples, select **Stats**, and enter the mean, standard deviation, and sample size for each sample. Choose the form of the alternative hypothesis. If you are assuming that the standard deviations are equal for the two populations, select **Yes** for **Pooled**; otherwise, select **No**. Select **Calculate** to find the  $p$ -value. (See Screens 10.1 and 10.2.)
2. To perform a hypothesis test about the proportions of two populations using independent samples, select **STAT >TESTS >2-PropZTest**. Enter the successes and trials (as  $x$  and  $n$  respectively) for each of the two samples. Select the alternative hypothesis, and then **Calculate** to find the  $p$ -value of the test. Be careful to distinguish between the  $p$ -value and the sample proportions, which have hats above them.
3. To find a confidence interval for the difference of the means of two populations using independent samples, select **STAT >TESTS >2-SampTInt**. If the data are stored in lists, select **Data**, and enter the names of the lists. If, instead, you have summary statistics for the two samples, select **Stats**, and enter the mean, standard deviation, and sample size for each sample. Enter the confidence level in decimal form as the **C-Level**. If you are assuming that the standard deviations are equal for the two populations, select **Yes** for **Pooled**; otherwise, select **No**. Select **Calculate** to find the confidence interval.

4. To find a confidence interval for the difference between two population proportions, select **STATS >TESTS >2-PropZInt**. Enter the successes and trials (as  $x$  and  $n$ , respectively) for each of the two samples. Enter the confidence level in decimal form, and then select **Calculate** to find the confidence interval.

## Minitab

1. To find a confidence interval for  $\mu_1 - \mu_2$  for two populations (using two independent samples) with unknown but equal standard deviations as discussed in Section 10.2, select **Stat >Basic Statistics >2-Sample t**. In the dialog box you obtain, select **Summarized data**, and enter the values of the **Sample sizes**, **Means**, and **Standard deviations** for the two samples. Check the box next to **Assume equal variances**. Click the **Options** button, and enter the value of the **Confidence level** in the new dialog box. Click **OK** in both boxes. The output containing the confidence interval will appear in the session window.

If instead of summary measures you have data from two samples, you can enter those data in two different formats. Format 1 involves entering each sample of data in a separate column, such as columns **C1** and **C2**. Format 2 involves entering all of the data in a single column and the corresponding group numbers or labels in a second column. Both formats are shown in **Screen 10.3**, with Format 2 shown in columns **C3** and **C4**. In the dialog box, click next to **Samples in different columns** if the two samples are in separate columns, and click next to **Samples in one column** if the data are in one column and the group numbers or labels are in a second column. (See **Screen 10.4**.) The rest of the procedure is the same as above. (See **Screen 10.5** for the output.)

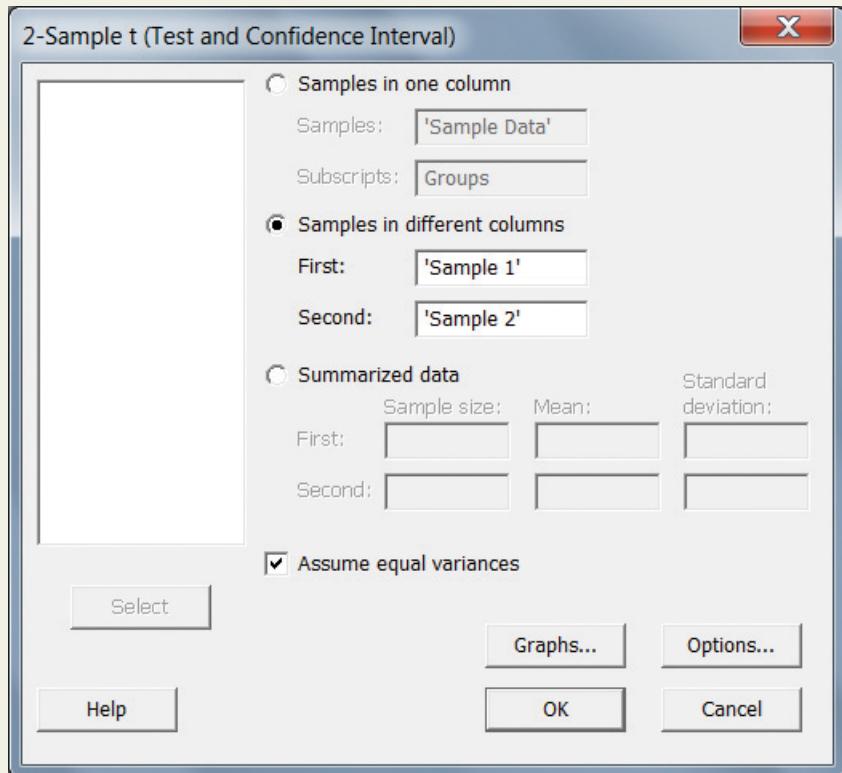
2. To perform a hypothesis test about  $\mu_1 - \mu_2$  for two populations (using two independent samples) with unknown but equal standard deviations as discussed in Section 10.2, select

**Stat >Basic Statistics >2-Sample t**. In the dialog box you obtain, select **Summarized data**, and enter the values of the **Sample sizes**, **Means**, and **Standard deviations** for the two samples. Check the box next to **Assume equal variances**. Click the **Options** button. In the new dialog box you obtain, enter **0** for the **Test difference**, and select the appropriate **Alternative** hypothesis. Click **OK** in both boxes. The output containing the *p*-value will appear in the Session window.

If instead of summary measures you have data from two samples, you can enter those data in two different formats. Format 1 involves entering each sample of data in a separate column, such as columns **C1** and **C2**. Format 2 involves entering all of the data in a single column and the corresponding group numbers or labels in a second column. Both formats are shown in **Screen 10.3**, with Format 2 shown in columns **C3** and **C4**. In the dialog box, click next to **Samples in different columns** if the two samples are in separate columns, and click next to **Samples in one column** if the data are in one column and the group numbers or labels are in a second column. (See **Screen 10.4**.) The rest of the procedure is the same as above.

3. To find a confidence interval for  $\mu_1 - \mu_2$  or to perform a hypothesis test about  $\mu_1 - \mu_2$  for two populations (using two independent samples) with unknown and unequal standard deviations discussed in Section 10.3, the procedures are the same as in steps 1 and 2 above, respectively, except that you do not check next to **Assume equal variances**.

↓	C1	C2	C3	C4
	Sample 1	Sample 2	Sample Data	Groups
1	30.10	28.87	30.10	1
2	28.74	21.06	28.74	1
3	32.46	29.39	32.46	1
4	28.21	26.60	28.21	1
5	32.78	25.33	32.78	1
6	28.66	26.88	28.66	1
7	31.76	24.88	31.76	1
8	29.54	25.96	29.54	1
9	34.49	22.72	34.49	1
10	32.79	21.70	32.79	1
11	26.95	25.14	26.95	1
12	33.67	29.46	33.67	1
13	25.13		25.13	1
14			28.87	2
15			21.06	2
16			29.39	2
17			26.60	2
18			25.33	2



Screen 10.4

```

Session
Results for: Worksheet 2

Two-Sample T-Test and CI: Sample 1, Sample 2

Two-sample T for Sample 1 vs Sample 2

      N    Mean   StDev   SE Mean
Sample 1 13  30.41    2.83    0.78
Sample 2 12  25.67    2.82    0.81

Difference = mu (Sample 1) - mu (Sample 2)
Estimate for difference: 4.74
95% CI for difference: (2.40, 7.08)
T-Test of difference = 0 (vs not =): T-Value = 4.19  P-Value = 0.000  DF = 23
Both use Pooled StDev = 2.8251

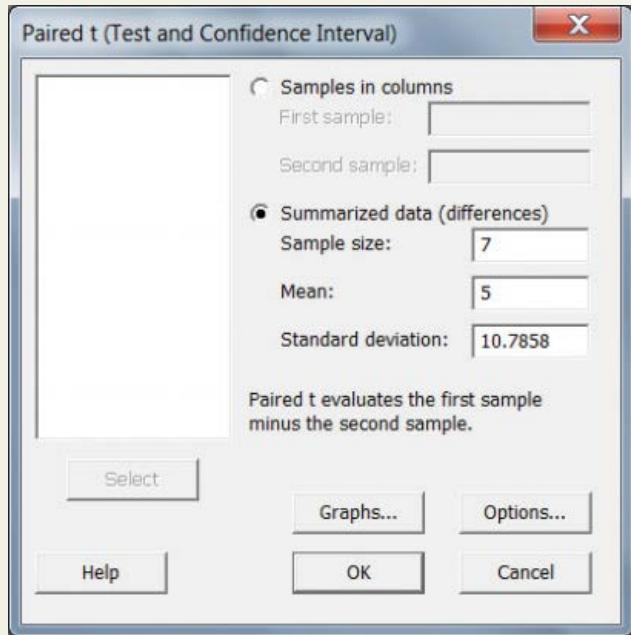
```

Screen 10.5

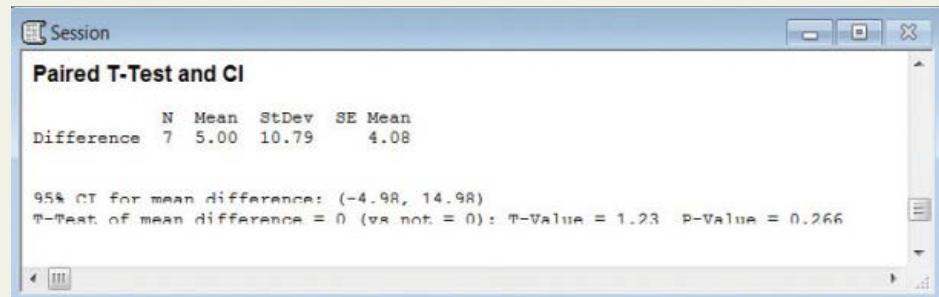
4. To find a confidence interval for  $\mu_d$  for paired data discussed in Section 10.4, enter the *Before* and *After* data into columns C1 and C2, respectively. Select **Stat >Basic Statistics >Paired t**. In the dialog box you obtain, select **Samples in columns**, and enter the column names C1 and C2 in the boxes next to **First sample** and **Second sample**. Click the **Options** button, and enter the value of the **Confidence level** in the new dialog box. Click **OK** in both boxes. The output containing the confidence interval will appear in the session window. Note that the confidence interval here is for the mean of the differences given by C1 – C2, which represents Before – After.

5. To perform a hypothesis test about  $\mu_d$  for paired data discussed in Section 10.4, enter the *Before* and *After* data into columns C1 and C2, respectively. Select **Stat >Basic Statistics >Paired t**. In the dialog box you obtain, select **Samples in columns**, and enter the column names C1 and C2 in the boxes next to **First sample** and **Second sample**. Click the **Options** button. In the new dialog box you obtain, enter 0 for the **Test mean**, and select the appropriate **Alternative hypothesis**. Click **OK** in both boxes. The output containing the *p*-value will appear in the session window. Note that the hypothesis test here is for the mean of the differences given by C1 – C2, which represents Before – After. You need to keep this in mind when determining your alternative hypothesis. (See Screens 10.6 and 10.7.)

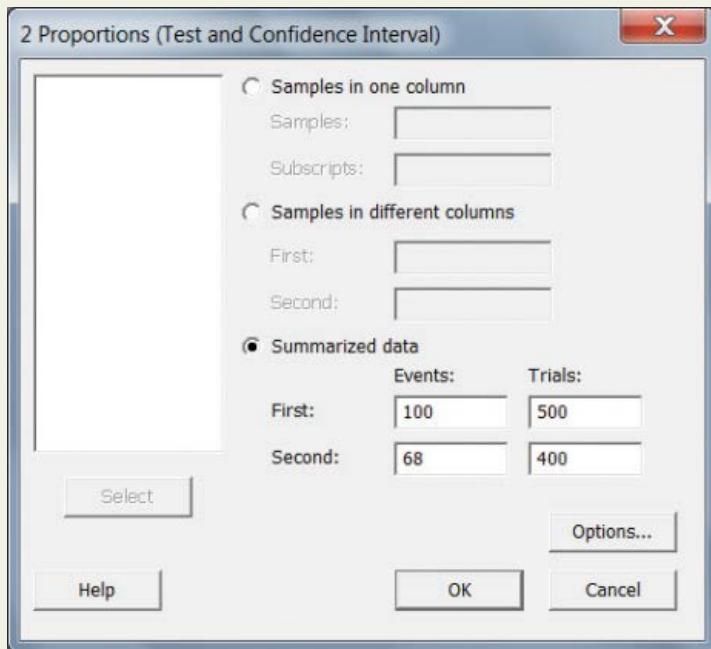
6. To find a confidence interval for  $p_1 - p_2$  using two large and independent samples as discussed in Section 10.5, select **Stat >Basic Statistics >2 Proportions**. In the dialog box you obtain, click on **Summarized data**, and enter the sample sizes and the numbers of successes in the boxes below **Trials** and **Events**, respectively, for the two samples. Click the **Options** button, and enter the value of the **Confidence Level** in the new dialog box. Click **OK** in both dialog boxes. The output containing the confidence interval for  $p_1 - p_2$  will appear in the session window.
7. To perform a hypothesis test about  $p_1 - p_2$  using two large and independent samples as discussed in Section 10.5, select **Stat >Basic Statistics >2 Proportions**. In the dialog box you obtain, select **Summarized data**, and then enter the sample sizes and the



Screen 10.6

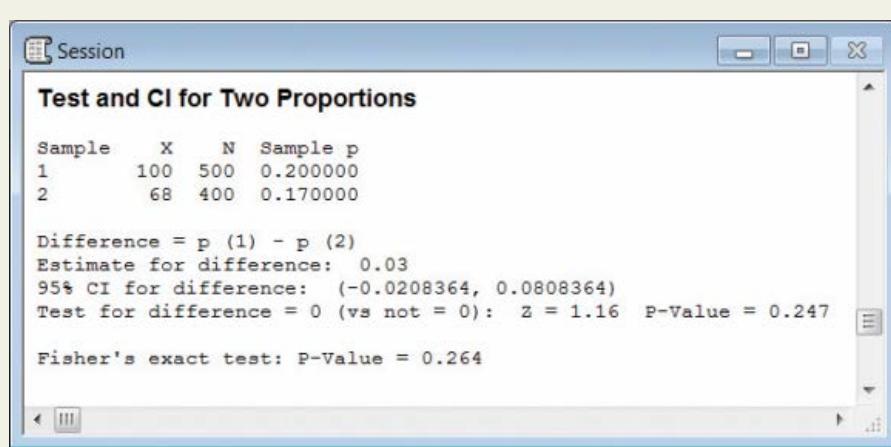


Screen 10.7



Screen 10.8

numbers of successes in the boxes below **Trials** and **Events**, respectively, for the two samples. Click the **Options** button. Set **Test difference** to **0**, select the appropriate **Alternative hypothesis**, and check next to **Use pooled estimate of p for test** in the new dialog box. Click **OK** in both dialog boxes. The output containing the p-value for the test will appear in the session window. (See Screens 10.8 and 10.9.)



Screen 10.9

**Excel**

The Data Analysis ToolPak contains preprogrammed functions for performing the following tests

- The paired *t*-test
- The two-independent-sample *t*-test for means, assuming equal variances
- The two-independent-sample *t*-test for means, assuming unequal variances

The dialog boxes for all three tests are set up in exactly the same fashion. Hence, no matter which test you are using, the processes of entering the data ranges, the hypothesized difference, and so on, are the same for all three tests. Although there is no restriction on the location of the data in the spreadsheet, the instructions will be provided assuming that the data are in adjacent columns.

1. Click the **Data** tab. Click the **Data Analysis** button within the **Analysis** group. From the **Data Analysis** window that appears, select the appropriate test from the list:
  - **t-test: Paired Two Sample for Means**
  - **t-test: Two-Sample Assuming Equal Variances**
  - **t-test: Two-Sample Assuming Unequal Variances**
2. Enter the location of first set of paired data in the **Variable 1 Range** box. Enter the location of the second set of paired data in the **Variable 2 Range** box. Excel will always create differences in the order “variable 1 – variable 2.” Enter the value for the hypothesized difference from the null hypothesis in the **Hypothesized Mean Difference** box. Enter the significance level, as a decimal, in the **Alpha** box. If your columns of data have labels in the top row, click the **Labels** box. Choose how you wish for the output to appear. (See Screen 10.10.) Click **OK**.

t-Test: Paired Two Sample for Means

**Input**

Variable 1 Range: \$A\$1:\$A\$7

Variable 2 Range: \$B\$1:\$B\$7

Hypothesized Mean Difference: 0

Labels

Alpha: 0.05

**Output options**

Output Range: [empty box]  
 New Worksheet Ply: [empty box]  
 New Workbook

Screen 10.10

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		With	Without
4	Mean	25.11667	23.4
5	Variance	34.50567	29.4
6	Observations	6	6
7	Pearson Correlation	0.971219	
8	Hypothesized Mean Difference	0	
9	df	5	
10	t Stat	2.945744	
11	P(T<=t) one-tail	0.016021	
12	t Critical one-tail	2.015048	
13	P(T<=t) two-tail	0.032043	
14	t Critical two-tail	2.570582	

Screen 10.11

3. The two lines in the output that you will need to determine the *p*-value are the lines labeled **t Stat** and **P(T<=t) two-tail**. (See Screen 10.11.) If the alternative hypothesis is two-tailed, the value in the **P(T<=t) two-tail** box is the *p*-value for the test. If the alternative hypothesis is one-tailed, use the following set of rules:

- a. If the hypothesis test is left-tailed and the value of **t Stat** is negative OR the hypothesis test is right-tailed and the value of **t Stat** is positive, the *p*-value of the test is equal to one-half the value in the **P(T<=t) two-tail** box.
- b. If the hypothesis test is left-tailed and the value of **t Stat** is positive OR the hypothesis test is right-tailed and the value of **t Stat** is negative, the *p*-value of the test is equal to 1 minus one-half the value in the **P(T<=t) two-tail** box.

*Note:* Screen 10.11 shows the output for a paired *t*-test. The output windows for the independent samples tests are very similar. More important, the instructions given in step 3 hold for all three types of tests.

## TECHNOLOGY ASSIGNMENTS

**TA10.1** Refer to Data Set IV, Population Data on the 2011 Beach to Beacon 10K Road Race. Select random samples of 50 runners from Maine and Away each. Assume that the population standard deviations of the race times are equal for runners from Maine and runners from Away.

- a. Construct a 99% confidence interval for the difference in the average race times for all Maine runners and for all runners from Away.
- b. Test at a 1% significance level whether the average race time for all Maine runners is higher than the average race time for all runners from Away.

**TA10.2** A company recently opened two supermarkets in two different areas. The management wants to know if the mean sales per day for these two supermarkets are different. A sample of 10 days for Supermarket A produced the following data on daily sales (in thousand dollars).

47.56	57.66	51.23	58.29	43.71
49.33	52.35	50.13	47.45	53.86

A sample of 12 days for Supermarket B produced the following data on daily sales (in thousand dollars).

56.34	63.55	61.64	63.75	54.78	58.19
55.40	59.44	62.33	67.82	56.65	67.90

Assume that the daily sales of the two supermarkets are both normally distributed with equal but unknown standard deviations.

- a. Construct a 99% confidence interval for the difference between the mean daily sales for these two supermarkets.

- b. Test at a 1% significance level whether the mean daily sales for these two supermarkets are different.

**TA10.3** Refer to Technology Assignment TA 10.1. Now do that assignment without assuming that the population standard deviations are the same.

**TA10.4** Refer to Technology Assignment TA10.2. Now do that assignment assuming the daily sales of the two supermarkets are both normally distributed with unequal and unknown standard deviations.

**TA10.5** Refer to Exercise 10.55. The following data, reproduced from that exercise, represent the times (in seconds) taken by each contestant to eat the first Whoopie Pie and the ninth (last) Whoopie Pie. Thirteen contestants actually finished eating all nine Whoopie Pies.

Contestant	1	2	3	4	5	6	7	8	9	10	11	12	13
First pie	49	59	66	49	63	70	77	59	64	69	60	58	71
Last pie	49	74	92	93	91	73	103	59	85	94	84	87	111

- a. Make a 95% confidence interval for the mean of the population paired differences, where a paired difference is equal to the time needed to eat the ninth pie minus the time needed to eat the first pie.
- b. Using a 10% significance level, can you conclude that it takes at least 15 more seconds, on average, to eat the ninth pie than to eat the first pie?

Assume that the population of paired differences is (approximately) normally distributed.

**TA10.6** A company is considering installing new machines to assemble its products. The company is considering two types of machines, but it will buy only one type. The company selected eight assembly workers and asked them to use these two types of machines to assemble products. The following table gives the time taken (in minutes) to assemble one unit of the product on each type of machine for each of these eight workers.

Machine I	23	26	19	24	27	22	20	18
Machine II	21	24	23	25	24	28	24	23

- a. Construct a 98% confidence interval for the mean  $\mu_d$  of the population paired differences, where a paired difference is equal to the time taken to assemble a unit of the product on Machine I minus the time taken to assemble a unit of the product on Machine II by the same worker.
- b. Test at a 5% significance level whether the mean time taken to assemble a unit of the product is different for the two types of machines.

Assume that the population of paired differences is (approximately) normally distributed.

**TA10.7** A company has two restaurants in two different areas of New York City. The company wants to estimate the percentages of patrons who think that the food and service at each of these restaurants are excellent. A sample of 200 patrons taken from the restaurant in Area A showed that 118 of them think that the food and service are excellent at this restaurant. Another sample of 250 patrons selected from the restaurant in Area B showed that 160 of them think that the food and service are excellent at this restaurant.

- a. Construct a 97% confidence interval for the difference between the two population proportions.
- b. Testing at a 2.5% significance level, can you conclude that the proportion of patrons at the restaurant in Area A who think that the food and service are excellent is lower than the corresponding proportion at the restaurant in Area B?

**TA10.8** The management of a supermarket wanted to investigate whether the percentages of all men and all women who prefer to buy national brand products over the store brand products are different. A sample of 600 men shoppers at the company's supermarkets showed that 246 of them prefer to buy national brand products over the store brand products. Another sample of 700 women shoppers at the company's supermarkets showed that 266 of them prefer to buy national brand products over the store brand products.

- a. Construct a 99% confidence interval for the difference between the proportions of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products.
- b. Testing at a 2% significance level, can you conclude that the proportions of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products are different?



## Chi-Square Tests

Are you a fan of people who work on Wall Street? Do you think that people who work on Wall Street are as honest and moral as the general public? In a Harris poll conducted in 2012, 28% of the U.S. adults polled agreed with the statement, "In general, people on Wall Street are as honest and moral as other people." Sixty-eight percent of the adults polled disagreed with this statement. (See Case Study 11-1.)

The tests of hypothesis about the mean, the difference between two means, the proportion, and the difference between two proportions were discussed in Chapters 9 and 10. The tests about proportions dealt with countable or categorical data. In the case of a proportion and the difference between two proportions in Chapters 9 and 10, the tests concerned experiments with only two categories. Recall from Chapter 5 that such experiments are called binomial experiments. This chapter describes three types of tests:

1. Tests of hypothesis for experiments with more than two categories, called goodness-of-fit tests
2. Tests of hypothesis about contingency tables, called independence and homogeneity tests
3. Tests of hypothesis about the variance and standard deviation of a single population

All of these tests are performed by using the **chi-square distribution**, which is sometimes written as  $\chi^2$  distribution and is read as "chi-square distribution." The symbol  $\chi$  is the Greek letter *chi*, pronounced "ki." The values of a chi-square distribution are denoted by the symbol  $\chi^2$  (read as "chi-square"), just as the values of the standard normal distribution and the *t* distribution are denoted by *z* and *t*, respectively. Section 11.1 describes the chi-square distribution.

### 11.1 The Chi-Square Distribution

### 11.2 A Goodness-of-Fit Test

#### Case Study 11-1 Are People On Wall Street Honest And Moral?

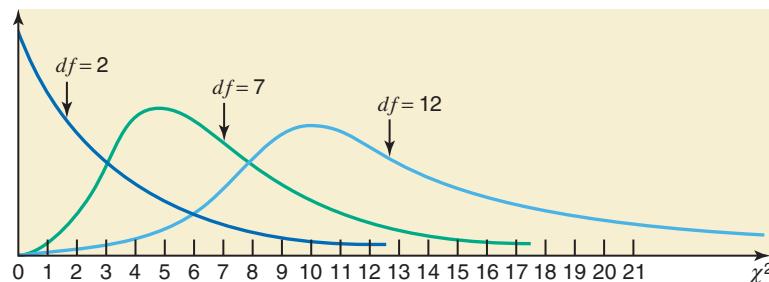
### 11.3 A Test of Independence or Homogeneity

### 11.4 Inferences About the Population Variance

## 11.1 The Chi-Square Distribution

Like the  $t$  distribution, the chi-square distribution has only one parameter, called the degrees of freedom ( $df$ ). The shape of a specific chi-square distribution depends on the number of degrees of freedom.<sup>1</sup> (The degrees of freedom for a chi-square distribution are calculated by using different formulas for different tests. This will be explained when we discuss those tests.) The random variable  $\chi^2$  assumes nonnegative values only. Hence, a chi-square distribution curve starts at the origin (zero point) and lies entirely to the right of the vertical axis. Figure 11.1 shows three chi-square distribution curves. They are for 2, 7, and 12 degrees of freedom, respectively.

**Figure 11.1** Three chi-square distribution curves.



As we can see from Figure 11.1, the shape of a chi-square distribution curve is skewed for very small degrees of freedom, and it changes drastically as the degrees of freedom increase. Eventually, for large degrees of freedom, the chi-square distribution curve looks like a normal distribution curve. The peak (or mode) of a chi-square distribution curve with 1 or 2 degrees of freedom occurs at zero and for a curve with 3 or more degrees of freedom at  $df - 2$ . For instance, the peak of the chi-square distribution curve with  $df = 2$  in Figure 11.1 occurs at zero. The peak for the curve with  $df = 7$  occurs at  $7 - 2 = 5$ . Finally, the peak for the curve with  $df = 12$  occurs at  $12 - 2 = 10$ . Like all other continuous distribution curves, the total area under a chi-square distribution curve is 1.0.

### Definition

**The Chi-Square Distribution** The *chi-square distribution* has only one parameter, called the degrees of freedom. The shape of a chi-square distribution curve is skewed to the right for small  $df$  and becomes symmetric for large  $df$ . The entire chi-square distribution curve lies to the right of the vertical axis. The chi-square distribution assumes nonnegative values only, and these are denoted by the symbol  $\chi^2$  (read as “chi-square”).

If we know the degrees of freedom and the area in the right tail of a chi-square distribution curve, we can find the value of  $\chi^2$  from Table VI of Appendix C. Examples 11–1 and 11–2 show how to read that table.

### ■ EXAMPLE 11–1

Reading the chi-square distribution table: area in the right tail known.

Find the value of  $\chi^2$  for 7 degrees of freedom and an area of .10 in the right tail of the chi-square distribution curve.

**Solution** To find the required value of  $\chi^2$ , we locate 7 in the column for  $df$  and .100 in the top row in Table VI of Appendix C. The required  $\chi^2$  value is given by the entry at the

<sup>1</sup>The mean of a chi-square distribution is equal to its  $df$ , and the standard deviation is equal to  $\sqrt{2 df}$ .

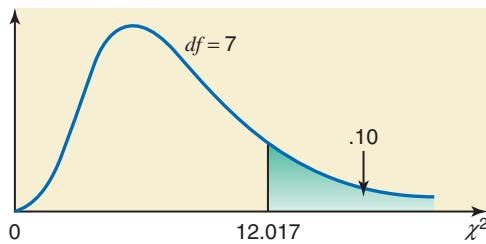
intersection of the row for 7 and the column for .100. This value is 12.017. The relevant portion of Table VI is presented as Table 11.1 here.

**Table 11.1**  $\chi^2$  for  $df = 7$  and .10 Area in the Right Tail

$df$	Area in the Right Tail Under the Chi-Square Distribution Curve				
	.995	...	.100	...	.005
1	0.000	...	2.706	...	7.879
2	0.010	...	4.605	...	10.597
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
7	0.989	...	12.017	...	20.278
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
100	67.328	...	118.498	...	140.169

Required value of  $\chi^2$

As shown in Figure 11.2, for  $df = 7$  and an area of .10 in the right tail of the chi-square distribution curve, the  $\chi^2$  value is **12.017**.



**Figure 11.2** The  $\chi^2$  value.

## ■ EXAMPLE 11-2

Find the value of  $\chi^2$  for 12 degrees of freedom and an area of .05 in the left tail of the chi-square distribution curve.

**Solution** We can read Table VI of Appendix C only when an area in the right tail of the chi-square distribution curve is known. When the given area is in the left tail, as in this example, the first step is to find the area in the right tail of the chi-square distribution curve as follows.

$$\text{Area in the right tail} = 1 - \text{Area in the left tail}$$

Therefore, for our example,

$$\text{Area in the right tail} = 1 - .05 = .95$$

Next, we locate 12 in the column for  $df$  and .950 in the top row in Table VI of Appendix C. The required value of  $\chi^2$ , given by the entry at the intersection of the row for 12 and the column for .950, is 5.226. The relevant portion of Table VI is presented as Table 11.2 here.

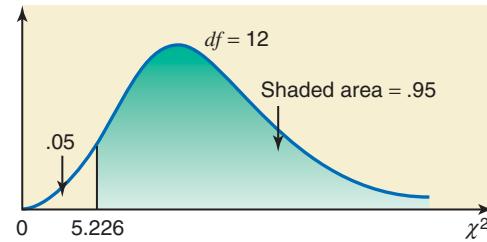
Reading the chi-square distribution table: area in the left tail known.

**Table 11.2**  $\chi^2$  for  $df = 12$  and .95 Area in the Right Tail

df	Area in the Right Tail Under the Chi-Square Distribution Curve				
	.995	...	.950	...	.005
1	0.000	...	0.004	...	7.879
2	0.010	...	0.103	...	10.597
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
12	3.074	...	5.226	...	28.300
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
100	67.328	...	77.929	...	140.169

Required value of  $\chi^2$ 

As shown in Figure 11.3, for  $df = 12$  and .05 area in the left tail, the  $\chi^2$  value is **5.226**.

**Figure 11.3** The  $\chi^2$  value.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 11.1 Describe the chi-square distribution. What is the parameter (parameters) of such a distribution?
- 11.2 Find the value of  $\chi^2$  for 12 degrees of freedom and an area of .025 in the right tail of the chi-square distribution curve.
- 11.3 Find the value of  $\chi^2$  for 28 degrees of freedom and an area of .05 in the right tail of the chi-square distribution curve.
- 11.4 Determine the value of  $\chi^2$  for 14 degrees of freedom and an area of .10 in the left tail of the chi-square distribution curve.
- 11.5 Determine the value of  $\chi^2$  for 23 degrees of freedom and an area of .990 in the left tail of the chi-square distribution curve.
- 11.6 Find the value of  $\chi^2$  for 4 degrees of freedom and
  - a. .005 area in the right tail of the chi-square distribution curve
  - b. .05 area in the left tail of the chi-square distribution curve
- 11.7 Determine the value of  $\chi^2$  for 13 degrees of freedom and
  - a. .025 area in the left tail of the chi-square distribution curve
  - b. .995 area in the right tail of the chi-square distribution curve

## 11.2 A Goodness-of-Fit Test

This section explains how to make tests of hypothesis about experiments with more than two possible outcomes (or categories). Such experiments, called **multinomial experiments**, possess four characteristics. Note that a binomial experiment is a special case of a multinomial experiment.

### Definition

**A Multinomial Experiment** An experiment with the following characteristics is called a *multinomial experiment*.

1. It consists of  $n$  identical trials (repetitions).
2. Each trial results in one of  $k$  possible outcomes (or categories), where  $k > 2$ .
3. The trials are independent.
4. The probabilities of the various outcomes remain constant for each trial.

An experiment of many rolls of a die is an example of a multinomial experiment. It consists of many identical rolls (trials); each roll (trial) results in one of the six possible outcomes; each roll is independent of the other rolls; and the probabilities of the six outcomes remain constant for each roll.

As a second example of a multinomial experiment, suppose we select a random sample of people and ask them whether or not the quality of American cars is better than that of Japanese cars. The response of a person can be *yes*, *no*, or *does not know*. Each person included in the sample can be considered as one trial (repetition) of the experiment. There will be as many trials for this experiment as the number of persons selected. Each person can belong to any of the three categories—*yes*, *no*, or *does not know*. The response of each selected person is independent of the responses of other persons. Given that the population is large, the probabilities of a person belonging to the three categories remain the same for each trial. Consequently, this is an example of a multinomial experiment.

The frequencies obtained from the actual performance of an experiment are called the **observed frequencies**. In a **goodness-of-fit test**, we test the null hypothesis that the observed frequencies for an experiment follow a certain pattern or theoretical distribution. The test is called a goodness-of-fit test because the hypothesis tested is how *good* the observed frequencies *fit* a given pattern.

For our first example involving the experiment of many rolls of a die, we may test the null hypothesis that the given die is fair. The die will be fair if the observed frequency for each outcome is close to one-sixth of the total number of rolls.

For our second example involving opinions of people on the quality of American cars, suppose such a survey was conducted in 2012, and in that survey 41% of the people said *yes*, 48% said *no*, and 11% said *do not know*. We want to test if these percentages still hold true. Suppose we take a random sample of 1000 adults and observe that 536 of them think that the quality of American cars is better than that of Japanese cars, 362 say it is worse, and 102 have no opinion. The frequencies 536, 362, and 102 are the observed frequencies. These frequencies are obtained by actually performing the survey. Now, assuming that the 2012 percentages are still true (which will be our null hypothesis), in a sample of 1000 adults we will expect 410 to say *yes*, 480 to say *no*, and 110 to say *do not know*. These frequencies are obtained by multiplying the sample size (1000) by the 2012 proportions. These frequencies are called the **expected frequencies**. Then, we will make a decision to reject or not to reject the null hypothesis based on how large the difference between the observed frequencies and the expected frequencies is. To perform this test, we will use the chi-square distribution. Note that in this case we are testing the null hypothesis that all three percentages (or proportions) are unchanged.

However, if we want to make a test for only one of the three proportions, we use the procedure learned in Section 9.4 of Chapter 9. For example, if we are testing the hypothesis that the current percentage of people who think the quality of American cars is better than that of the Japanese cars is different from 41%, then we will test the null hypothesis  $H_0: p = .41$  against the alternative hypothesis  $H_1: p \neq .41$ . This test will be conducted using the procedure discussed in Section 9.4 of Chapter 9.

As mentioned earlier, the frequencies obtained from the performance of an experiment are called the observed frequencies. They are denoted by  $O$ . To make a goodness-of-fit test, we calculate the expected frequencies for all categories of the experiment. The expected frequency for a category, denoted by  $E$ , is given by the product of  $n$  and  $p$ , where  $n$  is the total number of trials and  $p$  is the probability for that category.

### Definition

**Observed and Expected Frequencies** The frequencies obtained from the performance of an experiment are called the *observed frequencies* and are denoted by  $O$ . The *expected frequencies*, denoted by  $E$ , are the frequencies that we expect to obtain if the null hypothesis is true. The expected frequency for a category is obtained as

$$E = np$$

where  $n$  is the sample size and  $p$  is the probability that an element belongs to that category if the null hypothesis is true.

**Degrees of Freedom for a Goodness-of-Fit Test** In a goodness-of-fit test, the *degrees of freedom* are

$$df = k - 1$$

where  $k$  denotes the number of possible outcomes (or categories) for the experiment.

The procedure to make a goodness-of-fit test involves the same five steps that we used in the preceding chapters. *The chi-square goodness-of-fit test is always a right-tailed test.*

**Test Statistic for a Goodness-of-Fit Test** The *test statistic for a goodness-of-fit test* is  $\chi^2$ , and its value is calculated as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where

$O$  = observed frequency for a category

$E$  = expected frequency for a category =  $np$

Remember that a chi-square goodness-of-fit test is always a right-tailed test.

Whether or not the null hypothesis is rejected depends on how much the observed and expected frequencies differ from each other. To find how large the difference between the observed frequencies and the expected frequencies is, we do not look at just  $\Sigma(O - E)$ , because some of the  $O - E$  values will be positive and others will be negative. The net result of the sum of these differences will always be zero. Therefore, we square each of the  $O - E$  values to obtain  $(O - E)^2$ , and then we weight them according to the reciprocals of their expected frequencies. The sum of the resulting numbers gives the computed value of the test statistic  $\chi^2$ .

To make a goodness-of-fit test, the sample size should be large enough so that the expected frequency for each category is at least 5. If there is a category with an expected frequency of less than 5, either increase the sample size or combine two or more categories to make each expected frequency at least 5.

Examples 11–3 and 11–4 describe the procedure for performing goodness-of-fit tests using the chi-square distribution.

### ■ EXAMPLE 11–3

A bank has an ATM installed inside the bank, and it is available to its customers only from 7 AM to 6 PM Monday through Friday. The manager of the bank wanted to investigate if the number of transactions made on this ATM are the same for each of the 5 days (Monday through Friday) of the week. She randomly selected one week and counted the number of transactions made on this ATM on each of the 5 days during this week. The information she obtained is given in the following table, where the number of users represents the number of transactions on this ATM on these days. For convenience, we will refer to these transactions as “people” or “users.”

*Conducting a goodness-of-fit test: equal proportions for all categories.*

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Number of users	253	197	204	279	267

At a 1% level of significance, can we reject the null hypothesis that the number of people who use this ATM each of the 5 days of the week is the same? Assume that this week is typical of all weeks in regard to the use of this ATM.

**Solution** To conduct this test of hypothesis, we proceed as follows.

**Step 1.** *State the null and alternative hypotheses.*

Because there are 5 categories (days) as listed in the table, the number of ATM users will be the same for each of these 5 days if 20% of all users use the ATM each day. The null and alternative hypotheses are as follows.

$H_0$ : The number of people using the ATM is the same for all 5 days of the week.

$H_1$ : The number of people using the ATM is not the same for all 5 days of the week.

If the number of people using this ATM is the same for all 5 days of the week, then .20 of the users will use this ATM on any of the 5 days of the week. Let  $p_1, p_2, p_3, p_4$ , and  $p_5$  be the proportions of people who use this ATM on Monday, Tuesday, Wednesday, Thursday, and Friday, respectively. Then, the null and alternative hypotheses can also be written as

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = .20$$

$$H_1: \text{At least two of the five proportions are not equal to } .20$$

**Step 2.** *Select the distribution to use.*

Because there are 5 categories (i.e., 5 days on which the ATM is used), this is a multinomial experiment. Consequently, we use the chi-square distribution to make this test.

**Step 3.** *Determine the rejection and nonrejection regions.*

The significance level is given to be .01, and the goodness-of-fit test is always right-tailed. Therefore, the area in the right tail of the chi-square distribution curve is

$$\text{Area in the right tail} = \alpha = .01$$

The degrees of freedom are calculated as follows:

$$k = \text{number of categories} = 5$$

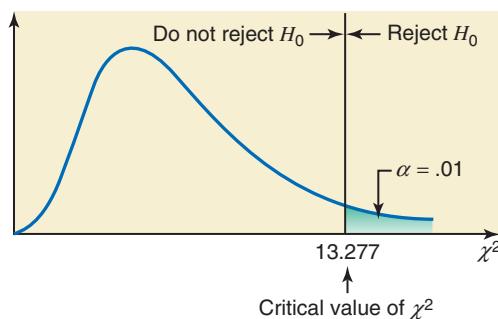
$$df = k - 1 = 5 - 1 = 4$$



Photodisc/Getty Images, Inc.

From the chi-square distribution table (Table VI of Appendix C), for  $df = 4$  and .01 area in the right tail of the chi-square distribution curve, the critical value of  $\chi^2$  is 13.277, as shown in Figure 11.4.

**Figure 11.4** Rejection and nonrejection regions.



**Step 4.** Calculate the value of the test statistic.

**Table 11.3** Calculating the Value of the Test Statistic

Category (Day)	Observed Frequency <i>O</i>	Expected Frequency <i>E</i> = <i>np</i>	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
Monday	253	.20	1200(.20) = 240	13	169
Tuesday	197	.20	1200(.20) = 240	-43	1849
Wednesday	204	.20	1200(.20) = 240	-36	1296
Thursday	279	.20	1200(.20) = 240	39	1521
Friday	267	.20	1200(.20) = 240	27	729
$n = 1200$					Sum = 23.184

All the required calculations to find the value of the test statistic  $\chi^2$  are shown in Table 11.3. The calculations made in Table 11.3 are explained next.

1. The first two columns of Table 11.3 list the 5 categories (days) and the observed frequencies for the 1200 persons who used the ATM during each of the 5 days of the selected week. The third column contains the probabilities for the 5 categories assuming that the null hypothesis is true.
2. The fourth column contains the expected frequencies. These frequencies are obtained by multiplying the total users ( $n = 1200$ ) by the probabilities listed in the third column. If the null hypothesis is true (i.e., the ATM users are equally distributed over all 5 days), then we will expect 240 out of 1200 persons to use the ATM each day. Consequently, each category in the fourth column has the same expected frequency.
3. The fifth column lists the differences between the observed and expected frequencies, that is,  $O - E$ . These values are squared and recorded in the sixth column.
4. Finally, we divide the squared differences (that appear in the sixth column) by the corresponding expected frequencies (listed in the fourth column) and write the resulting numbers in the seventh column.

5. The sum of the seventh column gives the value of the test statistic  $\chi^2$ . Thus,

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 23.184$$

**Step 5.** *Make a decision.*

The value of the test statistic  $\chi^2 = 23.184$  is larger than the critical value of  $\chi^2 = 13.277$ , and it falls in the rejection region. Hence, we reject the null hypothesis and state that the number of persons who use this ATM is not the same for the 5 days of the week. In other words, we conclude that a higher number of users of this ATM use this machine on one or more of these days.

If you make this chi-square test using any of the statistical software packages, you will obtain a  $p$ -value for the test. In this case you can compare the  $p$ -value obtained in the computer output with the level of significance and make a decision. As you know from Chapter 9, you will reject the null hypothesis if  $\alpha$  (significance level) is greater than or equal to the  $p$ -value and not reject it otherwise. ■

### ■ EXAMPLE 11-4

In a 2011 *Time/Money Magazine* survey, Americans age 18 years and older were asked if “we are less sure that our children will achieve the American Dream.” Of the respondents, 65% said yes, 29% said no, and 6% said that they did not know (*Time*, October 10, 2011). Assume that these percentages hold true for the 2011 population of Americans age 18 years and older. Recently 1000 randomly selected Americans age 18 years and older were asked the same question. The following table lists the number of Americans in this sample who made the respective response.

*Conducting a goodness-of-fit test: testing if results of a survey fit a given distribution.*

Response	Yes	No	Do Not Know
Frequency	624	306	70

Test at a 2.5% level of significance whether the current distribution of opinions is different from that for 2011.

**Solution** We perform the following five steps for this test of hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

The null and alternative hypotheses are

$H_0$ : The current percentage distribution of opinions is the same as for 2011.

$H_1$ : The current percentage distribution of opinions is different from that for 2011.

**Step 2.** *Select the distribution to use.*

Because this experiment has three categories as listed in the table, it is a multinomial experiment. Consequently we use the chi-square distribution to make this test.

**Step 3.** *Determine the rejection and nonrejection regions.*

The significance level is given to be .025, and because the goodness-of-fit test is always right-tailed, the area in the right tail of the chi-square distribution curve is

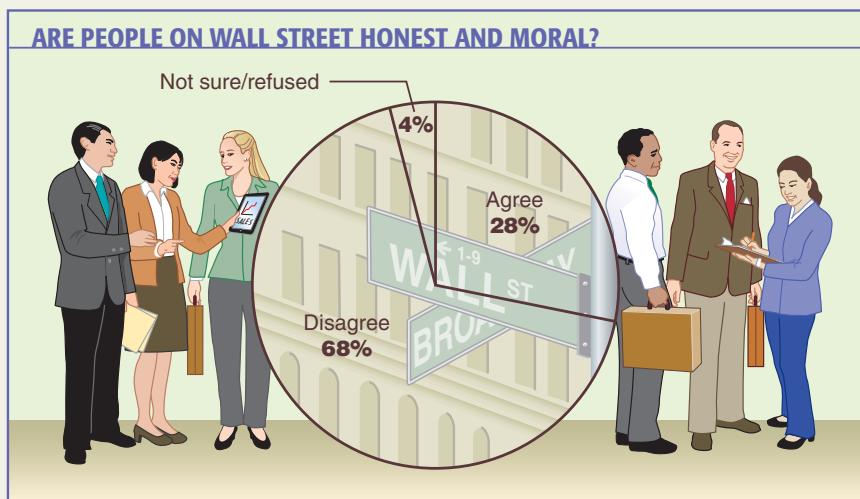
$$\text{Area in the right tail} = \alpha = .025$$

The degrees of freedom are calculated as follows:

$$k = \text{number of categories} = 3$$

$$df = k - 1 = 3 - 1 = 2$$

## ARE PEOPLE ON WALL STREET HONEST AND MORAL?



Data source: Harris Interactive telephone poll of U.S. adults conducted April 10-17, 2012.

In a Harris poll conducted by Harris Interactive between April 10 and April 17, 2012, U.S. adults aged 18 years and older were asked whether they agreed with the statement, "In general, people on Wall Street are as honest and moral as other people." (<http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/ctl/ReadCustom%20Default/mid/1508/ArticleId/1018/Default.aspx>.) The accompanying chart shows the percentage distribution of the responses of these adults. Twenty-eight percent of the adults polled said that they agree with this statement, 68% disagreed, and 4% were not sure or refused to answer. Assume that these percentages were true for the population of U.S. adults in 2012. Suppose that we want to test the hypothesis whether these percentages with respect to the foregoing statement are still true. Then the two hypotheses are as follows:

$$H_0: \text{The current percentage distribution of opinions is the same as in 2012}$$

$$H_1: \text{The current percentage distribution of opinions is not the same as in 2012}$$

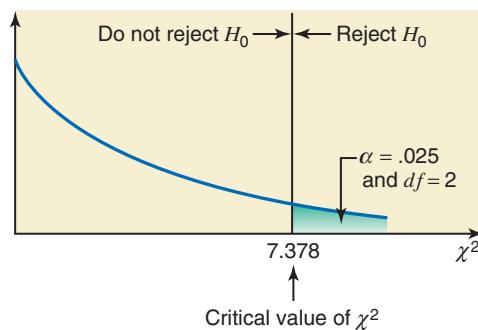
To test this hypothesis, suppose that we currently take a sample of 2000 U.S. adults and ask them whether they agree with the foregoing statement. Suppose that 488 of them say that they agree with the statement, 1444 say that they disagree, and 68 are not sure or refuse to give an answer. Using the given information, we calculate the value of the test statistic as shown in the following table:

Category	Observed Frequency <i>O</i>	Expected Frequency <i>E</i> = <i>np</i>	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
Agree	488	.28	2000(.28) = 560	-72	5184
Disagree	1444	.68	2000(.68) = 1360	84	7056
Not sure/refuse	68	.04	2000(.04) = 80	-12	144
<i>n</i> = 2000					Sum = 16.245

Suppose that we use a 1% significance level to perform this test. Then for  $df = 3 - 1 = 2$  and .01 area in the right tail, the critical value of  $\chi^2$  is 9.210 from Table VI of Appendix C. Since the observed value of  $\chi^2$  is 16.245 and it is larger than the critical value of  $\chi^2 = 9.210$ , we reject the null hypothesis. Thus, we conclude that the current percentage distribution of opinions of U.S. adults in response to the given statement is significantly different from the distribution of those in 2012.

We can also use the *p*-value approach to make this decision. In Table VI of Appendix C, for  $df = 2$ , the largest value of  $\chi^2$  is 10.597, and the area to the right of  $\chi^2 = 10.597$  is .005. Thus, the *p*-value for  $\chi^2 = 16.245$  will be less than .005. (By using technology, we obtain the *p*-value of .0003.) Since  $\alpha = .01$  in this example is greater than .005 (or .0003), we reject the null hypothesis and conclude that the current percentage distribution of opinions of U.S. adults in response to the given statement is significantly different from the distribution of those in 2012.

From the chi-square distribution table (Table VI of Appendix C), for  $df = 2$  and .025 area in the right tail of the chi-square distribution curve, the critical value of  $\chi^2$  is 7.378, as shown in Figure 11.5.



**Figure 11.5** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

All the required calculations to find the value of the test statistic  $\chi^2$  are shown in Table 11.4. Note that the three percentages for 2011 have been converted into probabilities and recorded in the third column of Table 11.4. The value of the test statistic  $\chi^2$  is given by the sum of the last column. Thus,

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 3.590$$

**Table 11.4** Calculating the Value of the Test Statistic

Category (Response)	Observed Frequency <i>O</i>	Expected Frequency <i>E</i> = <i>np</i>	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$	
Yes	624	.65	1000(.65) = 650	-26	676	1.040
No	306	.29	1000(.29) = 290	16	256	.883
Do not know	70	.06	1000(.06) = 60	10	100	1.667
<i>n</i> = 1000					Sum = 3.590	

**Step 5.** Make a decision.

The observed value of the test statistic  $\chi^2 = 3.590$  is smaller than the critical value of  $\chi^2 = 7.378$ , and it falls in the nonrejection region. Hence, we fail to reject the null hypothesis, and state that the current percentage distribution of opinions is the same as for 2011.

If you make this chi-square test using any of the statistical software packages, you will obtain a *p*-value for the test. In this case you can compare the *p*-value obtained in the computer output with the level of significance and make a decision. As you know from Chapter 9, you will reject the null hypothesis if  $\alpha$  (significance level) is greater than or equal to the *p*-value and not reject it otherwise. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

- 11.8 Describe the four characteristics of a multinomial experiment.
- 11.9 What is a goodness-of-fit test and when is it applied? Explain.
- 11.10 Explain the difference between the observed and expected frequencies for a goodness-of-fit test.

**11.11** How is the expected frequency of a category calculated for a goodness-of-fit test? What are the degrees of freedom for such a test?

**11.12** To make a goodness-of-fit test, what should be the minimum expected frequency for each category? What are the alternatives if this condition is not satisfied?

**11.13** The following table lists the frequency distribution for 60 rolls of a die.

Outcome	1-spot	2-spot	3-spot	4-spot	5-spot	6-spot
Frequency	7	12	8	15	11	7

Test at a 5% significance level whether the null hypothesis that the given die is fair is true.

## ■ APPLICATIONS

**11.14** In March 2012, the Gallup-Healthways Well-Being Index (<http://www.gallup.com/poll/153251/No-Major-Change-Americans-Exercise-Habits-2011.aspx>) reported on exercise habits of Americans. Specifically, they reported that during 2011, 51.6% of Americans exercised for 30 minutes or more on 3 or more days per week (A), 18.8% exercised for 30 minutes or more on 1 or 2 days per week (B), and 29.7% did not exercise for 30 minutes or more on at least 1 day per week (C). Assume that the Gallup-Healthways results were true for all Americans in 2011. Suppose a recent random sample of 520 Americans produced the frequencies given in the following table.

Exercise category	A	B	C
Number of people	141	131	248

Test at a 5% significance level whether the current distribution of exercise frequency differs from that of 2011.

**11.15** The October 2011 ISACA Shopping on the Job Survey asked employees, “During the holiday season (November and December), how much total time do you think an average employee at your enterprise spends shopping online using a work-supplied computer or smartphone?” Among those who responded, 3% said 0 hours, 24% said 1 to 2 hours, 22% said 3 to 5 hours, and 51% said 6 or more hours ([www.isaca.org/SiteCollectionDocuments/2011-ISACA-Shopping-on-the-Job-Survey-US.pdf](http://www.isaca.org/SiteCollectionDocuments/2011-ISACA-Shopping-on-the-Job-Survey-US.pdf)). Suppose that another poll conducted recently asked the same question of 215 randomly selected business executives, which produced the frequencies listed in the following table.

Response/category	0 hours	1–2 hours	3–5 hours	6 or more hours
Frequency	2	41	55	117

Test at a 2.5% significance level whether the distribution of responses for the executive survey differs from that of October 2011 survey of employees.

**11.16** Clasp your hands together. Which thumb is on top? Believe it or not, the thumb that you place on top is determined by genetics. If either of your parents has the gene that *tells* you to place your left thumb on top and passes it on to you, you will place your left thumb on top. The *left-thumb* gene is called the dominant gene, which means that if either parent passes it on to you, you will have that trait. If you place your right thumb on top, you received the recessive gene from both parents. If both parents have both the left and right thumb genes (the case denoted Lr), Mendelian genetics gives the probabilities listed in the following table about the children’s genes.

Child's genes	Lr (Left-Thumbed, but Also Received a Right-Thumbed Gene)		
	LL (Left-Thumbed)	rr (Right-Thumbed)	
Probability	.25	.50	.25

Source: [http://humangenetics.suite101.com/article.cfm/dominant\\_human\\_genetic\\_traits](http://humangenetics.suite101.com/article.cfm/dominant_human_genetic_traits).

Suppose that a random sample of 65 children whose both parents had Lr genes were tested for the genes. The following table lists the results of this experiment.

Child's genes	LL	Lr	rr
Frequency	14	31	20

Test at a 5% significance level whether the genes received by the sample of children are significantly different from what Mendelian genetics predicts.

- 11.17** A drug company is interested in investigating whether the color of their packaging has any impact on sales. To test this, they used five different colors (blue, green, orange, red, and yellow) for the boxes of an over-the-counter pain reliever, instead of their traditional white box. The following table shows the number of boxes of each color sold during the first month.

Box color	Blue	Green	Orange	Red	Yellow
Number of boxes sold	310	292	280	216	296

Using a 1% significance level, test the null hypothesis that the number of boxes sold of each of these five colors is the same.

- 11.18** Over the last 3 years, Art's Supermarket has observed the following distribution of modes of payment in the express lines: cash (C) 41%, check (CK) 24%, credit or debit card (D) 26%, and other (N) 9%. In an effort to make express checkout more efficient, Art's has just begun offering a 1% discount for cash payment in the express checkout line. The following table lists the frequency distribution of the modes of payment for a sample of 500 express-line customers after the discount went into effect.

Mode of payment	C	CK	D	N
Number of customers	240	104	111	45

Test at a 1% significance level whether the distribution of modes of payment in the express checkout line changed after the discount went into effect.

- 11.19** Home Mail Corporation sells products by mail. The company's management wants to find out if the number of orders received at the company's office on each of the 5 days of the week is the same. The company took a sample of 400 orders received during a 4-week period. The following table lists the frequency distribution for these orders by the day of the week.

Day of the week	Mon	Tue	Wed	Thu	Fri
Number of orders received	92	71	65	83	89

Test at a 5% significance level whether the null hypothesis that the orders are evenly distributed over all days of the week is true.

- 11.20** Of all students enrolled at a large undergraduate university, 19% are seniors, 23% are juniors, 27% are sophomores, and 31% are freshmen. A sample of 200 students taken from this university by the student senate to conduct a survey includes 50 seniors, 46 juniors, 55 sophomores, and 49 freshmen. Using a 2.5% significance level, test the null hypothesis that this sample is a random sample. (*Hint:* This sample will be a random sample if it includes approximately 19% seniors, 23% juniors, 27% sophomores, and 31% freshmen.)

- 11.21** Chance Corporation produces beauty products. Two years ago the quality control department at the company conducted a survey of users of one of the company's products. The survey revealed that 53% of the users said the product was excellent, 31% said it was satisfactory, 7% said it was unsatisfactory, and 9% had no opinion. Assume that these percentages were true for the population of all users of this product at that time. After this survey was conducted, the company redesigned this product. A recent survey of 800 users of the redesigned product conducted by the quality control department at the company

showed that 495 of the users think the product is excellent, 255 think it is satisfactory, 35 think it is unsatisfactory, and 15 have no opinion. Is the percentage distribution of the opinions of users of the redesigned product different from the percentage distribution of users of this product before it was redesigned? Use  $\alpha = .025$ .

**11.22** Henderson Corporation makes metal sheets, among other products. When the process that is used to make metal sheets works properly, 92% of the metal sheets contain no defects, 5% have one defect each, and 3% have two or more defects each. The quality control inspectors at the company take samples of metal sheets quite often and check them for defects. If the distribution of defects for a sample is significantly different from the above-mentioned percentage distribution, the process is stopped and adjusted. A recent sample of 300 sheets produced the frequency distribution of defects listed in the following table.

Number of defects	None	One	Two or More
Number of metal sheets	262	24	14

Does the evidence from this sample suggest that the process needs an adjustment? Use  $\alpha = .01$ .

## 11.3 A Test of Independence or Homogeneity

This section is concerned with tests of independence and homogeneity, which are performed using contingency tables. Except for a few modifications, the procedure used to make such tests is almost the same as the one applied in Section 11.2 for a goodness-of-fit test.

### 11.3.1 A Contingency Table

Often we may have information on more than one variable for each element. Such information can be summarized and presented using a two-way classification table, which is also called a *contingency table* or *cross-tabulation*. Suppose a university has a total of 20,758 students enrolled. By classifying these students based on gender and whether these students are full-time or part-time, we can prepare Table 11.5, which provides an example of a contingency table. Table 11.5 has two rows (one for males and the second for females) and two columns (one for full-time and the second for part-time students). Hence, it is also called a  $2 \times 2$  (read as “two by two”) contingency table.

**Table 11.5 Total Enrollment at a University**

	Full-Time	Part-Time	
Male	6768	2615	Students who are male and enrolled part-time
Female	7658	3717	

A contingency table can be of any size. For example, it can be  $2 \times 3$ ,  $3 \times 2$ ,  $3 \times 3$ , or  $4 \times 2$ . Note that in these notations, the first digit refers to the number of rows in the table, and the second digit refers to the number of columns. For example, a  $3 \times 2$  table will contain three rows and two columns. In general, an  $R \times C$  table contains  $R$  rows and  $C$  columns.

Each of the four boxes that contain numbers in Table 11.5 is called a *cell*. The number of cells in a contingency table is obtained by multiplying the number of rows by the number of columns. Thus, Table 11.5 contains  $2 \times 2 = 4$  cells. The subjects that belong to a cell of a contingency table possess two characteristics. For example, 2615 students listed in the second cell of the first row in Table 11.5 are *male* and *part-time*. The numbers written inside the cells are usually called the *joint frequencies*. For example, 2615 students belong to the joint category of *male* and *part-time*. Hence, it is referred to as the joint frequency of this category.

### 11.3.2 A Test of Independence

In a **test of independence** for a contingency table, we test the null hypothesis that the two attributes (characteristics) of the elements of a given population are not related (that is, they are independent) against the alternative hypothesis that the two characteristics are related (that is, they are dependent). For example, we may want to test if the affiliation of people with the Democratic and Republican parties is independent of their income levels. We perform such a test by using the chi-square distribution. As another example, we may want to test if there is an association between being a man or a woman and having a preference for watching sports or soap operas on television.

#### Definition

**Degrees of Freedom for a Test of Independence** A test of independence involves a test of the null hypothesis that two attributes of a population are not related. The *degrees of freedom for a test of independence* are

$$df = (R - 1)(C - 1)$$

where  $R$  and  $C$  are the number of rows and the number of columns, respectively, in the given contingency table.

The value of the test statistic  $\chi^2$  in a test of independence is obtained using the same formula as in the goodness-of-fit test described in Section 11.2.

**Test Statistic for a Test of Independence** The value of the *test statistic  $\chi^2$  for a test of independence* is calculated as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where  $O$  and  $E$  are the observed and expected frequencies, respectively, for a cell.

The null hypothesis in a test of independence is always that the two attributes are not related. The alternative hypothesis is that the two attributes are related.

The frequencies obtained from the performance of an experiment for a contingency table are called the **observed frequencies**. The procedure to calculate the **expected frequencies** for a contingency table for a test of independence is different from the one for a goodness-of-fit test. Example 11–5 describes this procedure.

### ■ EXAMPLE 11–5

Violence and lack of discipline have become major problems in schools in the United States. A random sample of 300 adults was selected, and these adults were asked if they favor giving more freedom to schoolteachers to punish students for violence and lack of discipline. The two-way classification of the responses of these adults is presented in the following table.

*Calculating expected frequencies for a test of independence.*

	In Favor (F)	Against (A)	No Opinion (N)
Men ( $M$ )	93	70	12
Women ( $W$ )	87	32	6

Calculate the expected frequencies for this table, assuming that the two attributes, gender and opinions on the issue, are independent.

**Solution** The preceding table is reproduced as Table 11.6 here. Note that Table 11.6 includes the row and column totals.

**Table 11.6 Observed Frequencies**

	In Favor (F)	Against (A)	No Opinion (N)	Row Totals
Men (M)	93	70	12	175
Women (W)	87	32	6	125
Column Totals	180	102	18	300

The numbers 93, 70, 12, 87, 32, and 6 listed inside the six cells of Table 11.6 are called the *observed frequencies* of the respective cells.

As mentioned earlier, the null hypothesis in a test of independence is that the two attributes (or classifications) are independent. In an independence test of hypothesis, first we assume that the null hypothesis is true and that the two attributes are independent. Assuming that the null hypothesis is true and that gender and opinions are not related in this example, we calculate the expected frequency for the cell corresponding to *Men* and *In Favor* as shown next. From Table 11.6,

$$P(\text{a person is a } \textit{Man}) = P(M) = 175/300$$

$$P(\text{a person is } \textit{In Favor}) = P(F) = 180/300$$

Because we are assuming that *M* and *F* are independent (by assuming that the null hypothesis is true), from the formula learned in Chapter 4, the joint probability of these two events is

$$P(M \text{ and } F) = P(M) \times P(F) = (175/300) \times (180/300)$$

Then, assuming that *M* and *F* are independent, the number of persons expected to be *Men* and *In Favor* in a sample of 300 is

$$\begin{aligned} E \text{ for Men and In Favor} &= 300 \times P(M \text{ and } F) \\ &= 300 \times \frac{175}{300} \times \frac{180}{300} = \frac{175 \times 180}{300} \\ &= \frac{(\text{Row total})(\text{Column total})}{\text{Sample size}} \end{aligned}$$

Thus, the rule for obtaining the expected frequency for a cell is to divide the product of the corresponding row and column totals by the sample size.

**Expected Frequencies for a Test of Independence** The expected frequency *E* for a cell is calculated as

$$E = \frac{(\text{Row total})(\text{Column total})}{\text{Sample size}}$$

Using this rule, we calculate the expected frequencies of the six cells of Table 11.6 as follows:

$$E \text{ for Men and In Favor cell} = (175)(180)/300 = \mathbf{105.00}$$

$$E \text{ for Men and Against cell} = (175)(102)/300 = \mathbf{59.50}$$

$$E \text{ for Men and No Opinion cell} = (175)(18)/300 = \mathbf{10.50}$$

$$E \text{ for Women and In Favor cell} = (125)(180)/300 = 75.00$$

$$E \text{ for Women and Against cell} = (125)(102)/300 = 42.50$$

$$E \text{ for Women and No Opinion cell} = (125)(18)/300 = 7.50$$

The expected frequencies are usually written in parentheses below the observed frequencies within the corresponding cells, as shown in Table 11.7.

**Table 11.7** Observed and Expected Frequencies

	In Favor (F)	Against (A)	No Opinion (N)	Row Totals
Men (M)	93 (105.00)	70 (59.50)	12 (10.50)	175
Women (W)	87 (75.00)	32 (42.50)	6 (7.50)	125
Column Totals	180	102	18	300

Like a goodness-of-fit test, a *test of independence is always right-tailed*. To apply a chi-square test of independence, the sample size should be large enough so that the expected frequency for each cell is at least 5. If the expected frequency for a cell is not at least 5, we either increase the sample size or combine some categories. Examples 11–6 and 11–7 describe the procedure to make tests of independence using the chi-square distribution.

## ■ EXAMPLE 11–6

Reconsider the two-way classification table given in Example 11–5. In that example, a random sample of 300 adults was selected, and they were asked if they favor giving more freedom to schoolteachers to punish students for violence and lack of discipline. Based on the results of the survey, a two-way classification table was prepared and presented in Example 11–5. Does the sample provide sufficient evidence to conclude that the two attributes, gender and opinions of adults, are dependent? Use a 1% significance level.

Making a test of independence:  $2 \times 3$  table.

**Solution** The test involves the following five steps.

**Step 1.** State the null and alternative hypotheses.

As mentioned earlier, the null hypothesis must be that the two attributes are independent. Consequently, the alternative hypothesis is that these attributes are dependent.

$H_0$ : Gender and opinions of adults are independent.

$H_1$ : Gender and opinions of adults are dependent.

**Step 2.** Select the distribution to use.

We use the chi-square distribution to make a test of independence for a contingency table.

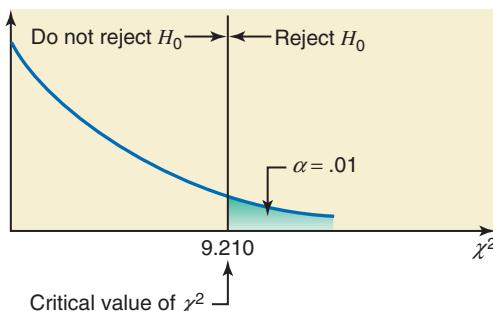
**Step 3.** Determine the rejection and nonrejection regions.

The significance level is 1%. Because a test of independence is always right-tailed, the area of the rejection region is .01, and it falls in the right tail of the chi-square distribution curve. The contingency table contains two rows (*Men* and *Women*) and three columns (*In Favor*, *Against*, and *No Opinion*). Note that we do not count the row and column of totals. The degrees of freedom are

$$df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$$

From Table VI of Appendix C, for  $df = 2$  and  $\alpha = .01$ , the critical value of  $\chi^2$  is 9.210. This value is shown in Figure 11.6.

**Figure 11.6** Rejection and nonrejection regions.



**Step 4.** Calculate the value of the test statistic.

Table 11.7, with the observed and expected frequencies constructed in Example 11–5, is reproduced as Table 11.8.

**Table 11.8** Observed and Expected Frequencies

	In Favor (F)	Against (A)	No Opinion (N)	Row Totals
Men (M)	93 (105.00)	70 (59.50)	12 (10.50)	175
Women (W)	87 (75.00)	32 (42.50)	6 (7.50)	125
Column Totals	180	102	18	300

To compute the value of the test statistic  $\chi^2$ , we take the difference between each pair of observed and expected frequencies listed in Table 11.8, square those differences, and then divide each of the squared differences by the respective expected frequency. The sum of the resulting numbers gives the value of the test statistic  $\chi^2$ . All these calculations are made as follows:

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(93 - 105.00)^2}{105.00} + \frac{(70 - 59.50)^2}{59.50} + \frac{(12 - 10.50)^2}{10.50} \\ &\quad + \frac{(87 - 75.00)^2}{75.00} + \frac{(32 - 42.50)^2}{42.50} + \frac{(6 - 7.50)^2}{7.50} \\ &= 1.371 + 1.853 + .214 + 1.920 + 2.594 + .300 = 8.252\end{aligned}$$

**Step 5.** Make a decision.

The value of the test statistic  $\chi^2 = 8.252$  is less than the critical value of  $\chi^2 = 9.210$ , and it falls in the nonrejection region. Hence, we fail to reject the null hypothesis and state that there is not enough evidence from the sample to conclude that the two characteristics, *gender* and *opinions of adults*, are dependent for this issue. ■

Making a test of independence:  $2 \times 2$  table.

## ■ EXAMPLE 11–7

A researcher wanted to study the relationship between gender and owning cell phones. She took a sample of 2000 adults and obtained the information given in the following table.

	Own Cell Phones	Do Not Own Cell Phones
Men	640	450
Women	440	470

At a 5% level of significance, can you conclude that gender and owning a cell phone are related for all adults?

**Solution** We perform the following five steps to make this test of hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

The null and alternative hypotheses are, respectively,

$H_0$ : Gender and owning a cell phone are not related.

$H_1$ : Gender and owning a cell phone are related.

**Step 2.** *Select the distribution to use.*

Because we are performing a test of independence, we use the chi-square distribution to make the test.

**Step 3.** *Determine the rejection and nonrejection regions.*

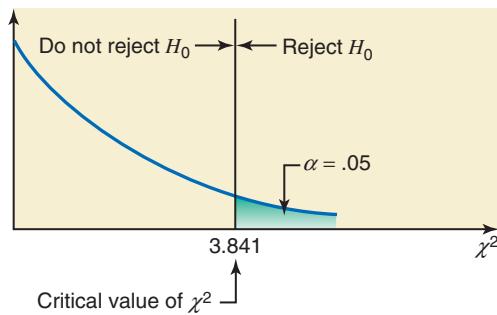
With a significance level of 5%, the area of the rejection region is .05, and it falls under the right tail of the chi-square distribution curve. The contingency table contains two rows (*men* and *women*) and two columns (*own cell phones* and *do not own cell phones*). The degrees of freedom are

$$df = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$$

From Table VI of Appendix C, the critical value of  $\chi^2$  for  $df = 1$  and  $\alpha = .05$  is 3.841. This value is shown in Figure 11.7.



© THEGIFT777/iStockphoto



**Figure 11.7** Rejection and nonrejection regions.

**Step 4.** *Calculate the value of the test statistic.*

The expected frequencies for the various cells are calculated as follows, and are listed within parentheses in Table 11.9.

**Table 11.9** Observed and Expected Frequencies

	Own Cell Phones (Y)	Do Not Own Cell Phones (N)	Row Totals
Men (M)	640 (588.60)	450 (501.40)	1090
Women (W)	440 (491.40)	470 (418.60)	910
Column Totals	1080	920	2000

$$E \text{ for men and own cell phones cell} = (1090)(1080)/2000 = 588.60$$

$$E \text{ for men and do not own cell phones cell} = (1090)(920)/2000 = 501.40$$

$$E \text{ for women and own cell phones cell} = (910)(1080)/2000 = 491.40$$

$$E \text{ for women and do not own cell phones cell} = (910)(920)/2000 = 418.60$$

The value of the test statistic  $\chi^2$  is calculated as follows:

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(640 - 588.60)^2}{588.60} + \frac{(450 - 501.40)^2}{501.40} + \frac{(440 - 491.40)^2}{491.40} + \frac{(470 - 418.60)^2}{418.60} \\ &= 4.489 + 5.269 + 5.376 + 6.311 = 21.445\end{aligned}$$

#### Step 5. Make a decision.

The value of the test statistic  $\chi^2 = 21.445$  is larger than the critical value of  $\chi^2 = 3.841$ , and it falls in the rejection region. Hence, we reject the null hypothesis and state that there is strong evidence from the sample to conclude that the two characteristics, *gender* and *owning cell phones*, are related for all adults. ■

### 11.3.3 A Test of Homogeneity

In a **test of homogeneity**, we test if two (or more) populations are homogeneous (similar) with regard to the distribution of a certain characteristic. For example, we might be interested in testing the null hypothesis that the proportions of households that belong to different income groups are the same in California and Wisconsin, or we may want to test whether or not the preferences of people in Florida, Arizona, and Vermont are similar with regard to Coke, Pepsi, and 7-Up.

#### Definition

**A Test of Homogeneity** A *test of homogeneity* involves testing the null hypothesis that the proportions of elements with certain characteristics in two or more different populations are the same against the alternative hypothesis that these proportions are not the same.

Let us consider the example of testing the null hypothesis that the proportions of households in California and Wisconsin who belong to various income groups are the same. (Note that in a test of homogeneity, the null hypothesis is always that the proportions of elements with certain characteristics are the same in two or more populations. The alternative hypothesis is that these proportions are not the same.) Suppose we define three income strata: high-income group (with an income of more than \$150,000), medium-income group (with an income of \$60,000 to \$150,000), and low-income group (with an income of less than \$60,000). Furthermore, assume that we take one sample of 250 households from California and another sample of 150 households from Wisconsin, collect the information on the incomes of these households, and prepare the contingency table as in Table 11.10.

**Table 11.10 Observed Frequencies**

	California	Wisconsin	Row Totals
High income	70	34	104
Medium income	80	40	120
Low income	100	76	176
Column Totals	250	150	400

Note that in this example the column totals are fixed. That is, we decided in advance to take samples of 250 households from California and 150 from Wisconsin. However, the row totals (of 104, 120, and 176) are determined randomly by the outcomes of the two samples. If we compare this example to the one about violence and lack of discipline in schools in the previous section, we will notice that neither the column nor the row totals were fixed in that example. Instead, the researcher took just one sample of 300 adults, collected the information on gender and opinions, and prepared the contingency table. Thus, in that example, the row and column totals were all determined randomly. Thus, when both the row and column totals are determined randomly, we perform a test of independence. However, when either the column totals or the row totals are fixed, we perform a test of homogeneity. In the case of income groups in California and Wisconsin, we will perform a test of homogeneity to test for the similarity of income groups in the two states.

The procedure to conduct a test of homogeneity is similar to the procedure used to make a test of independence discussed earlier. Like a test of independence, a test of homogeneity is right-tailed. Example 11–8 illustrates the procedure to make a homogeneity test.

## ■ EXAMPLE 11–8

Consider the data on income distributions for households in California and Wisconsin given in Table 11.10. Using a 2.5% significance level, test the null hypothesis that the distribution of households with regard to income levels is similar (homogeneous) for the two states.

*Performing a test of homogeneity.*

**Solution** We perform the following five steps to make this test of hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

The two hypotheses are, respectively,<sup>2</sup>

$H_0$ : The proportions of households that belong to different income groups are the same in both states.

$H_1$ : The proportions of households that belong to different income groups are not the same in both states.

**Step 2.** *Select the distribution to use.*

We use the chi-square distribution to make a homogeneity test.

**Step 3.** *Determine the rejection and nonrejection regions.*

The significance level is 2.5%. Because the homogeneity test is right-tailed, the area of the rejection region is .025, and it lies in the right tail of the chi-square distribution curve. The contingency table for income groups in California and Wisconsin contains three rows and two columns. Hence, the degrees of freedom are

$$df = (R - 1)(C - 1) = (3 - 1)(2 - 1) = 2$$

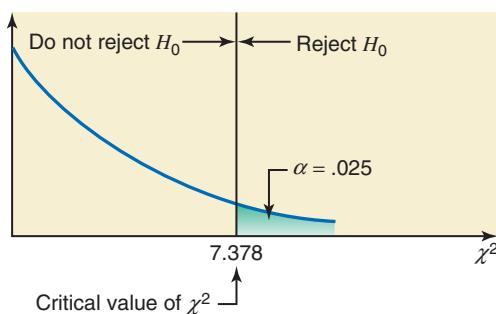
From Table VI of Appendix C, the value of  $\chi^2$  for  $df = 2$  and .025 area in the right tail of the chi-square distribution curve is 7.378. This value is shown in Figure 11.8.

<sup>2</sup>Let  $p_{HC}$ ,  $p_{MC}$ , and  $p_{LC}$  be the proportions of households in California who belong to high-, middle-, and low-income groups, respectively. Let  $p_{HW}$ ,  $p_{MW}$ , and  $p_{LW}$  be the corresponding proportions for Wisconsin. Then we can also write the null hypothesis as

$$H_0: p_{HC} = p_{HW}, p_{MC} = p_{MW}, \text{ and } p_{LC} = p_{LW}$$

and the alternative hypothesis as

$$H_1: \text{At least two of the equalities mentioned in } H_0 \text{ are not true.}$$

**Figure 11.8** Rejection and nonrejection regions.**Step 4.** Calculate the value of the test statistic.

To compute the value of the test statistic  $\chi^2$ , we need to calculate the expected frequencies first. Table 11.11 lists the observed and the expected frequencies. The numbers in parentheses in this table are the expected frequencies, which are calculated using the formula

$$E = \frac{(\text{Row total})(\text{Column total})}{\text{Total of both samples}}$$

Thus, for instance,

$$E \text{ for } \text{High income} \text{ and } \text{California cell} = \frac{(104)(250)}{400} = 65$$

**Table 11.11** Observed and Expected Frequencies

	California	Wisconsin	Row Totals
<b>High income</b>	70 (65)	34 (39)	104
<b>Medium income</b>	80 (75)	40 (45)	120
<b>Low income</b>	100 (110)	76 (66)	176
<b>Column Totals</b>	250	150	400

The remaining expected frequencies are calculated in the same way. Note that the expected frequencies in a test of homogeneity are calculated in the same way as in a test of independence. The value of the test statistic  $\chi^2$  is computed as follows:

$$\begin{aligned} \chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(70 - 65)^2}{65} + \frac{(34 - 39)^2}{39} + \frac{(80 - 75)^2}{75} + \frac{(40 - 45)^2}{45} \\ &\quad + \frac{(100 - 110)^2}{110} + \frac{(76 - 66)^2}{66} \\ &= .385 + .641 + .333 + .556 + .909 + 1.515 = \mathbf{4.339} \end{aligned}$$

**Step 5.** Make a decision.

The value of the test statistic  $\chi^2 = 4.339$  is less than the critical value of  $\chi^2 = 7.378$ , and it falls in the nonrejection region. Hence, we fail to reject the null hypothesis and state that the distribution of households with regard to income appears to be similar (homogeneous) in California and Wisconsin. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**11.23** Describe in your own words a test of independence and a test of homogeneity. Give one example of each.

**11.24** Explain how the expected frequencies for cells of a contingency table are calculated in a test of independence or homogeneity. How do you find the degrees of freedom for such tests?

**11.25** To make a test of independence or homogeneity, what should be the minimum expected frequency for each cell? What are the alternatives if this condition is not satisfied?

**11.26** Consider the following contingency table, which is based on a sample survey.

	Column 1	Column 2	Column 3
Row 1	137	64	105
Row 2	98	71	65
Row 3	115	81	115

- Write the null and alternative hypotheses for a test of independence for this table.
- Calculate the expected frequencies for all cells, assuming that the null hypothesis is true.
- For  $\alpha = .01$ , find the critical value of  $\chi^2$ . Show the rejection and nonrejection regions on the chi-square distribution curve.
- Find the value of the test statistic  $\chi^2$ .
- Using  $\alpha = .01$ , would you reject the null hypothesis?

**11.27** Consider the following contingency table, which records the results obtained for four samples of fixed sizes selected from four populations.

	Sample Selected From			
	Population 1	Population 2	Population 3	Population 4
Row 1	24	81	60	121
Row 2	46	64	91	72
Row 3	20	37	105	93

- Write the null and alternative hypotheses for a test of homogeneity for this table.
- Calculate the expected frequencies for all cells assuming that the null hypothesis is true.
- For  $\alpha = .025$ , find the critical value of  $\chi^2$ . Show the rejection and nonrejection regions on the chi-square distribution curve.
- Find the value of the test statistic  $\chi^2$ .
- Using  $\alpha = .025$ , would you reject the null hypothesis?

### APPLICATIONS

**11.28** During the recent economic recession, many families faced hard times financially. Some studies observed that more people stopped buying name brand products and started buying less expensive store brand products instead. Data produced by a recent sample of 700 adults on whether they usually buy store brand or name brand products are recorded in the following table.

	More Often Buy	
	Name Brand	Store Brand
Men	150	165
Women	160	225

Using a 1% significance level, can you reject the null hypothesis that the two attributes, gender and buying name or store brand products, are independent?

**11.29** One hundred auto drivers who were stopped by police for some violations were also checked to see if they were wearing seat belts. The following table records the results of this survey.

	Wearing Seat Belt	Not Wearing Seat Belt
Men	40	15
Women	38	7

Test at a 2.5% significance level whether being a man or a woman and wearing or not wearing a seat belt are related.

**11.30** Many students graduate from college deeply in debt from student loans, credit card debts, and so on. A sociologist took a random sample of 401 single persons, classified them by gender, and asked, "Would you consider marrying someone who was \$25,000 or more in debt?" The results of this survey are shown in the following table.

	Yes	No	Uncertain
Women	125	59	21
Men	101	79	16

Test at a 1% significance level whether gender and response are related.

**11.31** During the Bush and Obama administrations, there has been a great deal of discussion about partisanship. Did partisanship have an impact on approval ratings in public polls? The following tables display the approval ratings of both presidents during November of their third year in office (<http://www.gallup.com/poll/124922/Presidential-Approval-Center.aspx?ref=interactive>). The first table shows the approval numbers for each president from voters of his own party (Democrats for Obama, Republicans for Bush). The second table shows the approval numbers for each president from voters of the opposition party (Republicans for Obama, Democrats for Bush). The numbers are comparable to the percentages reported by [www.gallup.com](http://www.gallup.com).

President's Own Party		President's Opposition Party	
Obama (April 2011)	Bush (April 2003)	Obama (April 2011)	Bush (April 2003)
Approve	693	775	99
Not sure/ disapprove	207	125	171

Obama (April 2011)	Bush (April 2003)	Obama (April 2011)	Bush (April 2003)
Approve	99	801	729
Not sure/ disapprove	171	729	99

- a. Test at a 1% significance level whether the approval ratings by their own party voters are related.
- b. Test at a 1% significance level whether the approval ratings by the opposition party voters are related.

**11.32** The game show *Deal or No Deal* involves a series of opportunities for the contestant to either accept an amount of money from the show's *banker* or to decline it and open a specific number of briefcases in the hope of exposing and, thereby eliminating, low amounts of money from the game, which would lead the banker to increase the amount of the next offer. Suppose that 700 people aged 21 years and older were selected at random. Each of them watched an episode of the show until exactly four briefcases were left unopened. The money amounts in these four briefcases were \$750, \$5000, \$50,000, and \$400,000, respectively. The banker's offer to the contestant was \$81,600 if the contestant would stop the game and accept the offer. If the contestant were to decline the offer, he or she would choose one briefcase out of these four to open, and then there would be a new offer. All 700 persons were asked whether they would accept the offer (Deal) for \$81,600 or turn it down (No Deal), as well as their ages. The responses of these 700 persons are listed in the following table.

	Age Group (years)				
	21–29	30–39	40–49	50–59	60 and Over
Deal	78	82	89	92	63
No Deal	56	70	60	63	47

Test at a 5% significance level whether the decision to accept or not to accept the offer (Deal or No Deal) and age group are dependent.

- 11.33** A forestry official is comparing the causes of forest fires in two regions, A and B. The following table shows the causes of fire for 76 randomly selected recent fires in these two regions.

	Arson	Accident	Lightning	Unknown
Region A	6	9	6	10
Region B	7	14	15	9

Test at a 5% significance level whether causes of fire and regions of fires are related.

- 11.34** National Electronics Company buys parts from two subsidiaries. The quality control department at this company wanted to check if the distribution of good and defective parts is the same for the supplies of parts received from both subsidiaries. The quality control inspector selected a sample of 300 parts received from Subsidiary A and a sample of 400 parts received from Subsidiary B. These parts were checked for being good or defective. The following table records the results of this investigation.

	Subsidiary A	Subsidiary B
Good	284	381
Defective	16	19

Using a 5% significance level, test the null hypothesis that the distributions of good and defective parts are the same for both subsidiaries.

- 11.35** Two drugs were administered to two groups of randomly assigned 60 and 40 patients, respectively, to cure the same disease. The following table gives information about the number of patients who were cured and not cured by each of the two drugs.

	Cured	Not Cured
Drug I	44	16
Drug II	18	22

Test at a 1% significance level whether or not the two drugs are similar in curing and not curing the patients.

- 11.36** Four hundred people were selected from each of the four geographic regions (Midwest, Northeast, South, West) of the United States, and they were asked which form of camping they prefer. The choices were pop-up camper/trailer, family style (tenting with sanitary facilities), rustic (tenting, no sanitary facilities), or none. The results of the survey are shown in the following table.

	Midwest	Northeast	South	West
Camper/trailer	132	129	129	135
Family style	180	175	168	146
Rustic	46	50	59	68
None	42	46	44	51

Based on the evidence from these samples, can you conclude that the distributions of favorite forms of camping are different for at least two of the regions? Use  $\alpha = .01$ .

- 11.37** A December 2011 FOX News poll asked, “Which of the following comes closest to your view about what government policy should be toward illegal immigrants currently in the United States?” The three options were (A) Send all illegal immigrants back to their home country, (B) Have a guest worker program that allows immigrants to remain in the United States to work but only for a limited amount of time, and (C) Allow illegal immigrants to remain in the country and eventually qualify for U.S. citizenship, but only if they meet certain requirements, such as paying back taxes, learning English, and passing a background check. The following table lists the party affiliations of the respondents and their responses. The numbers in the table are approximately the same as reported in percentages in the poll.

	A	B	C	Unsure
Democrat	55	43	288	8
Independent	19	25	107	6
Republican	89	52	196	7

Source: [www.foxnews.com/interactive/us/2011/12/09/fox-news-poll-immigration/](http://www.foxnews.com/interactive/us/2011/12/09/fox-news-poll-immigration/).

Test at a 5% significance level whether the distributions of responses are significantly different for at least two of the political affiliations.

**11.38** The following table gives the distributions of grades for three professors for a few randomly selected classes that each of them taught during the last 2 years.

		Professor		
		Miller	Smith	Moore
Grade	A	18	36	20
	B	25	44	15
	C	85	73	82
	D and F	17	12	8

Using a 2.5% significance level, test the null hypothesis that the grade distributions are homogeneous for these three professors.

**11.39** Two random samples, one of 95 blue-collar workers and a second of 50 white-collar workers, were taken from a large company. These workers were asked about their views on a certain company issue. The following table gives the results of the survey.

		Opinion		
		Favor	Oppose	Uncertain
Blue-collar workers		44	39	12
White-collar workers		21	26	3

Using a 2.5% significance level, test the null hypothesis that the distributions of opinions are homogeneous for the two groups of workers.

## 11.4 Inferences About the Population Variance

Earlier chapters explained how to make inferences (confidence intervals and hypothesis tests) about the population mean and population proportion. However, we may often need to control the variance (or standard deviation). Consequently, there may be a need to estimate and to test a hypothesis about the population variance  $\sigma^2$ . Section 11.4.1 describes how to make a confidence interval for the population variance (or standard deviation). Section 11.4.2 explains how to test a hypothesis about the population variance.

As an example, suppose a machine is set up to fill packages of cookies so that the net weight of cookies per package is 32 ounces. Note that the machine will not put exactly 32 ounces of cookies into each package. Some of the packages will contain less and some will contain more than 32 ounces. However, if the variance (and, hence, the standard deviation) is too large, some of the packages will contain quite a bit less than 32 ounces of cookies, and some others will contain quite a bit more than 32 ounces. The manufacturer will not want a large variation in the amounts of cookies put into different packages. To keep this variation within some specified acceptable limit, the machine will be adjusted from time to time. Before the manager decides to adjust the machine at any time, he or she must estimate the variance or test a hypothesis or do both to find out if the variance exceeds the maximum acceptable value.

Like every sample statistic, the sample variance is a random variable, and it possesses a sampling distribution. If all the possible samples of a given size are taken from a population and their variances are calculated, the probability distribution of these variances is called the *sampling distribution of the sample variance*.

**Sampling Distribution of  $(n - 1)s^2/\sigma^2$**  If the population from which the sample is selected is (approximately) normally distributed, then

$$\frac{(n - 1)s^2}{\sigma^2}$$

has a chi-square distribution with  $n - 1$  degrees of freedom. Note that it is not  $s^2$  but the above expression that has a chi-square distribution.

Thus, the chi-square distribution is used to construct a confidence interval and test a hypothesis about the population variance  $\sigma^2$ .

### 11.4.1 Estimation of the Population Variance

The value of the sample variance  $s^2$  gives a point estimate of the population variance  $\sigma^2$ . The  $(1 - \alpha)100\%$  confidence interval for  $\sigma^2$  is given by the following formula.

**Confidence Interval for the Population Variance  $\sigma^2$**  Assuming that the population from which the sample is selected is (approximately) normally distributed, we obtain the  $(1 - \alpha)100\%$  confidence interval for the population variance  $\sigma^2$  as

$$\frac{(n - 1)s^2}{\chi_{\alpha/2}^2} \text{ to } \frac{(n - 1)s^2}{\chi_{1-\alpha/2}^2}$$

where  $\chi_{\alpha/2}^2$  and  $\chi_{1-\alpha/2}^2$  are obtained from the chi-square distribution table for  $\alpha/2$  and  $1 - \alpha/2$  areas in the right tail of the chi-square distribution curve, respectively, and for  $n - 1$  degrees of freedom.

The confidence interval for the population standard deviation can be obtained by simply taking the positive square roots of the two limits of the confidence interval for the population variance.

The procedure for making a confidence interval for  $\sigma^2$  involves the following three steps.

1. Take a sample of size  $n$  and compute  $s^2$  using the formula learned in Chapter 3. However, if  $n$  and  $s^2$  are given, then perform only steps 2 and 3.
2. Calculate  $\alpha/2$  and  $1 - \alpha/2$ . Find two values of  $\chi^2$  from the chi-square distribution table (Table VI of Appendix C): one for  $\alpha/2$  area in the right tail of the chi-square distribution curve and  $df = n - 1$ , and the second for  $1 - \alpha/2$  area in the right tail and  $df = n - 1$ .
3. Substitute all the values in the formula for the confidence interval for  $\sigma^2$  and simplify.

Example 11–9 illustrates the estimation of the population variance and population standard deviation.

#### ■ EXAMPLE 11–9

One type of cookie manufactured by Haddad Food Company is Cocoa Cookies. The machine that fills packages of these cookies is set up in such a way that the average net weight of these packages is 32 ounces with a variance of .015 square ounce. From time to time the quality control inspector at the company selects a sample of a few such packages, calculates the variance of the net weights of these packages, and constructs a 95% confidence interval for the population variance. If either both or one of the two limits of this confidence interval is not in the interval .008 to .030, the machine is stopped and adjusted. A recently taken random sample of 25 packages from the production line gave a sample variance of .029 square ounce. Based on this sample information, do you think the machine needs an adjustment? Assume that the net weights of cookies in all packages are normally distributed.

Constructing confidence intervals for  $\sigma^2$  and  $\sigma$ .

**Solution** The following three steps are performed to estimate the population variance and to make a decision.

**Step 1.** From the given information,  $n = 25$  and  $s^2 = .029$

**Step 2.** The confidence level is  $1 - \alpha = .95$ . Hence,  $\alpha = 1 - .95 = .05$ . Therefore,

$$\alpha/2 = .05/2 = .025$$

$$1 - \alpha/2 = 1 - .025 = .975$$

$$df = n - 1 = 25 - 1 = 24$$

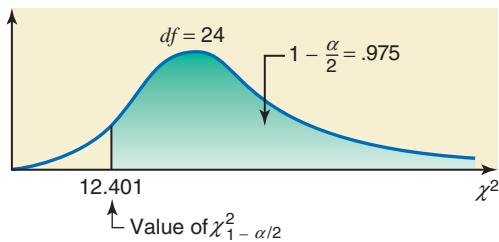
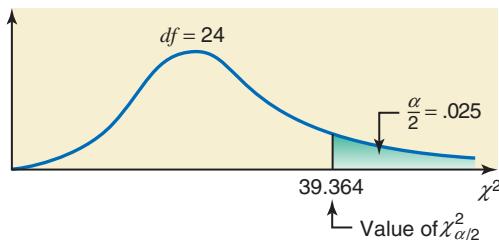
From Table VI of Appendix C,

$$\chi^2 \text{ for } 24 \text{ df and .025 area in the right tail} = 39.364$$

$$\chi^2 \text{ for } 24 \text{ df and .975 area in the right tail} = 12.401$$

These values are shown in Figure 11.9.

**Figure 11.9** The values of  $\chi^2$ .



**Step 3.** The 95% confidence interval for  $\sigma^2$  is

$$\begin{aligned} \frac{(n-1)s^2}{\chi_{\alpha/2}^2} &\text{ to } \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \\ \frac{(25-1)(.029)}{39.364} &\text{ to } \frac{(25-1)(.029)}{12.401} \\ .0177 &\text{ to } .0561 \end{aligned}$$

Thus, with 95% confidence, we can state that the variance for all packages of Cocoa Cookies lies between .0177 and .0561 square ounce. Note that the lower limit (.0177) of this confidence interval is between .008 and .030, but the upper limit (.0561) is larger than .030 and falls outside the interval .008 to .030. Because the upper limit is larger than .030, we can state that the machine needs to be stopped and adjusted.

We can obtain the confidence interval for the population standard deviation  $\sigma$  by taking the positive square roots of the two limits of the above confidence interval for the population variance. Thus, a 95% confidence interval for the population standard deviation is

$$\sqrt{.0177} \text{ to } \sqrt{.0561} \quad \text{or } .133 \text{ to } .237$$

Hence, the standard deviation of all packages of Cocoa Cookies is between .133 and .237 ounce at a 95% confidence level. ■

### 11.4.2 Hypothesis Tests About the Population Variance

A test of hypothesis about the population variance can be one-tailed or two-tailed. To make a test of hypothesis about  $\sigma^2$ , we perform the same five steps we used earlier in hypothesis-testing examples. The procedure to test a hypothesis about  $\sigma^2$  discussed in this section is applied only when the population from which a sample is selected is (approximately) normally distributed.

**Test Statistic for a Test of Hypothesis About  $\sigma^2$**  The value of the *test statistic*  $\chi^2$  is calculated as

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

where  $s^2$  is the sample variance,  $\sigma^2$  is the hypothesized value of the population variance, and  $n - 1$  represents the degrees of freedom. The population from which the sample is selected is assumed to be (approximately) normally distributed.

Examples 11–10 and 11–11 illustrate the procedure for making tests of hypothesis about  $\sigma^2$ .

### ■ EXAMPLE 11–10

One type of cookie manufactured by Haddad Food Company is Cocoa Cookies. The machine that fills packages of these cookies is set up in such a way that the average net weight of these packages is 32 ounces with a variance of .015 square ounce. From time to time the quality control inspector at the company selects a sample of a few such packages, calculates the variance of the net weights of these packages, and makes a test of hypothesis about the population variance. She always uses  $\alpha = .01$ . The acceptable value of the population variance is .015 square ounce or less. If the conclusion from the test of hypothesis is that the population variance is not within the acceptable limit, the machine is stopped and adjusted. A recently taken random sample of 25 packages from the production line gave a sample variance of .029 square ounce. Based on this sample information, do you think the machine needs an adjustment? Assume that the net weights of cookies in all packages are normally distributed.

Performing a right-tailed test of hypothesis about  $\sigma^2$ .

**Solution** From the given information,

$$n = 25, \quad \alpha = .01, \quad \text{and} \quad s^2 = .029$$

The population variance should not exceed .015 square ounce.

**Step 1.** State the null and alternative hypotheses.

We are to test whether or not the population variance is within the acceptable limit. The population variance is within the acceptable limit if it is less than or equal to .015; otherwise, it is not. Thus, the two hypotheses are

$$H_0: \sigma^2 \leq .015 \quad (\text{The population variance is within the acceptable limit.})$$

$$H_1: \sigma^2 > .015 \quad (\text{The population variance exceeds the acceptable limit.})$$

**Step 2.** Select the distribution to use.

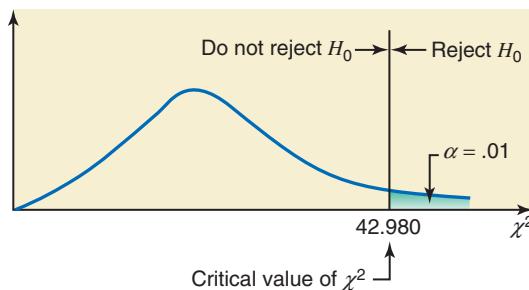
Since the population is normally distributed, we will use the chi-square distribution to test a hypothesis about  $\sigma^2$ .

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is 1% and, because of the  $>$  sign in  $H_1$ , the test is right-tailed. The rejection region lies in the right tail of the chi-square distribution curve with its area equal to .01. The degrees of freedom for a chi-square test about  $\sigma^2$  are  $n - 1$ ; that is,

$$df = n - 1 = 25 - 1 = 24$$

From Table VI of Appendix C, the critical value of  $\chi^2$  for 24 degrees of freedom and .01 area in the right tail is 42.980. This value is shown in Figure 11.10.



**Figure 11.10** Rejection and nonrejection regions.

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $\chi^2$  for the sample variance is calculated as follows:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} = \frac{(25 - 1)(.029)}{.015} = 46.400$$

↑  
From  $H_0$

**Step 5.** Make a decision.

The value of the test statistic  $\chi^2 = 46.400$  is greater than the critical value of  $\chi^2 = 42.980$ , and it falls in the rejection region. Consequently, we reject  $H_0$  and conclude that the population variance is not within the acceptable limit. The machine should be stopped and adjusted. ■

**EXAMPLE 11-11**

*Conducting a two-tailed test of hypothesis about  $\sigma^2$ .*



Corbis Digital Stock/©Corbis

The variance of scores on a standardized mathematics test for all high school seniors was 150 in 2011. A sample of scores for 20 high school seniors who took this test this year gave a variance of 170. Test at the 5% significance level if the variance of current scores of all high school seniors on this test is different from 150. Assume that the scores of all high school seniors on this test are (approximately) normally distributed.

**Solution** From the given information,

$$n = 20, \quad \alpha = .05, \quad \text{and} \quad s^2 = 170$$

The population variance was 150 in 2011.

**Step 1.** State the null and alternative hypotheses.

The null and alternative hypotheses are, respectively,

$$H_0: \sigma^2 = 150 \quad (\text{The population variance is not different from 150.})$$

$$H_1: \sigma^2 \neq 150 \quad (\text{The population variance is different from 150.})$$

**Step 2.** Select the distribution to use.

Since the population of scores is normally distributed, we will use the chi-square distribution to test a hypothesis about  $\sigma^2$ .

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is 5%. The  $\neq$  sign in  $H_1$  indicates that the test is two-tailed. The rejection region lies in both tails of the chi-square distribution curve with its total area equal to .05. Consequently, the area in each tail of the distribution curve is .025. The values of  $\alpha/2$  and  $1 - \alpha/2$  are, respectively,

$$\frac{\alpha}{2} = \frac{.05}{2} = .025 \quad \text{and} \quad 1 - \frac{\alpha}{2} = 1 - .025 = .975$$

The degrees of freedom are

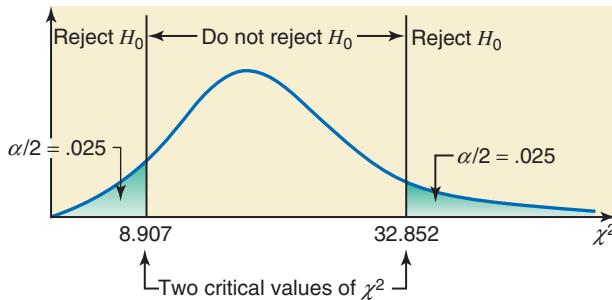
$$df = n - 1 = 20 - 1 = 19$$

From Table VI of Appendix C, the critical values of  $\chi^2$  for 19 degrees of freedom and for  $\alpha/2$  and  $1 - \alpha/2$  areas in the right tail are

$$\chi^2 \text{ for } 19 \text{ df and } .025 \text{ area in the right tail} = 32.852$$

$$\chi^2 \text{ for } 19 \text{ df and } .975 \text{ area in the right tail} = 8.907$$

These two values are shown in Figure 11.11.

**Figure 11.11** Rejection and nonrejection regions.**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $\chi^2$  for the sample variance is calculated as follows:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} = \frac{(20 - 1)(170)}{150} = 21.533$$

↑  
From  $H_0$

**Step 5.** Make a decision.

The value of the test statistic  $\chi^2 = 21.533$  is between the two critical values of  $\chi^2$ , 8.907 and 32.852, and it falls in the nonrejection region. Consequently, we fail to reject  $H_0$  and conclude that the population variance of the current scores of high school seniors on this standardized mathematics test does not appear to be different from 150. ■

Note that we can make a test of hypothesis about the population standard deviation  $\sigma$  using the same procedure as that for the population variance  $\sigma^2$ . To make a test of hypothesis about  $\sigma$ , the only change will be mentioning the values of  $\sigma$  in  $H_0$  and  $H_1$ . The rest of the procedure remains the same as in the case of  $\sigma^2$ .

**EXERCISES****■ CONCEPTS AND PROCEDURES**

- 11.40** A sample of certain observations selected from a normally distributed population produced a sample variance of 46. Construct a 95% confidence interval for  $\sigma^2$  for each of the following cases and comment on what happens to the confidence interval of  $\sigma^2$  when the sample size increases.

- a.  $n = 12$
- b.  $n = 16$
- c.  $n = 25$

- 11.41** A sample of 25 observations selected from a normally distributed population produced a sample variance of 35. Construct a confidence interval for  $\sigma^2$  for each of the following confidence levels and comment on what happens to the confidence interval of  $\sigma^2$  when the confidence level decreases.

- a.  $1 - \alpha = .99$
- b.  $1 - \alpha = .95$
- c.  $1 - \alpha = .90$

- 11.42** A sample of 22 observations selected from a normally distributed population produced a sample variance of 18.

- a. Write the null and alternative hypotheses to test whether the population variance is different from 14.
- b. Using  $\alpha = .05$ , find the critical values of  $\chi^2$ . Show the rejection and nonrejection regions on a chi-square distribution curve.
- c. Find the value of the test statistic  $\chi^2$ .
- d. Using a 5% significance level, will you reject the null hypothesis stated in part a?

**11.43** A sample of 21 observations selected from a normally distributed population produced a sample variance of 1.97.

- Write the null and alternative hypotheses to test whether the population variance is greater than 1.75.
- Using  $\alpha = .025$ , find the critical value of  $\chi^2$ . Show the rejection and nonrejection regions on a chi-square distribution curve.
- Find the value of the test statistic  $\chi^2$ .
- Using a 2.5% significance level, will you reject the null hypothesis stated in part a?

**11.44** A sample of 30 observations selected from a normally distributed population produced a sample variance of 5.8.

- Write the null and alternative hypotheses to test whether the population variance is different from 6.0.
- Using  $\alpha = .05$ , find the critical value of  $\chi^2$ . Show the rejection and nonrejection regions on a chi-square distribution curve.
- Find the value of the test statistic  $\chi^2$ .
- Using a 5% significance level, will you reject the null hypothesis stated in part a?

**11.45** A sample of 18 observations selected from a normally distributed population produced a sample variance of 4.6.

- Write the null and alternative hypotheses to test whether the population variance is different from 2.2.
- Using  $\alpha = .05$ , find the critical values of  $\chi^2$ . Show the rejection and nonrejection regions on a chi-square distribution curve.
- Find the value of the test statistic  $\chi^2$ .
- Using a 5% significance level, will you reject the null hypothesis stated in part a?

## ■ APPLICATIONS

**11.46** Sandpaper is rated by the coarseness of the grit on the paper. Sandpaper that is more coarse will remove material faster. Jobs such as the final sanding of bare wood prior to painting or sanding in between coats of paint require sandpaper that is much finer. A manufacturer of sandpaper rated 220, which is used for the final preparation of bare wood, wants to make sure that the variance of the diameter of the particles in their 220 sandpaper does not exceed 2.0 micrometers. Fifty-one randomly selected particles are measured. The variance of the particle diameters is 2.13 micrometers. Assume that the distribution of particle diameter is approximately normal.

- Construct the 95% confidence intervals for the population variance and standard deviation.
- Test at a 2.5% significance level whether the variance of the particle diameters of all particles in 220-rated sandpaper is greater than 2.0 micrometers.

**11.47** The makers of Flippin' Out Pancake Mix claim that one cup of their mix contains 11 grams of sugar. However, the mix is not uniform, so the amount of sugar varies from cup to cup. One cup of mix was taken from each of 24 randomly selected boxes. The sample variance of the sugar measurements from these 24 cups was 1.47 grams. Assume that the distribution of sugar content is approximately normal.

- Construct the 98% confidence intervals for the population variance and standard deviation.
- Test at a 1% significance level whether the variance of the sugar content per cup is greater than 1.0 gram.

**11.48** An auto manufacturing company wants to estimate the variance of miles per gallon for its auto model AST727. A random sample of 22 cars of this model showed that the variance of miles per gallon for these cars is .62. Assume that the miles per gallon for all such cars are (approximately) normally distributed.

- Construct the 95% confidence intervals for the population variance and standard deviation.
- Test at a 1% significance level whether the sample result indicates that the population variance is different from .30.

**11.49** The manufacturer of a certain brand of lightbulbs claims that the variance of the lives of these bulbs is 4200 square hours. A consumer agency took a random sample of 25 such bulbs and tested them. The variance of the lives of these bulbs was found to be 5200 square hours. Assume that the lives of all such bulbs are (approximately) normally distributed.

- Make the 99% confidence intervals for the variance and standard deviation of the lives of all such bulbs.
- Test at a 5% significance level whether the variance of such bulbs is different from 4200 square hours.

## USES AND MISUSES...

### 1. DO NOT FEED THE ANIMALS

You are a wildlife enthusiast studying African wildlife: gnus, zebras, and gazelles. You know that a herd of each species will visit one of three watering places in a region every day, but you do not know the distribution of choices that the animals make or whether these choices are dependent. You have observed that the animals sometimes drink together and sometimes do not. A statistician offers to help and says that he will perform a test for independence of watering place choices based on your observations of the animals' behavior over the past several months. The statistician performs some calculations and says that he has answered your question because his chi-square test of the independence of watering place choices, at a 5% significance level, told him to reject the null hypothesis. He has also performed a goodness-of-fit test on the hypothesis that the animals are equally likely to choose any watering place, and he has rejected that hypothesis as well.

The statistician barely helped you. In the first case, you know a single piece of information: the choice of a watering place for the three groups of animals is dependent. Another way of stating the result is that your data indicate that the choice of watering places for at least one of the animals is not independent of the others. Perhaps the zebras get up early, and the gnus and gazelles follow, making the gnus and gazelles dependent on the choice of the zebras. Or perhaps the animals choose the watering place of the day independent of the other animals, but always avoid the watering place at which the lions are drinking. Regarding the goodness-of-fit test, all you know is that the hypothesis that the animals equally favor the three watering places was wrong. But you do not know what the expected distribution should be. In short, the rejection of the null hypothesis raises more questions than it answers.

### 2. IS THERE A GENDER BIAS IN ADMISSIONS?

Categorical data analysis methods, such as a chi-square test for independence, are used quite often in analyzing employment and admissions data in discrimination cases. One of the more famous discrimination cases involved graduate admissions at the University of California, Berkeley, in 1973. The claim was that UC Berkeley was discriminating against women in their admissions decisions, as would seem to be the case based on the data in the following table.

	Applicants	Percentage Admitted
Male	8442	44%
Female	4321	35%

The  $p$ -value for the corresponding chi-square test of independence is approximately  $1.1 \times 10^{-22}$ , so it would seem clear that there is statistical dependence between gender and graduate admission. However, although it is true that admission rates differed by gender, the fact that the University was found not guilty of discrimination against women might shock you.

In order to find out why the University was found not guilty, we need to introduce another variable into the study: program of study. When the various programs were considered separately, the following admission rates for the six largest programs were observed.

Department	Men		Women	
	Applicants	Admitted (%)	Applicants	Admitted (%)
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	272	6	341	7

As one can see, four of the six programs had nominally higher admission rates for women than for men. So, why is the overall acceptance rate higher for men than for women? Looking at Departments A and B, which had the highest acceptance rates, we notice that 1385 men (16.4% of all men) applied to those programs, while 133 women (3.1% of all women) applied to these programs. On the other hand, almost twice as many women as men applied to the programs with the lowest acceptance rates. Hence, the overall acceptance rate for each gender was affected by the programs to which they applied and the number who applied to those programs. Since a much higher percentage of men applied to programs with higher (overall) acceptance rates, the overall acceptance rate for men ended up being higher than the overall acceptance rate for women.

This result is an example of what is known as Simpson's Paradox, which occurs when the inclusion of an additional variable (characteristic) reverses the conclusion made without that variable.

Source: P. J. Bickel, E. A. Hammel, and J. W. O'Connell (1975): Sex Bias in Graduate Admissions: Data From Berkeley. *Science* 187(4175): 398–404.

## Glossary

**Chi-square distribution** A distribution, with degrees of freedom as the only parameter, that is skewed to the right for small  $df$  and looks like a normal curve for large  $df$ .

**Expected frequencies** The frequencies for different categories of a multinomial experiment or for different cells of a contingency table that are expected to occur when a given null hypothesis is true.

**Goodness-of-fit test** A test of the null hypothesis that the observed frequencies for an experiment follow a certain pattern or theoretical distribution.

**Multinomial experiment** An experiment with  $n$  trials for which (1) the trials are identical, (2) there are more than two possible outcomes per trial, (3) the trials are independent, and (4) the probabilities of the various outcomes remain constant for each trial.

**Observed frequencies** The frequencies actually obtained from the performance of an experiment.

**Test of homogeneity** A test of the null hypothesis that the proportions of elements that belong to different groups in two (or more) populations are similar.

**Test of independence** A test of the null hypothesis that two attributes of a population are not related.

## Supplementary Exercises

**11.50** According to a report in the *Wall Street Journal* ([online.wsj.com/mdc/public/page/2\\_3022-autosales.html](http://online.wsj.com/mdc/public/page/2_3022-autosales.html)), the distribution of all auto sales by segment (type of vehicle) in the United States during November 2011 was as follows:

Segment	Cars	Light-Duty Trucks	SUVs	Crossovers
November 2011 percentage	46.59	20.23	11.52	21.66

A recent survey of 730 new auto sales had the following distribution:

Segment	Cars	Light-Duty Trucks	SUVs	Crossovers
Number of sales	374	138	65	153

Test at a 2.5% significance level whether the distribution of recent auto sales is significantly different from the November 2011 distribution.

**11.51** One of the products produced by Branco Food Company is Total-Bran Cereal, which competes with three other brands of similar total-bran cereals. The company's research office wants to investigate if the percentage of people who consume total-bran cereal is the same for each of these four brands. Let us denote the four brands of cereal by A, B, C, and D. A sample of 1000 persons who consume total-bran cereal was taken, and they were asked which brand they most often consume. Of the respondents, 212 said they usually consume Brand A, 284 consume Brand B, 254 consume Brand C, and 250 consume Brand D. Does the sample provide enough evidence to reject the null hypothesis that the percentage of people who consume total-bran cereal is the same for all four brands? Use  $\alpha = .05$ .

**11.52** The percentage distribution of birth weights for all children in cases of multiple births (twins, triplets, etc.) in North Carolina during 2009 was as given in the following table:

Weight (grams)	0–500	501–1500	1501–2500	2501–8165
Percentage	1.45	11.02	49.23	38.30

Source: <http://www.schs.state.nc.us/SCHS/data/births/bd.cfm>.

The frequency distribution of birth weights of a sample of 587 children who shared multiple births and were born in North Carolina in 2012 is as shown in the following table:

Weight (grams)	0–500	501–1500	1501–2500	2501–8165
Frequency	2	60	305	220

Test at a 2.5% significance level whether the 2012 distribution of birth weights for all children born in North Carolina who shared multiple births is significantly different from the one for 2009.

**11.53** A 2010 poll by Marist University asked people to choose their favorite classic Christmas movie from a list of five choices. The following table shows the frequencies for the various movies:

Movie	It's A Wonderful Life	A Christmas Story	Miracle on 34th Street	White Christmas	A Christmas Carol
Frequency	247	237	226	124	134

Source: [www.maristpoll.marist.edu/1221-no-consensus-on-favorite-holiday-film/](http://www.maristpoll.marist.edu/1221-no-consensus-on-favorite-holiday-film/).

Test at a 1% significance level whether these five movies are equally preferred.

**11.54** During a bear market, 140 investors were asked how they were adjusting their portfolios to protect themselves. Some of these investors were keeping most of their money in stocks, whereas others were shifting large amounts of money to bonds, real estate, or cash (such as money market accounts). The results of the survey are shown in the following table.

Favored choice	Stocks	Bonds	Real Estate	Cash
Number of investors	46	41	32	21

Using a 2.5% significance level, test the null hypothesis that the percentages of investors favoring the four choices are all equal.

**11.55** A randomly selected sample of 100 persons who suffer from allergies were asked during what season they suffer the most. The results of the survey are recorded in the following table.

Season	Fall	Winter	Spring	Summer
Persons allergic	18	13	31	38

Using a 1% significance level, test the null hypothesis that the proportions of all allergic persons are equally distributed over the four seasons.

**11.56** All shoplifting cases in the town of Seven Falls are randomly assigned to either Judge Stark or Judge Rivera. A citizens group wants to know whether either of the two judges is more likely to sentence the offenders to jail time. A sample of 180 recent shoplifting cases produced the following two-way table.

	Jail	Other Sentence
Judge Stark	27	65
Judge Rivera	31	57

Test at a 5% significance level whether the type of sentence for shoplifting depends on which judge tries the case.

**11.57** A November 2011 Kaiser Family Foundation study asked a random sample of Americans what they would like to see done with the American Healthcare Act. The following table lists the frequencies of the results that are comparable to those found in the study:

	Political Affiliation		
	Democrat	Independent	Republican
Expand it	216	143	99
Keep it as is	126	149	129
Repeal it or repeal and replace it	121	141	101

Source: [www.kff.org/kaiserpolls/upload/8259-C.pdf](http://www.kff.org/kaiserpolls/upload/8259-C.pdf).

Test at a 5% significance level whether political affiliation and desired action regarding the American Healthcare Act are dependent.

**11.58** In the National Survey on Drug Use and Health, one of the questions asked is about illicit drug use by Americans age 18 to 25 years ([www.samhsa.gov/data/NSDUH/2k10NSDUH/2k10Results.htm#Fig2-7](http://www.samhsa.gov/data/NSDUH/2k10NSDUH/2k10Results.htm#Fig2-7)). On the assumption that the data were based on random samples of 3600 people in each of the years 2008 to 2010, the percentages reported in the survey would yield the numbers given in the following table.

	Used Illicit Drugs		
	2008	2009	2010
Yes	706	763	774
No	2894	2837	2826

Test at a 1% significance level whether the proportion of Americans age 18 to 25 years who used illicit drugs is the same in each of the three years 2008 to 2010.

**11.59** Recent recession and bad economic conditions forced many people to hold more than one job to make ends meet. A sample of 500 persons who held more than one job produced the following two-way table.

	Single	Married	Other
Male	72	209	39
Female	33	102	45

Test at a 1% significance level whether gender and marital status are related for all people who hold more than one job.

**11.60** ATVs (all-terrain vehicles) have become a source of controversy. Some people feel that their use should be tightly regulated, while others prefer fewer restrictions. Suppose a survey consisting of a random sample of 200 people aged 18 to 27 and another survey of a random sample of 210 people aged 28 to 37 was conducted, and these people were asked whether they favored more restrictions on ATVs, fewer restrictions, or no change. The results of this survey are summarized in the following table.

Age (years)	More Restrictions	Fewer Restrictions	No Change
18 to 27	40	92	68
28 to 37	55	68	87

Test at a 2.5% significance level whether the distribution of opinions in regard to ATVs are the same for both age groups.

**11.61** A random sample of 100 persons was selected from each of four regions in the United States. These people were asked whether or not they support a certain farm subsidy program. The results of the survey are summarized in the following table.

	Favor	Oppose	Uncertain
Northeast	56	33	11
Midwest	73	23	4
South	67	28	5
West	59	35	6

Using a 1% significance level, test the null hypothesis that the percentages of people with different opinions are similar for all four regions.

**11.62** Construct the 98% confidence intervals for the population variance and standard deviation for the following data, assuming that the respective populations are (approximately) normally distributed.

a.  $n = 21, s^2 = 9.2$       b.  $n = 17, s^2 = 1.7$

**11.63** Construct the 95% confidence intervals for the population variance and standard deviation for the following data, assuming that the respective populations are (approximately) normally distributed.

a.  $n = 10, s^2 = 7.2$       b.  $n = 18, s^2 = 14.8$

**11.64** Refer to Exercise 11.62a. Test at a 5% significance level if the population variance is different from 6.5.

**11.65** Refer to Exercise 11.62b. Test at a 2.5% significance level if the population variance is greater than 1.1.

**11.66** Refer to Exercise 11.63a. Test at a 1% significance level if the population variance is greater than 4.2.

**11.67** Refer to Exercise 11.63b. Test at a 5% significance level if the population variance is different from 10.4.

**11.68** Usually people do not like waiting in line for a long time for service. A bank manager does not want the variance of the waiting times for her customers to be greater than 4.0 square minutes. A random sample of 25 customers taken from this bank gave the variance of the waiting times equal to 8.3 square minutes.

a. Test at a 1% significance level whether the variance of the waiting times for all customers at this bank is greater than 4.0 square minutes. Assume that the waiting times for all customers are normally distributed.

b. Construct a 99% confidence interval for the population variance.

**11.69** The variance of the SAT scores for all students who took that test this year is 5000. The variance of the SAT scores for a random sample of 20 students from one school is equal to 3175.

a. Test at a 2.5% significance level whether the variance of the SAT scores for students from this school is lower than 5000. Assume that the SAT scores for all students at this school are (approximately) normally distributed.

- b.** Construct the 98% confidence intervals for the variance and the standard deviation of SAT scores for all students at this school.

**11.70** A company manufactures ball bearings that are supplied to other companies. The machine that is used to manufacture these ball bearings produces them with a variance of diameters of .025 square millimeter or less. The quality control officer takes a sample of such ball bearings quite often and checks, using confidence intervals and tests of hypotheses, whether or not the variance of these bearings is within .025 square millimeter. If it is not, the machine is stopped and adjusted. A recently taken random sample of 23 ball bearings gave a variance of the diameters equal to .034 square millimeter.

- a.** Using a 5% significance level, can you conclude that the machine needs an adjustment?

Assume that the diameters of all ball bearings have a normal distribution.

- b.** Construct a 95% confidence interval for the population variance.

**11.71** A random sample of 25 students taken from a university gave the variance of their GPAs equal to .19.

- a.** Construct the 99% confidence intervals for the population variance and standard deviation. Assume that the GPAs of all students at this university are (approximately) normally distributed.  
**b.** The variance of GPAs of all students at this university was .13 two years ago. Test at a 1% significance level whether the variance of current GPAs at this university is different from .13.

**11.72** A sample of seven passengers boarding a domestic flight produced the following data on weights (in pounds) of their carry-on bags.

46.3    41.5    39.7    31.0    40.6    35.8    43.2

- a.** Using the formula from Chapter 3, find the sample variance,  $s^2$ , for these data.  
**b.** Make the 98% confidence intervals for the population variance and standard deviation. Assume that the population from which this sample is selected is normally distributed.  
**c.** Test at a 5% significance level whether the population variance is larger than 20 square pounds.

**11.73** The following are the prices (in dollars) of the same brand of camcorder found at eight stores in Los Angeles.

568    628    602    642    550    688    615    604

- a.** Using the formula from Chapter 3, find the sample variance,  $s^2$ , for these data.  
**b.** Make the 95% confidence intervals for the population variance and standard deviation. Assume that the prices of this camcorder at all stores in Los Angeles follow a normal distribution.  
**c.** Test at a 5% significance level whether the population variance is different from 750 square dollars.

## Advanced Exercises

**11.74** A 2009 survey reported in *USA TODAY* asked U.S. households who cooks in their homes on Mother's Day. The results from the survey are reported in the following table. Assume that these results are true for the population of all U.S. households in 2009.

Person who cooks	Mom	Mom with help from the family	Mom's spouse	The guests bring food
Percentage of households	38	34	19	9

Suppose that recently a random sample of 300 U.S. households were asked the same question and the number of households with two responses are shown in the following table, and the numbers are missing for the other two responses.

Person who cooks	Mom	Mom with help from the family	Mom's spouse	The guests bring food
Number of households	114	102	?	?

- a.** Suppose you were to perform a hypothesis test to compare the sample data to the *USA TODAY* percentages. What would the counts for the categories *Mom's spouse* and *The guests bring food* have to be in order for the value of the test statistic to be as small as possible?  
Note: There is only one correct pair of values for this question.
- b.** By how much would the count for *Mom's spouse* have to increase from the value in part a in order to reject the null hypothesis at a 10% significance level?

- c. Suppose you were to reduce the count for *Mom's spouse* in part a by the same amount by which you increased it in part b. Calculate the value of the test statistic. How does this compare to the value of the test statistic you calculated in part b?

**11.75** A chemical manufacturing company wants to locate a hazardous waste disposal site near a city of 50,000 residents and has offered substantial financial inducements to the city. Two hundred adults (110 women and 90 men) who are residents of this city are chosen at random. Sixty percent of these adults oppose the site, 32% are in favor, and 8% are undecided. Of those who oppose the site, 65% are women; of those in favor, 62.5% are men. Using a 5% level of significance, can you conclude that opinions on the disposal site are dependent on gender?

**11.76** A student who needs to pass an elementary statistics course wonders whether it will make a difference if she takes the course with instructor A rather than instructor B. Observing the final grades given by each instructor in a recent elementary statistics course, she finds that Instructor A gave 48 passing grades in a class of 52 students and Instructor B gave 44 passing grades in a class of 54 students. Assume that these classes and grades make simple random samples of all classes and grades of these instructors.

- Compute the value of the standard normal test statistic  $z$  of Section 10.5.3 for the data and use it to find the  $p$ -value when testing for the difference between the proportions of passing grades given by these instructors.
- Construct a  $2 \times 2$  contingency table for these data. Compute the value of the  $\chi^2$  test statistic for the test of independence and use it to find the  $p$ -value.
- How do the test statistics in parts a and b compare? How do the  $p$ -values for the tests in parts a and b compare? Do you think this is a coincidence, or do you think this will always happen?

**11.77** Each of five boxes contains a large (but unknown) number of red and green marbles. You have been asked to find if the proportions of red and green marbles are the same for each of the five boxes. You sample 50 times, with replacement, from each of the five boxes and observe 20, 14, 23, 30, and 18 red marbles, respectively. Can you conclude that the five boxes have the same proportions of red and green marbles? Use a .05 level of significance.

**11.78** Suppose that you have a two-way table with the following row and column totals.

		Variable 1			Total
		A	B	C	
Variable 2	X				120
	Y				205
	Z				175
	Total	165	140	195	500

The observed values in the cells must be counts, which are nonnegative integers. Calculate the expected counts for the cells under the assumption that the two variables are independent. Based on your calculations, explain why it is impossible for the test statistic to have a value of zero.

**11.79** You have collected data on a variable, and you want to determine if a normal distribution is a reasonable model for these data. The following table shows how many of the values fall within certain ranges of  $z$  values for these data.

Category	Count
$z$ score below $-2$	48
$z$ score from $-2$ to less than $-1.5$	67
$z$ score from $-1.5$ to less than $-1$	146
$z$ score from $-1$ to less than $-0.5$	248
$z$ score from $-0.5$ to less than $0$	187
$z$ score from $0$ to less than $0.5$	125
$z$ score from $0.5$ to less than $1$	88
$z$ score from $1$ to less than $1.5$	47
$z$ score from $1.5$ to less than $2$	25
$z$ score of $2$ or above	19
Total	1000

Perform a hypothesis test to determine if a normal distribution is an appropriate model for these data. Use a significance level of 5%.

**11.80** Refer to Problem 11.61. Explain why the hypothesis test in that problem is a test of homogeneity as opposed to a test of independence. What feature of the data would change if you were to collect data in order to test for independence?

**11.81** You are performing a goodness-of-fit test with four categories, all of which are supposed to be equally likely. You have a total of 100 observations. The observed frequencies are 21, 26, 31, and 22, respectively, for the four categories.

- Show that you would fail to reject the null hypothesis for these data for any reasonable significance level.
- The sum of the absolute differences (between the expected and the observed frequencies) for these data is 14 (i.e.,  $4 + 1 + 6 + 3 = 14$ ). Is it possible to have different observed frequencies keeping the sum at 14 so that you get a  $p$ -value of .10 or less?

## Self-Review Test

- The random variable  $\chi^2$  assumes only
  - positive
  - nonnegative
  - nonpositive values
- The parameter(s) of the chi-square distribution is (are)
  - degrees of freedom
  - $df$  and  $n$
  - $\chi^2$
- Which of the following is *not* a characteristic of a multinomial experiment?
  - It consists of  $n$  identical trials.
  - There are  $k$  possible outcomes for each trial and  $k > 2$ .
  - The trials are random.
  - The trials are independent.
  - The probabilities of outcomes remain constant for each trial.
- The observed frequencies for a goodness-of-fit test are
  - the frequencies obtained from the performance of an experiment
  - the frequencies given by the product of  $n$  and  $p$
  - the frequencies obtained by adding the results of a and b
- The expected frequencies for a goodness-of-fit test are
  - the frequencies obtained from the performance of an experiment
  - the frequencies given by the product of  $n$  and  $p$
  - the frequencies obtained by adding the results of a and b
- The degrees of freedom for a goodness-of-fit test are
  - $n - 1$
  - $k - 1$
  - $n + k - 1$
- The chi-square goodness-of-fit test is always
  - two-tailed
  - left-tailed
  - right-tailed
- To apply a goodness-of-fit test, the expected frequency of each category must be at least
  - 10
  - 5
  - 8
- The degrees of freedom for a test of independence are
  - $(R - 1)(C - 1)$
  - $n - 2$
  - $(n - 1)(k - 1)$
- According to the Henry J. Kaiser Family Foundation ([www.statehealthfacts.org](http://www.statehealthfacts.org)), the percentage distribution of the source of health insurance in the United States in 2010 was as listed in the following table.

Source	Employer	Individual	Medicaid	Medicare	Other Public	Uninsured
Percentage	49.12	4.89	15.86	12.49	1.29	16.35

Recently 15,000 randomly selected Americans were asked about the source of their health insurance. The following table contains the frequency distribution that resulted from this survey.

Source	Employer	Individual	Medicaid	Medicare	Other Public	Uninsured
Frequency	7286	698	2402	1927	171	2516

Test at a 5% significance level whether the distribution of health insurance source in the recent survey differs from the 2010 distribution.

11. The following table gives the two-way classification of 1000 persons who have been married at least once. They are classified by educational level and marital status.

	Educational Level			
	Less Than High School	High School Degree	Some College	College Degree
Divorced	173	158	95	53
Never divorced	162	126	110	123

Test at a 1% significance level whether educational level and ever being divorced are dependent.

12. A researcher wanted to investigate if people who belong to different income groups are homogeneous with regard to playing lotteries. She took a sample of 600 people from the low-income group, another sample of 500 people from the middle-income group, and a third sample of 400 people from the high-income group. All these people were asked whether they play the lottery often, sometimes, or never. The results of the survey are summarized in the following table.

	Income Group		
	Low	Middle	High
Play often	174	163	90
Play sometimes	286	217	120
Never play	140	120	190

Using a 5% significance level, can you reject the null hypothesis that the percentages of people who play the lottery often, sometimes, and never are the same for each income group?

13. The owner of an ice cream parlor is concerned about consistency in the amount of ice cream his servers put in each cone. He would like the variance of all such cones to be no more than .25 square ounce. He decides to weigh each double-dip cone just before it is given to the customer. For a sample of 20 double-dip cones, the weights were found to have a variance of .48 square ounce. Assume that the weights of all such cones are (approximately) normally distributed.

- a. Construct the 99% confidence intervals for the population variance and the population standard deviation.
- b. Test at a 1% significance level whether the variance of the weights of all such cones exceeds .25 square ounce.

## Mini-Projects

### ■ MINI-PROJECT 11-1

In recent years drivers have become careless about signaling their turns. To study this problem, go to a busy intersection and observe at least 75 vehicles that make left turns. Divide these vehicles into three or four classes. For example, you might use cars, trucks, and others, where “others” include minivans and sport-utility vehicles, as classes. For each left turn made by a vehicle, record the type of vehicle and whether or not the driver used the left turn signal before making this turn. It would be better to avoid intersections that have designated left-turn lanes or green arrows for left turns because drivers in these situations often assume that their intent to turn left is obvious. Carry out an appropriate test at the 1% level of significance to determine if signaling behavior and vehicle type are dependent.

### ■ MINI-PROJECT 11-2

One day during lunch, visit your school cafeteria, observe at least 100 people, and write down what they are drinking. Categorize the drinks as soft drink (soda, fruit punch, or lemonade), iced tea, milk or juice, hot drink, and water. Also identify the gender of each person. Perform a hypothesis test to determine if the type of drink and gender are independent.

### ■ MINI-PROJECT 11-3

Many studies have been performed to determine the sources that people use to get their news. Survey at least 50 people at random from your class or dorm and ask the following question:

Which of the following would you classify as being your primary source for news?

- Network news broadcasts
- Cable news broadcasts
- Newspapers
- Internet-based news sources
- Radio news broadcasts

Use the data to test the null hypothesis that college students are equally likely to classify the five options as being their primary source for news. Use a 5% significance level.

### ■ MINI-PROJECT 11-4

As reported in *USA Today*, August 20, 2009, a Baby Oracl survey of 1004 adults conducted by Kelton Research asked these adults which of a set of four noises (car alarm, jackhammer, baby crying, or dog barking) they find to be the most frustrating to hear. Suppose that the survey also included information on the gender of the respondents as listed in the following table.

Sound That Is Most Frustrating to Hear	Females	Males
Car alarm	225	197
Jackhammer	154	131
Baby crying	79	143
Dog barking	44	31

- Perform a hypothesis test to determine whether the sound that is most frustrating to hear and gender are independent. Use a 1% significance level.
- Suppose that it was noticed that the surveyor incorrectly marked 30 of the survey sheets. Specifically, thirty of the male respondents were mistakenly recorded as saying “baby crying.” Furthermore, it was determined that all 30 of the mistakes should have one of the other three responses but the same response. That is, all these 30 responses should have been “car alarm,” or all 30 should have been “jackhammer,” or all 30 should have been “dog barking.” Determine which of these three changes would result in a different conclusion than the one obtained in part a.

## DECIDE FOR YOURSELF

## TESTING FOR THE FAIRNESS OF GAMBLING EQUIPMENT

Casino gambling has grown rapidly in the United States. Native American tribes have opened casinos on reservations, many horse racing tracks have been allowed to add slot machines on site, and riverboat/lakefront casinos have also been opened in recent years. States with casino gambling have state agencies that are responsible for verifying and making sure that the games and equipment are fair and not fixed. In many states, such an agency is called the Division of Gaming Enforcement. New Jersey and Nevada have two of the largest such agencies, given the presence of Atlantic City and Las Vegas in these states. The chi-square procedures that you have learned in this chapter can be used to test the validity of the *fairness* assumption in regard to the gaming equipment.

A simple example would involve checking to see whether or not a given die is balanced. Under the null hypothesis, we would assume that the probability of a specific side coming up when we roll this die is  $1/6$ . To test this notion, we can roll the given die a specific number of times and observe the frequency for each outcome.

Suppose we roll this die 180 times and obtain the frequencies for various outcomes as listed in the following table.

Outcome	1-spot	2-spots	3-spots	4-spots	5-spots	6-spots
Frequency	26	31	29	33	26	35

- Theoretically, how often would you expect each outcome to occur if we roll this die 180 times, assuming it is a fair die?
- Perform the appropriate hypothesis test to determine the *p*-value with the null hypothesis that the die is fair. What is your conclusion?
- How much do you have to change the frequencies for various outcomes in the above table to obtain a conclusion for the hypothesis test of question 2 that is the opposite of the one you obtained above? Does your conclusion switch faster if you make a big change to one frequency and small changes to the others or if you make moderate changes to all of the categories? (Remember that the sum of all frequencies has to remain 180.)

# TECHNOLOGY INSTRUCTION

## Chi-Square Tests

### TI-84

**MATRIX[A]**  $3 \times 2$

20	80
40	160
40	60

Screen 11.1

**X<sup>2</sup>-Test**  
Observed: [A]  
Expected: [B]  
Calculate Draw

Screen 11.2

- To perform an independence or homogeneity test on a contingency table, enter the actual data and the expected values as matrices. To do so, select **2nd > MATRIX >EDIT**, and use the arrow key to select the name of your matrix. Press **ENTER**, and then type in the number of rows, the number of columns, and the entries for each matrix. (See **Screen 11.1**.)
- Select **STAT>TESTS> $\chi^2$ -Test**. You will need to enter the names of the **Observed** and the **Expected** data matrices. For each entry, position the cursor, and then select **2nd > MATRIX >NAMES**. Use the arrow keys to choose the appropriate name and then press **ENTER**. (See **Screen 11.2**.) After entering the matrix names, press **ENTER**. The result includes the value of  $\chi^2$ , the *p*-value, and the degrees of freedom. (See **Screen 11.3**.)
- To perform a goodness-of-fit test using the TI-84, you will need to calculate the expected counts for entry into the calculator. Select **STAT > EDIT > 1. Edit**. Enter the observed counts into one list, such as **L1**, and the corresponding expected counts into another list, such as **L2**, in the same way you would enter data into these lists. (See **Screen 11.4**.)
- Select **STAT > TESTS >  $\chi^2$ GOF-Test**. You will need to enter the names of the **Observed** and **Expected** counts lists as you did for the **1-Var Stats** function in Chapter 3. Enter the number of degrees of freedom (**df**), which is the number of groups minus one. (See **Screen 11.5**.) Select **Calculate**, and press **ENTER**. (See **Screen 11.6**.)

**X<sup>2</sup>-Test**  
 $\chi^2=16$   
 $P=3.3546263E^{-4}$   
 $df=2$

Screen 11.3

L1	L2	L3	Z
27	31.56		
117	92.7		
102	113.94		
268	239.46		
86	122.34		
-----			
L2(6) =			

Screen 11.4

**X<sup>2</sup>GOF-Test**  
Observed:L1  
Expected:L2  
df:4  
Calculate Draw

Screen 11.5

**X<sup>2</sup>GOF-Test**  
 $\chi^2=22.47598492$   
 $P=1.6110613E^{-4}$   
 $df=4$   
CNTRB=.658859...

Screen 11.6

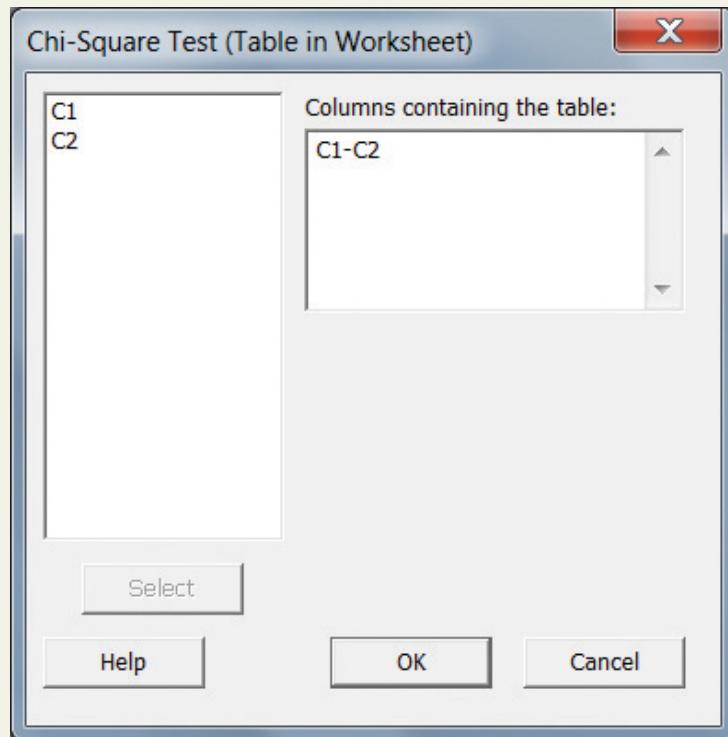
### Minitab

- To perform a test of homogeneity or independence on a contingency table, enter the table into columns (see **Screen 11.7**), then select **Stat > Tables > Chi-Square Test (Two-Way Table In Worksheet)**. Enter the names of the columns containing the table, and select **OK**.

Screen 11.7

	C1	C2
1	20	80
2	40	160
3	40	60

(See Screen 11.8.) The result includes the expected counts, the degrees of freedom, the value of the test statistic, and the *p*-value. (See Screen 11.9.)



Screen 11.8

The session window is titled "Session" and shows the output for a Chi-Square Test. The output includes the observed and expected counts for three categories (1, 2, 3) across two columns (C1, C2), the Chi-Square statistic (16.000), degrees of freedom (2), and the p-value (0.000).

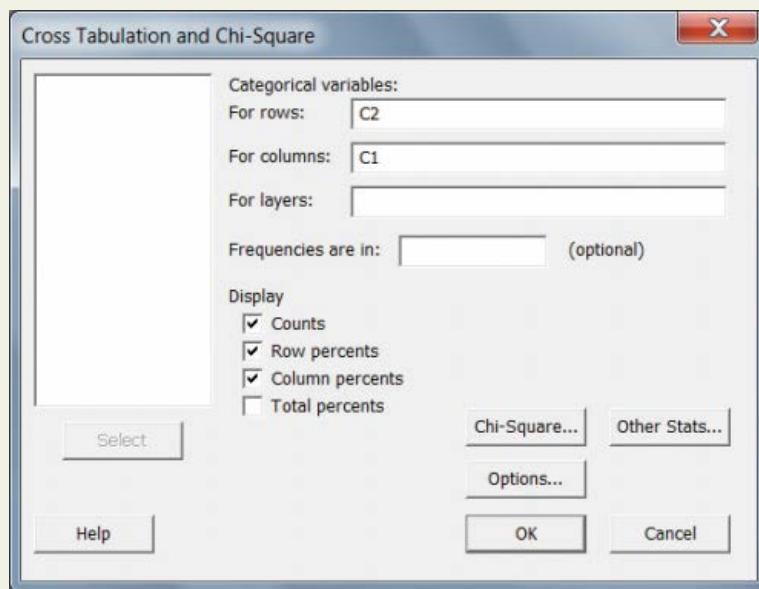
	C1	C2	Total
1	20	80	100
	25.00	75.00	
	1.000	0.333	
2	40	160	200
	50.00	150.00	
	2.000	0.667	
3	40	60	100
	25.00	75.00	
	9.000	3.000	
Total	100	300	400
Chi-Sq = 16.000, DF = 2, P-Value = 0.000			

Screen 11.9

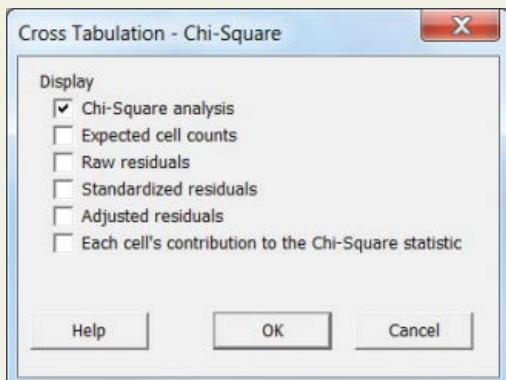
2. To perform a test of homogeneity or independence on categorical data entered in columns (see Screen 11.10), select **Stat > Tables > Cross Tabulation and Chi-Square Test**. Enter the name of the column for the row variable in the output and for the column variable in the output in **For rows:** and **For columns:**, respectively, and check the **Counts**, **Row percents**, and **Column percents** boxes. (See Screen 11.11.) Click the **Chi-Square** button, and check the **Chi-Square analysis** box. (See Screen 11.12.) Select **OK** to close both dialog boxes. The result includes the expected counts, the degrees of freedom, the value of the test statistic, and the *p*-value. (See Screen 11.13.)

↓	C1-T	C2-T
1	N	B
2	Y	A
3	Y	A
4	Y	B
5	Y	A
6	N	B
7	Y	A
8	Y	A
9	Y	C
10	N	A
11	N	A
12	Y	A

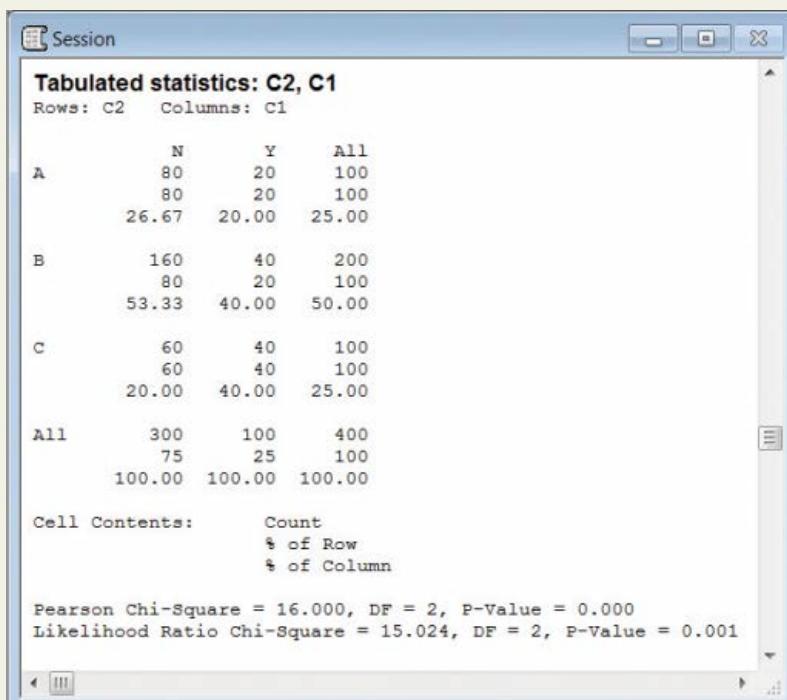
Screen 11.10



Screen 11.11



Screen 11.12



Screen 11.13

**EXCEL**

A	B	C	D	E
Actual		Expected		
20	80	25	75	
40	160	50	150	
40	60	25	75	
p-value	=CHISQ.TEST(A3:B5,D3:E5)			

Screen 11.14

1. To perform a goodness-of-fit or independence test on a contingency table, enter the actual data in a range of cells and the expected data in another range of cells with the same number of rows and columns.

2. Type =CHISQ.TEST(actual range, expected range), and press ENTER. The result is the *p*-value of the test. (See Screen 11.14.)

## TECHNOLOGY ASSIGNMENTS

**TA11.1** Air Quality Index (AQI) data for the city of Kitchener, Ontario, Canada, during the period January 1, 2010, to June 21, 2012, produced the following percentage distribution:

AQI	Very good	Good	Moderate	Poor
Percentage	13.10	71.81	14.76	0.33

Source: www.airqualityontario.com.

The following table gives the AQI data for a sample of 1100 readings from cities similar to Kitchener (population 150,000 to 250,000, metropolitan area population of 400,000 to 500,000):

AQI	Very good	Good	Moderate	Poor
Number of readings	127	816	150	7

Test at a 5% significance level whether the distribution of AQI for the sample data differs from the distribution for Kitchener, Ontario.

**TA11.2** A sample of 4000 persons aged 18 years and older produced the following two-way classification table:

	Men	Women
Single	531	357
Married	1375	1179
Widowed	55	195
Divorced	139	169

Test at a 1% significance level whether gender and marital status are dependent for all persons aged 18 years and older.

**TA11.3** Two samples, one of 3000 students from urban high schools and another of 2000 students from rural high schools, were taken. These students were asked if they have ever smoked. The following table lists the summary of the results.

	Urban	Rural
Have never smoked	1448	1228
Have smoked	1552	772

Using a 5% significance level, test the null hypothesis that the proportions of urban and rural students who have smoked and who have never smoked are homogeneous.

**TA11.4** Using the Data Set V, which contains a random sample of 500 runners from the 2011 Beach to Beacon 10K Road Race, perform a hypothesis test with the null hypothesis that the gender of a participant is independent of whether the person is from Maine or somewhere else.

# CHAPTER 12



JGI/Jamie Grill/Getty Images, Inc.

## Analysis of Variance

### 12.1 The F Distribution

### 12.2 One-Way Analysis of Variance

Trying something new can be risky, and there can be uncertainty about the results. Suppose a school district plans to test three different methods for teaching arithmetic. After teachers implement these different methods for a semester, administrators want to know if the mean scores of students taught with these three different methods are all the same. What data will they require and how will they test for this equality of more than two means? (See Examples 12-2 and 12-3.)

Chapter 10 described the procedures that are used to test hypotheses about the difference between two population means using the normal and *t* distributions. Also described in that chapter were the hypothesis-testing procedures for the difference between two population proportions using the normal distribution. Then, Chapter 11 explained the procedures that are used to test hypotheses about the equality of more than two population proportions using the chi-square distribution.

This chapter explains how to test the null hypothesis that the means of more than two populations are equal. For example, suppose that teachers at a school have devised three different methods to teach arithmetic. They want to find out if these three methods produce different mean scores. Let  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$  be the mean scores of all students who will be taught by Methods I, II, and III, respectively. To test whether or not the three teaching methods produce the same mean, we test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad (\text{All three population means are equal.})$$

against the alternative hypothesis

$$H_1: \text{Not all three population means are equal.}$$

We use the analysis of variance procedure to perform such a test of hypothesis.

Note that the analysis of variance procedure can be used to compare two population means. However, the procedures learned in Chapter 10 are more efficient for performing tests of hypothesis

about the difference between two population means; the analysis of variance procedure, to be discussed in this chapter, is used to compare three or more population means.

An *analysis of variance* test is performed using the *F* distribution. First, the *F* distribution is described in Section 12.1 of this chapter. Then, Section 12.2 discusses the application of the one-way analysis of variance procedure to perform tests of hypothesis.

## 12.1 The *F* Distribution

Like the chi-square distribution, the shape of a particular ***F* distribution**<sup>1</sup> curve depends on the number of degrees of freedom. However, the *F* distribution has *two* numbers of degrees of freedom: *degrees of freedom for the numerator* and *degrees of freedom for the denominator*. These two numbers representing two types of degrees of freedom are the *parameters of the F distribution*. Each combination of degrees of freedom for the numerator and for the denominator gives a different *F* distribution curve. The units of an *F* distribution are denoted by *F*, which assumes only nonnegative values. Like the normal, *t*, and chi-square distributions, the *F* distribution is a continuous distribution. The shape of an *F* distribution curve is skewed to the right, but the skewness decreases as the number of degrees of freedom increases.

### Definition

#### The *F* Distribution

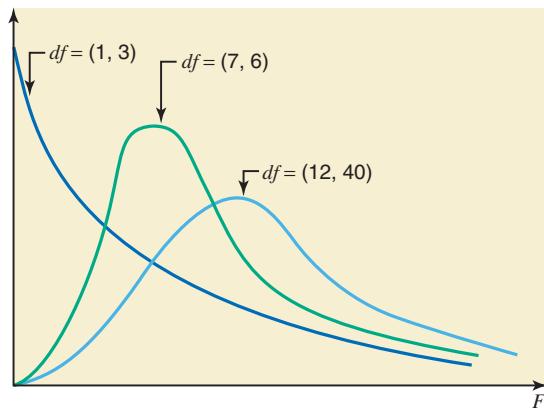
1. The *F distribution* is continuous and skewed to the right.
2. The *F* distribution has two numbers of degrees of freedom: *df* for the numerator and *df* for the denominator.
3. The units (the values of the *F*-variable) of an *F* distribution, denoted by *F*, are nonnegative.

For an *F* distribution, degrees of freedom for the numerator and degrees of freedom for the denominator are usually written as follows:

$$df = (8, 14)$$

↑                   ↑  
First number denotes the      Second number denotes the  
*df* for the numerator      *df* for the denominator

Figure 12.1 shows three *F* distribution curves for three sets of degrees of freedom for the numerator and for the denominator. In the figure, the first number gives the degrees of freedom associated with the numerator, and the second number gives the degrees of freedom associated with the denominator. We can observe from this figure that as the degrees of freedom increase, the peak of the curve moves to the right; that is, the skewness decreases.



**Figure 12.1** Three *F* distribution curves.

<sup>1</sup>The *F* distribution is named after Sir Ronald Fisher.

Table VII in Appendix C lists the values of  $F$  for the  $F$  distribution. To read Table VII, we need to know three quantities: the degrees of freedom for the numerator, the degrees of freedom for the denominator, and an area in the right tail of an  $F$  distribution curve. Note that the  $F$  distribution table (Table VII) is read only for an area in the right tail of the  $F$  distribution curve. Also note that Table VII has four parts. These four parts give the  $F$  values for areas of .01, .025, .05, and .10, respectively, in the right tail of the  $F$  distribution curve. We can make the  $F$  distribution table for other values in the right tail. Example 12–1 illustrates how to read Table VII.

### ■ EXAMPLE 12–1

*Reading the F distribution table.*

Find the  $F$  value for 8 degrees of freedom for the numerator, 14 degrees of freedom for the denominator, and .05 area in the right tail of the  $F$  distribution curve.

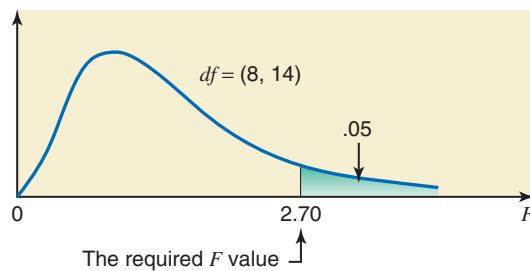
**Solution** To find the required value of  $F$ , we use the portion of Table VII of Appendix C that corresponds to .05 area in the right tail of the  $F$  distribution curve. The relevant portion of that table is shown here as Table 12.1. To find the required  $F$  value, we locate 8 in the row for degrees of freedom for the numerator (at the top of Table VII) and 14 in the column for degrees of freedom for the denominator (the first column on the left side in Table VII). The entry where the column for 8 and the row for 14 intersect gives the required  $F$  value. This value of  $F$  is **2.70**, as shown in Table 12.1 and Figure 12.2. The  $F$  value obtained from this table for a test of hypothesis is called the critical value of  $F$ .

**Table 12.1** Obtaining the  $F$  Value From Table VII

		Degrees of Freedom for the Numerator					
		1	2	...	8	...	100
Degrees of Freedom for the Denominator	1	161.5	199.5	...	238.9	...	253.0
	2	18.51	19.00	...	19.37	...	19.49
	.	...	...	...	...	...	...
	.	...	...	...	...	...	...
	14	4.60	3.74	...	2.70	...	2.19
	.	...	...	...	...	...	...
	.	...	...	...	...	...	...
	100	3.94	3.09	...	2.03	...	1.39

The  $F$  value for 8  $df$  for the numerator, 14  $df$  for the denominator, and .05 area in the right tail

**Figure 12.2** The value of  $F$  from Table VII for 8  $df$  for the numerator, 14  $df$  for the denominator, and .05 area in the right tail.



## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**12.1** Describe the main characteristics of an  $F$  distribution.

**12.2** Find the critical value of  $F$  for the following.

- a.  $df = (3, 3)$  and area in the right tail = .05
- b.  $df = (3, 10)$  and area in the right tail = .05
- c.  $df = (3, 30)$  and area in the right tail = .05

- 12.3** Find the critical value of  $F$  for the following.
- $df = (2, 6)$  and area in the right tail = .025
  - $df = (6, 6)$  and area in the right tail = .025
  - $df = (15, 6)$  and area in the right tail = .025
- 12.4** Determine the critical value of  $F$  for the following.
- $df = (6, 12)$  and area in the right tail = .01
  - $df = (6, 40)$  and area in the right tail = .01
  - $df = (6, 100)$  and area in the right tail = .01
- 12.5** Determine the critical value of  $F$  for the following.
- $df = (2, 2)$  and area in the right tail = .10
  - $df = (8, 8)$  and area in the right tail = .10
  - $df = (20, 20)$  and area in the right tail = .10
- 12.6** Find the critical value of  $F$  for an  $F$  distribution with  $df = (3, 12)$  and
- area in the right tail = .05
  - area in the right tail = .10
- 12.7** Find the critical value of  $F$  for an  $F$  distribution with  $df = (11, 5)$  and
- area in the right tail = .01
  - area in the right tail = .025
- 12.8** Find the critical value of  $F$  for an  $F$  distribution with .025 area in the right tail and
- $df = (4, 11)$
  - $df = (15, 3)$
- 12.9** Find the critical value of  $F$  for an  $F$  distribution with .01 area in the right tail and
- $df = (10, 10)$
  - $df = (9, 25)$

## 12.2 One-Way Analysis of Variance

As mentioned in the beginning of this chapter, the analysis of variance procedure is used to test the null hypothesis that the means of three or more populations are the same against the alternative hypothesis that not all population means are the same. The analysis of variance procedure can be used to compare two population means. However, the procedures learned in Chapter 10 are more efficient for performing tests of hypotheses about the difference between two population means; the analysis of variance procedure is used to compare three or more population means.

Reconsider the example of teachers at a school who have devised three different methods to teach arithmetic. They want to find out if these three methods produce different mean scores. Let  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$  be the mean scores of all students who are taught by Methods I, II, and III, respectively. To test if the three teaching methods produce different means, we test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad (\text{All three population means are equal.})$$

against the alternative hypothesis

$$H_1: \text{Not all three population means are equal.}$$

One method to test such a hypothesis is to test the three hypotheses  $H_0: \mu_1 = \mu_2$ ,  $H_0: \mu_1 = \mu_3$ , and  $H_0: \mu_2 = \mu_3$  separately using the procedure discussed in Chapter 10. Besides being time consuming, such a procedure has other disadvantages. First, if we reject even one of these three hypotheses, then we must reject the null hypothesis  $H_0: \mu_1 = \mu_2 = \mu_3$ . Second, combining the Type I error probabilities for the three tests (one for each test) will give a very large Type I error probability for the test  $H_0: \mu_1 = \mu_2 = \mu_3$ . Hence, we should prefer a procedure that can test the equality of three means in one test. The **ANOVA**, short for **analysis of variance**, provides such a procedure. It is used to compare three or more population means in a single test. Note that if the null hypothesis is rejected, it does not necessarily imply that all three of the means are different or unequal. It could imply that one mean is different from the other two means, or that all three means are different, or that two means are significantly different from each other, but neither is significantly different from the third mean.

### Definition

**ANOVA** ANOVA is a procedure that is used to test the null hypothesis that the means of three or more populations are all equal.

This section discusses the **one-way ANOVA** procedure to make tests by comparing the means of several populations. By using a one-way ANOVA test, we analyze only **one factor or variable**. For instance, in the example of testing for the equality of mean arithmetic scores of students taught by each of the three different methods, we are considering only one factor, which is the effect of different teaching methods on the scores of students. Sometimes we may analyze the effects of two factors. For example, if different teachers teach arithmetic using these three methods, we can analyze the effects of teachers and teaching methods on the scores of students. This is done by using a two-way ANOVA. The procedure under discussion in this chapter is called the analysis of variance because the test is based on the analysis of variation in the data obtained from different samples. The application of one-way ANOVA requires that the following assumptions hold true.

**Assumptions of One-Way ANOVA** The following assumptions must hold true to use *one-way ANOVA*.

1. The populations from which the samples are drawn are (approximately) normally distributed.
2. The populations from which the samples are drawn have the same variance (or standard deviation).
3. The samples drawn from different populations are random and independent.

For instance, in the example about three methods of teaching arithmetic, we first assume that the scores of all students taught by each method are (approximately) normally distributed. Second, the means of the distributions of scores for the three teaching methods may or may not be the same, but all three distributions have the same variance,  $\sigma^2$ . Third, when we take samples to make an ANOVA test, these samples are drawn independently and randomly from three different populations.

The ANOVA test is applied by calculating two estimates of the variance,  $\sigma^2$ , of population distributions: the **variance between samples** and the **variance within samples**. The variance between samples is also called the **mean square between samples** or **MSB**. The variance within samples is also called the **mean square within samples** or **MSW**.

The variance between samples, MSB, gives an estimate of  $\sigma^2$  based on the variation among the means of samples taken from different populations. For the example of three teaching methods, MSB will be based on the values of the mean scores of three samples of students taught by three different methods. If the means of all populations under consideration are equal, the means of the respective samples will still be different, but the variation among them is expected to be small, and, consequently, the value of MSB is expected to be small. However, if the means of populations under consideration are not all equal, the variation among the means of respective samples is expected to be large, and, consequently, the value of MSB is expected to be large.

The variance within samples, MSW, gives an estimate of  $\sigma^2$  based on the variation within the data of different samples. For the example of three teaching methods, MSW will be based on the scores of individual students included in the three samples taken from three populations. The concept of MSW is similar to the concept of the pooled standard deviation,  $s_p$ , for two samples discussed in Section 10.2 of Chapter 10.

*The one-way ANOVA test is always right-tailed with the rejection region in the right tail of the F distribution curve.* The hypothesis-testing procedure using ANOVA involves the same five steps that were used in earlier chapters. The next subsection explains how to calculate the value of the test statistic  $F$  for an ANOVA test.

### 12.2.1 Calculating the Value of the Test Statistic

The value of the test statistic  $F$  for a test of hypothesis using ANOVA is given by the ratio of two variances, the variance between samples (MSB) and the variance within samples (MSW).

**Test Statistic  $F$  for a One-Way ANOVA Test** The value of the *test statistic  $F$*  for an ANOVA test is calculated as

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} \quad \text{or} \quad \frac{\text{MSB}}{\text{MSW}}$$

The calculation of MSB and MSW is explained in Example 12–2.

Example 12–2 describes the calculation of MSB, MSW, and the value of the test statistic  $F$ . Since the basic formulas are laborious to use, they are not presented here. We have used only the short-cut formulas to make calculations in this chapter.

## ■ EXAMPLE 12–2

Fifteen fourth-grade students were randomly assigned to three groups to experiment with three different methods of teaching arithmetic. At the end of the semester, the same test was given to all 15 students. The following table gives the scores of students in the three groups.

Method I	Method II	Method III
48	55	84
73	85	68
51	70	95
65	69	74
87	90	67

*Calculating the value of the test statistic  $F$ .*



PhotoDisc, Inc./Getty Images

Calculate the value of the test statistic  $F$ . Assume that all the required assumptions mentioned in Section 12.2 hold true.

**Solution** In ANOVA terminology, the three methods used to teach arithmetic are called **treatments**. The table contains data on the scores of fourth-graders included in the three samples. Each sample of students is taught by a different method. Let

$x$  = the score of a student

$k$  = the number of different samples (or treatments)

$n_i$  = the size of sample  $i$

$T_i$  = the sum of the values in sample  $i$

$n$  = the number of values in all samples =  $n_1 + n_2 + n_3 + \dots$

$\Sigma x$  = the sum of the values in all samples =  $T_1 + T_2 + T_3 + \dots$

$\Sigma x^2$  = the sum of the squares of the values in all samples

To calculate MSB and MSW, we first compute the **between-samples sum of squares**, denoted by **SSB**, and the **within-samples sum of squares**, denoted by **SSW**. The sum of SSB and SSW is called the **total sum of squares** and is denoted by **SST**; that is,

$$\text{SST} = \text{SSB} + \text{SSW}$$

The values of SSB and SSW are calculated using the following formulas.

**Between- and Within-Samples Sums of Squares** The *between-samples sum of squares*, denoted by **SSB**, is calculated as

$$\text{SSB} = \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \dots \right) - \frac{(\Sigma x)^2}{n}$$

The *within-samples sum of squares*, denoted by **SSW**, is calculated as

$$\text{SSW} = \Sigma x^2 - \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \dots \right)$$

Table 12.2 lists the scores of 15 students who were taught arithmetic by each of the three different methods; the values of  $T_1$ ,  $T_2$ , and  $T_3$ ; and the values of  $n_1$ ,  $n_2$ , and  $n_3$ .

**Table 12.2**

Method I	Method II	Method III
48	55	84
73	85	68
51	70	95
65	69	74
87	90	67
$T_1 = 324$	$T_2 = 369$	$T_3 = 388$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$

In Table 12.2,  $T_1$  is obtained by adding the five scores of the first sample. Thus,  $T_1 = 48 + 73 + 51 + 65 + 87 = 324$ . Similarly, the sums of the values in the second and third samples give  $T_2 = 369$  and  $T_3 = 388$ , respectively. Because there are five observations in each sample,  $n_1 = n_2 = n_3 = 5$ . The values of  $\Sigma x$  and  $n$  are, respectively,

$$\begin{aligned}\Sigma x &= T_1 + T_2 + T_3 = 324 + 369 + 388 = 1081 \\ n &= n_1 + n_2 + n_3 = 5 + 5 + 5 = 15\end{aligned}$$

To calculate  $\Sigma x^2$ , we square all the scores included in all three samples and then add them. Thus,

$$\begin{aligned}\Sigma x^2 &= (48)^2 + (73)^2 + (51)^2 + (65)^2 + (87)^2 + (55)^2 + (85)^2 + (70)^2 \\ &\quad + (69)^2 + (90)^2 + (84)^2 + (68)^2 + (95)^2 + (74)^2 + (67)^2 \\ &= 80,709\end{aligned}$$

Substituting all the values in the formulas for SSB and SSW, we obtain the following values of SSB and SSW:

$$\begin{aligned}SSB &= \left( \frac{(324)^2}{5} + \frac{(369)^2}{5} + \frac{(388)^2}{5} \right) - \frac{(1081)^2}{15} = 432.1333 \\ SSW &= 80,709 - \left( \frac{(324)^2}{5} + \frac{(369)^2}{5} + \frac{(388)^2}{5} \right) = 2372.8000\end{aligned}$$

The value of SST is obtained by adding the values of SSB and SSW. Thus,

$$SST = 432.1333 + 2372.8000 = 2804.9333$$

The variance between samples (MSB) and the variance within samples (MSW) are calculated using the following formulas.

**Calculating the Values of MSB and MSW** MSB and MSW are calculated as, respectively,

$$MSB = \frac{SSB}{k - 1} \quad \text{and} \quad MSW = \frac{SSW}{n - k}$$

where  $k - 1$  and  $n - k$  are, respectively, the  $df$  for the numerator and the  $df$  for the denominator for the  $F$  distribution. Remember,  $k$  is the number of different samples.

Consequently, the variance between samples is

$$MSB = \frac{SSB}{k - 1} = \frac{432.1333}{3 - 1} = 216.0667$$

The variance within samples is

$$MSW = \frac{SSW}{n - k} = \frac{2372.8000}{15 - 3} = 197.7333$$

The value of the test statistic  $F$  is given by the ratio of MSB and MSW. Therefore,

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{216.0667}{197.7333} = 1.09$$

For convenience, all these calculations are often recorded in a table called the *ANOVA table*. Table 12.3 gives the general form of an ANOVA table.

**Table 12.3** ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	$k - 1$	SSB	MSB	$F = \frac{\text{MSB}}{\text{MSW}}$
Within	$n - k$	SSW	MSW	
Total	$n - 1$	SST		

Substituting the values of the various quantities into Table 12.3, we write the ANOVA table for our example as Table 12.4.

**Table 12.4** ANOVA Table for Example 12–2

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	2	432.1333	216.0667	$F = \frac{216.0667}{197.7333} = 1.09$
Within	12	2372.8000	197.7333	
Total	14	2804.9333		

## 12.2.2 One-Way ANOVA Test

Now suppose we want to test the null hypothesis that the mean scores are equal for all three groups of fourth-graders taught by three different methods of Example 12–2 against the alternative hypothesis that the mean scores of all three groups are not equal. Note that in a one-way ANOVA test, the null hypothesis is that the means for all populations are equal. The alternative hypothesis is that not all population means are equal. In other words, the alternative hypothesis states that at least one of the population means is different from the others. Example 12–3 demonstrates how we use the one-way ANOVA procedure to make such a test.

### EXAMPLE 12–3

Reconsider Example 12–2 about the scores of 15 fourth-grade students who were randomly assigned to three groups in order to experiment with three different methods of teaching arithmetic. At a 1% significance level, can we reject the null hypothesis that the mean arithmetic score of all fourth-grade students taught by each of these three methods is the same? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Performing a one-way ANOVA test: all samples the same size.

**Solution** To make a test about the equality of the means of three populations, we follow our standard procedure with five steps.

**Step 1.** *State the null and alternative hypotheses.*

Let  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$  be the mean arithmetic scores of all fourth-grade students who are taught, respectively, by Methods I, II, and III. The null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad (\text{The mean scores of the three groups are all equal.})$$

$$H_1: \text{Not all three means are equal.}$$

Note that the alternative hypothesis states that at least one population mean is different from the other two.

**Step 2. Select the distribution to use.**

Because we are comparing the means for three normally distributed populations and all of the assumption required to apply ANOVA procedure are satisfied, we use the  $F$  distribution to make this test.

**Step 3. Determine the rejection and nonrejection regions.**

The significance level is .01. Because a one-way ANOVA test is always right-tailed, the area in the right tail of the  $F$  distribution curve is .01, which is the rejection region in Figure 12.3.

Next we need to know the degrees of freedom for the numerator and the denominator. In our example, the students were assigned to three different methods. As mentioned earlier, these methods are called treatments. The number of treatments is denoted by  $k$ . The total number of observations in all samples taken together is denoted by  $n$ . Then, the number of degrees of freedom for the numerator is equal to  $k - 1$  and the number of degrees of freedom for the denominator is equal to  $n - k$ . In our example, there are 3 treatments (methods of teaching) and 15 total observations (total number of students) in all 3 samples. Thus,

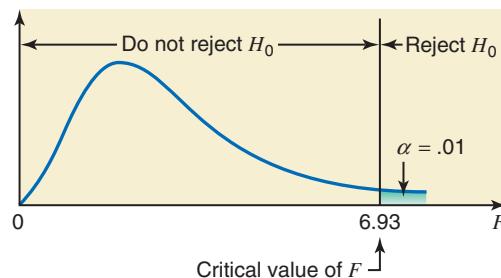
$$\text{Degrees of freedom for the numerator} = k - 1 = 3 - 1 = 2$$

$$\text{Degrees of freedom for the denominator} = n - k = 15 - 3 = 12$$

From Table VII of Appendix C, we find the critical value of  $F$  for 2  $df$  for the numerator, 12  $df$  for the denominator, and .01 area in the right tail of the  $F$  distribution curve. This value of  $F$  is 6.93, as shown in Figure 12.3.

Thus, we will fail to reject  $H_0$  if the calculated value of the test statistic  $F$  is less than 6.93, and we will reject  $H_0$  if it is 6.93 or larger.

**Figure 12.3** Critical value of  $F$  for  $df = (2, 12)$  and  $\alpha = .01$ .



**Step 4. Calculate the value of the test statistic.**

We computed the value of the test statistic  $F$  for these data in Example 12–2. This value is

$$F = 1.09$$

**Step 5. Make a decision.**

Because the value of the test statistic  $F = 1.09$  is less than the critical value of  $F = 6.93$ , it falls in the nonrejection region. Hence, we fail to reject the null hypothesis, and conclude that the means of the three populations are equal. In other words, the three different methods of teaching arithmetic do not seem to affect the mean scores of students. The difference in the three mean scores in the case of our three samples occurred only because of sampling error. ■

In Example 12–3, the sample sizes were the same for all treatments. Example 12–4 describes a case in which the sample sizes are not the same for all treatments.

## ■ EXAMPLE 12–4

*Performing a one-way ANOVA test: all samples not the same size.*

From time to time, unknown to its employees, the research department at Post Bank observes various employees for their work productivity. Recently this department wanted to check whether the four tellers at a branch of this bank serve, on average, the same number of customers per hour. The research manager observed each of the four tellers for a certain number

of hours. The following table gives the number of customers served by the four tellers during each of the observed hours.

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	



© YinYang/iStockphoto

At a 5% significance level, test the null hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

**Solution** To make a test about the equality of means of four populations, we follow our standard procedure with five steps.

**Step 1. State the null and alternative hypotheses.**

Let  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$ , and  $\mu_4$  be the mean number of customers served per hour by tellers A, B, C, and D, respectively. The null and alternative hypotheses are, respectively,

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  (The mean number of customers served per hour by each of the four tellers is the same.)

$H_1:$  Not all four population means are equal.

**Step 2. Select the distribution to use.**

Because we are testing for the equality of four means for four normally distributed populations and all of the assumptions required to apply ANOVA procedure hold true, we use the  $F$  distribution to make the test.

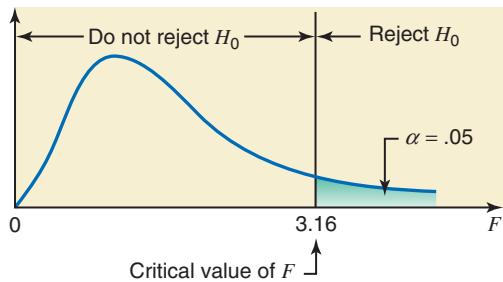
**Step 3. Determine the rejection and nonrejection regions.**

The significance level is .05, which means the area in the right tail of the  $F$  distribution curve is .05. In this example, there are 4 treatments (tellers) and 22 total observations in all four samples. Thus,

Degrees of freedom for the numerator =  $k - 1 = 4 - 1 = 3$

Degrees of freedom for the denominator =  $n - k = 22 - 4 = 18$

The critical value of  $F$  from Table VII for 3  $df$  for the numerator, 18  $df$  for the denominator, and .05 area in the right tail of the  $F$  distribution curve is 3.16. This value is shown in Figure 12.4.



**Figure 12.4** Critical value of  $F$  for  $df = (3, 18)$  and  $\alpha = .05$ .

**Step 4. Calculate the value of the test statistic.**

First we calculate SSB and SSW. Table 12.5 lists the numbers of customers served by the four tellers during the selected hours; the values of  $T_1$ ,  $T_2$ ,  $T_3$ , and  $T_4$ ; and the values of  $n_1$ ,  $n_2$ ,  $n_3$ , and  $n_4$ .

The values of  $\Sigma x$  and  $n$  are, respectively,

$$\Sigma x = T_1 + T_2 + T_3 + T_4 = 108 + 87 + 93 + 110 = 398$$

$$n = n_1 + n_2 + n_3 + n_4 = 5 + 6 + 6 + 5 = 22$$

**Table 12.5**

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	
$T_1 = 108$	$T_2 = 87$	$T_3 = 93$	$T_4 = 110$
$n_1 = 5$	$n_2 = 6$	$n_3 = 6$	$n_4 = 5$

The value of  $\Sigma x^2$  is calculated as follows:

$$\begin{aligned}\Sigma x^2 &= (19)^2 + (21)^2 + (26)^2 + (24)^2 + (18)^2 + (14)^2 + (16)^2 + (14)^2 \\ &\quad + (13)^2 + (17)^2 + (13)^2 + (11)^2 + (14)^2 + (21)^2 + (13)^2 \\ &\quad + (16)^2 + (18)^2 + (24)^2 + (19)^2 + (21)^2 + (26)^2 + (20)^2 \\ &= 7614\end{aligned}$$

Substituting all the values in the formulas for SSB and SSW, we obtain the following values of SSB and SSW:

$$\begin{aligned}SSB &= \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} \right) - \frac{(\Sigma x)^2}{n} \\ &= \left( \frac{(108)^2}{5} + \frac{(87)^2}{6} + \frac{(93)^2}{6} + \frac{(110)^2}{5} \right) - \frac{(398)^2}{22} = 255.6182 \\ SSW &= \Sigma x^2 - \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} \right) \\ &= 7614 - \left( \frac{(108)^2}{5} + \frac{(87)^2}{6} + \frac{(93)^2}{6} + \frac{(110)^2}{5} \right) = 158.2000\end{aligned}$$

Hence, the variance between samples MSB and the variance within samples MSW are, respectively,

$$MSB = \frac{SSB}{k-1} = \frac{255.6182}{4-1} = 85.2061$$

$$MSW = \frac{SSW}{n-k} = \frac{158.2000}{22-4} = 8.7889$$

The value of the test statistic  $F$  is given by the ratio of MSB and MSW, which is

$$F = \frac{MSB}{MSW} = \frac{85.2061}{8.7889} = 9.69$$

Writing the values of the various quantities in the ANOVA table, we obtain Table 12.6.

**Table 12.6** ANOVA Table for Example 12-4

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	3	255.6182	85.2061	
Within	18	158.2000	8.7889	
Total	21	413.8182		$F = \frac{85.2061}{8.7889} = 9.69$

**Step 5.** *Make a decision.*

Because the value of the test statistic  $F = 9.69$  is greater than the critical value of  $F = 3.16$ , it falls in the rejection region. Consequently, we reject the null hypothesis, and conclude that the mean number of customers served per hour by each of the four tellers is not the same. In other words, at least one of the four means is different from the other three. ■

**Note: What if the Sample Size is Large and the Number of df Are Not in the F Distribution Table?**

In this chapter, we used the  $F$  distribution to perform tests of hypothesis about the equality of population means for three or more populations. If we use technology to perform such tests, it does not matter how large the  $df$  (degrees of freedom) for the numerator and denominator are. However, when we use the  $F$  distribution table (Table VII in Appendix C), sometime we may not find the exact  $df$  for the numerator and/or for the denominator in this table, especially when either of these  $df$  are large. In such cases, we use the following alternative.

If the number of  $df$  is not given in the table, use the closest number of  $df$  that falls below the actual value of  $df$ . For example, if an ANOVA problem has 4  $df$  for the numerator and 47  $df$  for the denominator, we will use 4  $df$  for the numerator and 40  $df$  for the denominator to obtain the critical value of  $F$  from the table. As long as the number of  $df$  for the denominator is 3 or larger, the critical values of  $F$  become smaller as the numbers of  $df$  increase. Hence, whenever the observed value of  $F$  falls in the rejection region for a smaller number of  $df$ , it will fall in the rejection region for the larger number of  $df$  also.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 12.10** Briefly explain when a one-way ANOVA procedure is used to make a test of hypothesis.  
**12.11** Describe the assumptions that must hold true to apply the one-way analysis of variance procedure to test hypotheses.  
**12.12** Consider the following data obtained for two samples selected at random from two populations that are independent and normally distributed with equal variances.

Sample I	Sample II
32	27
26	35
31	33
20	40
27	38
34	31

- a. Calculate the means and standard deviations for these samples using the formulas from Chapter 3.  
b. Using the procedure learned in Section 10.2 of Chapter 10, test at a 1% significance level whether the means of the populations from which these samples are drawn are equal.  
c. Using the one-way ANOVA procedure, test at a 1% significance level whether the means of the populations from which these samples are drawn are equal.  
d. Are the conclusions reached in parts b and c the same?
- 12.13** Consider the following data obtained for two samples selected at random from two populations that are independent and normally distributed with equal variances.

Sample I	Sample II
14	11
21	8
11	12
9	18
13	15
20	7
17	6

- Calculate the means and standard deviations for these samples using the formulas from Chapter 3.
- Using the procedure learned in Section 10.2 of Chapter 10, test at a 5% significance level whether the means of the populations from which these samples are drawn are equal.
- Using the one-way ANOVA procedure, test at a 5% significance level whether the means of the populations from which these samples are drawn are equal.
- Are the conclusions reached in parts b and c the same?

**12.14** The following ANOVA table, based on information obtained for three samples selected from three independent populations that are normally distributed with equal variances, has a few missing values.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	2		19.2813	
Within		89.3677		$F = \text{_____} =$
Total	12			

- Find the missing values and complete the ANOVA table.
- Using  $\alpha = .01$ , what is your conclusion for the test with the null hypothesis that the means of the three populations are all equal against the alternative hypothesis that the means of the three populations are not all equal?

**12.15** The following ANOVA table, based on information obtained for four samples selected from four independent populations that are normally distributed with equal variances, has a few missing values.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between				
Within	15		9.2154	$F = \text{_____} = 4.07$
Total	18			

- Find the missing values and complete the ANOVA table.
- Using  $\alpha = .05$ , what is your conclusion for the test with the null hypothesis that the means of the four populations are all equal against the alternative hypothesis that the means of the four populations are not all equal?

## ■ APPLICATIONS

*For the following exercises assume that all the assumptions required to apply the one-way ANOVA procedure hold true.*

**12.16** A clothing store chain is having a sale based on the use of a coupon. The company is interested in knowing whether the wording of the coupon affects the number of units of the product purchased by customers. The company created four coupons for the same product, each with different wording. Four groups of 50 customers each were selected at random. Group 1 received the first version of the coupon; Group 2 received the second version; and so on. The units of the product purchased by each customer were recorded. The following ANOVA table contains some of the values from the analysis.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between				
Within		75127.856		$F = \text{_____} =$
Total		77478.291		

Assume that the four populations are normally distributed with equal variances.

- Find the missing values and complete the ANOVA table.
- What are the appropriate null and alternative hypotheses for this analysis? Using  $\alpha = .05$ , what is your conclusion about the equality of population means for all four coupons?

**12.17** People who have home gaming systems, such as Wii™, Playstation™, and Xbox™, are well aware of how quickly they need to replace the batteries in the remote controls. A consumer agency decided to test three major brands of alkaline batteries to determine whether they differ in their average lifetimes in these remotes. For each of the three brands, 10 sets of batteries were placed in the remotes, and people played games until the batteries died. The following ANOVA table contains some of the values from the analysis.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between		25711.60		
Within		22388.25		$F = \text{_____} =$
Total				

Assume that the three populations are normally distributed with equal variances.

- Find the missing values, and complete the ANOVA table.
- What are the appropriate hypotheses (null and alternative) for this analysis? Using  $\alpha = .05$ , what is your conclusion about the equality of the three population means?

**12.18** The recommended acidity levels for sweet white wines (e.g., certain Rieslings, Port, Eiswein, Muscat) is .70% to .85% ([www.grapestompers.com/articles/measure\\_acidity.htm](http://www.grapestompers.com/articles/measure_acidity.htm)). A vintner (winemaker) takes three random samples of Riesling from casks that are 15, 20, and 25 years old, respectively, and measures the acidity of each sample. The sample results are given in the table below.

15 years	20 years	25 years
.8036	.8109	.7735
.8001	.8246	.7813
.8291	.8245	.8052
.8077	.8070	.8000
.8298	.8023	.8091
.8126	.8182	.7952
.8169	.8265	.7882
.8066	.8262	.7789
.8142	.8048	.7976
.8197	.7995	.7918
.8129	.8102	.7850
.8133	.7957	.7801
.8251	.8164	.7843

- We are to test whether the mean acidity levels for all casks of Riesling are the same for the three different ages. Write the null and alternative hypotheses.
- Show the rejection and nonrejection regions on the  $F$  distribution curve for  $\alpha = .025$ .
- Calculate SSB, SSW, and SST.
- What are the degrees of freedom for the numerator and the denominator?
- Calculate the between-samples and within-samples variances.
- What is the critical value of  $F$  for  $\alpha = .025$ ?
- What is the calculated value of the test statistic  $F$ ?
- Write the ANOVA table for this exercise.
- Will you reject the null hypothesis stated in part a at a significance level of 2.5%?

**12.19** A local “pick-your-own” farmer decided to grow blueberries. The farmer purchased and planted eight plants of each of the four different varieties of highbush blueberries. The yield (in pounds) of each plant was measured in the upcoming year to determine whether the average yields were different for at least two of the four plant varieties. The yields of these plants of the four varieties are given in the following table.

Berkeley	5.13	5.36	5.20	5.15	4.96	5.14	5.54	5.22
Duke	5.31	4.89	5.09	5.57	5.36	4.71	5.13	5.30
Jersey	5.20	4.92	5.44	5.20	5.17	5.24	5.08	5.13
Sierra	5.08	5.30	5.43	4.99	4.89	5.30	5.35	5.26

- a. We are to test whether the mean yields for all such bushes of the four varieties are the same.  
Write the null and alternative hypotheses.
- b. What are the degrees of freedom for the numerator and the denominator?
- c. Calculate SSB, SSW, and SST.
- d. Show the rejection and nonrejection regions on the  $F$  distribution curve for  $\alpha = .01$ .
- e. Calculate the between-samples and within-samples variances.
- f. What is the critical value of  $F$  for  $\alpha = .01$ ?
- g. What is the calculated value of the test statistic  $F$ ?
- h. Write the ANOVA table for this exercise.
- i. Will you reject the null hypothesis stated in part a at a significance level of 1%?

**12.20** Surfer Dude swimsuit company plans to produce a new line of quick-dry swimsuits. Three textile companies are competing for the company's quick-dry fabric contract. To check the fabrics of the three companies, Surfer Dude selected 10 random swatches of fabric from each company, soaked them with water, and then measured the amount of time (in seconds) each swatch took to dry when exposed to sun and a temperature of 80°F. The following table contains the amount of time (in seconds) each of these swatches took to dry.

Company A	756	801	750	777	772	768	812	770	743	824
Company B	791	696	761	760	741	810	770	823	815	845
Company C	773	794	733	740	780	801	794	719	766	743

Using a 5% significance level, test the null hypothesis that the mean drying times for all such fabric produced by the three companies are the same.

**12.21** A university employment office wants to compare the time taken by graduates with three different majors to find their first full-time job after graduation. The following table lists the time (in days) taken to find their first full-time job after graduation for a random sample of eight business majors, seven computer science majors, and six engineering majors who graduated in May 2011.

Business	Computer Science	Engineering
208	156	126
162	113	275
240	281	363
180	128	146
148	305	298
312	147	392
176	232	
292		

At a 5% significance level, can you conclude that the mean time taken to find their first full-time job for all May 2011 graduates in these fields is the same?

**12.22** A consumer agency wanted to find out if the mean time taken by each of three brands of medicines to provide relief from a headache is the same. The first drug was administered to six randomly selected patients, the second to four randomly selected patients, and the third to five randomly selected patients. The following table gives the time (in minutes) taken by each patient to get relief from a headache after taking the medicine.

Drug I	Drug II	Drug III
25	15	44
38	21	39
42	19	54
65	25	58
47		73
52		

At a 2.5% significance level, will you conclude that the mean time taken to provide relief from a headache is the same for each of the three drugs?

**12.23** A large company buys thousands of lightbulbs every year. The company is currently considering four brands of lightbulbs to choose from. Before the company decides which lightbulbs to buy, it wants to investigate if the mean lifetimes of the four types of lightbulbs are the same. The company's research department randomly selected a few bulbs of each type and tested them. The following table lists the number of hours (in thousands) that each of the bulbs in each brand lasted before being burned out.

Brand I	Brand II	Brand III	Brand IV
23	19	23	26
24	23	27	24
19	18	25	21
26	24	26	29
22	20	23	28
23	22	21	24
25	19	27	28

At a 2.5% significance level, test the null hypothesis that the mean lifetime of bulbs for each of these four brands is the same.

## USES AND MISUSES... DO NOT BE LATE

Imagine that working at your company requires that staff travel frequently. You want to determine if the on-time performance of any one airline is sufficiently different from that of the remaining airlines to warrant a preferred status with your company. The local airport Web site publishes the scheduled and actual departure and arrival times for the four airlines that service it. You decide to perform an ANOVA test on the mean delay times for all airline carriers at the airport. The null hypothesis here is that the mean delay times for Airlines A, B, C, and D are all the same. The results of the ANOVA test tell you to accept the null hypothesis: All airline carriers have the same mean departure and arrival delay times, so that adopting a preferred status based on the on-time performance is not warranted.

When your boss tells you to redo your analysis, you should not be surprised. The choice to study flights only at the local air-

port was a good one because your company should be concerned about the performance of an airline at the most convenient airport. A regional airport will have a much different on-time performance profile than a large hub airport. By mixing both arrival and departure data, however, you violated the assumption that the populations are normally distributed. For arrival data, this assumption could be valid: The influence of high-altitude winds, local weather, and the fact that the arrival time is an estimate in the first place result in a distribution of arrival times around the predicted arrival times. However, departure delays are not normally distributed. Because a flight does not leave before its departure time but can leave after, departure delays are skewed to the right. As the statistical methods become more sophisticated, so do the assumptions regarding the characteristics of the data. Careful attention to these assumptions is required.

## Glossary

**Analysis of variance (ANOVA)** A statistical technique used to test whether the means of three or more populations are all equal.

**F distribution** A continuous distribution that has two parameters:  $df$  for the numerator and  $df$  for the denominator.

**Mean square between samples or MSB** A measure of the variation among the means of samples taken from different populations.

**Mean square within samples or MSW** A measure of the variation within the data of all samples taken from different populations.

**One-way ANOVA** The analysis of variance technique that analyzes one variable only.

**SSB** The sum of squares between samples. Also called the sum of squares of the factor or treatment.

**SST** The total sum of squares given by the sum of SSB and SSW.

**SSW** The sum of squares within samples. Also called the sum of squares of errors.

## Supplementary Exercises

*For the following exercises, assume that all the assumptions required to apply the one-way ANOVA procedure hold true.*

- 12.24** The following table lists the numbers of violent crimes reported to police on randomly selected days for this year. The data are taken from three large cities of about the same size.

City A	City B	City C
5	2	8
9	4	12
12	1	10
3	13	3
9	7	9
7	6	14
13		

Using a 5% significance level, test the null hypothesis that the mean number of violent crimes reported per day is the same for each of these three cities.

- 12.25** A music company collects data from customers who purchase CDs and MP3 downloads from them. Each person is asked to state his or her favorite musical genre from the following list: Classic Rock, Country, Hip-Hop/Rap, Jazz, Pop, and R&B. Random samples of customers were selected from each genre. Each customer was asked how much he or she spent (in dollars) on music purchases in the last month. The following table gives the information (in dollars) obtained from these customers.

Classic Rock	22	35	62	17	11	59	43
Country	60	36	59	27	32	56	
Hip-Hop/Rap	35	52	35	55	71	75	
Jazz	13	40	27	38	31	28	22
Pop	40	17	52	59	56	24	55
R&B	24	45	36	65	58	44	51

- a. At a 10% significance level, will you reject the null hypothesis that the average monthly expenditures of all customers in each of the six genres are the same?
- b. What is the Type I error in this case, and what is the probability of committing such an error? Explain.

- 12.26** A local car dealership is interested in determining how successful their salespeople are in turning a profit when selling a car. Specifically, they are interested in the average percentage of price markups earned on various car sales. The following table lists the percentages of price markups for a random sample of car sales by three salespeople at this dealership. Note that here the markups are calculated as follows. Suppose an auto dealer pays \$14,000 for a car and lists the sale price as \$20,000, which gives a markup of \$6000. If the car is sold for \$17,000, the markup percentage earned on this sale is 50% (\$3000 is half of \$6000).

Ira	23.2	26.9	27.3	34.1	30.7	31.6	43.8
Jim	19.6	41.2	60.3	34.3	52.0	23.3	39.1
Kelly	52.3	50.0	53.4	37.9	26.4	41.1	25.2

- a. Test at a 5% significance level whether the average markup percentage earned on all car sales is the same for Ira, Jim, and Kelly.
- b. What is the Type I error in this case, and what is the probability of committing such an error? Explain.

- 12.27** A farmer wants to test three brands of weight-gain diets for chickens to determine if the mean weight gain for each of these brands is the same. He selected 15 chickens and randomly put each of them

on one of these three brands of diet. The following table lists the weights (in pounds) gained by these chickens after a period of 1 month.

Brand A	Brand B	Brand C
.8	.6	1.2
1.3	1.3	.8
1.7	.6	.7
.9	.4	1.5
.6	.7	.9

- a. At a 1% significance level, can you conclude that the mean weight gain for all chickens is the same for each of these three diets?
- b. If you did not reject the null hypothesis in part a, explain the Type II error that you may have made in this case. Note that you cannot calculate the probability of committing a Type II error without additional information.

**12.28** An ophthalmologist is interested in determining whether a golfer's type of vision (far-sightedness, near-sightedness, no prescription) impacts how well he or she can judge distance. Random samples of golfers from these three groups (far-sightedness, near-sightedness, no prescription) were selected, and these golfers were blindfolded and taken to the same location on a golf course. Then each of them was asked to estimate the distance from this location to the pin at the end of the hole. The data (in yards) given in the following table represent how far off the estimates (let us call these errors) of these golfers were from the actual distance. A negative value implies that the person underestimated the distance, and a positive value implies that a person overestimated the distance.

Far-sighted	-11	-9	-8	-10	-3	-11	-8	1	-4
Near-sighted	-2	-5	-7	-8	-6	-9	2	-10	-10
No prescription	-5	1	0	4	3	-2	0	-8	

Test at a 1% significance level whether the average errors in predicting distance for all golfers of the three different vision types are the same.

**12.29** A resort area has three seafood restaurants, which employ students during the summer season. The local chamber of commerce took a random sample of five servers from each restaurant and recorded the tips they received on a recent Friday night. The results (in dollars) of the survey are shown in the table below. Assume that the Friday night for which the data were collected is typical of all Friday nights of the summer season.

Barzini's	Hwang's	Jack's
97	67	93
114	85	102
105	92	98
85	78	80
120	90	91

- a. Would a student seeking a server's job at one of these three restaurants conclude that the mean tips on a Friday night are the same for all three restaurants? Use a 5% level of significance.
- b. What will your decision be in part a if the probability of making a Type I error is zero? Explain.

**12.30** A student who has a 9 A.M. class on Monday, Wednesday, and Friday mornings wants to know if the mean time taken by students to find parking spaces just before 9 A.M. is the same for each of these three days of the week. He randomly selects five weeks and records the time taken to find a parking space on Monday, Wednesday, and Friday of each of these five weeks. These times (in minutes) are given in the following table. Assume that this student is representative of all students who need to find a parking space just before 9 A.M. on these three days.

Monday	Wednesday	Friday
6	9	3
12	12	2
15	5	10
14	14	7
10	13	5

At a 5% significance level, test the null hypothesis that the mean time taken to find a parking space just before 9 A.M. on Monday, Wednesday, and Friday is the same for all students.

## Advanced Exercises

**12.31** A billiards parlor in a small town is open just 4 days per week—Thursday through Sunday. Revenues vary considerably from day to day and week to week, so the owner is not sure whether some days of the week are more profitable than others. He takes random samples of 5 Thursdays, 5 Fridays, 5 Saturdays, and 5 Sundays from last year's records and lists the revenues for these 20 days. His bookkeeper finds the average revenue for each of the four samples, and then calculates  $\Sigma x^2$ . The results are shown in the following table. The value of the  $\Sigma x^2$  came out to be 2,890,000.

Day	Mean Revenue (\$)	Sample Size
Thursday	295	5
Friday	380	5
Saturday	405	5
Sunday	345	5

Assume that the revenues for each day of the week are normally distributed and that the standard deviations are equal for all four populations. At a 1% level of significance, can you conclude that the mean revenue is the same for each of the four days of the week?

**12.32** Suppose that you are a reporter for a newspaper whose editor has asked you to compare the hourly wages of carpenters, plumbers, electricians, and masons in your city. Since many of these workers are not union members, the wages vary considerably among individuals in the same trade.

- a. What data should you gather, and how would you collect them? What statistics would you present in your article, and how would you calculate them? Assume that your newspaper is not intended for technical readers.
- b. Suppose that you must submit your findings to a technical journal that requires statistical analysis of your data. If you want to determine whether or not the mean hourly wages are the same for all four trades, briefly describe how you would analyze the data. Assume that hourly wages in each trade are normally distributed and that the four variances are equal.

**12.33** The editor of an automotive magazine has asked you to compare the mean gas mileages of city driving for three makes of compact cars. The editor has made available to you one car of each of the three makes, three drivers, and a budget sufficient to buy gas and pay the drivers for approximately 500 miles of city driving for each car.

- a. Explain how you would conduct an experiment and gather the data for a magazine article comparing the gas mileage.
- b. Suppose that you wish to test the null hypothesis that the mean gas mileages of city driving are the same for all three makes. Outline the procedure for using your data to conduct this test. Assume that the assumptions for applying analysis of variance are satisfied.

**12.34** Do rock music CDs and country music CDs give the consumers the same amount of music listening time? A sample of 12 randomly selected single rock music CDs and a sample of 14 randomly selected single country music CDs have the following total lengths (in minutes).

Rock Music	Country Music
43.0	45.3
44.3	40.2
63.8	42.8
32.8	33.0
54.2	33.5
51.3	37.7
64.8	36.8
36.1	34.6
33.9	33.4
51.7	36.5
36.5	43.3
59.7	31.7
	44.0
	42.7

Assume that the two populations are normally distributed with equal standard deviations.

- Compute the value of the test statistic  $t$  for testing the null hypothesis that the mean lengths of the rock and country music single CDs are the same against the alternative hypothesis that these mean lengths are not the same. Use the value of this  $t$  statistic to compute the (approximate)  $p$ -value.
- Compute the value of the (one-way ANOVA) test statistic  $F$  for performing the test of equality of the mean lengths of the rock and country music single CDs and use it to find the (approximate)  $p$ -value.
- How do the test statistics in parts a and b compare? How do the  $p$ -values computed in parts a and b compare? Do you think that this is a coincidence, or will this always happen?

**12.35** Suppose you are performing a one-way ANOVA test with only the information given in the following table.

Source of Variation	Degrees of Freedom	Sum of Squares
Between	4	200
Within	45	3547

- Suppose the sample sizes for all groups are equal. How many groups are there? What are the group sample sizes?
- The  $p$ -value for the test of the equality of the means of all populations is calculated to be .6406. Suppose you plan to increase the sample sizes for all groups but keep them all equal. However, when you do this, the sum of squares within samples and the sum of squares between samples (magically) remain the same. What are the smallest sample sizes for groups that would make this result significant at a 5% significance level?

## Self-Review Test

- The  $F$  distribution is
  - continuous
  - discrete
  - neither
- The  $F$  distribution is always
  - symmetric
  - skewed to the right
  - skewed to the left
- The units of the  $F$  distribution, denoted by  $F$ , are always
  - nonpositive
  - positive
  - nonnegative
- The one-way ANOVA test analyzes only one
  - variable
  - population
  - sample

5. The one-way ANOVA test is always
  - a. right-tailed
  - b. left-tailed
  - c. two-tailed
6. For a one-way ANOVA with  $k$  treatments and  $n$  observations in all samples taken together, the degrees of freedom for the numerator are
  - a.  $k - 1$
  - b.  $n - k$
  - c.  $n - 1$
7. For a one-way ANOVA with  $k$  treatments and  $n$  observations in all samples taken together, the degrees of freedom for the denominator are
  - a.  $k - 1$
  - b.  $n - k$
  - c.  $n - 1$
8. The ANOVA test can be applied to compare
  - a. two or more population means
  - b. more than four population means only
  - c. more than three population means only
9. Briefly describe the assumptions that must hold true to apply the one-way ANOVA procedure as mentioned in this chapter.
10. A small college town has four pizza parlors that make deliveries. A student doing a research paper for her business management class decides to compare how promptly the four parlors deliver. On six randomly chosen nights, she orders a large pepperoni pizza from each establishment, then records the elapsed time until the pizza is delivered to her apartment. Assume that her apartment is approximately the same distance from the four pizza parlors. The following table shows the times (in minutes) for these deliveries. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Tony's	Luigi's	Angelo's	Kowalski's
20.0	22.1	22.3	23.9
24.0	27.0	26.0	24.1
18.3	20.2	24.0	25.8
22.0	32.0	30.1	29.0
20.8	26.0	28.0	25.0
19.0	24.8	25.8	24.2

- a. Using a 5% significance level, test the null hypothesis that the mean delivery time is the same for each of the four pizza parlors.
- b. Is it a Type I error or a Type II error that may have been committed in part a? Explain.

## Mini-Projects

### ■ MINI-PROJECT 12-1

Are some days of the week busier than others on the New York Stock Exchange (NYSE)? Record the number of shares traded on the NYSE each day for a period of 6 weeks (round the number of shares to the nearest million). You will have five samples—first for shares traded on 6 Mondays, second for shares traded on 6 Tuesdays, and so forth. Assume that these days make up random samples for the respective populations. Further assume that each of the five populations from which these five samples are taken follows a normal distribution with the same variance. Test if the mean number of shares traded is the same for each of the five populations. Use a 1% significance level.

### ■ MINI-PROJECT 12-2

Pick at least 30 students at random and divide them randomly into three groups (A, B, and C) of approximately equal size. Take the students one by one, ring a bell, and 17 seconds later ring another bell. Then ask the students to estimate the elapsed time between the first and second rings. For group A, tell each student before the experiment starts that people tend to underestimate the elapsed time. Tell each student in group B that people tend to overestimate the time. Do not make any such statement to the students in group C. Record the estimates for all students, and then conduct an appropriate hypothesis test to see if the mean estimates of elapsed time are all equal for the populations represented by these groups. Use the 5% level of significance, and assume that the three populations of elapsed time are normally distributed with equal standard deviations.

### MINI-PROJECT 12-3

Obtain a Wiffle™ ball, a plastic golf ball with dimples and no holes, and a plastic golf ball with holes instead of dimples. Throw each ball 20 times and measure the distances. Perform a hypothesis test to determine if the average distance is the same for each type of ball. Use a significance level of 5%.

### MINI-PROJECT 12-4

Using Data Set III (NFL Data) that is on the Web site of this text, take a random sample of 15 offensive linemen, 15 linebackers, and 15 defensive linemen.

- a. Perform an analysis of variance to test the null hypothesis that the mean heights of offensive linemen, linebackers, and defensive linemen are all the same versus the alternative hypothesis that at least two of the positions have different average heights. Use a 5% significance level.
- b. Create a stacked dotplot of the data. (See Chapter 2 in case you need a review of how to make a stacked dotplot.) Use this dotplot to explain the conclusion that you reached in part a.
- c. Using the stacked dotplot that you made in part b, discuss whether the underlying conditions are reasonable for this analysis. Specifically, discuss whether it seems reasonable to assume that the heights are normally distributed and that the variances of heights are equal for the three positions.

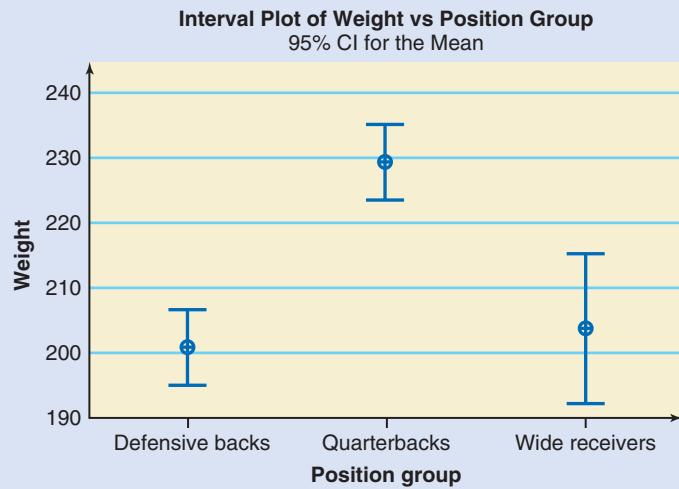
## DECIDE FOR YOURSELF DECIDING ABOUT WEIGHTS OF NFL PLAYERS

One-way ANOVA has given you a method/procedure to compare three or more means obtained from independent samples to make a decision about the corresponding population means. If you fail to reject the null hypothesis, you conclude that the assumption that the means of all populations under consideration are equal is a reasonable assumption. However, if you reject the null hypothesis, you conclude that at least two of the population means are different. Of course, there is still a glaring piece of information that you need in the latter case. If at least two means are different, which ones are different?

To determine which two means are different requires what is called a *pairwise comparison* procedure. This type of procedure compares each pair of means to determine whether or not they are equal. There are many such procedures available that can be used to make these pairwise comparisons. Some of these procedures are the Tukey HSD, Bonferroni, Scheffe, and Tamhane T2. To select the method that should be used depends on conditions such as whether or not the sample sizes are equal and whether or not using a pooled variance is reasonable.

There are a few informal (or *ad hoc*) methods that can be used to have an idea about what might happen with the *pairwise comparisons*. It is very important to note that the results from these procedures depend on how well the data meet the assumptions of an ANOVA, so these methods are not a substitute for a formal statistical process. These informal methods are simply graphical methods that can help you understand what is going on in a data set.

The accompanying figure gives a side-by-side plot of 95% confidence intervals for the mean weights of the groups of NFL players who play defensive back, quarterback, and wide receiver, respectively. The horizontal lines for each interval plot represent the ends of the interval, and the circle identifies the value of the sample mean for that group. Each of these confidence intervals is based on a random sample of 15 players selected from the corresponding group. It is important to note that the condition  $n/N \leq .05$  is not met for the



quarterback and wide receiver groups, but we will not address this issue at this time.

1. From the graph, we observe that players at one position seem to be significantly lighter or heavier, on average, than players at the other two positions. Identify the position for which this is the case, the specific difference (lighter or heavier), and what characteristic of the graph led you to make this conclusion.
2. The confidence interval for the wide receivers is much wider than the confidence intervals for the defensive backs and quarterbacks, yet the standard deviations for all players at these three positions are relatively close. What does this tell you about the effect of random sampling on summary statistics?
3. Suppose the confidence interval for the wide receivers remains of the same width but shifts up by 10 to 12 pounds. How would the results of the ANOVA change and why?

# TECHNOLOGY INSTRUCTION

## Analysis of Variance

### TI-84

- To perform a one-way analysis of variance on a collection of samples, store the sample data in lists.
- Select **STAT >TESTS >ANOVA(**.
- Enter the names of the lists, separated by commas, and then type a right parenthesis. Press **ENTER**. (See **Screen 12.1**.)
- The results include the *F* statistic for performing the test, as well as the *p*-value. (See **Screen 12.2**.)

ANOVA(L<sub>1</sub>,L<sub>2</sub>,L<sub>3</sub>)

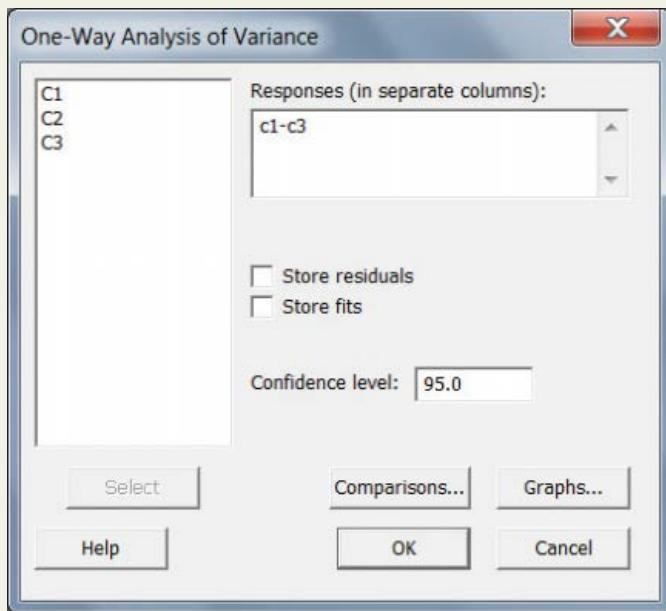
Screen 12.1

One-way ANOVA  
*F*=4.341586944  
*P*=.0325375737  
 Factor  
*df*=2  
*SS*=171.444444  
 $\downarrow$  *MS*=85.7222222

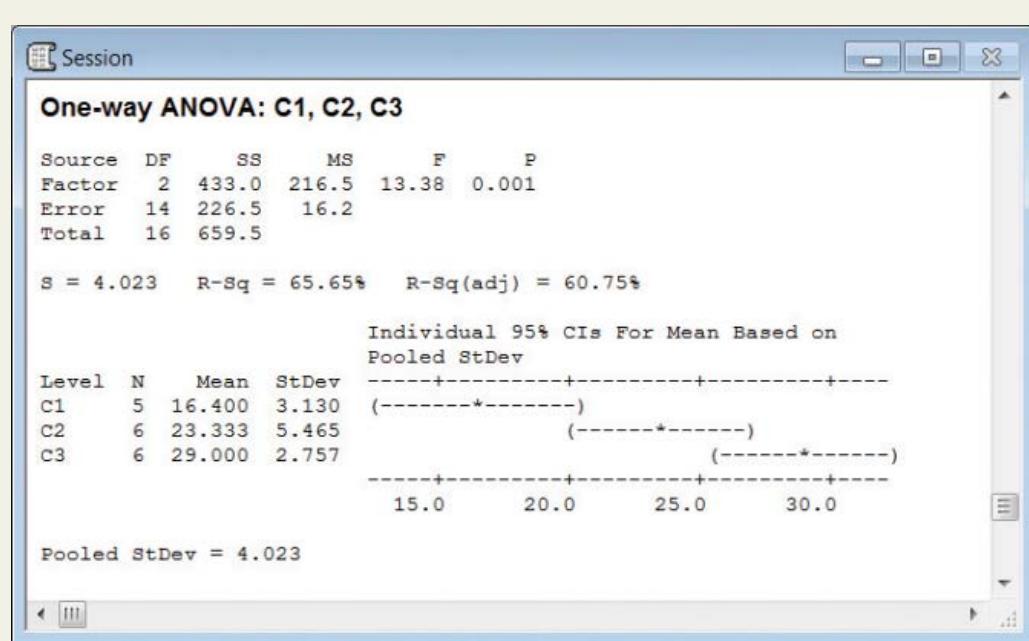
Screen 12.2

### Minitab

- To perform a one-way analysis of variance on a collection of samples, enter the data for samples into columns.
- Select **Stat>ANOVA>One-way (Unstacked)**.
- Enter the names of the columns and select **OK**. (See **Screen 12.3**.)
- The results include the components of the ANOVA, including the *p*-value, as well as the 95% confidence interval for each population mean using a pooled estimate of the variance. (See **Screen 12.4**.)



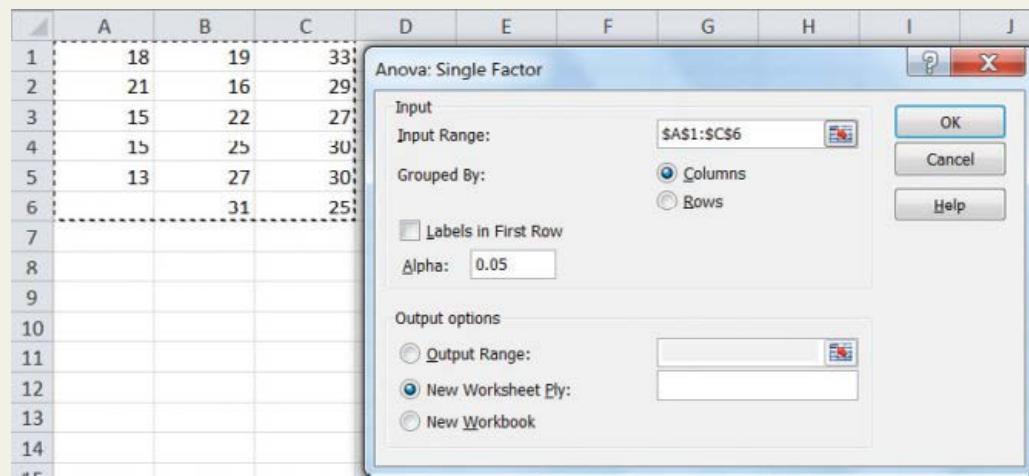
Screen 12.3



Screen 12.4

**Excel**

1. Click the **Data** tab. Click the **Data Analysis** button within the **Analysis** group. From the **Data Analysis** window that appears, select **Anova: Single Factor**.
2. Enter the location of the data in the **Input Range** box. Click the button to identify whether the data for each sample are given in columns or rows. Enter the significance level, as a decimal, in the **Alpha** box. If your data have labels in the top row (or in the left column), click the **Labels** box. Choose how you wish the output to appear. (See **Screen 12.5**.) Click **OK**.



Screen 12.5

3. The output contains the summary statistics for each group, as well as the ANOVA table. In addition to all of the standard items, the ANOVA table contains the critical value of F for the given significance level and degrees of freedom. (See **Screen 12.6**.)

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	Column 1	5	82	16.4	9.8		
6	Column 2	6	140	23.33333	29.86667		
7	Column 3	6	174	29	7.6		
8							
9							
10	ANOVA						
11	<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
12	Between Groups	432.9961	2	216.498	13.37981	0.000564	3.738892
13	Within Groups	226.5333	14	16.18095			
14							
15	Total	659.5294	16				

Screen 12.6

## TECHNOLOGY ASSIGNMENTS

---

**TA12.1** Solve Exercise 12.18.

**TA12.2** Solve Exercise 12.19.

**TA12.3** Solve Exercise 12.26.



CHAPTER  
**13**

## Simple Linear Regression

Are the heights and weights of persons related? Does a person's weight depend on his/her height? If yes, what is the change in the weight of a person, on average, for every one inch increase in height? What is this rate of change for National Football League players? (See Case Study 13-1.)

This chapter considers the relationship between two variables in two ways: (1) by using regression analysis and (2) by computing the correlation coefficient. By using the regression model, we can evaluate the magnitude of change in one variable due to a certain change in another variable. For example, an economist can estimate the amount of change in food expenditure due to a certain change in the income of a household by using the regression model. A sociologist may want to estimate the increase in the crime rate due to a particular increase in the unemployment rate. Besides answering these questions, a regression model also helps predict the value of one variable for a given value of another variable. For example, by using the regression line, we can predict the (approximate) food expenditure of a household with a given income.

The correlation coefficient, on the other hand, simply tells us how strongly two variables are related. It does not provide any information about the size of the change in one variable as a result of a certain change in the other variable. For example, the correlation coefficient tells us how strongly income and food expenditure or crime rate and unemployment rate are related.

**13.1 Simple Linear Regression**

**Case Study 13-1 Regression of Weights on Heights for NFL Players**

**13.2 Standard Deviation of Errors and Coefficient of Determination**

**13.3 Inferences About  $B$**

**13.4 Linear Correlation**

**13.5 Regression Analysis: A Complete Example**

**13.6 Using the Regression Model**

## 13.1 Simple Linear Regression

Only simple linear regression will be discussed in this chapter.<sup>1</sup> In the next two subsections the meaning of the words *simple* and *linear* as used in *simple linear regression* is explained.

### 13.1.1 Simple Regression

Let us return to the example of an economist investigating the relationship between food expenditure and income. What factors or variables does a household consider when deciding how much money it should spend on food every week or every month? Certainly, income of the household is one factor. However, many other variables also affect food expenditure. For instance, the assets owned by the household, the size of the household, the preferences and tastes of household members, and any special dietary needs of household members are some of the variables that influence a household's decision about food expenditure. These variables are called **independent** or **explanatory variables** because they all vary independently, and they explain the variation in food expenditures among different households. In other words, these variables explain why different households spend different amounts of money on food. Food expenditure is called the **dependent variable** because it depends on the independent variables. Studying the effect of two or more independent variables on a dependent variable using regression analysis is called **multiple regression**. However, if we choose only one (usually the most important) independent variable and study the effect of that single variable on a dependent variable, it is called a **simple regression**. Thus, a simple regression includes only two variables: one independent and one dependent. Note that whether it is a simple or a multiple regression analysis, it always includes one and only one dependent variable. It is the number of independent variables that changes in simple and multiple regressions.

#### Definition

**Simple Regression** A regression model is a mathematical equation that describes the relationship between two or more variables. A *simple regression* model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one that explains the variation in the dependent variable.

### 13.1.2 Linear Regression

The relationship between two variables in a regression analysis is expressed by a mathematical equation called a **regression equation** or **model**. A regression equation, when plotted, may assume one of many possible shapes, including a straight line. A regression equation that gives a straight-line relationship between two variables is called a **linear regression model**; otherwise, the model is called a **nonlinear regression model**. In this chapter, only linear regression models are studied.

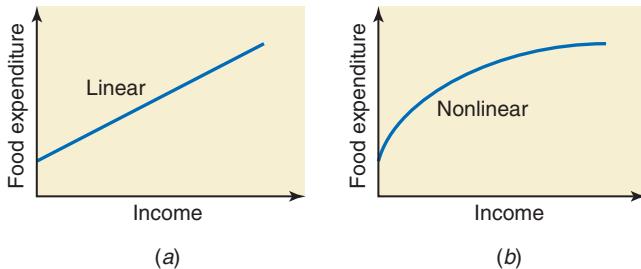
#### Definition

**Linear Regression** A (simple) regression model that gives a straight-line relationship between two variables is called a *linear regression* model.

The two diagrams in Figure 13.1 show a linear and a nonlinear relationship between the dependent variable food expenditure and the independent variable income. A linear relationship

<sup>1</sup>The term *regression* was first used by Sir Francis Galton (1822–1911), who studied the relationship between the heights of children and the heights of their parents.

between income and food expenditure, shown in Figure 13.1a, indicates that as income increases, the food expenditure always increases at a constant rate. A nonlinear relationship between income and food expenditure, as depicted in Figure 13.1b, shows that as income increases, the food expenditure increases, although, after a point, the rate of increase in food expenditure is lower for every subsequent increase in income.



**Figure 13.1** Relationship between food expenditure and income. (a) Linear relationship. (b) Nonlinear relationship.

The **equation of a linear relationship** between two variables  $x$  and  $y$  is written as

$$y = a + bx$$

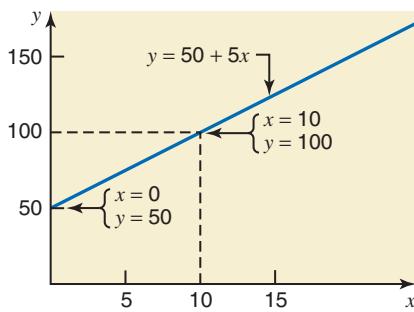
Each set of values of  $a$  and  $b$  gives a different straight line. For instance, when  $a = 50$  and  $b = 5$ , this equation becomes

$$y = 50 + 5x$$

To plot a straight line, we need to know two points that lie on that line. We can find two points on a line by assigning any two values to  $x$  and then calculating the corresponding values of  $y$ . For the equation  $y = 50 + 5x$ :

1. When  $x = 0$ , then  $y = 50 + 5(0) = 50$ .
2. When  $x = 10$ , then  $y = 50 + 5(10) = 100$ .

These two points are plotted in Figure 13.2. By joining these two points, we obtain the line representing the equation  $y = 50 + 5x$ .



**Figure 13.2** Plotting a linear equation.

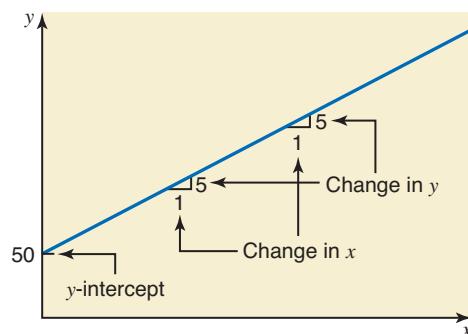
Note that in Figure 13.2 the line intersects the  $y$  (vertical) axis at 50. Consequently, 50 is called the  **$y$ -intercept**. The  $y$ -intercept is given by the constant term in the equation. It is the value of  $y$  when  $x$  is zero.

In the equation  $y = 50 + 5x$ , 5 is called the **coefficient of  $x$**  or the **slope** of the line. It gives the amount of change in  $y$  due to a change of one unit in  $x$ . For example:

$$\text{If } x = 10, \text{ then } y = 50 + 5(10) = 100.$$

$$\text{If } x = 11, \text{ then } y = 50 + 5(11) = 105.$$

Hence, as  $x$  increases by 1 unit (from 10 to 11),  $y$  increases by 5 units (from 100 to 105). This is true for any value of  $x$ . Such changes in  $x$  and  $y$  are shown in Figure 13.3.

**Figure 13.3** *y*-intercept and slope of a line.

In general, when an equation is written in the form

$$y = a + bx$$

$a$  gives the *y*-intercept and  $b$  represents the slope of the line. In other words,  $a$  represents the point where the line intersects the *y*-axis, and  $b$  gives the amount of change in *y* due to a change of one unit in *x*. Note that  $b$  is also called the coefficient of *x*.

### 13.1.3 Simple Linear Regression Model

In a regression model, the independent variable is usually denoted by *x*, and the dependent variable is usually denoted by *y*. The *x* variable, with its coefficient, is written on the right side of the  $=$  sign, whereas the *y* variable is written on the left side of the  $=$  sign. The *y*-intercept and the slope, which we earlier denoted by  $a$  and  $b$ , respectively, can be represented by any of the many commonly used symbols. Let us denote the *y*-intercept (which is also called the *constant term*) by  $A$ , and the slope (or the coefficient of the *x* variable) by  $B$ . Then, our simple linear regression model is written as

$$\begin{array}{c} \text{Constant term or } y\text{-intercept} \quad \text{Slope} \\ \downarrow \qquad \qquad \downarrow \\ y = A + Bx \\ \uparrow \qquad \qquad \uparrow \\ \text{Dependent variable} \qquad \text{Independent variable} \end{array} \quad (1)$$

In model (1),  $A$  gives the value of *y* for  $x = 0$ , and  $B$  gives the change in *y* due to a change of one unit in *x*.

Model (1) is called a **deterministic model**. It gives an **exact relationship** between *x* and *y*. This model simply states that *y* is determined exactly by *x*, and for a given value of *x* there is one and only one (unique) value of *y*.

However, in many cases the relationship between variables is not exact. For instance, if *y* is food expenditure and *x* is income, then model (1) would state that food expenditure is determined by income only and that all households with the same income spend the same amount on food. As mentioned earlier, however, food expenditure is determined by many variables, only one of which is included in model (1). In reality, different households with the same income spend different amounts of money on food because of the differences in the sizes of the household, the assets they own, and their preferences and tastes. Hence, to take these variables into consideration and to make our model complete, we add another term to the right side of model (1). This term is called the **random error term**. It is denoted by  $\epsilon$  (Greek letter *epsilon*). The complete regression model is written as

$$\begin{array}{c} y = A + Bx + \epsilon \\ \uparrow \\ \text{Random error term} \end{array} \quad (2)$$

The regression model (2) is called a **probabilistic model** or a **statistical relationship**.

### Definition

**Equation of a Regression Model** In the *regression model*  $y = A + Bx + \epsilon$ ,  $A$  is called the *y-intercept* or *constant term*,  $B$  is the *slope*, and  $\epsilon$  is the *random error term*. The dependent and independent variables are  $y$  and  $x$ , respectively.

The random error term  $\epsilon$  is included in the model to represent the following two phenomena:

1. *Missing or omitted variables.* As mentioned earlier, food expenditure is affected by many variables other than income. The random error term  $\epsilon$  is included to capture the effect of all those missing or omitted variables that have not been included in the model.
2. *Random variation.* Human behavior is unpredictable. For example, a household may have many parties during one month and spend more than usual on food during that month. The same household may spend less than usual during another month because it spent quite a bit of money to buy furniture. The variation in food expenditure for such reasons may be called random variation.

In model (2),  $A$  and  $B$  are the **population parameters**. The regression line obtained for model (2) by using the population data is called the **population regression line**. The values of  $A$  and  $B$  in the population regression line are called the **true values of the y-intercept and slope**, respectively.

However, population data are difficult to obtain. As a result, we almost always use sample data to estimate model (2). The values of the *y-intercept* and *slope* calculated from sample data on  $x$  and  $y$  are called the **estimated values of  $A$  and  $B$**  and are denoted by  $a$  and  $b$ , respectively. Using  $a$  and  $b$ , we write the estimated regression model as

$$\hat{y} = a + bx \quad (3)$$

where  $\hat{y}$  (read as *y hat*) is the **estimated or predicted value of  $y$**  for a given value of  $x$ . Equation (3) is called the **estimated regression model**; it gives the **regression of  $y$  on  $x$** .

### Definition

**Estimates of  $A$  and  $B$**  In the model  $\hat{y} = a + bx$ ,  $a$  and  $b$ , which are calculated using sample data, are called the *estimates of  $A$  and  $B$* , respectively.

#### 13.1.4 Scatter Diagram

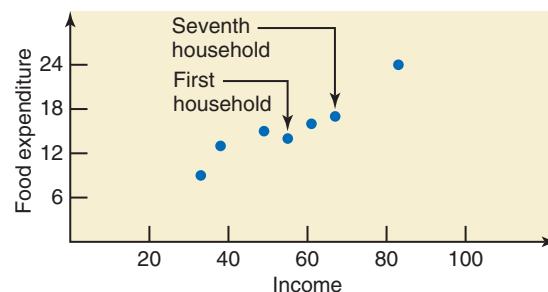
Suppose we take a sample of seven households from a small city and collect information on their incomes and food expenditures for the last month. The information obtained (in hundreds of dollars) is given in Table 13.1.

**Table 13.1** Incomes and Food Expenditures of Seven Households

Income	Food Expenditure
55	14
83	24
38	13
61	16
33	9
49	15
67	17

In Table 13.1, we have a pair of observations for each of the seven households. Each pair consists of one observation on income and a second on food expenditure. For example, the first household's income for the last month was \$5500 and its food expenditure was \$1400. By plotting all seven pairs of values, we obtain a **scatter diagram** or **scatterplot**. Figure 13.4 gives the scatter diagram for the data of Table 13.1. Each dot in this diagram represents one household. A scatter diagram is helpful in detecting a relationship between two variables. For example, by looking at the scatter diagram of Figure 13.4, we can observe that there exists a strong linear relationship between food expenditure and income. If a straight line is drawn through the points, the points will be scattered closely around the line.

**Figure 13.4** Scatter diagram.



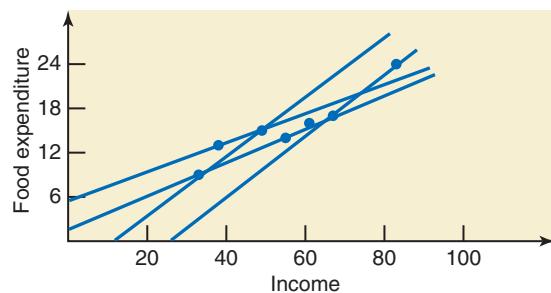
### Definition

**Scatter Diagram** A plot of paired observations is called a *scatter diagram*.

As shown in Figure 13.5, a large number of straight lines can be drawn through the scatter diagram of Figure 13.4. Each of these lines will give different values for  $a$  and  $b$  of model (3).

In regression analysis, we try to find a line that best fits the points in the scatter diagram. Such a line provides a best possible description of the relationship between the dependent and independent variables. The **least squares method**, discussed in the next section, gives such a line. The line obtained by using the least squares method is called the **least squares regression line**.

**Figure 13.5** Scatter diagram and straight lines.



### 13.1.5 Least Squares Regression Line

The value of  $y$  obtained for a member from the survey is called the **observed or actual value of  $y$** . As mentioned earlier in this section, the value of  $y$ , denoted by  $\hat{y}$ , obtained for a given  $x$  by using the regression line is called the **predicted value of  $y$** . The random error  $\epsilon$  denotes the difference between the actual value of  $y$  and the predicted value of  $y$  for population data. For example, for a given household,  $\epsilon$  is the difference between what this household actually spent on food during the last month and what is predicted using the population regression line. The  $\epsilon$  is also called the *residual* because it measures the surplus (positive or negative) of actual food expenditure over what is predicted by using the regression model. If we estimate model (2) by

using sample data, the difference between the actual  $y$  and the predicted  $\hat{y}$  based on this estimation cannot be denoted by  $\epsilon$ . The random error for the sample regression model is denoted by  $e$ . Thus,  $e$  is an estimator of  $\epsilon$ . If we estimate model (2) using sample data, then the value of  $e$  is given by

$$e = \text{Actual food expenditure} - \text{Predicted food expenditure} = y - \hat{y}$$

In Figure 13.6,  $e$  is the vertical distance between the actual position of a household and the point on the regression line. Note that in such a diagram, we always measure the dependent variable on the vertical axis and the independent variable on the horizontal axis.

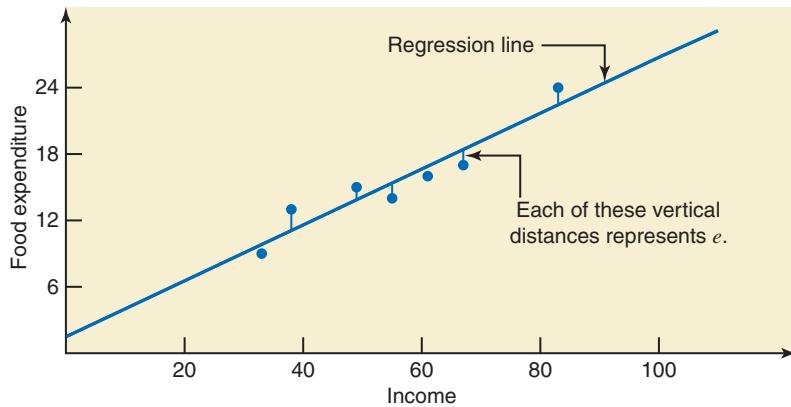


Figure 13.6 Regression line and random errors.

The value of an error is positive if the point that gives the actual food expenditure is above the regression line and negative if it is below the regression line. *The sum of these errors is always zero.* In other words, the sum of the actual food expenditures for seven households included in the sample will be the same as the sum of the food expenditures predicted by the regression model. Thus,

$$\sum e = \sum(y - \hat{y}) = 0$$

Hence, to find the line that best fits the scatter of points, we cannot minimize the sum of errors. Instead, we minimize the **error sum of squares**, denoted by **SSE**, which is obtained by adding the squares of errors. Thus,

$$\text{SSE} = \sum e^2 = \sum(y - \hat{y})^2$$

The least squares method gives the values of  $a$  and  $b$  for model (3) such that the sum of squared errors (SSE) is minimum.

**Error Sum of Squares (SSE)** The *error sum of squares*, denoted by SSE, is

$$\text{SSE} = \sum e^2 = \sum(y - \hat{y})^2$$

The values of  $a$  and  $b$  that give the minimum SSE are called the *least squares estimates* of  $A$  and  $B$ , and the regression line obtained with these estimates is called the *least squares line*.

**The Least Squares Line** For the least squares regression line  $\hat{y} = a + bx$ ,

$$b = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

$$\text{where } \text{SS}_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \quad \text{and} \quad \text{SS}_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

and SS stands for “sum of squares.” The least squares regression line  $\hat{y} = a + bx$  is also called the regression of  $y$  on  $x$ .

The least squares values of  $a$  and  $b$  are computed using the formulas just given.<sup>2</sup> These formulas are for estimating a sample regression line. Suppose we have access to a population data set. We can find the population regression line by using the same formulas with a little adaptation. If we have access to population data, we replace  $a$  by  $A$ ,  $b$  by  $B$ , and  $n$  by  $N$  in these formulas, and use the values of  $\Sigma x$ ,  $\Sigma y$ ,  $\Sigma xy$ , and  $\Sigma x^2$  calculated for population data to make the required computations. The population regression line is written as

$$\mu_{y|x} = A + Bx$$

where  $\mu_{y|x}$  is read as “the mean value of  $y$  for a given  $x$ .” When plotted on a graph, the points on this population regression line give the average values of  $y$  for the corresponding values of  $x$ . These average values of  $y$  are denoted by  $\mu_{y|x}$ .

Example 13–1 illustrates how to estimate a regression line for sample data.

### ■ EXAMPLE 13–1

*Estimating the least squares regression line.*



© Troels Graugaard/iStockphoto

Find the least squares regression line for the data on incomes and food expenditures of the seven households given in Table 13.1. Use income as an independent variable and food expenditure as a dependent variable.

**Solution** We are to find the values of  $a$  and  $b$  for the regression model  $\hat{y} = a + bx$ . Table 13.2 shows the calculations required for the computation of  $a$  and  $b$ . We denote the independent variable (income) by  $x$  and the dependent variable (food expenditure) by  $y$ , both in hundreds of dollars.

**Table 13.2**

Income $x$	Food Expenditure $y$	$xy$	$x^2$
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma xy = 6403$	$\Sigma x^2 = 23,058$

The following steps are performed to compute  $a$  and  $b$ .

**Step 1.** Compute  $\Sigma x$ ,  $\Sigma y$ ,  $\bar{x}$ , and  $\bar{y}$ .

$$\Sigma x = 386, \quad \Sigma y = 108$$

$$\bar{x} = \Sigma x/n = 386/7 = 55.1429$$

$$\bar{y} = \Sigma y/n = 108/7 = 15.4286$$

**Step 2.** Compute  $\Sigma xy$  and  $\Sigma x^2$ .

To calculate  $\Sigma xy$ , we multiply the corresponding values of  $x$  and  $y$ . Then, we sum all the products. The products of  $x$  and  $y$  are recorded in the third column of Table 13.2. To compute  $\Sigma x^2$ , we square each of the  $x$  values and then add them. The squared values of  $x$  are listed in the fourth column of Table 13.2. From these calculations,

$$\Sigma xy = 6403 \quad \text{and} \quad \Sigma x^2 = 23,058$$

<sup>2</sup>The values of  $SS_{xy}$  and  $SS_{xx}$  can also be obtained by using the following basic formulas:

$$SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) \quad \text{and} \quad SS_{xx} = \Sigma(x - \bar{x})^2$$

However, these formulas take longer to make calculations.

**Step 3.** Compute  $SS_{xy}$  and  $SS_{xx}$ :

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 6403 - \frac{(386)(108)}{7} = 447.5714$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 23,058 - \frac{(386)^2}{7} = 1772.8571$$

**Step 4.** Compute  $a$  and  $b$ :

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{447.5714}{1772.8571} = .2525$$

$$a = \bar{y} - b\bar{x} = 15.4286 - (.2525)(55.1429) = 1.5050$$

Thus, our estimated regression model  $\hat{y} = a + bx$  is

$$\hat{y} = 1.5050 + .2525x$$

This regression line is called the least squares regression line. It gives the *regression of food expenditure on income*.

Note that we have rounded all calculations to four decimal places. We can round the values of  $a$  and  $b$  in the regression equation to two decimal places, but we do not do this here because we will use this regression equation for prediction and estimation purposes later. ■

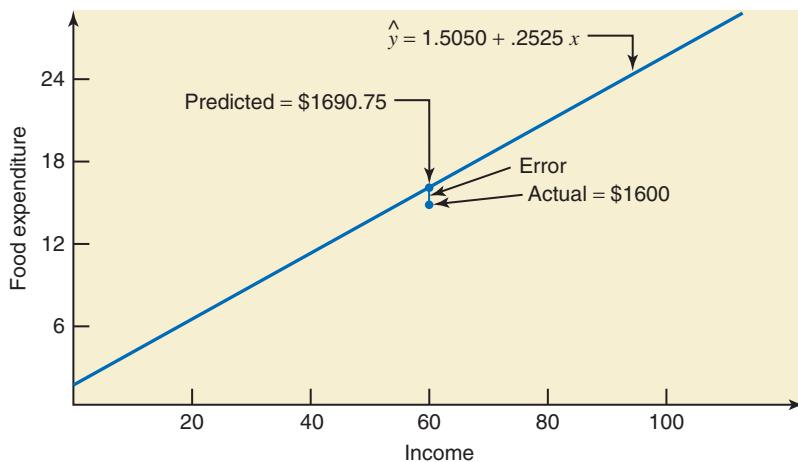
Using this estimated regression model, we can find the predicted value of  $y$  for any specific value of  $x$ . For instance, suppose we randomly select a household whose monthly income is \$6100, so that  $x = 61$  (recall that  $x$  denotes income in hundreds of dollars). The predicted value of food expenditure for this household is

$$\hat{y} = 1.5050 + (.2525)(61) = \$16.9075 \text{ hundred} = \$1690.75$$

In other words, based on our regression line, we predict that a household with a monthly income of \$6100 is expected to spend \$1690.75 per month on food. This value of  $\hat{y}$  can also be interpreted as a point estimator of the mean value of  $y$  for  $x = 61$ . Thus, we can state that, on average, all households with a monthly income of \$6100 spend about \$1690.75 per month on food.

In our data on seven households, there is one household whose income is \$6100. The actual food expenditure for that household is \$1600 (see Table 13.1). The difference between the actual and predicted values gives the error of prediction. Thus, the error of prediction for this household, which is shown in Figure 13.7, is

$$e = y - \hat{y} = 16 - 16.9075 = -\$90.75 \text{ hundred} = -\$90.75$$



**Figure 13.7** Error of prediction.

Therefore, the error of prediction is  $-\$90.75$ . The negative error indicates that the predicted value of  $y$  is greater than the actual value of  $y$ . Thus, if we use the regression model, this household's food expenditure is overestimated by \$90.75.

### 13.1.6 Interpretation of $a$ and $b$

How do we interpret  $a = 1.5050$  and  $b = .2525$  obtained in Example 13–1 for the regression of food expenditure on income? A brief explanation of the  $y$ -intercept and the slope of a regression line was given in Section 13.1.2. Below we explain the meaning of  $a$  and  $b$  in more detail.

#### Interpretation of $a$

Consider a household with zero income. Using the estimated regression line obtained in Example 13–1, we get the predicted value of  $y$  for  $x = 0$  as

$$\hat{y} = 1.5050 + .2525(0) = \$1.5050 \text{ hundred} = \$150.50$$

Thus, we can state that a household with no income is expected to spend \$150.50 per month on food. Alternatively, we can also state that the point estimate of the average monthly food expenditure for all households with zero income is \$150.50. Note that here we have used  $\hat{y}$  as a point estimate of  $\mu_{y|x}$ . Thus,  $a = 150.50$  gives the predicted or mean value of  $y$  for  $x = 0$  based on the regression model estimated for the sample data.

However, we should be very careful when making this interpretation of  $a$ . In our sample of seven households, the incomes vary from a minimum of \$3300 to a maximum of \$8300. (Note that in Table 13.1, the minimum value of  $x$  is 33 and the maximum value is 83.) Hence, our regression line is valid only for the values of  $x$  between 33 and 83. If we predict  $y$  for a value of  $x$  outside this range, the prediction usually will not hold true. Thus, since  $x = 0$  is outside the range of household incomes that we have in the sample data, the prediction that a household with zero income spends \$150.50 per month on food does not carry much credibility. The same is true if we try to predict  $y$  for an income greater than \$8300, which is the maximum value of  $x$  in Table 13.1.

#### Interpretation of $b$

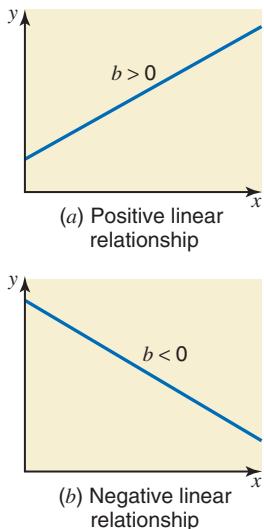
The value of  $b$  in a regression model gives the change in  $y$  (dependent variable) due to a change of one unit in  $x$  (independent variable). For example, by using the regression equation obtained in Example 13–1, we see:

$$\text{When } x = 50, \quad \hat{y} = 1.5050 + .2525(50) = 14.1300$$

$$\text{When } x = 51, \quad \hat{y} = 1.5050 + .2525(51) = 14.3825$$

Hence, when  $x$  increased by one unit, from 50 to 51,  $\hat{y}$  increased by  $14.3825 - 14.1300 = .2525$ , which is the value of  $b$ . Because our unit of measurement is hundreds of dollars, we can state that, on average, a \$100 increase in income will result in a \$25.25 increase in food expenditure. We can also state that, on average, a \$1 increase in income of a household will increase the food expenditure by \$.2525. Note the phrase “on average” in these statements. The regression line is seen as a measure of the mean value of  $y$  for a given value of  $x$ . If one household’s income is increased by \$100, that household’s food expenditure may or may not increase by \$25.25. However, if the incomes of all households are increased by \$100 each, the average increase in their food expenditures will be very close to \$25.25.

Note that when  $b$  is positive, an increase in  $x$  will lead to an increase in  $y$ , and a decrease in  $x$  will lead to a decrease in  $y$ . In other words, when  $b$  is positive, the movements in  $x$  and  $y$  are in the same direction. Such a relationship between  $x$  and  $y$  is called a **positive linear relationship**. The regression line in this case slopes upward from left to right. On the other hand, if the value of  $b$  is negative, an increase in  $x$  will lead to a decrease in  $y$ , and a decrease in  $x$  will cause an increase in  $y$ . The changes in  $x$  and  $y$  in this case are in opposite directions. Such a relationship between  $x$  and  $y$  is called a **negative linear relationship**. The regression line in this case slopes downward from left to right. The two diagrams in Figure 13.8 show these two cases.



**Figure 13.8** Positive and negative linear relationships between  $x$  and  $y$ .

#### Remember ▶

For a regression model,  $b$  is computed as  $b = SS_{xy}/SS_{xx}$ . The value of  $SS_{xx}$  is always positive, and that of  $SS_{xy}$  can be positive or negative. Hence, the sign of  $b$  depends on the sign of  $SS_{xy}$ . If  $SS_{xy}$  is positive (as in our example on the incomes and food expenditures of seven households), then  $b$  will be positive, and if  $SS_{xy}$  is negative, then  $b$  will be negative.

Case Study 13–1 illustrates the difference between the population regression line and a sample regression line.

## REGRESSION OF WEIGHTS ON HEIGHTS FOR NFL PLAYERS

Data Set III that accompanies this text lists many characteristics of National Football League (NFL) players who were on the rosters of all NFL teams as of October 31, 2011. These data comprise the population of NFL players for that point in time. We postulate the following simple linear regression model for these data:

$$y = A + Bx + \epsilon$$

where  $y$  is the weight (in pounds) and  $x$  is the height (in inches) of an NFL player.

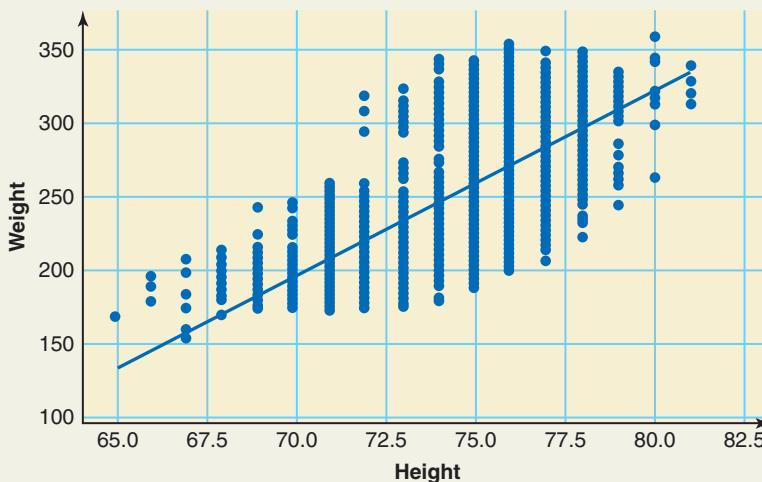
Using the population data that contain 1874 players, we obtain the following regression line:

$$\mu_{y|x} = -690 + 12.7x$$

This equation gives the population regression line because it is obtained by using the population data. (Note that in the population regression line we write  $\mu_{y|x}$  instead of  $\hat{y}$ .) Thus, the true values of  $A$  and  $B$  are, respectively,

$$A = -690 \quad \text{and} \quad B = 12.7$$

The value of  $B$  indicates that for every 1-inch increase in the height of an NFL player, weight increases on average by 12.7 pounds. However,  $A = -690$  does not make any sense. It states that the weight of a player with zero height is  $-690$  pounds. (Recall from Section 13.1.6 that we must be very careful if and when we apply the regression equation to predict  $y$  for values of  $x$  outside the range of data used to find the regression line.) Figure 13.9 gives the scatter diagram and the regression line for the heights and weights of all NFL players.



**Figure 13.9** Scatter diagram for the data on heights and weights of all NFL players.

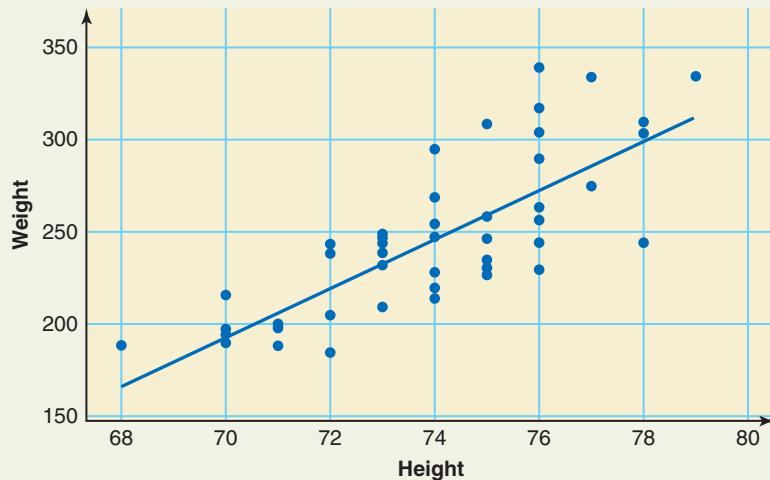
Next, we selected a random sample of 50 players and estimated the regression model for this sample. The estimated regression line for this sample is

$$\hat{y} = -739 + 13.3x$$

The values of  $a$  and  $b$  are

$$a = -739 \quad \text{and} \quad b = 13.3$$

These values of  $a$  and  $b$  give the estimates of  $A$  and  $B$  based on sample data. The scatter diagram and the regression line for the sample observations on heights and weights is given in Figure 13.10. Note that this figure does not show exactly 50 dots because some points/dots may be exactly the same or very close to each other.



**Figure 13.10** Scatter diagram for the data on heights and weights of 50 NFL players.

As we can observe from Figures 13.9 and 13.10, the scatter diagrams for population and sample data both show a (positive) linear relationship between the heights and weights of NFL players, although not a very strong positive relationship.

Source: [www.sportscity.com/NFL-salaries](http://www.sportscity.com/NFL-salaries) and [www.nfl.com/teams](http://www.nfl.com/teams)

### 13.1.7 Assumptions of the Regression Model

Like any other theory, the linear regression analysis is also based on certain assumptions. Consider the population regression model

$$y = A + Bx + \epsilon \quad (4)$$

Four assumptions are made about this model. These assumptions are explained next with reference to the example on incomes and food expenditures of households. Note that these assumptions are made about the population regression model and not about the sample regression model.

**Assumption 1:** The random error term  $\epsilon$  has a mean equal to zero for each  $x$ . In other words, among all households with the same income, some spend more than the predicted food expenditure (and, hence, have positive errors) and others spend less than the predicted food expenditure (and, consequently, have negative errors). This assumption simply states that the sum of the positive errors is equal to the sum of the negative errors, so that the mean of errors for all households with the same income is zero. Thus, when the mean value of  $\epsilon$  is zero, the mean value of  $y$  for a given  $x$  is equal to  $A + Bx$ , and it is written as

$$\mu_{y|x} = A + Bx$$

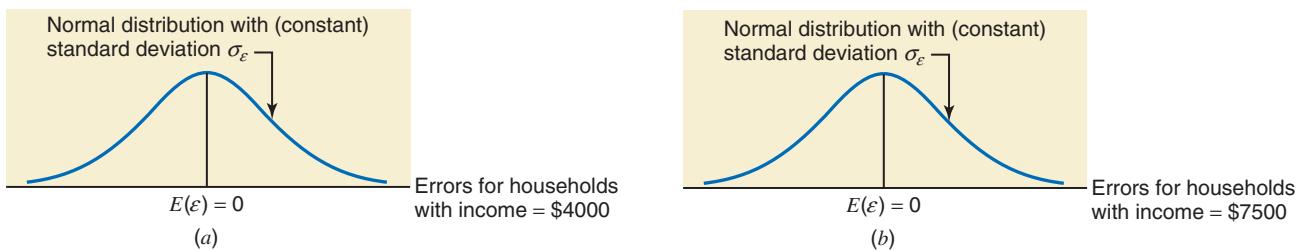
As mentioned earlier in this chapter,  $\mu_{y|x}$  is read as “the mean value of  $y$  for a given value of  $x$ .” When we find the values of  $A$  and  $B$  for model (4) using the population data, the points on the regression line give the average values of  $y$ , denoted by  $\mu_{y|x}$ , for the corresponding values of  $x$ .

**Assumption 2:** The errors associated with different observations are independent. According to this assumption, the errors for any two households in our example are independent. In other words, all households decide independently how much to spend on food.

**Assumption 3:** For any given  $x$ , the distribution of errors is normal. The corollary of this assumption is that the food expenditures for all households with the same income are normally distributed.

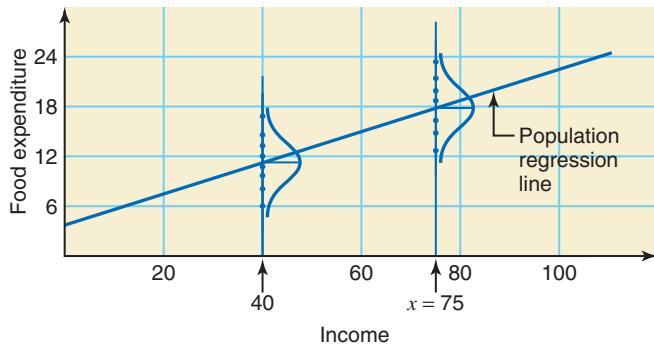
**Assumption 4:** The distribution of population errors for each  $x$  has the same (constant) standard deviation, which is denoted by  $\sigma_\epsilon$ . This assumption indicates that the spread of points around the regression line is similar for all  $x$  values.

Figure 13.11 illustrates the meanings of the first, third, and fourth assumptions for households with incomes of \$4000 and \$7500 per month. The same assumptions hold true for any other income level. In the population of all households, there will be many households with a monthly income of \$4000. Using the population regression line, if we calculate the errors for all these households and prepare the distribution of these errors, it will look like the distribution given in Figure 13.11a. Its standard deviation will be  $\sigma_\epsilon$ . Similarly, Figure 13.11b gives the distribution of errors for all those households in the population whose monthly income is \$7500. Its standard deviation is also  $\sigma_\epsilon$ . Both of these distributions are identical. Note that the mean of both of these distributions is  $E(\epsilon) = 0$ .



**Figure 13.11** (a) Errors for households with an income of \$4000 per month. (b) Errors for households with an income of \$7500 per month.

Figure 13.12 shows how the distributions given in Figure 13.11 look when they are plotted on the same diagram with the population regression line. The points on the vertical line through  $x = 40$  give the food expenditures for various households in the population, each of which has the same monthly income of \$4000. The same is true about the vertical line through  $x = 75$  or any other vertical line for some other value of  $x$ .



**Figure 13.12** Distribution of errors around the population regression line.

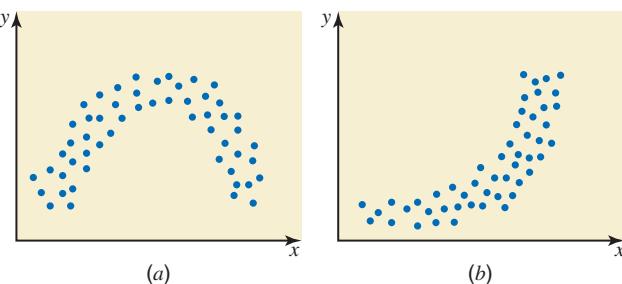
### 13.1.8 Cautions in Using Regression

When carefully applied, regression is a very helpful technique for making predictions and estimations about one variable for a certain value of another variable. However, we need to be cautious when using the regression analysis, for it can give us misleading results and predictions. The following are the two most important points to remember when using regression.

#### (a) A Note on the Use of Simple Linear Regression

We should apply linear regression with caution. When we use simple linear regression, we assume that the relationship between two variables is described by a straight line. In the real world, the relationship between variables may not be linear. Hence, before we use a simple linear regression, it is better to construct a scatter diagram and look at the plot of the data points. We should estimate a linear regression model only if the scatter diagram indicates such a relationship. The scatter diagrams of Figure 13.13 give two examples for which the relationship between  $x$  and  $y$  is not linear. Consequently, using linear regression in such cases would be wrong.

**Figure 13.13** Nonlinear relationship between  $x$  and  $y$ .



### (b) Extrapolation

The regression line estimated for the sample data is reliable only for the range of  $x$  values observed in the sample. For example, the values of  $x$  in our example on incomes and food expenditures vary from a minimum of 33 to a maximum of 83. Hence, our estimated regression line is applicable only for values of  $x$  between 33 and 83; that is, we should use this regression line to estimate the mean food expenditure or to predict the food expenditure of a single household only for income levels between \$3300 and \$8300. If we estimate or predict  $y$  for a value of  $x$  either less than 33 or greater than 83, it is called *extrapolation*. This does not mean that we should never use the regression line for extrapolation. Instead, we should interpret such predictions cautiously and not attach much importance to them.

Similarly, if the data used for the regression estimation are time-series data (see Exercises 13.100 and 13.101), the predicted values of  $y$  for periods outside the time interval used for the estimation of the regression line should be interpreted very cautiously. When using the estimated regression line for extrapolation, we are assuming that the same linear relationship between the two variables holds true for values of  $x$  outside the given range. It is possible that the relationship between the two variables may not be linear outside that range. Nonetheless, even if it is linear, adding a few more observations at either end will probably give a new estimation of the regression line.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

- 13.1 Explain the meaning of the words *simple* and *linear* as used in *simple linear regression*.
- 13.2 Explain the meaning of independent and dependent variables for a regression model.
- 13.3 Explain the difference between exact and nonexact relationships between two variables. Give one example of each.
- 13.4 Explain the difference between linear and nonlinear relationships between two variables.
- 13.5 Explain the difference between a simple and a multiple regression model.
- 13.6 Briefly explain the difference between a deterministic and a probabilistic regression model.
- 13.7 Why is the random error term included in a regression model?
- 13.8 Explain the least squares method and least squares regression line. Why are they called by these names?
- 13.9 Explain the meaning and concept of SSE. You may use a graph for illustration purposes.
- 13.10 Explain the difference between  $y$  and  $\hat{y}$ .
- 13.11 Two variables  $x$  and  $y$  have a positive linear relationship. Explain what happens to the value of  $y$  when  $x$  increases. Give one example of a positive relationship between two variables.
- 13.12 Two variables  $x$  and  $y$  have a negative linear relationship. Explain what happens to the value of  $y$  when  $x$  increases. Give one example of a negative relationship between two variables.
- 13.13 Explain the following.
  - a. Population regression line
  - b. Sample regression line
  - c. True values of  $A$  and  $B$
  - d. Estimated values of  $A$  and  $B$  that are denoted by  $a$  and  $b$ , respectively
- 13.14 Briefly explain the assumptions of the population regression model.
- 13.15 Plot the following straight lines. Give the values of the  $y$ -intercept and slope for each of these lines and interpret them. Indicate whether each of the lines gives a positive or a negative relationship between  $x$  and  $y$ .
  - a.  $y = 100 + 5x$
  - b.  $y = 400 - 4x$

**13.16** Plot the following straight lines. Give the values of the  $y$ -intercept and slope for each of these lines and interpret them. Indicate whether each of the lines gives a positive or a negative relationship between  $x$  and  $y$ .

a.  $y = -60 + 8x$     b.  $y = 300 - 6x$

**13.17** A population data set produced the following information.

$$N = 250, \quad \Sigma x = 9880, \quad \Sigma y = 1456, \quad \Sigma xy = 85,080, \quad \Sigma x^2 = 485,870$$

Find the population regression line.

**13.18** A population data set produced the following information.

$$N = 460, \quad \Sigma x = 3920, \quad \Sigma y = 2650, \quad \Sigma xy = 26,570, \quad \Sigma x^2 = 48,530$$

Find the population regression line.

**13.19** The following information is obtained from a sample data set.

$$n = 10, \quad \Sigma x = 100, \quad \Sigma y = 220, \quad \Sigma xy = 3680, \quad \Sigma x^2 = 1140$$

Find the estimated regression line.

**13.20** The following information is obtained from a sample data set.

$$n = 12, \quad \Sigma x = 66, \quad \Sigma y = 588, \quad \Sigma xy = 2244, \quad \Sigma x^2 = 396$$

Find the estimated regression line.

## ■ APPLICATIONS

**13.21** A car rental company charges \$50 a day and 20 cents per mile for renting a car. Let  $y$  be the total rental charges (in dollars) for a car for one day and  $x$  be the miles driven. The equation for the relationship between  $x$  and  $y$  is

$$y = 50 + .20x$$

- a. How much will a person pay who rents a car for one day and drives it 100 miles?
- b. Suppose each of 20 persons rents a car from this agency for one day and drives it 100 miles. Will each of them pay the same amount for renting a car for a day or do you expect each person to pay a different amount? Explain.
- c. Is the relationship between  $x$  and  $y$  exact or nonexact?

**13.22** Bob's Pest Removal Service specializes in removing wild creatures (skunks, bats, reptiles, etc.) from private homes. He charges \$70 to go to a house plus \$20 per hour for his services. Let  $y$  be the total amount (in dollars) paid by a household using Bob's services and  $x$  the number of hours Bob spends capturing and removing the animal(s). The equation for the relationship between  $x$  and  $y$  is

$$y = 70 + 20x$$

- a. Bob spent 3 hours removing a coyote from under Alice's house. How much will he be paid?
- b. Suppose nine persons called Bob for assistance during a week. Strangely enough, each of these jobs required exactly 3 hours. Will each of these clients pay Bob the same amount, or do you expect each one to pay a different amount? Explain.
- c. Is the relationship between  $x$  and  $y$  exact or nonexact?

**13.23** A researcher took a sample of 25 electronics companies and found the following relationship between  $x$  and  $y$ , where  $x$  is the amount of money (in millions of dollars) spent on advertising by a company in 2011 and  $y$  represents the total gross sales (in millions of dollars) of that company for 2011.

$$\hat{y} = 3.6 + 11.75x$$

- a. An electronics company spent \$2 million on advertising in 2011. What are its expected gross sales for 2011?
- b. Suppose four electronics companies spent \$2 million each on advertising in 2011. Do you expect these four companies to have the same actual gross sales for 2011? Explain.
- c. Is the relationship between  $x$  and  $y$  exact or nonexact?

**13.24** A researcher took a sample of 10 years and found the following relationship between  $x$  and  $y$ , where  $x$  is the number of major natural calamities (such as tornadoes, hurricanes, earthquakes, floods, etc.) that occurred during a year and  $y$  represents the average annual total profits (in millions of dollars) of a sample of insurance companies in the United States.

$$\hat{y} = 342.6 - 2.10x$$

- a. A randomly selected year had 24 major calamities. What are the expected average profits of U.S. insurance companies for that year?

- b. Suppose the number of major calamities was the same for each of 3 years. Do you expect the average profits for all U.S. insurance companies to be the same for each of these 3 years? Explain.  
 c. Is the relationship between  $x$  and  $y$  exact or nonexact?

**13.25** An auto manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The research department at the company took a sample of eight cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

- a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between ages and prices of cars?  
 b. Find the regression line with price as a dependent variable and age as an independent variable.  
 c. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.  
 d. Plot the regression line on the scatter diagram of part a and show the errors by drawing vertical lines between scatter points and the regression line.  
 e. Predict the price of a 7-year-old car of this model.  
 f. Estimate the price of an 18-year-old car of this model. Comment on this finding.

**13.26** The following table gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Source: kelloggs.com.

- a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between the amount of sugar and the number of calories per serving?  
 b. Find the predictive regression equation of the number of calories on the amount of sugar.  
 c. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.  
 d. Plot the predictive regression line on the scatter diagram of part a and show the errors by drawing vertical lines between scatter points and the predictive regression line.  
 e. Calculate the predicted calorie count for a cereal with 16 grams of sugar per serving.  
 f. Estimate the calorie count for a cereal with 52 grams of sugar per serving. Comment on this finding.

**13.27** The following table contains information on the amount of time that each of 12 students spends each day (on average) on social networks (Facebook, Twitter, etc.) and the Internet for social or entertainment purposes and his or her grade point average (GPA).

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

- a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between grade point average and time spent on social networks and the Internet?  
 b. Find the predictive regression line of GPA on time.  
 c. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.  
 d. Plot the predictive regression line on the scatter diagram of part a, and show the errors by drawing vertical lines between scatter points and the predictive regression line.  
 e. Calculate the predicted GPA for a college student who spends 3.8 hours per day on social networks and the Internet for social or entertainment purposes.  
 f. Calculate the predicted GPA for a college student who spends 16 hours per day on social networks and the Internet for social or entertainment purposes. Comment on this finding.

**13.28** While browsing through the magazine rack at a bookstore, a statistician decides to examine the relationship between the price of a magazine and the percentage of the magazine space that contains advertisements. The data collected for eight magazines are given in the following table.

Percentage containing ads	37	43	58	49	70	28	65	32
Price (\$)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75

- a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between the percentage of a magazine's space containing ads and the price of the magazine?

- b. Find the estimated regression equation of price on the percentage of space containing ads.
- c. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.
- d. Plot the estimated regression line on the scatter diagram of part a, and show the errors by drawing vertical lines between scatter points and the predictive regression line.
- e. Predict the price of a magazine with 50% of its space containing ads.
- f. Estimate the price of a magazine with 99% of its space containing ads. Comment on this finding.

**13.29** The following table gives the total payroll (in millions of dollars) on the opening day of the 2011 season and the percentage of games won during the 2011 season by each of the National League baseball teams.

Team	Total Payroll (millions of dollars)	Percentage of Games Won
Arizona Diamondbacks	53.60	58.0
Atlanta Braves	87.00	54.9
Chicago Cubs	125.50	43.8
Cincinnati Reds	76.20	48.8
Colorado Rockies	88.00	45.1
Houston Astros	70.70	34.6
Los Angeles Dodgers	103.80	50.9
Miami Marlins	56.90	44.4
Milwaukee Brewers	85.50	59.3
New York Mets	120.10	47.5
Philadelphia Phillies	173.00	63.0
Pittsburgh Pirates	46.00	44.4
San Diego Padres	45.90	43.8
San Francisco Giants	118.20	53.1
St. Louis Cardinals	105.40	55.6
Washington Nationals	63.70	49.7

Source: <http://baseball.about.com/od/newsrumors/a/2011-Baseball-Team-Payrolls.htm>.

- a. Find the least squares regression line with total payroll as the independent variable and percentage of games won as the dependent variable.
- b. Is the equation of the regression line obtained in part a the population regression line? Why or why not? Do the values of the  $y$ -intercept and the slope of the regression line give  $A$  and  $B$  or  $a$  and  $b$ ?
- c. Give a brief interpretation of the values of the  $y$ -intercept and the slope obtained in part a.
- d. Predict the percentage of games won by a team with a total payroll of \$100 million.

**13.30** The following table gives the total payroll (in millions of dollars) on the opening day of the 2011 season and the percentage of games won during the 2011 season by each of the American League baseball teams.

Team	Total Payroll (millions of dollars)	Percentage of Games Won
Baltimore Orioles	85.30	42.6
Boston Red Sox	161.40	55.6
Chicago White Sox	129.30	48.8
Cleveland Indians	49.20	49.4
Detroit Tigers	105.70	58.6
Kansas City Royals	36.10	43.8
Los Angeles Angels	139.00	53.1
Minnesota Twins	112.70	38.9
New York Yankees	201.70	59.9
Oakland Athletics	66.60	45.7
Seattle Mariners	86.40	41.4
Tampa Bay Rays	41.90	56.2
Texas Rangers	92.30	59.3
Toronto Blue Jays	62.50	50.0

Source: <http://baseball.about.com/od/newsrumors/a/2011-Baseball-Team-Payrolls.htm>.

- Find the least squares regression line with total payroll as the independent variable and percentage of games won as the dependent variable.
- Is the equation of the regression line obtained in part a the population regression line? Why or why not? Do the values of the  $y$ -intercept and the slope of the regression line give  $A$  and  $B$  or  $a$  and  $b$ ?
- Give a brief interpretation of the values of the  $y$ -intercept and the slope obtained in part a.
- Predict the percentage of games won by a team with a total payroll of \$100 million.

## 13.2

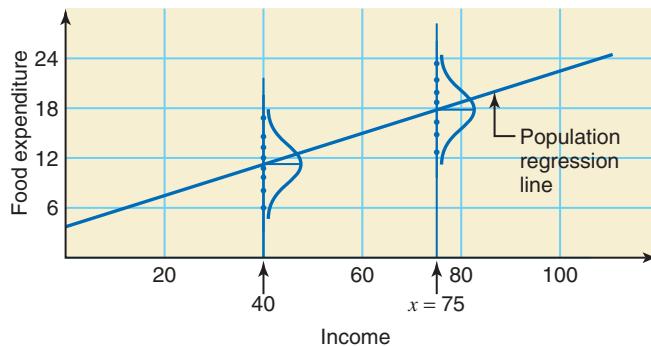
# Standard Deviation of Errors and Coefficient of Determination

In this section we discuss two concepts related to regression analysis. First we discuss the concept of the standard deviation of random errors and its calculation. Then we learn about the concept of the coefficient of determination and its calculation.

### 13.2.1 Standard Deviation of Errors

When we consider incomes and food expenditures, all households with the same income are expected to spend different amounts on food. Consequently, the random error  $\epsilon$  will assume different values for these households. The standard deviation  $\sigma_\epsilon$  measures the spread of these errors around the population regression line. The **standard deviation of errors** tells us how widely the errors are, and, hence, the values of  $y$  are spread for a given  $x$ . In Figure 13.12, which is reproduced as Figure 13.14, the points on the vertical line through  $x = 40$  give the monthly food expenditures for all households with a monthly income of \$4000. The distance of each dot from the point on the regression line gives the value of the corresponding error. The standard deviation of errors  $\sigma_\epsilon$  measures the spread of such points around the population regression line. The same is true for  $x = 75$  or any other value of  $x$ .

**Figure 13.14** Spread of errors for  $x = 40$  and  $x = 75$ .



Note that  $\sigma_\epsilon$  denotes the standard deviation of errors for the population. However, usually  $\sigma_\epsilon$  is unknown. In such cases, it is estimated by  $s_e$ , which is the standard deviation of errors for the sample data. The following is the basic formula to calculate  $s_e$ :

$$s_e = \sqrt{\frac{SSE}{n - 2}} \quad \text{where} \quad SSE = \sum(y - \hat{y})^2$$

In this formula,  $n - 2$  represents the **degrees of freedom** for the regression model. The reason  $df = n - 2$  is that we lose one degree of freedom to calculate  $\bar{x}$  and one for  $\bar{y}$ .

**Degrees of Freedom for a Simple Linear Regression Model** The *degrees of freedom* for a simple linear regression model are

$$df = n - 2$$

For computational purposes, it is more convenient to use the following formula to calculate the standard deviation of errors  $s_e$ .

**Standard Deviation of Errors** The *standard deviation of errors* is calculated as<sup>3</sup>

$$s_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{n - 2}}$$

where

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

The calculation of  $SS_{xy}$  was discussed earlier in this chapter.<sup>4</sup>

Like the value of  $SS_{xx}$ , the value of  $SS_{yy}$  is always positive.

Example 13–2 illustrates the calculation of the standard deviation of errors for the data of Table 13.1.

## ■ EXAMPLE 13–2

Compute the standard deviation of errors  $s_e$  for the data on monthly incomes and food expenditures of the seven households given in Table 13.1.

*Calculating the standard deviation of errors.*

**Solution** To compute  $s_e$ , we need to know the values of  $SS_{yy}$ ,  $SS_{xy}$ , and  $b$ . In Example 13–1, we computed  $SS_{xy}$  and  $b$ . These values are

$$SS_{xy} = 447.5714 \quad \text{and} \quad b = .2525$$

To compute  $SS_{yy}$ , we calculate  $\sum y^2$  as shown in Table 13.3.

**Table 13.3**

Income <i>x</i>	Food Expenditure <i>y</i>	<i>y</i> <sup>2</sup>
55	14	196
83	24	576
38	13	169
61	16	256
33	9	81
49	15	225
67	17	289
$\sum x = 386$	$\sum y = 108$	$\sum y^2 = 1792$

The value of  $SS_{yy}$  is

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 1792 - \frac{(108)^2}{7} = 125.7143$$

Hence, the standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{n - 2}} = \sqrt{\frac{125.7143 - .2525(447.5714)}{7 - 2}} = 1.5939$$

### 13.2.2 Coefficient of Determination

We may ask the question: How good is the regression model? In other words: How well does the independent variable explain the dependent variable in the regression model? The *coefficient of determination* is one concept that answers this question.

<sup>3</sup>If we have access to population data, the value of  $\sigma_\epsilon$  is calculated using the formula

$$\sigma_\epsilon = \sqrt{\frac{SS_{yy} - B SS_{xy}}{N}}$$

<sup>4</sup>The basic formula to calculate  $SS_{yy}$  is  $\sum(y - \bar{y})^2$ .

For a moment, assume that we possess information only on the food expenditures of households and not on their incomes. Hence, in this case, we cannot use the regression line to predict the food expenditure for any household. As we did in earlier chapters, in the absence of a regression model, we use  $\bar{y}$  to estimate or predict every household's food expenditure. Consequently, the error of prediction for each household is now given by  $y - \bar{y}$ , which is the difference between the actual food expenditure of a household and the mean food expenditure. If we calculate such errors for all households in the sample and then square and add them, the resulting sum is called the **total sum of squares** and is denoted by **SST**. Actually SST is the same as  $SS_{yy}$  and is defined as

$$SST = SS_{yy} = \sum(y - \bar{y})^2$$

However, for computational purposes, SST is calculated using the following formula.

**Total Sum of Squares (SST)** The *total sum of squares*, denoted by SST, is calculated as

$$SST = \sum y^2 - \frac{(\sum y)^2}{n}$$

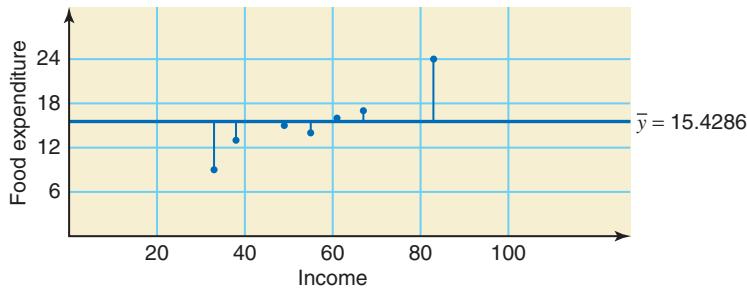
Note that this is the same formula that we used to calculate  $SS_{yy}$ .

The value of  $SS_{yy}$ , which is 125.7143, was calculated in Example 13–2. Consequently, the value of SST is

$$SST = 125.7143$$

From Example 13–1,  $\bar{y} = 15.4286$ . Figure 13.15 shows the error for each of the seven households in our sample using the scatter diagram of Figure 13.4 and using  $\bar{y}$ .

**Figure 13.15** Total errors.



Now suppose we use the simple linear regression model to predict the food expenditure of each of the seven households in our sample. In this case, we predict each household's food expenditure by using the regression line we estimated earlier in Example 13–1, which is

$$\hat{y} = 1.5050 + .2525x$$

The predicted food expenditures, denoted by  $\hat{y}$ , for the seven households are listed in Table 13.4. Also given are the errors and error squares.

**Table 13.4**

x	y	$\hat{y} = 1.5050 + .2525x$	$e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
55	14	15.3925	-1.3925	1.9391
83	24	22.4625	1.5375	2.3639
38	13	11.1000	1.9000	3.6100
61	16	16.9075	-0.9075	.8236
33	9	9.8375	-0.8375	.7014
49	15	13.8775	1.1225	1.2600
67	17	18.4225	-1.4225	2.0235
				$\Sigma e^2 = \Sigma (y - \hat{y})^2 = 12.7215$

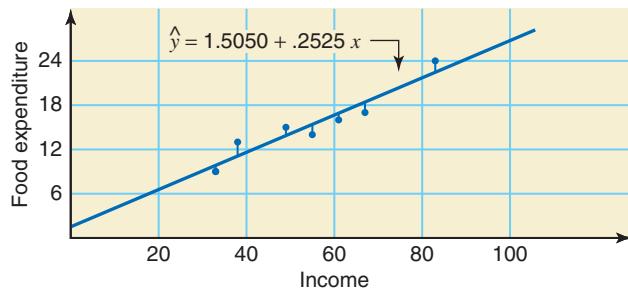
We calculate the values of  $\hat{y}$  (given in the third column of Table 13.4) by substituting the values of  $x$  in the estimated regression model. For example, the value of  $x$  for the first household is 55. Substituting this value of  $x$  in the regression equation, we obtain

$$\hat{y} = 1.5050 + .2525(55) = 15.3925$$

Similarly we find the other values of  $\hat{y}$ . The error sum of squares SSE is given by the sum of the fifth column in Table 13.4. Thus,

$$SSE = \sum(y - \hat{y})^2 = 12.7215$$

The errors of prediction for the regression model for the seven households are shown in Figure 13.16.



**Figure 13.16** Errors of prediction when regression model is used.

Thus, from the foregoing calculations,

$$SST = 125.7143 \quad \text{and} \quad SSE = 12.7215$$

These values indicate that the sum of squared errors decreased from 125.7143 to 12.7215 when we used  $\hat{y}$  in place of  $\bar{y}$  to predict food expenditures. This reduction in squared errors is called the **regression sum of squares** and is denoted by **SSR**. Thus,

$$SSR = SST - SSE = 125.7143 - 12.7215 = 112.9928$$

The value of SSR can also be computed by using the formula

$$SSR = \sum(\hat{y} - \bar{y})^2$$

**Regression Sum of Squares (SSR)** The *regression sum of squares*, denoted by SSR, is

$$SSR = SST - SSE$$

Thus, SSR is the portion of SST that is explained by the use of the regression model, and SSE is the portion of SST that is not explained by the use of the regression model. The sum of SSR and SSE is always equal to SST. Thus,

$$SST = SSR + SSE$$

The ratio of SSR to SST gives the **coefficient of determination**. The coefficient of determination calculated for population data is denoted by  $\rho^2$  ( $\rho$  is the Greek letter *rho*), and the one calculated for sample data is denoted by  $r^2$ . The coefficient of determination gives the proportion of SST that is explained by the use of the regression model. The value of the coefficient of determination always lies in the range zero to one. The coefficient of determination can be calculated by using the formula

$$r^2 = \frac{SSR}{SST} \quad \text{or} \quad \frac{SST - SSE}{SST}$$

However, for computational purposes, the following formula is more efficient to use to calculate the coefficient of determination.

**Coefficient of Determination** The *coefficient of determination*, denoted by  $r^2$ , represents the proportion of SST that is explained by the use of the regression model. The computational formula for  $r^2$  is<sup>5</sup>

$$r^2 = \frac{b \text{ SS}_{xy}}{\text{SS}_{yy}}$$

and

$$0 \leq r^2 \leq 1$$

Example 13–3 illustrates the calculation of the coefficient of determination for a sample data set.

### ■ EXAMPLE 13–3

*Calculating the coefficient of determination.*

For the data of Table 13.1 on monthly incomes and food expenditures of seven households, calculate the coefficient of determination.

**Solution** From earlier calculations made in Examples 13–1 and 13–2,

$$b = .2525, \quad \text{SS}_{xy} = 447.5714, \quad \text{and} \quad \text{SS}_{yy} = 125.7143$$

Hence,

$$r^2 = \frac{b \text{ SS}_{xy}}{\text{SS}_{yy}} = \frac{(.2525)(447.5714)}{125.7143} = .8990 = .90$$

Thus, we can state that SST is reduced by approximately 90% (from 125.7143 to 12.7215) when we use  $\hat{y}$ , instead of  $\bar{y}$ , to predict the food expenditures of households. Note that  $r^2$  is usually rounded to two decimal places. ■

The total sum of squares SST is a measure of the total variation in food expenditures, the regression sum of squares SSR is the portion of total variation explained by the regression model (or by income), and the error sum of squares SSE is the portion of total variation not explained by the regression model. Hence, for Example 13–3 we can state that 90% of the total variation in food expenditures of households occurs because of the variation in their incomes, and the remaining 10% is due to randomness and other variables.

Usually, the higher the value of  $r^2$ , the better is the regression model. This is so because if  $r^2$  is larger, a greater portion of the total errors is explained by the included independent variable, and a smaller portion of errors is attributed to other variables and randomness.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**13.31** What are the degrees of freedom for a simple linear regression model?

**13.32** Explain the meaning of coefficient of determination.

**13.33** Explain the meaning of SST and SSR. You may use graphs for illustration purposes.

**13.34** A population data set produced the following information.

$$\begin{aligned} N &= 250, \quad \Sigma x = 9880, \quad \Sigma y = 1456, \quad \Sigma xy = 85,080, \\ \Sigma x^2 &= 485,870, \quad \text{and} \quad \Sigma y^2 = 135,675 \end{aligned}$$

Find the values of  $\sigma_e$  and  $\rho^2$ .

<sup>5</sup>If we have access to population data, the value of  $\rho^2$  is calculated using the formula

$$\rho^2 = \frac{B \text{ SS}_{xy}}{\text{SS}_{yy}}$$

The values of  $\text{SS}_{xy}$  and  $\text{SS}_{yy}$  used here are calculated for the population data set.

- 13.35** A population data set produced the following information.

$$\begin{aligned}N &= 460, \quad \Sigma x = 3920, \quad \Sigma y = 2650, \quad \Sigma xy = 26,570, \\ \Sigma x^2 &= 48,530, \quad \text{and} \quad \Sigma y^2 = 39,347\end{aligned}$$

Find the values of  $\sigma_e$  and  $r^2$ .

- 13.36** The following information is obtained from a sample data set.

$$\begin{aligned}n &= 10, \quad \Sigma x = 100, \quad \Sigma y = 220, \quad \Sigma xy = 3680, \\ \Sigma x^2 &= 1140, \quad \text{and} \quad \Sigma y^2 = 25,272\end{aligned}$$

Find the values of  $s_e$  and  $r^2$ .

- 13.37** The following information is obtained from a sample data set.

$$\begin{aligned}n &= 12, \quad \Sigma x = 66, \quad \Sigma y = 588, \quad \Sigma xy = 2244, \\ \Sigma x^2 &= 396, \quad \text{and} \quad \Sigma y^2 = 58,734\end{aligned}$$

Find the values of  $s_e$  and  $r^2$ .

## ■ APPLICATIONS

- 13.38** The following table gives information on the calorie count and grams of fat for the 11 types of bagels produced by Panera Bread.

Bagel	Calories	Fat (grams)
Asiago Cheese	330	6.0
Blueberry	330	1.5
Chocolate Chip	370	6.0
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	2.5
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.5
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

With calories as the dependent variable and fat content as the independent variable, find the following:

- a.  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$
- b. Standard deviation of errors
- c. SST, SSE, and SSR
- d. Coefficient of determination

- 13.39** The following table provides information on the speed at takeoff (in meters per second) and distance traveled (in meters) by a random sample of 10 world-class long jumpers.

Speed	8.5	8.8	9.3	8.9	8.2	8.6	8.7	9.0	8.7	9.1
Distance	7.72	7.91	8.33	7.93	7.39	7.65	7.95	8.28	7.86	8.14

With distance traveled as the dependent variable and speed at takeoff as the independent variable, find the following:

- a.  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$
- b. Standard deviation of errors
- c. SST, SSE, and SSR
- d. Coefficient of determination

- 13.40** Refer to Exercise 13.25. The following table, which gives the ages (in years) and prices (in hundreds of dollars) of eight cars of a specific model, is reproduced from that exercise.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

- a. Calculate the standard deviation of errors.
- b. Compute the coefficient of determination and give a brief interpretation of it.

**13.41** The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the caloric count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Source: kelloggs.com.

- a. Determine the standard deviation of errors.
- b. Find the coefficient of determination and give a brief interpretation of it.

**13.42** The following table, reproduced from Exercise 13.27, contains information on the amount of time spent each day (on average) on social networks and the Internet for social or entertainment purposes and the grade point average for a random sample of 12 college students.

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

- a. Calculate the standard deviation of errors.
- b. Compute the coefficient of determination, and give a brief interpretation of it. What percentage of the variation in GPA is explained by the least squares regression line of GPA on time? What percentage is not explained?

**13.43** The following table, reproduced from Exercise 13.28, lists the percentages of space for eight magazines that contain advertisements and the prices of these magazines.

Percentage containing ads	37	43	58	49	70	28	65	32
Price (\$)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75

- a. Find the standard deviation of errors.
- b. Compute the coefficient of determination. What percentage of the variation in price is explained by the least squares regression of price on the percentage of magazine space containing ads? What percentage of this variation is not explained?

**13.44** Refer to data given in Exercise 13.29 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the National League baseball teams.

- a. Find the standard deviation of errors,  $\sigma_e$ . (Note that this data set belongs to a population.)
- b. Compute the coefficient of determination,  $\rho^2$ .

**13.45** Refer to data given in Exercise 13.30 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the American League baseball teams.

- a. Find the standard deviation of errors,  $\sigma_e$ . (Note that this data set belongs to a population.)
- b. Compute the coefficient of determination,  $\rho^2$ .

## 13.3 Inferences About $B$

This section is concerned with estimation and tests of hypotheses about the population regression slope  $B$ . We can also make confidence intervals and test hypotheses about the  $y$ -intercept  $A$  of the population regression line. However, making inferences about  $A$  is beyond the scope of this text.

### 13.3.1 Sampling Distribution of $b$

One of the main purposes for determining a regression line is to find the true value of the slope  $B$  of the population regression line. However, in almost all cases, the regression line is estimated using sample data. Then, based on the sample regression line, inferences are made about the population regression line. The slope  $b$  of a sample regression line is a point estimator of the slope  $B$  of the population regression line. The different sample regression lines estimated for different samples taken from the same population will give different values of  $b$ . If only one sample is taken and the regression line for that sample is estimated, the value of  $b$  will depend on which elements are included in the sample. Thus,  $b$  is a random variable, and it possesses a

probability distribution that is more commonly called its sampling distribution. The shape of the sampling distribution of  $b$ , its mean, and standard deviation are given next.

**Mean, Standard Deviation, and Sampling Distribution of  $b$**  Because of the assumption of normally distributed random errors, the sampling distribution of  $b$  is normal. The mean and standard deviation of  $b$ , denoted by  $\mu_b$  and  $\sigma_b$ , respectively, are

$$\mu_b = B \quad \text{and} \quad \sigma_b = \frac{\sigma_\epsilon}{\sqrt{SS_{xx}}}$$

However, usually the standard deviation of population errors  $\sigma_\epsilon$  is not known. Hence, the sample standard deviation of errors  $s_e$  is used to estimate  $\sigma_\epsilon$ . In such a case, when  $\sigma_\epsilon$  is unknown, the standard deviation of  $b$  is estimated by  $s_b$ , which is calculated as

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

If  $\sigma_\epsilon$  is known, the normal distribution can be used to make inferences about  $B$ . However, if  $\sigma_\epsilon$  is not known, the normal distribution is replaced by the  $t$  distribution to make inferences about  $B$ .

### 13.3.2 Estimation of $B$

The value of  $b$  obtained from the sample regression line is a point estimate of the slope  $B$  of the population regression line. As mentioned in Section 13.3.1, if  $\sigma_\epsilon$  is not known, the  $t$  distribution is used to make a confidence interval for  $B$ .

**Confidence Interval for  $B$**  The  $(1 - \alpha)100\%$  confidence interval for  $B$  is given by

$$b \pm ts_b$$

where

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

and the value of  $t$  is obtained from the  $t$  distribution table for  $\alpha/2$  area in the right tail of the  $t$  distribution and  $n - 2$  degrees of freedom.

Example 13–4 describes the procedure for making a confidence interval for  $B$ .

#### ■ EXAMPLE 13–4

Construct a 95% confidence interval for  $B$  for the data on incomes and food expenditures of seven households given in Table 13.1.

Constructing a confidence interval for  $B$ .

**Solution** From the given information and earlier calculations in Examples 13–1 and 13–2,

$$n = 7, \quad b = .2525, \quad SS_{xx} = 1772.8571, \quad \text{and} \quad s_e = 1.5939$$

The confidence level is 95%. We have

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{1.5939}{\sqrt{1772.8571}} = .0379$$

$$df = n - 2 = 7 - 2 = 5$$

$$\alpha/2 = (1 - .95)/2 = .025$$

From the  $t$  distribution table, the value of  $t$  for 5  $df$  and .025 area in the right tail of the  $t$  distribution curve is 2.571. The 95% confidence interval for  $B$  is

$$b \pm ts_b = .2525 \pm 2.571(.0379) = .2525 \pm .0974 = .155 \text{ to } .350$$

Thus, we are 95% confident that the slope  $B$  of the population regression line is between .155 and .350. ■

### 13.3.3 Hypothesis Testing About $B$

Testing a hypothesis about  $B$  when the null hypothesis is  $B = 0$  (that is, the slope of the regression line is zero) is equivalent to testing that  $x$  does not determine  $y$  and that the regression line is of no use in predicting  $y$  for a given  $x$ . However, we should remember that we are testing for a linear relationship between  $x$  and  $y$ . It is possible that  $x$  may determine  $y$  nonlinearly. Hence, a nonlinear relationship may exist between  $x$  and  $y$ .

To test the hypothesis that  $x$  does not determine  $y$  linearly, we will test the null hypothesis that the slope of the regression line is zero; that is,  $B = 0$ . The alternative hypothesis can be: (1)  $x$  determines  $y$ , that is,  $B \neq 0$ ; (2)  $x$  determines  $y$  positively, that is,  $B > 0$ ; or (3)  $x$  determines  $y$  negatively, that is,  $B < 0$ .

The procedure used to make a hypothesis test about  $B$  is similar to the one used in earlier chapters. It involves the same five steps. Of course, we can use the  $p$ -value approach too.

**Test Statistic for  $b$**  The value of the *test statistic  $t$  for  $b$*  is calculated as

$$t = \frac{b - B}{s_b}$$

The value of  $B$  is substituted from the null hypothesis.

Example 13–5 illustrates the procedure for testing a hypothesis about  $B$ .

### ■ EXAMPLE 13–5

Conducting a test of hypothesis about  $B$ .

Test at the 1% significance level whether the slope of the regression line for the example on incomes and food expenditures of seven households is positive.

**Solution** From the given information and earlier calculations in Examples 13–1 and 13–4,

$$n = 7, \quad b = .2525, \quad \text{and} \quad s_b = .0379$$

**Step 1. State the null and alternative hypotheses.**

We are to test whether or not the slope  $B$  of the population regression line is positive. Hence, the two hypotheses are

$$H_0: B = 0 \quad (\text{The slope is zero})$$

$$H_1: B > 0 \quad (\text{The slope is positive})$$

Note that we can also write the null hypothesis as  $H_0: B \leq 0$ , which states that the slope is either zero or negative.

**Step 2. Select the distribution to use.**

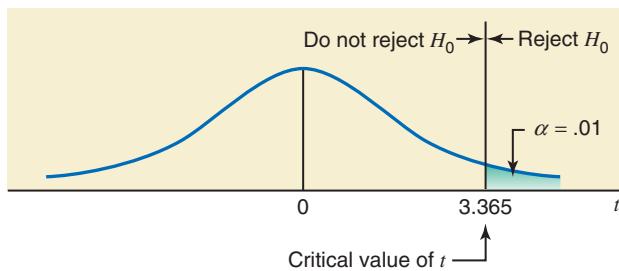
Here,  $\sigma_\epsilon$  is not known. All assumptions for the population regression model are assumed to hold true. Hence, we will use the  $t$  distribution to make the test about  $B$ .

**Step 3. Determine the rejection and nonrejection regions.**

The significance level is .01. The  $>$  sign in the alternative hypothesis indicates that the test is right-tailed. Therefore,

$$\text{Area in the right tail of the } t \text{ distribution} = \alpha = .01$$

$$df = n - 2 = 7 - 2 = 5$$



**Figure 13.17** Rejection and nonrejection regions.

From the  $t$  distribution table, the critical value of  $t$  for  $5\ df$  and  $.01$  area in the right tail of the  $t$  distribution is  $3.365$ , as shown in Figure 13.17.

**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $b$  is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{.2525 - 0}{.0379} = 6.662$$

From  $H_0$

**Step 5.** Make a decision.

The value of the test statistic  $t = 6.662$  is greater than the critical value of  $t = 3.365$ , and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that  $x$  (income) determines  $y$  (food expenditure) positively. That is, food expenditure increases with an increase in income and it decreases with a decrease in income.

### Using the $p$ -Value to Make a Decision

We can find the range for the  $p$ -value (as we did in Chapters 9 and 10) from the  $t$  distribution table, Table V of Appendix C, and make a decision by comparing that  $p$ -value with the significance level. For this example,  $df = 5$ , and the observed value of  $t$  is  $6.662$ . From Table V (the  $t$  distribution table) in the row of  $df = 5$ , the largest value of  $t$  is  $5.893$  for which the area in the right tail of the  $t$  distribution is  $.001$ . Since our observed value of  $t = 6.662$  is larger than  $5.893$ , the  $p$ -value for  $t = 6.662$  is less than  $.001$ , that is,

$$p\text{-value} < .001$$

Note that if we use technology to find this  $p$ -value, we will obtain a  $p$ -value of  $.000$ . Thus, we can state that for any  $\alpha$  equal to or higher than  $.001$  (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For our example,  $\alpha = .01$ , which is larger than the  $p$ -value of  $.001$ . As a result, we reject the null hypothesis. ■

Note that the null hypothesis does not always have to be  $B = 0$ . We may test the null hypothesis that  $B$  is equal to a certain value. See Exercises 13.47 to 13.50, 13.54, 13.57, and 13.58 for such cases.

### A Note on Regression and Causality

The regression line does not prove causality between two variables; that is, it does not predict that a change in  $y$  is caused by a change in  $x$ . The information about causality is based on theory or common sense. A regression line describes only whether or not a significant quantitative relationship between  $x$  and  $y$  exists. Significant relationship means that we reject the null hypothesis  $H_0: B = 0$  at a given significance level. The estimated regression line gives the change in  $y$  due to a change of one unit in  $x$ . Note that it does not indicate that the reason  $y$  has changed is that  $x$  has changed. In our example on incomes and food expenditures, it is economic theory and common sense, not the regression line, that tell us that food expenditure depends on income. The regression analysis simply helps determine whether or not this dependence is significant.

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**13.46** Describe the mean, standard deviation, and shape of the sampling distribution of the slope  $b$  of the simple linear regression model.

**13.47** The following information is obtained for a sample of 16 observations taken from a population.

$$SS_{xx} = 340.700, \quad s_e = 1.951, \quad \text{and} \quad \hat{y} = 12.45 + 6.32x$$

- a. Make a 99% confidence interval for  $B$ .
- b. Using a significance level of .025, can you conclude that  $B$  is positive?
- c. Using a significance level of .01, can you conclude that  $B$  is different from zero?
- d. Using a significance level of .02, test whether  $B$  is different from 4.50. (*Hint:* The null hypothesis here will be  $H_0: B = 4.50$ , and the alternative hypothesis will be  $H_1: B \neq 4.50$ . Notice that the value of  $B = 4.50$  will be used to calculate the value of the test statistic  $t$ .)

**13.48** The following information is obtained for a sample of 25 observations taken from a population.

$$SS_{xx} = 274.600, \quad s_e = .932, \quad \text{and} \quad \hat{y} = 280.56 - 3.77x$$

- a. Make a 95% confidence interval for  $B$ .
- b. Using a significance level of .01, test whether  $B$  is negative.
- c. Testing at the 5% significance level, can you conclude that  $B$  is different from zero?
- d. Test if  $B$  is different from  $-5.20$ . Use  $\alpha = .01$ .

**13.49** The following information is obtained for a sample of 100 observations taken from a population.

$$SS_{xx} = 524.884 \quad s_e = 1.464, \quad \text{and} \quad \hat{y} = 5.48 + 2.50x$$

- a. Make a 98% confidence interval for  $B$ .
- b. Test at the 2.5% significance level whether  $B$  is positive.
- c. Can you conclude that  $B$  is different from zero? Use  $\alpha = .01$ .
- d. Using a significance level of .01, test whether  $B$  is greater than 1.75.

**13.50** The following information is obtained for a sample of 80 observations taken from a population.

$$SS_{xx} = 380.592, \quad s_e = .961, \quad \text{and} \quad \hat{y} = 160.24 - 2.70x$$

- a. Make a 97% confidence interval for  $B$ .
- b. Test at the 1% significance level whether  $B$  is negative.
- c. Can you conclude that  $B$  is different from zero? Use  $\alpha = .01$ .
- d. Using a significance level of .025, test whether  $B$  is less than  $-1.25$ .

### ■ APPLICATIONS

**13.51** Refer to Exercise 13.25. The data on ages (in years) and prices (in hundreds of dollars) for eight cars of a specific model are reproduced from that exercise.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

- a. Construct a 95% confidence interval for  $B$ . You can use results obtained in Exercises 13.25 and 13.40 here.
- b. Test at the 5% significance level whether  $B$  is negative.

**13.52** The data given in the table below are the midterm scores in a course for a sample of 10 students and the scores of student evaluations of the instructor. (In the instructor evaluation scores, 1 is the lowest and 4 is the highest score.)

Instructor score	3	2	3	1	2	4	3	4	4	2
Midterm score	90	75	97	64	47	99	75	88	93	81

- a. Find the regression of instructor scores on midterm scores.
- b. Construct a 99% confidence interval for  $B$ .
- c. Test at the 1% significance level whether  $B$  is positive.

- 13.53** The following data give the experience (in years) and monthly salaries (in hundreds of dollars) of nine randomly selected secretaries.

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	62	29	37	43	35	60	67	32	60

- a. Find the least squares regression line with experience as an independent variable and monthly salary as a dependent variable.
- b. Construct a 98% confidence interval for  $B$ .
- c. Test at the 2.5% significance level whether  $B$  is greater than zero.

- 13.54** The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Source: kelloggs.com.

- a. Make a 95% confidence interval for  $B$ . You can use the calculations made in Exercises 13.26 and 13.41 here.
- b. It is well known that each additional gram of carbohydrate adds 4 calories. Sugar is one type of carbohydrate. Using regression equation for the data in the table, test at the 1% significance level whether  $B$  is different from 4.

- 13.55** Refer to Exercise 13.27. The following table, reproduced from that exercise, contains information on the amount of time spent each day (on average) on social networks and the Internet for social or entertainment purposes and the grade point average for a random sample of 12 college students.

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

- a. Construct a 98% confidence interval for  $B$ . You can use the results obtained in Exercises 13.27 and 13.42.
- b. Test at the 1% significance level whether  $B$  is negative.

- 13.56** The following table, reproduced from Exercise 13.28, lists the percentages of space for eight magazines that contain advertisements and the prices of these magazines.

Percentage containing ads	37	43	58	49	70	28	65	32
Price (\$)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75

- a. Construct a 98% confidence interval for  $B$ . You can use the calculations made in Exercises 13.28 and 13.43 here.
- b. Testing at the 5% significance level, can you conclude that  $B$  is different from zero?

- 13.57** The following table, reproduced from Exercise 13.38, gives information on the calorie count and grams of fat for the 11 types of bagels produced by Panera Bread.

Bagel	Calories	Fat (grams)
Asiago Cheese	330	6.0
Blueberry	330	1.5
Chocolate Chip	370	6.0
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	2.5
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.5
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

- Find the least squares regression line with calories as the dependent variable and fat content as the independent variable.
- Make a 95% confidence interval for  $B$ . You may use the results obtained in Exercise 13.38.
- Test at the 5% significance level whether  $B$  is different from 14.

**13.58** The following table, reproduced from Exercise 13.39, provides information on the speed at takeoff (in meters per second) and distance traveled (in meters) by a random sample of 10 world-class long jumpers.

Speed	8.5	8.8	9.3	8.9	8.2	8.6	8.7	9.0	8.7	9.1
Distance	7.72	7.91	8.33	7.93	7.39	7.65	7.95	8.28	7.86	8.14

- Find the predictive regression line of the distance traveled on the speed at takeoff.
- Make a 98% confidence interval for  $B$ . You may use the results obtained in Exercise 13.39.
- Test at the 1% significance level whether  $B$  is less than 1.2.

## 13.4 Linear Correlation

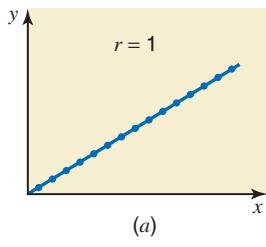
This section describes the meaning and calculation of the linear correlation coefficient and the procedure to conduct a test of hypothesis about it.

### 13.4.1 Linear Correlation Coefficient

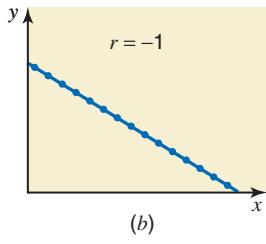
Another measure of the relationship between two variables is the correlation coefficient. This section describes the simple linear correlation, for short **linear correlation**, which measures the strength of the linear association between two variables. In other words, the linear correlation coefficient measures how closely the points in a scatter diagram are spread around the regression line. The correlation coefficient calculated for the population data is denoted by  $\rho$  (Greek letter *rho*) and the one calculated for sample data is denoted by  $r$ . (Note that the square of the correlation coefficient is equal to the coefficient of determination.)

**Value of the Correlation Coefficient** The value of the correlation coefficient always lies in the range  $-1$  to  $1$ ; that is,

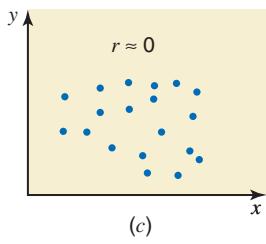
$$-1 \leq \rho \leq 1 \quad \text{and} \quad -1 \leq r \leq 1$$



(a)



(b)



(c)

Although we can explain the linear correlation using the population correlation coefficient  $\rho$ , we will do so using the sample correlation coefficient  $r$ .

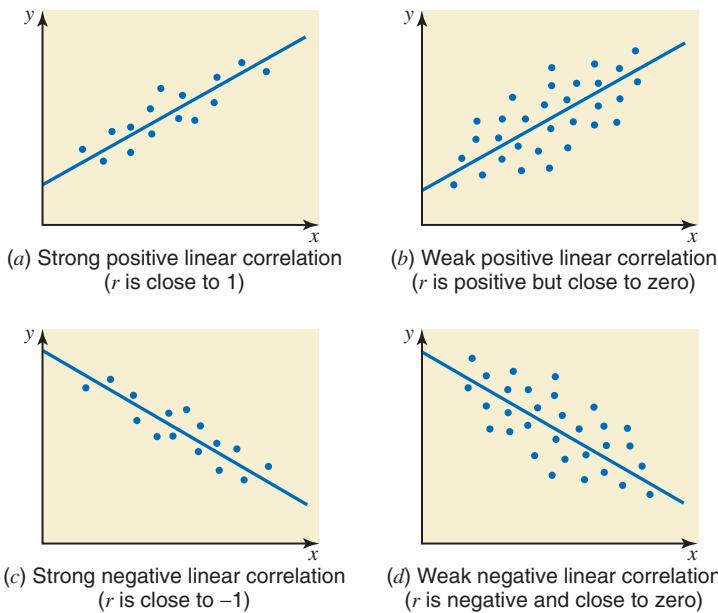
If  $r = 1$ , it is said to be a *perfect positive linear correlation*. In such a case, all points in the scatter diagram lie on a straight line that slopes upward from left to right, as shown in Figure 13.18a. If  $r = -1$ , the correlation is said to be a *perfect negative linear correlation*. In this case, all points in the scatter diagram fall on a straight line that slopes downward from left to right, as shown in Figure 13.18b. If the points are scattered all over the diagram, as shown in Figure 13.18c, then there is *no linear correlation* between the two variables, and consequently  $r$  is close to 0. Note that here  $r$  is not equal to zero but is very *close* to zero.

We do not usually encounter an example with perfect positive or perfect negative correlation (unless the relationship between variables is exact). What we observe in real-world problems is either a positive linear correlation with  $0 < r < 1$  (that is, the correlation coefficient is greater than zero but less than 1) or a negative linear correlation with  $-1 < r < 0$  (that is, the correlation coefficient is greater than  $-1$  but less than zero).

If the correlation between two variables is positive and close to 1, we say that the variables have a *strong positive linear correlation*. If the correlation between two variables is positive but close to zero, then the variables have a *weak positive linear correlation*. In contrast, if the correlation between two variables is negative and close to  $-1$ , then the variables are said to have a *strong negative linear correlation*. If the correlation between two variables is negative but close to zero, there exists a *weak negative linear correlation* between the variables. Graphically, a strong correlation indicates that the points in the scatter diagram are very close to the regression line,

**Figure 13.18** Linear correlation between two variables.  
(a) Perfect positive linear correlation,  $r = 1$ . (b) Perfect negative linear correlation,  $r = -1$ .  
(c) No linear correlation,  $r \approx 0$ .

and a weak correlation indicates that the points in the scatter diagram are widely spread around the regression line. These four cases are shown in Figure 13.19a–d.



**Figure 13.19** Linear correlation between two variables.

The linear correlation coefficient is calculated by using the following formula. (This correlation coefficient is also called the *Pearson product moment correlation coefficient*.)

**Linear Correlation Coefficient** The *simple linear correlation coefficient*, denoted by  $r$ , measures the strength of the linear relationship between two variables for a sample and is calculated as<sup>6</sup>

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Because both  $SS_{xx}$  and  $SS_{yy}$  are always positive, the sign of the correlation coefficient  $r$  depends on the sign of  $SS_{xy}$ . If  $SS_{xy}$  is positive, then  $r$  will be positive, and if  $SS_{xy}$  is negative, then  $r$  will be negative. Another important observation to remember is that  $r$  and  $b$ , calculated for the same sample, will always have the same sign. That is, both  $r$  and  $b$  are either positive or negative. This is so because both  $r$  and  $b$  provide information about the relationship between  $x$  and  $y$ . Likewise, the corresponding population parameters  $\rho$  and  $B$  will always have the same sign.

Example 13–6 illustrates the calculation of the linear correlation coefficient  $r$ .

## ■ EXAMPLE 13–6

Calculate the correlation coefficient for the example on incomes and food expenditures of seven households.

*Calculating the linear correlation coefficient.*

**Solution** From earlier calculations made in Examples 13–1 and 13–2,

$$SS_{xy} = 447.5714, \quad SS_{xx} = 1772.8571, \quad \text{and} \quad SS_{yy} = 125.7143$$

<sup>6</sup>If we have access to population data, the value of  $\rho$  is calculated using the formula

$$\rho = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Here the values of  $SS_{xy}$ ,  $SS_{xx}$ , and  $SS_{yy}$  are calculated using the population data.

Substituting these values in the formula for  $r$ , we obtain

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{447.5714}{\sqrt{(1772.8571)(125.7143)}} = .9481 = .95$$

Thus, the linear correlation coefficient is .95. The correlation coefficient is usually rounded to two decimal places. ■

The linear correlation coefficient simply tells us how strongly the two variables are (linearly) related. The correlation coefficient of .95 for incomes and food expenditures of seven households indicates that income and food expenditure are very strongly and positively correlated. This correlation coefficient does not, however, provide us with any more information.

*The square of the correlation coefficient gives the coefficient of determination*, which was explained in Section 13.4. Thus,  $(.95)^2$  is .90, which is the value of  $r^2$  calculated in Example 13–3.

Sometimes the calculated value of  $r$  may indicate that the two variables are very strongly linearly correlated, but in reality they may not be. For example, if we calculate the correlation coefficient between the price of haircut and the size of families in the United States using data for the last 30 years, we will find a strong negative linear correlation. Over time, the price of haircut has increased and the size of families has decreased. This finding does not mean that family size and the price of haircut are related. As a result, before we calculate the correlation coefficient, we must seek help from a theory or from common sense to postulate whether or not the two variables have a causal relationship.

Another point to note is that in a simple regression model, one of the two variables is categorized as an independent (also known as an explanatory or predictor) variable and the other is classified as a dependent (also known as a response) variable. However, no such distinction is made between the two variables when the correlation coefficient is calculated.

### 13.4.2 Hypothesis Testing About the Linear Correlation Coefficient

This section describes how to perform a test of hypothesis about the population correlation coefficient  $\rho$  using the sample correlation coefficient  $r$ . We can use the  $t$  distribution to make this test. However, to use the  $t$  distribution, both variables should be normally distributed.

Usually (although not always), the null hypothesis is that the linear correlation coefficient between the two variables is zero, that is,  $\rho = 0$ . The alternative hypothesis can be one of the following: (1) the linear correlation coefficient between the two variables is less than zero, that is,  $\rho < 0$ ; (2) the linear correlation coefficient between the two variables is greater than zero, that is,  $\rho > 0$ ; or (3) the linear correlation coefficient between the two variables is not equal to zero, that is,  $\rho \neq 0$ .

**Test Statistic for  $r$**  If both variables are normally distributed and the null hypothesis is  $H_0: \rho = 0$ , then the value of the test statistic  $t$  is calculated as

$$t = r \sqrt{\frac{n - 2}{1 - r^2}}$$

Here  $n - 2$  are the degrees of freedom.

Example 13–7 describes the procedure to perform a test of hypothesis about the linear correlation coefficient.

#### ■ EXAMPLE 13–7

*Performing a test of hypothesis about the correlation coefficient.*

Using a 1% level of significance and the data from Example 13–1, test whether the linear correlation coefficient between incomes and food expenditures is positive. Assume that the populations of both variables are normally distributed.

**Solution** From Examples 13–1 and 13–6,

$$n = 7 \quad \text{and} \quad r = .9481$$

Below we use the five steps to perform this test of hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

We are to test whether the linear correlation coefficient between incomes and food expenditures is positive. Hence, the null and alternative hypotheses are, respectively,

$$H_0: \rho = 0 \quad (\text{The linear correlation coefficient is zero.})$$

$$H_1: \rho > 0 \quad (\text{The linear correlation coefficient is positive.})$$

**Step 2.** *Select the distribution to use.*

The population distributions for both variables are normally distributed. Hence, we can use the  $t$  distribution to perform this test about the linear correlation coefficient.

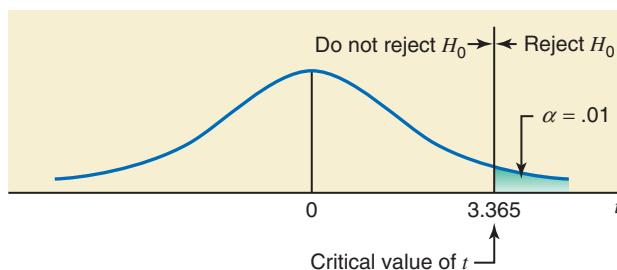
**Step 3.** *Determine the rejection and nonrejection regions.*

The significance level is 1%. From the alternative hypothesis we know that the test is right-tailed. Hence,

$$\text{Area in the right tail of the } t \text{ distribution} = .01$$

$$df = n - 2 = 7 - 2 = 5$$

From the  $t$  distribution table, the critical value of  $t$  is 3.365. The rejection and nonrejection regions for this test are shown in Figure 13.20.



**Figure 13.20** Rejection and nonrejection regions.

**Step 4.** *Calculate the value of the test statistic.*

The value of the test statistic  $t$  for  $r$  is calculated as follows:

$$t = r \sqrt{\frac{n-2}{1-r^2}} = .9481 \sqrt{\frac{7-2}{1-(.9481)^2}} = 6.667$$

**Step 5.** *Make a decision.*

The value of the test statistic  $t = 6.667$  is greater than the critical value of  $t = 3.365$ , and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that there is a positive linear relationship between incomes and food expenditures.

### Using the $p$ -Value to Make a Decision

We can find the range for the  $p$ -value from the  $t$  distribution table (Table V of Appendix C) and make a decision by comparing that  $p$ -value with the significance level. For this example,  $df = 5$ , and the observed value of  $t$  is 6.667. From Table V (the  $t$  distribution table) in the row of  $df = 5$ , the largest value of  $t$  is 5.893, for which the area in the right tail of the  $t$  distribution is .001. Since our observed value of  $t = 6.667$  is larger than 5.893, the  $p$ -value for  $t = 6.667$  is less than .001, that is,

$$p\text{-value} < .001$$

Thus, we can state that for any  $\alpha$  equal to or greater than .001 (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For our example,  $\alpha = .01$ , which is greater than the  $p$ -value of .001. As a result, we reject the null hypothesis. ■

## EXERCISES

### ■ CONCEPTS AND PROCEDURES

**13.59** What does a linear correlation coefficient tell about the relationship between two variables? Within what range can a correlation coefficient assume a value?

**13.60** What is the difference between  $\rho$  and  $r$ ? Explain.

**13.61** Explain each of the following concepts. You may use graphs to illustrate each concept.

- a. Perfect positive linear correlation
- b. Perfect negative linear correlation
- c. Strong positive linear correlation
- d. Strong negative linear correlation
- e. Weak positive linear correlation
- f. Weak negative linear correlation
- g. No linear correlation

**13.62** Can the values of  $B$  and  $\rho$  calculated for the same population data have different signs? Explain.

**13.63** For a sample data set, the linear correlation coefficient  $r$  has a positive value. Which of the following is true about the slope  $b$  of the regression line estimated for the same sample data?

- a. The value of  $b$  will be positive.
- b. The value of  $b$  will be negative.
- c. The value of  $b$  can be positive or negative.

**13.64** For a sample data set, the slope  $b$  of the regression line has a negative value. Which of the following is true about the linear correlation coefficient  $r$  calculated for the same sample data?

- a. The value of  $r$  will be positive.
- b. The value of  $r$  will be negative.
- c. The value of  $r$  can be positive or negative.

**13.65** For a sample data set on two variables, the value of the linear correlation coefficient is (close to) zero. Does this mean that these variables are not related? Explain.

**13.66** Will you expect a positive, zero, or negative linear correlation between the two variables for each of the following examples?

- a. Grade of a student and hours spent studying
- b. Incomes and entertainment expenditures of households
- c. Ages of women and makeup expenses per month
- d. Price of a computer and consumption of Coca-Cola
- e. Price and consumption of wine

**13.67** Will you expect a positive, zero, or negative linear correlation between the two variables for each of the following examples?

- a. SAT scores and GPAs of students
- b. Stress level and blood pressure of individuals
- c. Amount of fertilizer used and yield of corn per acre
- d. Ages and prices of houses
- e. Heights of husbands and incomes of their wives

**13.68** A population data set produced the following information.

$$N = 250, \quad \Sigma x = 9880, \quad \Sigma y = 1456, \quad \Sigma xy = 85,080, \\ \Sigma x^2 = 485,870, \quad \text{and} \quad \Sigma y^2 = 135,675$$

Find the linear correlation coefficient  $\rho$ .

**13.69** A population data set produced the following information.

$$N = 460, \quad \Sigma x = 3920, \quad \Sigma y = 2650, \quad \Sigma xy = 26,570, \\ \Sigma x^2 = 48,530, \quad \text{and} \quad \Sigma y^2 = 39,347$$

Find the linear correlation coefficient  $\rho$ .

**13.70** A sample data set produced the following information.

$$n = 10, \quad \Sigma x = 100, \quad \Sigma y = 220, \quad \Sigma xy = 3680, \\ \Sigma x^2 = 1140, \quad \text{and} \quad \Sigma y^2 = 25,272$$

- a. Calculate the linear correlation coefficient  $r$ .
- b. Using a 2% significance level, can you conclude that  $\rho$  is different from zero?

**13.71** A sample data set produced the following information.

$$n = 12, \quad \Sigma x = 66, \quad \Sigma y = 588, \quad \Sigma xy = 2244, \\ \Sigma x^2 = 396, \quad \text{and} \quad \Sigma y^2 = 58,734$$

- Calculate the linear correlation coefficient  $r$ .
- Using a 1% significance level, can you conclude that  $\rho$  is negative?

## ■ APPLICATIONS

**13.72** Refer to Exercise 13.25. The data on ages (in years) and prices (in hundreds of dollars) for eight cars of a specific model are reproduced from that exercise.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

- Do you expect the ages and prices of cars to be positively or negatively related? Explain.
- Calculate the linear correlation coefficient.
- Test at a 2.5% significance level whether  $\rho$  is negative.

**13.73** The following table, reproduced from Exercise 13.53, gives the experience (in years) and monthly salaries (in hundreds of dollars) of nine randomly selected secretaries.

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	62	29	37	43	35	60	67	32	60

- Do you expect the experience and monthly salaries to be positively or negatively related? Explain.
- Compute the linear correlation coefficient.
- Test at a 5% significance level whether  $\rho$  is positive.

**13.74** The following table lists the midterm and final exam scores for seven students in a statistics class.

Midterm score	79	95	81	66	87	94	59
Final exam score	85	97	78	76	94	84	67

- Do you expect the midterm and final exam scores to be positively or negatively related?
- Plot a scatter diagram. By looking at the scatter diagram, do you expect the correlation coefficient between these two variables to be close to zero, 1, or  $-1$ ?
- Find the correlation coefficient. Is the value of  $r$  consistent with what you expected in parts a and b?
- Using a 1% significance level, test whether the linear correlation coefficient is positive.

**13.75** The following data give the ages (in years) of husbands and wives for six couples.

Husband's age	43	57	28	19	35	39
Wife's age	37	51	32	20	33	38

- Do you expect the ages of husbands and wives to be positively or negatively related?
- Plot a scatter diagram. By looking at the scatter diagram, do you expect the correlation coefficient between these two variables to be close to zero, 1, or  $-1$ ?
- Find the correlation coefficient. Is the value of  $r$  consistent with what you expected in parts a and b?
- Using a 5% significance level, test whether the correlation coefficient is different from zero.

**13.76** The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Source: kelloggs.com.

- Find the correlation coefficient. Is its sign the same as that of  $b$  calculated in Exercise 13.26?
- Test at a 1% significance level whether the linear correlation coefficient between the two variables listed in the table is positive.

**13.77** The following table, reproduced from Exercises 13.38 and 13.57, gives information on the calorie count and grams of fat for 11 types of bagels produced by Panera Bread.

Bagel	Calories	Fat (grams)
Asiago Cheese	330	6.0
Blueberry	330	1.5
Chocolate Chip	370	6.0
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	2.5
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.5
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

- a. Find the correlation coefficient. Is the sign of the correlation coefficient the same as that of  $b$  calculated in Exercise 13.57?
- b. Test at a 1% significance level whether  $\rho$  is different from zero.

**13.78** Refer to data given in Exercise 13.29 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the National League baseball teams. Compute the linear correlation coefficient,  $\rho$ . Does it make sense to make a confidence interval and to test a hypothesis about  $\rho$  here? Explain.

**13.79** Refer to data given in Exercise 13.30 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the American League baseball teams. Compute the linear correlation coefficient,  $\rho$ . Does it make sense to make a confidence interval and to test a hypothesis about  $\rho$  here? Explain.

## 13.5 Regression Analysis: A Complete Example

This section works out an example that includes all the topics we have discussed so far in this chapter.

### ■ EXAMPLE 13–8

*A complete example of regression analysis.*



A random sample of eight drivers selected from a small town insured with a company and having similar minimum required auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums (in dollars):

Driving Experience (years)	Monthly Auto Insurance Premium (\$)
5	64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

- (a) Does the insurance premium depend on the driving experience, or does the driving experience depend on the insurance premium? Do you expect a positive or a negative relationship between these two variables?
- (b) Compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .

- (c) Find the least squares regression line by choosing appropriate dependent and independent variables based on your answer in part a.
- (d) Interpret the meaning of the values of  $a$  and  $b$  calculated in part c.
- (e) Plot the scatter diagram and the regression line.
- (f) Calculate  $r$  and  $r^2$ , and explain what they mean.
- (g) Predict the monthly auto insurance premium for a driver with 10 years of driving experience.
- (h) Compute the standard deviation of errors.
- (i) Construct a 90% confidence interval for  $B$ .
- (j) Test at a 5% significance level whether  $B$  is negative.
- (k) Using  $\alpha = .05$ , test whether  $\rho$  is different from zero.

### Solution

- (a) Based on theory and intuition, we expect the insurance premium to depend on driving experience. Consequently, the insurance premium is a dependent variable (variable  $y$ ) and driving experience is an independent variable (variable  $x$ ) in the regression model. A new driver is considered a high risk by the insurance companies, and he or she has to pay a higher premium for auto insurance. On average, the insurance premium is expected to decrease with an increase in the years of driving experience. Therefore, we expect a negative relationship between these two variables. In other words, both the population correlation coefficient  $\rho$  and the population regression slope  $B$  are expected to be negative.
- (b) Table 13.5 shows the calculation of  $\Sigma x$ ,  $\Sigma y$ ,  $\Sigma xy$ ,  $\Sigma x^2$ , and  $\Sigma y^2$ .

**Table 13.5**

Experience $x$	Premium $y$	$xy$	$x^2$	$y^2$
5	64	320	25	4096
2	87	174	4	7569
12	50	600	144	2500
9	71	639	81	5041
15	44	660	225	1936
6	56	336	36	3136
25	42	1050	625	1764
16	60	960	256	3600
$\Sigma x = 90$	$\Sigma y = 474$	$\Sigma xy = 4739$	$\Sigma x^2 = 1396$	$\Sigma y^2 = 29,642$

The values of  $\bar{x}$  and  $\bar{y}$  are

$$\bar{x} = \Sigma x/n = 90/8 = 11.25$$

$$\bar{y} = \Sigma y/n = 474/8 = 59.25$$

The values of  $SS_{xy}$ ,  $SS_{xx}$ , and  $SS_{yy}$  are computed as follows:

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4739 - \frac{(90)(474)}{8} = \mathbf{-593.5000}$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1396 - \frac{(90)^2}{8} = \mathbf{383.5000}$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 29,642 - \frac{(474)^2}{8} = \mathbf{1557.5000}$$

- (c) To find the regression line, we calculate  $a$  and  $b$  as follows:

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-593.5000}{383.5000} = -1.5476$$

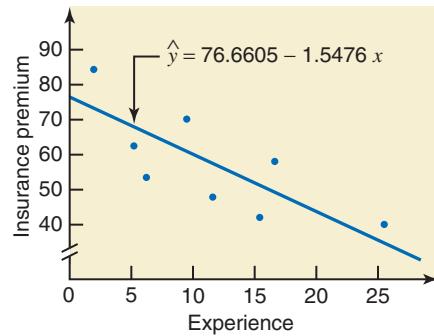
$$a = \bar{y} - b\bar{x} = 59.25 - (-1.5476)(11.25) = 76.6605$$

Thus, our estimated regression line  $\hat{y} = a + bx$  is

$$\hat{y} = 76.6605 - 1.5476x$$

- (d) The value of  $a = 76.6605$  gives the value of  $\hat{y}$  for  $x = 0$ ; that is, it gives the monthly auto insurance premium for a driver with no driving experience. However, as mentioned earlier in this chapter, we should not attach much importance to this statement because the sample contains drivers with only 2 or more years of experience. The value of  $b$  gives the change in  $\hat{y}$  due to a change of one unit in  $x$ . Thus,  $b = -1.5476$  indicates that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by \$1.55. Note that when  $b$  is negative,  $y$  decreases as  $x$  increases.
- (e) Figure 13.21 shows the scatter diagram and the regression line for the data on eight auto drivers. Note that the regression line slopes downward from left to right. This result is consistent with the negative relationship we anticipated between driving experience and insurance premium.

**Figure 13.21** Scatter diagram and the regression line.



- (f) The values of  $r$  and  $r^2$  are computed as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{-593.5000}{\sqrt{(383.5000)(1557.5000)}} = -.7679 = -.77$$

$$r^2 = \frac{b SS_{xy}}{SS_{yy}} = \frac{(-1.5476)(-593.5000)}{1557.5000} = .5897 = .59$$

The value of  $r = -.77$  indicates that the driving experience and the monthly auto insurance premium are negatively related. The (linear) relationship is strong but not very strong. The value of  $r^2 = .59$  states that 59% of the total variation in insurance premiums is explained by years of driving experience, and 41% is not. The low value of  $r^2$  indicates that there may be many other important variables that contribute to the determination of auto insurance premiums. For example, the premium is expected to depend on the driving record of a driver and the type and age of the car.

- (g) Using the estimated regression line, we find the predicted value of  $y$  for  $x = 10$  as:

$$\hat{y} = 76.6605 - 1.5476(10) = 76.6605 - 1.5476(10) = \$61.18$$

Thus, we expect the monthly auto insurance premium of a driver with 10 years of driving experience to be \$61.18.

- (h) The standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{n - 2}} = \sqrt{\frac{1557.5000 - (-1.5476)(-593.5000)}{8 - 2}} = 10.3199$$

- (i) To construct a 90% confidence interval for  $B$ , first we calculate the standard deviation of  $b$ :

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{10.3199}{\sqrt{383.5000}} = .5270$$

For a 90% confidence level, the area in each tail of the  $t$  distribution is

$$\alpha/2 = (1 - .90)/2 = .05$$

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

From the  $t$  distribution table, the  $t$  value for .05 area in the right tail of the  $t$  distribution and 6  $df$  is 1.943. The 90% confidence interval for  $B$  is

$$\begin{aligned} b \pm ts_b &= -1.5476 \pm 1.943(.5270) \\ &= -1.5476 \pm 1.0240 = -2.57 \text{ to } -.52 \end{aligned}$$

Thus, we can state with 90% confidence that  $B$  lies in the interval  $-2.57$  to  $-.52$ . That is, on average, the monthly auto insurance premium of a driver decreases by an amount between  $-.52$  and  $\$2.57$  for every extra year of driving experience.

- (j) We perform the following five steps to test the hypothesis about  $B$ .

**Step 1. State the null and alternative hypotheses.**

The null and alternative hypotheses are, respectively,

$$H_0: B = 0 \quad (B \text{ is not negative.})$$

$$H_1: B < 0 \quad (B \text{ is negative.})$$

Note that the null hypothesis can also be written as  $H_0: B \geq 0$ .

**Step 2. Select the distribution to use.**

Because  $\sigma_\epsilon$  is not known, we use the  $t$  distribution to make the hypothesis test.

**Step 3. Determine the rejection and nonrejection regions.**

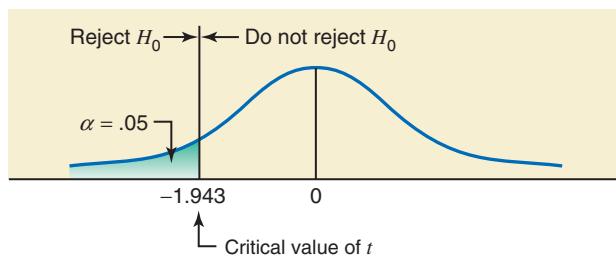
The significance level is .05. The  $<$  sign in the alternative hypothesis indicates that it is a left-tailed test.

Area in the left tail of the  $t$  distribution =  $\alpha = .05$

$$df = n - 2 = 8 - 2 = 6$$

From the  $t$  distribution table, the critical value of  $t$  for .05 area in the left tail of the  $t$  distribution and 6  $df$  is  $-1.943$ , as shown in Figure 13.22.

**Figure 13.22** Rejection and nonrejection regions.



**Step 4. Calculate the value of the test statistic.**

The value of the test statistic  $t$  for  $b$  is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{-1.5476 - 0}{.5270} = -2.937$$

From  $H_0$

**Step 5. Make a decision.**

The value of the test statistic  $t = -2.937$  falls in the rejection region. Hence, we reject the null hypothesis and conclude that  $B$  is negative. That is, the monthly auto insurance premium decreases with an increase in years of driving experience.

**Using the  $p$ -Value to Make a Decision**

We can find the range for the  $p$ -value from the  $t$  distribution table (Table V of Appendix C) and make a decision by comparing that  $p$ -value with the significance level. For this example,  $df = 6$  and the observed value of  $t$  is  $-2.937$ . From Table V (the  $t$  distribution table) in the row of  $df = 6$ ,  $2.937$  is between  $2.447$  and  $3.143$ . The corresponding areas in the right tail of the  $t$  distribution are  $.025$  and  $.01$ , respectively. Our test is left-tailed, however, and the observed value of  $t$  is negative. Thus,  $t = -2.937$  lies between  $-2.447$  and  $-3.143$ . The corresponding areas in the left tail of the  $t$  distribution are  $.025$  and  $.01$ . Therefore the range of the  $p$ -value is

$$.01 < p\text{-value} < .025$$

Thus, we can state that for any  $\alpha$  equal to or greater than  $.025$  (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For our example,  $\alpha = .05$ , which is greater than the upper limit of the  $p$ -value of  $.025$ . As a result, we reject the null hypothesis.

Note that if we use technology to find this  $p$ -value, we will obtain a  $p$ -value of  $.013$ . Then we can reject the null hypothesis for any  $\alpha \geq .013$ .

- (k) We perform the following five steps to test the hypothesis about the linear correlation coefficient  $\rho$ .

**Step 1. State the null and alternative hypotheses.**

The null and alternative hypotheses are, respectively,

$$H_0: \rho = 0 \quad (\text{The linear correlation coefficient is zero.})$$

$$H_1: \rho \neq 0 \quad (\text{The linear correlation coefficient is different from zero.})$$

**Step 2. Select the distribution to use.**

Assuming that variables  $x$  and  $y$  are normally distributed, we will use the  $t$  distribution to perform this test about the linear correlation coefficient.

**Step 3. Determine the rejection and nonrejection regions.**

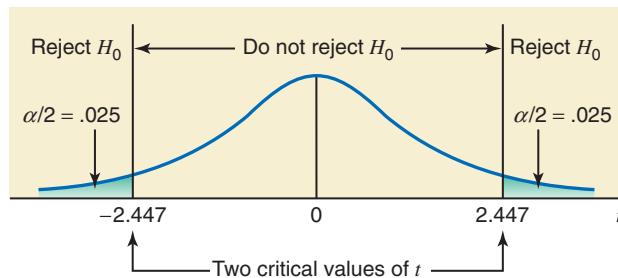
The significance level is  $5\%$ . From the alternative hypothesis we know that the test is two-tailed. Hence,

$$\text{Area in each tail of the } t \text{ distribution} = .05/2 = .025$$

$$df = n - 2 = 8 - 2 = 6$$

From the  $t$  distribution table, Table V of Appendix C, the critical values of  $t$  are  $-2.447$  and  $2.447$ . The rejection and nonrejection regions for this test are shown in Figure 13.23.

**Figure 13.23** Rejection and nonrejection regions.



**Step 4.** Calculate the value of the test statistic.

The value of the test statistic  $t$  for  $r$  is calculated as follows:

$$t = r \sqrt{\frac{n - 2}{1 - r^2}} = (-.7679) \sqrt{\frac{8 - 2}{1 - (-.7679)^2}} = -2.936$$

**Step 5.** Make a decision.

The value of the test statistic  $t = -2.936$  falls in the rejection region. Hence, we reject the null hypothesis and conclude that the linear correlation coefficient between driving experience and auto insurance premium is different from zero.

**Using the  $p$ -Value to Make a Decision**

We can find the range for the  $p$ -value from the  $t$  distribution table and make a decision by comparing that  $p$ -value with the significance level. For this example,  $df = 6$  and the observed value of  $t$  is  $-2.936$ . From Table V (the  $t$  distribution table) in the row of  $df = 6$ ,  $t = 2.936$  is between 2.447 and 3.143. The corresponding areas in the right tail of the  $t$  distribution curve are .025 and .01, respectively. Since the test is two tailed, the range of the  $p$ -value is

$$2(.01) < p\text{-value} < 2(.025) \quad \text{or} \quad .02 < p\text{-value} < .05$$

Thus, we can state that for any  $\alpha$  equal to or greater than .05 (the upper limit of the  $p$ -value range), we will reject the null hypothesis. For our example,  $\alpha = .05$ , which is equal to the upper limit of the  $p$ -value. As a result, we reject the null hypothesis. ■

**EXERCISES****APPLICATIONS**

- 13.80** The owner of a small factory that produces working gloves is concerned about the high cost of air conditioning in the summer but is afraid that keeping the temperature in the factory too high will lower productivity. During the summer, he experiments with temperature settings from 68°F to 81°F and measures each day's productivity. The following table gives the temperature and the number of pairs of gloves (in hundreds) produced on each of the 8 randomly selected days.

Temperature (°F)	72	71	78	75	81	77	68	76
Pairs of gloves	37	37	32	36	33	35	39	34

- Do the pairs of gloves produced depend on temperature, or does temperature depend on pairs of gloves produced? Do you expect a positive or a negative relationship between these two variables?
- Taking temperature as an independent variable and pairs of gloves produced as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- Find the least squares regression line.
- Interpret the meaning of the values of  $a$  and  $b$  calculated in part c.
- Plot the scatter diagram and the regression line.
- Calculate  $r$  and  $r^2$ , and explain what they mean.
- Compute the standard deviation of errors.
- Predict the number of pairs of gloves produced when  $x = 74$ .
- Construct a 99% confidence interval for  $B$ .
- Test at a 5% significance level whether  $B$  is negative.
- Using  $\alpha = .01$  can you conclude that  $\rho$  is negative?

- 13.81** The following table gives information on the limited tread warranties (in thousands of miles) and the prices of 12 randomly selected tires at a national tire retailer as of July 2012.

Warranty (thousands of miles)	60	70	75	50	80	55	65	65	70	65	60	65
Price per tire (\$)	95	135	94	90	121	70	140	80	92	125	160	155

- Taking warranty length as an independent variable and price per tire as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .

- b. Find the regression of price per tire on warranty length.
- c. Briefly explain the meaning of the values of  $a$  and  $b$  calculated in part b.
- d. Calculate  $r$  and  $r^2$  and explain what they mean.
- e. Plot the scatter diagram and the regression line.
- f. Predict the price of a tire with a warranty length of 73,000 miles.
- g. Compute the standard deviation of errors.
- h. Construct a 95% confidence interval for  $B$ .
- i. Test at a 5% significance level if  $B$  is positive.
- j. Using  $\alpha = .025$ , can you conclude that the linear correlation coefficient is positive?

**13.82** The recommended air pressure in a basketball is between 7 and 9 pounds per square inch (psi). When dropped from a height of 6 feet, a properly inflated basketball should bounce upward between 52 and 56 inches (<http://www.bestsoccerbuys.com/balls-basketball.html>). The basketball coach at a local high school purchased 10 new basketballs for the upcoming season, inflated the balls to pressures between 7 and 9 psi, and performed the *bounce test* mentioned above. The data obtained are given in the following table.

Pressure (psi)	7.8	8.1	8.3	7.4	8.9	7.2	8.6	7.5	8.1	8.5
Bounce height (inches)	54.1	54.3	55.2	53.3	55.4	52.2	55.7	54.6	54.8	55.3

- a. With the pressure as an independent variable and bounce height as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- b. Find the least squares regression line.
- c. Interpret the meaning of the values of  $a$  and  $b$  calculated in part b.
- d. Calculate  $r$  and  $r^2$  and explain what they mean.
- e. Compute the standard deviation of errors.
- f. Predict the bounce height of a basketball for  $x = 8.0$ .
- g. Construct a 98% confidence interval for  $B$ .
- h. Test at a 5% significance level whether  $B$  is different from zero.
- i. Using  $\alpha = .05$ , can you conclude that  $\rho$  is different from zero?

**13.83** The following table gives information on the incomes (in thousands of dollars) and charitable contributions (in hundreds of dollars) for the last year for a random sample of 10 households.

Income	Charitable Contributions
76	15
57	4
140	42
97	33
75	5
107	32
65	10
77	18
102	28
53	4

- a. With income as an independent variable and charitable contributions as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- b. Find the regression of charitable contributions on income.
- c. Briefly explain the meaning of the values of  $a$  and  $b$ .
- d. Calculate  $r$  and  $r^2$  and briefly explain what they mean.
- e. Compute the standard deviation of errors.
- f. Construct a 99% confidence interval for  $B$ .
- g. Test at a 1% significance level whether  $B$  is positive.
- h. Using a 1% significance level, can you conclude that the linear correlation coefficient is different from zero?

**13.84** The following data give information on the average ticket prices (in U.S. dollars) and the average percentage of capacity filled for seven hockey teams during the 2011–2012 National Hockey League regular season. (Note: Capacity levels exceeding 100.0% imply standing-room-only attendees.)

Team	Anaheim	Vancouver	Dallas	Edmonton	New Jersey	Toronto	Philadelphia
Average ticket price (\$)	36.94	68.38	29.95	70.13	45.86	123.27	66.89
Percentage capacity filled	86.4	102.5	76.8	100.0	87.4	103.7	107.4

Source: [http://espn.go.com/blog/dallas/stars/post/\\_id/13315/stars-have-cheapest-ticket-in-nhl](http://espn.go.com/blog/dallas/stars/post/_id/13315/stars-have-cheapest-ticket-in-nhl) and <http://espn.go.com/nhl/attendance>.

- Taking average ticket price as an independent variable and percentage of capacity filled as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- Find the least squares regression line.
- Briefly explain the meaning of the values of  $a$  and  $b$  calculated in part b.
- Calculate  $r$  and  $r^2$  and briefly explain what they mean.
- Compute the standard deviation of errors.
- Construct a 95% confidence interval for  $B$ .
- Test at a 2.5% significance level whether  $B$  is positive.
- Using a 2.5% significance level, test whether  $\rho$  is positive.

- 13.85** The following table gives information on GPAs and starting salaries (rounded to the nearest thousand dollars) of seven recent college graduates.

GPA	2.90	3.81	3.20	2.42	3.94	2.05	2.25
Starting salary	48	53	50	37	65	32	37

- With GPA as an independent variable and starting salary as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- Find the least squares regression line.
- Interpret the meaning of the values of  $a$  and  $b$  calculated in part b.
- Calculate  $r$  and  $r^2$  and briefly explain what they mean.
- Compute the standard deviation of errors.
- Construct a 95% confidence interval for  $B$ .
- Test at a 1% significance level whether  $B$  is different from zero.
- Test at a 1% significance level whether  $\rho$  is positive.

## 13.6 Using the Regression Model

Let us return to the example on incomes and food expenditures to discuss two major uses of a regression model:

- Estimating the mean value of  $y$  for a given value of  $x$ . For instance, we can use our food expenditure regression model to estimate the mean food expenditure of all households with a specific income (say, \$5500 per month).
- Predicting a particular value of  $y$  for a given value of  $x$ . For instance, we can determine the expected food expenditure of a randomly selected household with a particular monthly income (say, \$5500) using our food expenditure regression model.

### 13.6.1 Using the Regression Model for Estimating the Mean Value of $y$

Our population regression model is

$$y = A + Bx + \epsilon$$

As mentioned earlier in this chapter, the mean value of  $y$  for a given  $x$  is denoted by  $\mu_{y|x}$ , read as “the mean value of  $y$  for a given value of  $x$ .” Because of the assumption that the mean value of  $\epsilon$  is zero, the mean value of  $y$  is given by

$$\mu_{y|x} = A + Bx$$

Our objective is to estimate this mean value. The value of  $\hat{y}$ , obtained from the sample regression line by substituting the value of  $x$ , is the *point estimate of  $\mu_{y|x}$*  for that  $x$ .

For our example on incomes and food expenditures, the estimated sample regression line (from Example 13–1) is

$$\hat{y} = 1.5050 + .2525x$$

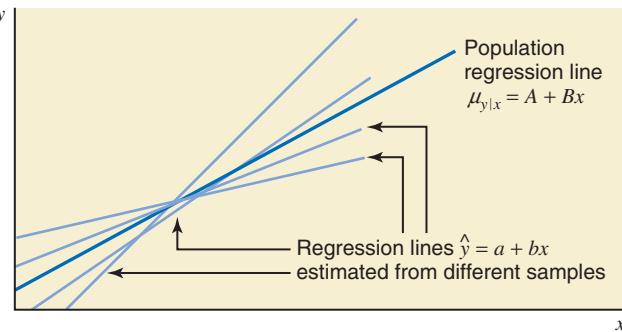
Suppose we want to estimate the mean food expenditure for all households with a monthly income of \$5500. We will denote this population mean by  $\mu_{y|x=55}$  or  $\mu_{y|55}$ . Note that we have written  $x = 55$  and not  $x = 5500$  in  $\mu_{y|55}$  because the units of measurement for the data used to estimate the above regression line in Example 13–1 were hundreds of dollars. Using the regression line, we find that the point estimate of  $\mu_{y|55}$  is

$$\hat{y} = 1.5050 + .2525(55) = \$15.3925 \text{ hundred}$$

Thus, based on the sample regression line, the point estimate for the mean food expenditure  $\mu_{y|55}$  for all households with a monthly income of \$5500 is \$1539.25 per month.

However, suppose we take a second sample of seven households from the same population and estimate the regression line for this sample. The point estimate of  $\mu_{y|55}$  obtained from the regression line for the second sample is expected to be different. All possible samples of the same size taken from the same population will give different regression lines as shown in Figure 13.24, and, consequently, a different point estimate of  $\mu_{y|x}$ . Therefore, a confidence interval constructed for  $\mu_{y|x}$  based on one sample will give a more reliable estimate of  $\mu_{y|x}$  than will a point estimate.

**Figure 13.24** Population and sample regression lines.



To construct a confidence interval for  $\mu_{y|x}$ , we must know the mean, the standard deviation, and the shape of the sampling distribution of its point estimator  $\hat{y}$ .

The point estimator  $\hat{y}$  of  $\mu_{y|x}$  is normally distributed with a mean of  $A + Bx$  and a standard deviation of

$$\sigma_{\hat{y}_m} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where  $\sigma_{\hat{y}_m}$  is the standard deviation of  $\hat{y}$  when it is used to estimate  $\mu_{y|x}$ ,  $x_0$  is the value of  $x$  for which we are estimating  $\mu_{y|x}$ , and  $\sigma_\epsilon$  is the population standard deviation of  $\epsilon$ .

However, usually  $\sigma_\epsilon$  is not known. Rather, it is estimated by the standard deviation of sample errors  $s_e$ . In this case, we replace  $\sigma_\epsilon$  by  $s_e$  and  $\sigma_{\hat{y}_m}$  by  $s_{\hat{y}_m}$  in the foregoing expression. To make a confidence interval for  $\mu_{y|x}$ , we use the  $t$  distribution because  $\sigma_\epsilon$  is not known.

**Confidence Interval for  $\mu_{y|x}$**  The  $(1 - \alpha)100\%$  confidence interval for  $\mu_{y|x}$  for  $x = x_0$  is

$$\hat{y} \pm ts_{\hat{y}_m}$$

where the value of  $t$  is obtained from the  $t$  distribution table for  $\alpha/2$  area in the right tail of the  $t$  distribution curve and  $df = n - 2$ . The value of  $s_{\hat{y}_m}$  is calculated as follows:

$$s_{\hat{y}_m} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Example 13–9 illustrates how to make a confidence interval for the mean value of  $y$ ,  $\mu_{y|x}$ .

### ■ EXAMPLE 13–9

Refer to Example 13–1 on incomes and food expenditures. Find a 99% confidence interval for the mean food expenditure for all households with a monthly income of \$5500.

**Solution** Using the regression line estimated in Example 13–1, we find the point estimate of the mean food expenditure for  $x = 55$  as

$$\hat{y} = 1.5050 + .2525(55) = \$15.3925 \text{ hundred}$$

The confidence level is 99%. Hence, the area in each tail of the  $t$  distribution is

$$\alpha/2 = (1 - .99)/2 = .005$$

The degrees of freedom are

$$df = n - 2 = 7 - 2 = 5$$

From the  $t$  distribution table, the  $t$  value for .005 area in the right tail of the  $t$  distribution and 5  $df$  is 4.032. From calculations in Examples 13–1 and 13–2, we know that

$$s_e = 1.5939, \bar{x} = 55.1429, \text{ and } SS_{xx} = 1772.8571$$

The standard deviation of  $\hat{y}$  as an estimate of  $\mu_{y|x}$  for  $x = 55$  is calculated as follows:

$$s_{\hat{y}_m} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = (1.5939) \sqrt{\frac{1}{7} + \frac{(55 - 55.1429)^2}{1772.8571}} = .6025$$

Hence, the 99% confidence interval for  $\mu_{y|55}$  is

$$\begin{aligned} \hat{y} \pm ts_{\hat{y}_m} &= 15.3925 \pm 4.032(.6025) \\ &= 15.3925 \pm 2.4293 = \mathbf{12.9632 \text{ to } 17.8218} \end{aligned}$$

Thus, with 99% confidence we can state that the mean food expenditure for all households with a monthly income of \$5500 is between \$1296.32 and \$1782.18. ■

*Constructing a confidence interval for the mean value of  $y$  for a given  $x$ .*

### 13.6.2 Using the Regression Model for Predicting a Particular Value of $y$

The second major use of a regression model is to predict a particular value of  $y$  for a given value of  $x$ —say,  $x_0$ . For example, we may want to predict the food expenditure of a randomly selected household with a monthly income of \$5500. In this case, we are not interested in the mean food expenditure of all households with a monthly income of \$5500 but in the food expenditure of one particular household with a monthly income of \$5500. This predicted value of  $y$  is denoted by  $y_p$ . Again, to predict a single value of  $y$  for  $x = x_0$  from the estimated sample regression line, we use the value of  $\hat{y}$  as a point estimate of  $y_p$ . Using the estimated regression line, we find that  $\hat{y}$  for  $x = 55$  is

$$\hat{y} = 1.5050 + .2525(55) = \$15.3925 \text{ hundred}$$

Thus, based on our regression line, the point estimate for the food expenditure of a given household with a monthly income of \$5500 is \$1539.25 per month. Note that  $\hat{y} = 1539.25$  is the point estimate for the mean food expenditure for all households with  $x = 55$  as well as for the predicted value of food expenditure of one household with  $x = 55$ .

Different regression lines estimated by using different samples of seven households each taken from the same population will give different values of the point estimator for the predicted value of  $y$  for  $x = 55$ . Hence, a confidence interval constructed for  $y_p$  based on one sample will give a more reliable estimate of  $y_p$  than will a point estimate. The confidence interval constructed for  $y_p$  is more commonly called a **prediction interval**.

The procedure for constructing a prediction interval for  $y_p$  is similar to that for constructing a confidence interval for  $\mu_{y|x}$  except that the standard deviation of  $\hat{y}$  is larger when we predict a single value of  $y$  than when we estimate  $\mu_{y|x}$ .

The point estimator  $\hat{y}$  of  $y_p$  is normally distributed with a mean of  $A + Bx$  and a standard deviation of

$$\sigma_{\hat{y}_p} = \sigma_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where  $\sigma_{\hat{y}_p}$  is the standard deviation of the predicted value of  $y$ ,  $x_0$  is the value of  $x$  for which we are predicting  $y$ , and  $\sigma_\epsilon$  is the population standard deviation of  $\epsilon$ .

However, usually  $\sigma_\epsilon$  is not known. In this case, we replace  $\sigma_\epsilon$  by  $s_e$  and  $\sigma_{\hat{y}_p}$  by  $s_{\hat{y}_p}$  in the foregoing expression. To make a prediction interval for  $y_p$ , we use the  $t$  distribution when  $\sigma_\epsilon$  is not known.

**Prediction Interval for  $y_p$**  The  $(1 - \alpha)100\%$  prediction interval for the predicted value of  $y$ , denoted by  $y_p$ , for  $x = x_0$  is

$$\hat{y} \pm ts_{\hat{y}_p}$$

where the value of  $t$  is obtained from the  $t$  distribution table for  $\alpha/2$  area in the right tail of the  $t$  distribution curve and  $df = n - 2$ . The value of  $s_{\hat{y}_p}$  is calculated as follows:

$$s_{\hat{y}_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Example 13–10 illustrates the procedure to make a prediction interval for a particular value of  $y$ .

### ■ EXAMPLE 13–10

Making a prediction interval for a particular value of  $y$  for a given  $x$ .

Refer to Example 13–1 on incomes and food expenditures. Find a 99% prediction interval for the predicted food expenditure for a randomly selected household with a monthly income of \$5500.

**Solution** Using the regression line estimated in Example 13–1, we find the point estimate of the predicted food expenditure for  $x = 55$ :

$$\hat{y} = 1.5050 + .2525(55) = \$15.3925 \text{ hundred}$$

The area in each tail of the  $t$  distribution for a 99% confidence level is

$$\alpha/2 = (1 - .99)/2 = .005$$

The degrees of freedom are

$$df = n - 2 = 7 - 2 = 5$$

From the  $t$  distribution table, the  $t$  value for .005 area in the right tail of the  $t$  distribution curve and 5  $df$  is 4.032. From calculations in Examples 13–1 and 13–2,

$$s_e = 1.5939, \quad \bar{x} = 55.1429, \quad \text{and} \quad SS_{xx} = 1772.8571$$

The standard deviation of  $\hat{y}$  as an estimator of  $y_p$  for  $x = 55$  is calculated as follows:

$$\begin{aligned} s_{\hat{y}_p} &= s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \\ &= (1.5939) \sqrt{1 + \frac{1}{7} + \frac{(55 - 55.1429)^2}{1772.8571}} = 1.7040 \end{aligned}$$

Hence, the 99% prediction interval for  $y_p$  for  $x = 55$  is

$$\begin{aligned} \hat{y} \pm ts_{\hat{y}_p} &= 15.3925 \pm 4.032(1.7040) \\ &= 15.3925 \pm 6.8705 = \mathbf{8.5220 \text{ to } 22.2630} \end{aligned}$$

Thus, with 99% confidence we can state that the predicted food expenditure of a household with a monthly income of \$5500 is between \$852.20 and \$2226.30. ■

As we can observe in Example 13–10, this interval is much wider than the one for the mean value of  $y$  for  $x = 55$  calculated in Example 13–9, which was \$1296.32 to \$1782.18. This is always true. The prediction interval for predicting a single value of  $y$  is always larger than the confidence interval for estimating the mean value of  $y$  for a certain value of  $x$ .

## EXERCISES

### CONCEPTS AND PROCEDURES

**13.86** Briefly explain the difference between estimating the mean value of  $y$  and predicting a particular value of  $y$  using a regression model.

**13.87** Construct a 99% confidence interval for the mean value of  $y$  and a 99% prediction interval for the predicted value of  $y$  for the following.

a.  $\hat{y} = 3.25 + .80x$  for  $x = 15$  given  $s_e = .954$ ,  $\bar{x} = 18.52$ ,  $SS_{xx} = 144.65$ , and  $n = 10$

b.  $\hat{y} = -27 + 7.67x$  for  $x = 12$  given  $s_e = 2.46$ ,  $\bar{x} = 13.43$ ,  $SS_{xx} = 369.77$ , and  $n = 10$

**13.88** Construct a 95% confidence interval for the mean value of  $y$  and a 95% prediction interval for the predicted value of  $y$  for the following.

a.  $\hat{y} = 13.40 + 2.58x$  for  $x = 8$  given  $s_e = 1.29$ ,  $\bar{x} = 11.30$ ,  $SS_{xx} = 210.45$ , and  $n = 12$

b.  $\hat{y} = -8.6 + 3.72x$  for  $x = 24$  given  $s_e = 1.89$ ,  $\bar{x} = 19.70$ ,  $SS_{xx} = 315.40$ , and  $n = 10$

### APPLICATIONS

**13.89** Refer to Exercise 13.53. Construct a 90% confidence interval for the mean monthly salary of all secretaries with 10 years of experience. Construct a 90% prediction interval for the monthly salary of a randomly selected secretary with 10 years of experience.

**13.90** Refer to the data on temperature settings and pairs of gloves produced for 8 days given in Exercise 13.80. Construct a 99% confidence interval for  $\mu_{y|x}$  for  $x = 77$  and a 99% prediction interval for  $y_p$  for  $x = 77$ .

**13.91** Refer to Exercise 13.81. Construct a 95% confidence interval for the mean price of all tires that have a 65,000-mile limited tread warranty. Construct a 95% prediction interval for the price of a randomly selected tire that has a 65,000-mile limited tread warranty.

**13.92** Refer to Exercise 13.82. Construct a 99% confidence interval for the mean bounce height of all basketballs that are inflated to 8.5 psi. Construct a 99% prediction interval for the bounce height of a randomly selected basketball that is inflated to 8.5 psi.

**13.93** Refer to Exercise 13.83. Construct a 95% confidence interval for the mean charitable contributions made by all households with an income of \$84,000. Make a 95% prediction interval for the charitable contributions made by a randomly selected household with an income of \$84,000.

**13.94** Refer to Exercise 13.85. Construct a 98% confidence interval for the mean starting salary of recent college graduates with a GPA of 3.15. Construct a 98% prediction interval for the starting salary of a randomly selected recent college graduate with a GPA of 3.15.

## USES AND MISUSES...

### 1. PROCESSING ERRORS

Stuck on the far right side of the linear regression model is the Greek letter epsilon,  $\epsilon$ . Despite its diminutive size, proper respect for the error term is critical to good linear regression modeling and analysis.

One interpretation of the error term is that it is a process. Imagine you are a chemist and you have to weigh a number of chemicals for an experiment. The balance that you use in your laboratory is very accurate—so accurate, in fact, that the shuffling of your feet,

your exhaling near it, or the rumbling of trucks on the road outside can cause the reading to fluctuate. Because the value of the measurement that you take will be affected by a number of factors out of your control, you must make several measurements for each chemical, note each measurement, and then take the means and standard deviations of your samples. The distribution of measurements around a mean is the result of a random error process dependent on a number of factors out of your control; each time you use the balance, the measurement you take is the sum of the actual mass of the

chemical and a “random” error. In this example, the measurements will most likely be normally distributed around the mean.

Linear regression analysis makes the same assumption about the two variables you are comparing: The value of the dependent variable is a linear function of the independent variable, plus a little bit of error that you cannot control. Unfortunately, when working with economic or survey data, you rarely can duplicate an experiment to identify the error model. As a statistician, however, you can use the errors to help you refine your model of the relationship among the variables and to guide your collection of new data. For example, if the errors are skewed to the right for moderate values of the independent variable and skewed to the left for small and large values of the independent variable, you can modify your model to account for this difference. Or you can think about other relationships among the variables that might explain this particular distribution of errors. A detailed analysis of the error in your model can be just as instructive as analysis of the slope and  $y$ -intercept of the identified model.

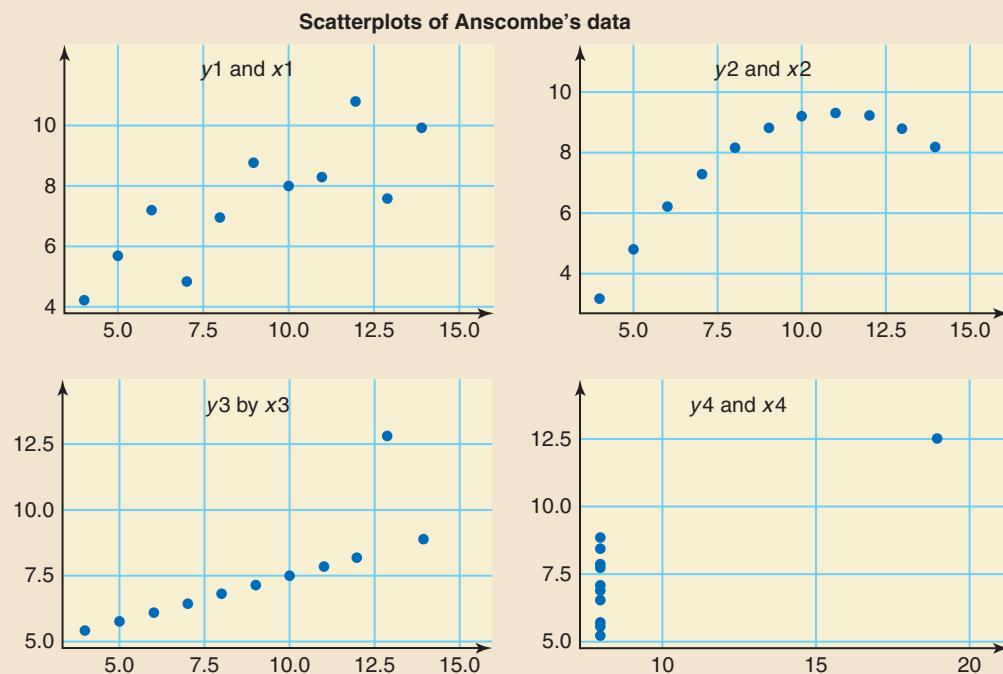
## 2. OUTLIERS AND CORRELATION

In Chapter 3 we learned that outliers can affect the values of some of the summary measures such as the mean, standard deviation, and range. Note that although outliers do affect many other summary measures, these three are affected substantially. Here we will see that just looking at a number that represents the correlation coefficient does not provide the entire story. A very famous data set for

demonstrating this concept was created by F. J. Anscombe (Anscombe, F. J., Graphs in Statistical Analysis, *American Statistician*, 27, pp. 17–21). He created four pairs of data sets on  $x$  and  $y$  variables, each of which has a correlation of .816. To the novice, it may seem that the scatterplots for these four data sets should look virtually the same, but that may not be true. Look at the four scatterplots shown in Figure 13.25.

No two of these scatterplots are even remotely close to being the same or even similar. The data used in the upper left plot are linearly associated, as are the data in the lower left plot. However the plot of  $y_3$  versus  $x_3$  contains an outlier. Without this outlier, the correlation between  $x_3$  and  $y_3$  would be 1. On the other hand, there is much more variability in the relationship between  $x_1$  and  $y_1$ . As far as  $x_4$  and  $y_4$  are concerned, the strong correlation is defined by the single point in the upper right corner of the scatterplot. Without this point, there would be no variability among the  $x_4$  values, and the correlation would be undefined. Lastly, the scatterplot of  $y_2$  versus  $x_2$  reveals that there is an extremely well-defined relationship between these variables, but it is not linear. Being satisfied that the correlation coefficient is close to 1.0 between variables  $x_2$  and  $y_2$  implies that there is a strong linear association between the variables when actually we are fitting a line to a set of data that should be represented by another type of mathematical function.

As we have mentioned before, the process of making a graph may seem trivial, but the importance of graphs in our analysis can never be overstated.



**Figure 13.25** Four scatterplots with the same correlation coefficient.

## Glossary

**Coefficient of determination** A measure that gives the proportion (or percentage) of the total variation in a dependent variable that is explained by a given independent variable.

**Degrees of freedom for a simple linear regression model** Sample size minus 2; that is,  $n - 2$ .

**Dependent variable** The variable to be predicted or explained.

**Deterministic model** A model in which the independent variable determines the dependent variable exactly. Such a model gives an exact relationship between two variables.

**Estimated or predicted value of  $y$**  The value of the dependent variable, denoted by  $\hat{y}$ , that is calculated for a given value of  $x$  using the estimated regression model.

**Independent or explanatory variable** The variable included in a model to explain the variation in the dependent variable.

**Least squares estimates of  $A$  and  $B$**  The values of  $a$  and  $b$  that are calculated by using the sample data.

**Least squares method** The method used to fit a regression line through a scatter diagram such that the error sum of squares is minimum.

**Least squares regression line** A regression line obtained by using the least squares method.

**Linear correlation coefficient** A measure of the strength of the linear relationship between two variables.

**Linear regression model** A regression model that gives a straight-line relationship between two variables.

**Multiple regression model** A regression model that contains two or more independent variables.

**Negative relationship between two variables** The value of the slope in the regression line and the correlation coefficient between two variables are both negative.

**Nonlinear (simple) regression model** A regression model that does not give a straight-line relationship between two variables.

**Population parameters for a simple regression model** The values of  $A$  and  $B$  for the regression model  $y = A + Bx + \epsilon$  that are obtained by using population data.

**Positive relationship between two variables** The value of the slope in the regression line and the correlation coefficient between two variables are both positive.

**Prediction interval** The confidence interval for a particular value of  $y$  for a given value of  $x$ .

**Probabilistic or statistical model** A model in which the independent variable does not determine the dependent variable exactly.

**Random error term ( $\epsilon$ )** The difference between the actual and predicted values of  $y$ .

**Scatter diagram or scatterplot** A plot of the paired observations of  $x$  and  $y$ .

**Simple linear regression** A regression model with one dependent and one independent variable that assumes a straight-line relationship.

**Slope** The coefficient of  $x$  in a regression model that gives the change in  $y$  for a change of one unit in  $x$ .

**SSE** (error sum of squares) The sum of the squared differences between the actual and predicted values of  $y$ . It is the portion of the SST that is not explained by the regression model.

**SSR** (regression sum of squares) The portion of the SST that is explained by the regression model.

**SST** (total sum of squares) The sum of the squared differences between actual  $y$  values and  $\bar{y}$ .

**Standard deviation of errors** A measure of spread for the random errors.

**y-intercept** The point at which the regression line intersects the vertical axis on which the dependent variable is marked. It is the value of  $y$  when  $x$  is zero.

## Supplementary Exercises

**13.95** The following data give information on the ages (in years) and the numbers of breakdowns during the last month for a sample of seven machines at a large company.

Age (years)	12	7	2	8	13	9	4
Number of breakdowns	10	5	1	4	12	7	2

- Taking age as an independent variable and number of breakdowns as a dependent variable, what is your hypothesis about the sign of  $B$  in the regression line? (In other words, do you expect  $B$  to be positive or negative?)
- Find the least squares regression line. Is the sign of  $b$  the same as you hypothesized for  $B$  in part a?
- Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.
- Compute  $r$  and  $r^2$  and explain what they mean.
- Compute the standard deviation of errors.
- Construct a 99% confidence interval for  $B$ .
- Test at a 2.5% significance level whether  $B$  is positive.
- At a 2.5% significance level, can you conclude that  $\rho$  is positive? Is your conclusion the same as in part g?

**13.96** The health department of a large city has developed an air pollution index that measures the level of several air pollutants that cause respiratory distress in humans. The accompanying table gives the pollution index (on a scale of 1 to 10, with 10 being the worst) for 7 randomly selected summer days and the number of patients with acute respiratory problems admitted to the emergency rooms of the city's hospitals.

Air pollution index	4.5	6.7	8.2	5.0	4.6	6.1	3.0
Emergency admissions	53	82	102	60	39	42	27

- a. Taking the air pollution index as an independent variable and the number of emergency admissions as a dependent variable, do you expect  $B$  to be positive or negative in the regression model  $y = A + Bx + \epsilon$ ?
- b. Find the least squares regression line. Is the sign of  $b$  the same as you hypothesized for  $B$  in part a?
- c. Compute  $r$  and  $r^2$ , and explain what they mean.
- d. Compute the standard deviation of errors.
- e. Construct a 90% confidence interval for  $B$ .
- f. Test at a 5% significance level whether  $B$  is positive.
- g. Test at a 5% significance level whether  $\rho$  is positive. Is your conclusion the same as in part f?

**13.97** The management of a supermarket wants to find if there is a relationship between the number of times a specific product is promoted on the intercom system in the store and the number of units of that product sold. To experiment, the management selected a product and promoted it on the intercom system for 7 days. The following table gives the number of times this product was promoted each day and the number of units sold.

Number of Promotions per Day	Number of Units Sold per Day (hundreds)
15	11
22	22
42	30
30	26
18	17
12	15
38	23

- a. With the number of promotions as an independent variable and the number of units sold as a dependent variable, what do you expect the sign of  $B$  in the regression line  $y = A + Bx + \epsilon$  will be?
- b. Find the least squares regression line  $\hat{y} = a + bx$ . Is the sign of  $b$  the same as you hypothesized for  $B$  in part a?
- c. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part b.
- d. Compute  $r$  and  $r^2$  and explain what they mean.
- e. Predict the number of units of this product sold on a day with 35 promotions.
- f. Compute the standard deviation of errors.
- g. Construct a 98% confidence interval for  $B$ .
- h. Testing at a 1% significance level, can you conclude that  $B$  is positive?
- i. Using  $\alpha = .02$ , can you conclude that the correlation coefficient is different from zero?

**13.98** The following table provides information on the living area (in square feet) and price (in thousands of dollars) of 10 randomly selected houses listed for sale in a city.

Living area	3008	2032	2272	1840	2579	2583	1650	3932	2978	2176
Price	275	220	255	189	260	284	172	370	295	260

- a. Find the least squares regression line  $\hat{y} = a + bx$ . Take living area as an independent variable and price as a dependent variable.

- b. Give a brief interpretation of the values of  $a$  and  $b$ .
- c. Compute  $r$  and  $r^2$  and explain what they mean.
- d. Predict the price of a house with 2700 square feet of living area.
- e. Compute the standard deviation of errors.
- f. Construct a 99% confidence interval for  $B$ .
- g. Testing at a 1% significance level, can you conclude that  $B$  is different from zero?
- h. Using  $\alpha = .01$ , can you conclude that the correlation coefficient is different from zero?

**13.99** A local ice cream parlor wants to determine whether the temperature has an effect on its business. The following table contains data on the temperature at 7 PM on 10 rain-free weekend days during the summer and the number of customers served by this ice cream parlor.

Temperature (°F)	68	63	74	72	79	78	71	71	69	66
Customers served	317	355	463	419	507	482	433	388	362	340

- a. With temperature as an independent variable and number of customers as a dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- b. Construct a scatter diagram for these data. Does the scatter diagram exhibit a positive linear relationship between temperature and the number of customers served?
- c. Find the regression equation  $\hat{y} = a + bx$ .
- d. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part c.
- e. Compute the correlation coefficient  $r$ .
- f. Predict the number of customers served on a rain-free weekend summer day when the temperature is 73°F. Returning to part b, how reliable do you think this prediction will be? Explain.

**13.100** The following table gives the average weekly retail price of a gallon of regular gasoline in the eastern United States over a 9-week period from December 19, 2011, through February 13, 2012. Consider these 9 weeks as a random sample.

Date	12/19/11	12/26/11	01/02/12	01/09/12	01/16/12	01/23/12	01/30/12	02/06/12	02/13/12
Price (\$)	3.26	3.264	3.322	3.419	3.436	3.455	3.523	3.559	3.617

Source: <http://www.eia.gov/petroleum/gasdiesel/xls/pswrgvwall.xls>.

- a. Assign a value of 0 to 12/19/11, of 1 to 12/16/11, of 2 to 01/02/12, and so on. Call this new variable *Time*. Make a new table with the variables *Time* and *Price*.
- b. With time as an independent variable and price as the dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- c. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear positive relationship between time and price?
- d. Find the least squares regression line  $\hat{y} = a + bx$ .
- e. Give a brief interpretation of the values of  $a$  and  $b$  calculated in part d.
- f. Compute the correlation coefficient  $r$ .
- g. Predict the average price of a gallon of regular gasoline in the eastern United States for  $Time = 26$ . Comment on this prediction.
- h. The following table gives the average weekly retail price of a gallon of regular gasoline in the eastern United States for the weeks 10/24/11 through 12/12/11.

Date	10/24/11	10/31/11	11/07/11	11/14/11	11/21/11	11/28/11	12/05/11	12/12/11
Price (\$)	3.447	3.424	3.401	3.414	3.364	3.308	3.286	3.298

Calculate the correlation coefficient and the least squares regression line for the 17-week period given in the two tables, assigning 10/24/11 a value of 0, 10/31/11 a value of 1, and so on. What happens to the value of the correlation coefficient? Create a scatter diagram along with the regression line for the data having time on the horizontal axis and price on the vertical axis. Use the diagram to explain how the values of  $r$  and  $b$  changed.

**13.101** The following table gives the completion times for the winners in the women's 200-meter dash finals in the Summer Olympic Games from 1972 to 2008. The times are in seconds rounded to the nearest 1/100 second.

Olympic Year	Time (seconds)
1972	22.40
1976	22.37
1980	22.03
1984	21.81
1988	21.34
1992	21.81
1996	22.12
2000	21.85
2004	22.05
2008	21.74

Source: Wikipedia.

- Assign a value of 0 to 1972, 1 to 1976, 2 to 1980, and so on. Call this new variable *Year*. Make a new table with the variables *Year* and *Time*.
- With year as an independent variable and time as the dependent variable, compute  $SS_{xx}$ ,  $SS_{yy}$ , and  $SS_{xy}$ .
- Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear negative relationship between year and time?
- Find the least squares regression line  $\hat{y} = a + bx$ .
- Give a brief interpretation of the values of  $a$  and  $b$  calculated in part d.
- Compute the correlation coefficient  $r$ .
- Predict the time for the year 2016. Comment on this prediction.

**13.102** Refer to the data on ages and numbers of breakdowns for seven machines given in Exercise 13.95. Construct a 99% confidence interval for the mean number of breakdowns per month for all machines with an age of 8 years. Find a 99% prediction interval for the number of breakdowns per month for a randomly selected machine with an age of 8 years.

**13.103** Refer to the data on the air pollution index and the number of emergency hospital admissions for acute respiratory problems given in Exercise 13.96. Determine a 95% confidence interval for the mean number of such emergency admissions on all days with an air pollution index of 7.0. Make a 95% prediction interval for the number of such emergency admissions on a day when the air pollution index is 7.0.

**13.104** Refer to the data given in Exercise 13.97 on the number of times a specific product is promoted on the intercom system in a supermarket and the number of units of that product sold. Make a 90% confidence interval for the mean number of units of that product sold on days with 35 promotions. Construct a 90% prediction interval for the number of units of that product sold on a randomly selected day with 35 promotions.

**13.105** Refer to the data given in Exercise 13.98 on the living area (in square feet) and price (in thousands of dollars) of 10 randomly selected houses listed for sale in a city. Construct a 98% confidence interval for the mean price of all houses with living areas of 2400 square feet. Construct a 98% prediction interval for the price of a randomly selected house with a living area of 2400 square feet.

## Advanced Exercises

**13.106** Consider the data given in the following table.

x	10	20	30	40	50	60
y	12	15	19	21	25	30

- Find the least squares regression line and the linear correlation coefficient  $r$ .
- Suppose that each value of  $y$  given in the table is increased by 5 and the  $x$  values remain unchanged. Would you expect  $r$  to increase, decrease, or remain the same? How do you expect the least squares regression line to change?
- Increase each value of  $y$  given in the table by 5 and find the new least squares regression line and the correlation coefficient  $r$ . Do these results agree with your expectation in part b?

**13.107** Suppose that you work part-time at a bowling alley that is open daily from noon to midnight. Although business is usually slow from noon to 6 PM, the owner has noticed that it is better on hotter days during the summer, perhaps because the premises are comfortably air-conditioned. The owner shows you some data that she gathered last summer. This data set includes the maximum temperature and the number of lines bowled between noon and 6 PM for each of 20 days. (The maximum temperatures ranged from 77°F to 95°F during this period.) The owner would like to know if she can estimate tomorrow's business from noon to 6 PM by looking at tomorrow's weather forecast. She asks you to analyze the data. Let  $x$  be the maximum temperature for a day and  $y$  the number of lines bowled between noon and 6 PM on that day. The computer output based on the data for 20 days provided the following results:

$$\hat{y} = -432 + 7.7x, \quad s_e = 28.17, \quad SS_{xx} = 607, \quad \text{and} \quad \bar{x} = 87.5$$

Assume that the weather forecasts are reasonably accurate.

- Does the maximum temperature seem to be a useful predictor of bowling activity between noon and 6 PM? Use an appropriate statistical procedure based on the information given. Use  $\alpha = .05$ .
- The owner wants to know how many lines of bowling she can expect, on average, for days with a maximum temperature of 90°. Answer using a 95% confidence level.
- The owner has seen tomorrow's weather forecast, which predicts a high of 90°F. About how many lines of bowling can she expect? Answer using a 95% confidence level.
- Give a brief commonsense explanation to the owner for the difference in the interval estimates of parts b and c.
- The owner asks you how many lines of bowling she could expect if the high temperature were 100°F. Give a point estimate, together with an appropriate warning to the owner.

**13.108** An economist is studying the relationship between the incomes of fathers and their sons or daughters. Let  $x$  be the annual income of a 30-year-old person and let  $y$  be the annual income of that person's father at age 30 years, adjusted for inflation. A random sample of 300 thirty-year-olds and their fathers yields a linear correlation coefficient of .60 between  $x$  and  $y$ . A friend of yours, who has read about this research, asks you several questions, such as: Does the positive value of the correlation coefficient suggest that the 30-year-olds tend to earn more than their fathers? Does the correlation coefficient reveal anything at all about the difference between the incomes of 30-year-olds and their fathers? If not, what other information would we need from this study? What does the correlation coefficient tell us about the relationship between the two variables in this example? Write a short note to your friend answering these questions.

**13.109** For the past 25 years Burton Hodge has been keeping track of how many times he mows his lawn and the average size of the ears of corn in his garden. Hearing about the Pearson correlation coefficient from a statistician buddy of his, Burton decides to substantiate his suspicion that the more often he mows his lawn, the bigger are the ears of corn. He does so by computing the correlation coefficient. Lo and behold, Burton finds a .93 coefficient of correlation! Elated, he calls his friend the statistician to thank him and announce that next year he will have prize-winning ears of corn because he plans to mow his lawn every day. Do you think Burton's logic is correct? If not, how would you explain to Burton the mistake he is making in his presumption (without eroding his new opinion of statistics)? Suggest what Burton could do next year to make the ears of corn large, and relate this to the Pearson correlation coefficient.

**13.110** It seems reasonable that the more hours per week a full-time college student works at a job, the less time he or she will have to study and, consequently, the lower his or her GPA would be.

- Assuming a linear relationship, suggest specifically what the equation relating  $x$  and  $y$  would be, where  $x$  is the average number of hours a student works per week and  $y$  represents the student's GPA. Try several values of  $x$  and see if your equation gives reasonable values of  $y$ .
- Using the following observations taken from 10 randomly selected students, compute the regression equation and compare it to yours of part a.

Average number of hours worked	20	28	10	35	5	14	0	40	8	23
GPA	2.8	2.5	3.1	2.1	3.4	3.3	2.8	2.5	3.6	1.8

**13.111** Consider the formulas for calculating a prediction interval for a new (specific) value of  $y$ . For each of the changes mentioned in parts a through c below, state the effect on the width of the confidence interval (increase, decrease, or no change) and why it happens. Note that besides the change mentioned in each part, everything else such as the values of  $a$ ,  $b$ ,  $\bar{x}$ ,  $s_e$ , and  $SS_{xx}$  remains unchanged.

- The confidence level is increased.
- The sample size is increased.

- c. The value of  $x_0$  is moved farther away from the value of  $\bar{x}$ .
- d. What will the value of the margin of error be if  $x_0$  equals  $\bar{x}$ ?

**13.112** For each of the regression lines in Exercises 13.53 through 13.56, interpret the slope in terms of the application of that exercise. In addition, state whether the value of the intercept is logical, and why it is or is not logical. If it is logical, state what the value of the intercept represents in terms of the specific application of that exercise.

**13.113** Consider the following data

$x$	-5	-4	-3	-2	-1	0	1	2	3	4	5
$y$	-125	-64	-27	-8	-1	0	1	8	27	64	125

- a. Calculate the correlation between  $x$  and  $y$ , and perform a hypothesis test to determine if the correlation is significantly greater than zero. Use a significance level of 5%.
- b. Are you willing to conclude that there is a strong linear association between the two variables? Use at least one graph to support your answer, and to explain why or why not.

## Self-Review Test

1. A simple regression is a regression model that contains
  - a. only one independent variable
  - b. only one dependent variable
  - c. more than one independent variable
  - d. both a and b
2. The relationship between independent and dependent variables represented by the (simple) linear regression is that of
  - a. a straight line
  - b. a curve
  - c. both a and b
3. A deterministic regression model is a model that
  - a. contains the random error term
  - b. does not contain the random error term
  - c. gives a nonlinear relationship
4. A probabilistic regression model is a model that
  - a. contains the random error term
  - b. does not contain the random error term
  - c. shows an exact relationship
5. The least squares regression line minimizes the sum of
  - a. errors
  - b. squared errors
  - c. predictions
6. The degrees of freedom for a simple regression model are
  - a.  $n - 1$
  - b.  $n - 2$
  - c.  $n - 5$
7. Is the following statement true or false?

The coefficient of determination gives the proportion of total squared errors (SST) that is explained by the use of the regression model.

8. Is the following statement true or false?

The linear correlation coefficient measures the strength of the linear association between two variables.

9. The value of the coefficient of determination is always in the range
  - a. 0 to 1
  - b. -1 to 1
  - c. -1 to 0
10. The value of the correlation coefficient is always in the range
  - a. 0 to 1
  - b. -1 to 1
  - c. -1 to 0
11. Explain why the random error term  $\epsilon$  is added to the regression model.
12. Explain the difference between  $A$  and  $a$  and between  $B$  and  $b$  for a regression model.
13. Briefly explain the assumptions of a regression model.

14. Briefly explain the difference between the population regression line and a sample regression line.
15. The following table gives the temperatures (in degrees Fahrenheit) at 6 PM and the attendance (rounded to hundreds) at a minor league baseball team's night games on 7 randomly selected evenings in May.

Temperature	61	70	50	65	48	75	53
Attendance	10	16	12	15	8	20	18

- a. Do you think temperature depends on attendance or attendance depends on temperature?
- b. With temperature as an independent variable and attendance as a dependent variable, what is your hypothesis about the sign of  $B$  in the regression model?
- c. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between the two variables?
- d. Find the least squares regression line. Is the sign of  $b$  the same as the one you hypothesized for  $B$  in part b?
- e. Give a brief interpretation of the values of the  $y$ -intercept and slope calculated in part d.
- f. Compute  $r$  and  $r^2$ , and explain what they mean.
- g. Predict the attendance at a night game in May for a temperature of 60°F.
- h. Compute the standard deviation of errors.
- i. Construct a 99% confidence interval for  $B$ .
- j. Testing at a 1% significance level, can you conclude that  $B$  is positive?
- k. Construct a 95% confidence interval for the mean attendance at a night game in May when the temperature is 60°F.
- l. Make a 95% prediction interval for the attendance at a night game in May when the temperature is 60°F.
- m. Using a 1% significance level, can you conclude that the linear correlation coefficient is positive?

## Mini-Projects

### ■ MINI-PROJECT 13-1

Using the weather sections from back issues of a local newspaper or some other source, do the following for a period of 30 or more days. For each day, record the predicted maximum temperature for the next day, and then find the actual maximum temperature in the next day's newspaper. Thus, you will have the predicted and actual maximum temperatures for 30 or more days.

- a. Make a scatter diagram for your data.
- b. Find the regression line with actual maximum temperature as a dependent variable and predicted maximum temperature as an independent variable.
- c. Using a 1% significance level, can you conclude that the slope of the regression line is different from zero?
- d. If the actual maximum temperatures were exactly the same as the predicted maximum temperatures for all days, what would the value of the correlation coefficient be?
- e. Find the correlation coefficient between the predicted and actual maximum temperatures for your data.
- f. Using a 1% significance level, can you conclude that the linear correlation coefficient is positive?

### ■ MINI-PROJECT 13-2

Two friends are arguing about the relationship between the prices of soft drinks and wine in U.S. cities. Justin thinks that the prices of any two types of beverages (a soft drink and wine) should be positively related. Ivan disagrees, arguing that the prices of alcoholic beverages in a city depend primarily on state and local taxes.

- a. Take a random sample of 15 U.S. cities from the City Data that accompany this text. Let  $x$  be the price of a 2-liter bottle of Coca-Cola and  $y$  the price of a 1.5-liter bottle of Livingston Cellars or Gallo Chablis or Chenin Blanc wine. Calculate the linear correlation coefficient between  $x$  and  $y$ .
- b. Does your value of  $r$  suggest a positive linear relationship between  $x$  and  $y$ ?
- c. Do you think finding a regression line makes sense here?
- d. Using a 1% level of significance, can you conclude that the linear correlation coefficient is positive?

### ■ MINI-PROJECT 13-3

Visit a grocery store and choose 30 different types of food items that include nutrition information on the packaging. For each food, identify the amount of fat (in grams) and the sodium content (in milligrams) per serving. Make sure that you pick a wide variety of foods in order to get a wide variety of values of these two variables. For example, selecting 30 different diet sodas would not make for an interesting analysis.

- Calculate the linear correlation coefficient between the two variables. Do you find a positive or a negative association between sodium content and the amount of fat?
- Create a scatterplot of these data using the amount of fat as the  $x$  variable. Does your scatterplot suggest that creating a regression line to represent these data makes sense?
- Find a regression line for your data. If it makes sense to fit a line, interpret the values of the slope and intercept. If it does not make sense, explain why these numbers could be misleading.

### ■ MINI-PROJECT 13-4

Using Data Set VIII (McDonald's Data) that is on the Web site of this book, take a random sample of 10 food items. Record the grams of carbohydrates, grams of fat, and the number of calories for each food.

- Calculate the correlation coefficient between the grams of carbohydrates and the number of calories. Do the same for the grams of fat and the number of calories. In each case, do you find a positive association or a negative association?
- Create a scatterplot of these data using the carbohydrate content as the  $x$  variable and number of calories as the  $y$  variable. Does your scatterplot suggest that fitting a regression line to these data makes sense? Repeat the process using the fat content as the  $x$  variable and number of calories as the  $y$  variable.
- Find the equation of the estimated regression line for each of the two comparisons mentioned in parts a and b. If it makes sense to estimate regression lines for these variables, interpret the values of the slope and intercept for each regression line. If it does not make sense, explain why the numbers could be misleading.

## DECIDE FOR YOURSELF

### DOES A REGRESSION EQUATION ALWAYS MAKE SENSE?

Regression is a very powerful statistical tool. However, like any other tool, a failure to understand both its uses as well as its limitations can lead to ridiculous, if not disastrous, results. To demonstrate this, we took the data on two variables—the year of the Olympics from 1928 to 2004 as the independent variable and winner's time (in seconds) in the men's 100 meter dash (race) as a dependent variable. Figure 13.26 shows the scatterplot and the regression line for these data.

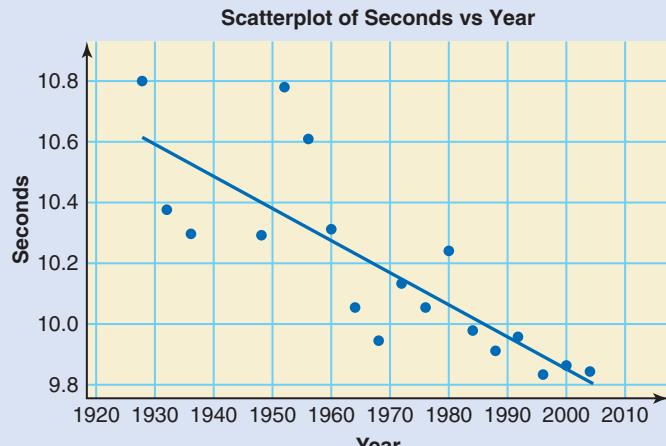


Figure 13.26 Scatterplot and regression line.

Looking at this scatterplot, it seems reasonable to use a regression line to explain the relationship between the year of Olympics and the winning time in the 100 meter dash. Specifically, the equation of that regression line is

$$\text{Seconds} = 31.1 - .0106 \text{ Year}$$

To calculate this regression line, we used the actual years of the Olympiad for the independent variable. Theoretically, we could use this regression equation to estimate the winning times for the years when Olympics are not held. We could also use it to predict the future winning times or to calculate what would have happened in the past. Answer the following questions to see how reasonable this process is.

- Based on this regression equation, what is the change in the winning time per Olympic period (4 years)? Does the change represent an increase or a decrease?
- Find the predicted winning times for the years 2200, 2600, and 3000. Using these predicted times, determine the winners' speeds (in miles per hour) for the years 2200 and 2600. Does it make sense to believe that this pattern will continue in the future? Explain.
- A similar analysis could be done in the reverse direction. A 2005 scientific discovery stated that fossils from 35,000-year-old *modern* humans were found in Transylvania (<http://www.theglobeandmail.com/servlet/story/RTGAM.20040306.wfossil0306/BNStory/specialScienceandHealth/>). Using the above regression equation, calculate the winning time for the 100 meter dash at this point in history. Does this number make sense? Why or why not?

# TECHNOLOGY INSTRUCTION

## Simple Linear Regression

**TI-84**

```
LinReg(a+bx)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:Y1
Calculate
```

Screen 13.1

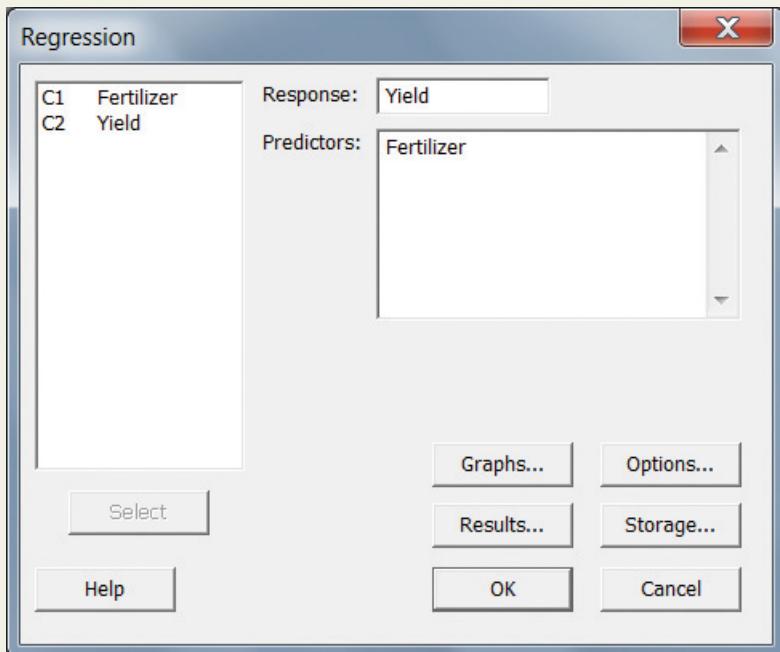
```
y=a+bx
a=37.2195602
b=.9027419574
r²=.9632095921
r=.9814324185
```

Screen 13.2

- To construct a simple linear regression equation, enter the independent and dependent variable values into lists. Select **STAT >CALC >LinReg(a+bx)** and then press **ENTER**. Enter the list that contains the independent variable at the **Xlist** prompt and then enter the list that contains the dependent variable at the **Ylist** prompt. Leave the **FreqList** prompt blank unless you have a separate list that gives the frequencies of each data point. At the **Store RegEQ** prompt, enter the function name where the regression equation will be stored, such as **Y1**. (**Y1** can be found by selecting **VARS >Y-VARS > Function >Y1**.) Then select **Calculate**. (See Screen 13.1 and Screen 13.2.) The results include the slope and intercept of the regression equation.
- To find the correlation coefficient, select **VARS >Statistics >EQ >r**. To find the coefficient of determination, square the correlation coefficient.
- To find a fitted value for a given value of  $x$ , type **Y1(x)**.
- To test that the slope of the line is nonzero, select **STAT >TESTS >LinRegTTest**. (Note that this set of commands will give you the output obtained under 1 and 2 above.) Enter the names of the lists. Leave **Freq:1**. Choose the alternative hypothesis. Leave **RegEQ:Y1**. Select **Calculate**. The results include a  $t$ -statistic value and a  $p$ -value.

**Note:** To see  $r$  and  $r^2$  using the **LinReg** command, select **CATALOG (2ND > 0)**, scroll to **DiagnosticOn**, and press **ENTER** twice. You will need to do this only one time.

## Minitab



Screen 13.3

- To construct and analyze a simple linear regression equation, enter the independent and dependent variable values into columns.
- Select **Stat >Regression >Regression**.
- Enter the dependent variable's column name in the **Response** box.
- Enter the independent variable's column name in the **Predictors** box. (See Screen 13.3.)
- Select **Options** if you wish to predict a value with the equation, and enter the value of the independent variable in the entry marked **Prediction intervals for new observations**. Enter the **Confidence level** and select **OK**.
- Select **Results**, and choose **Regression equation, . . .**. Select **OK** for each dialog box.
- The output includes the regression equation,  $t$  statistics and  $p$ -values for tests on both the slope and intercept to find out if they are zero, the coefficient of determination (as **R-sq**), and, if requested, the fitted value, as well as confidence and prediction intervals for the fitted value.

**Excel**

1. Click the **Data** tab. Click the **Data Analysis** button within the **Analysis** group.
2. To calculate the linear correlation coefficient, select **Correlation**. Enter the location of the data in the **Input Range** box. Click the button to identify whether the data for each sample are given in columns or rows. If your data have labels in the top row (or in the left column), click the **Labels** box. Choose how you wish the output to appear. (See **Screen 13.4**.) Click **OK**.

Fertilizer	Yield
120	142
80	112
100	132
70	96
88	119
75	104
110	136

Screen 13.4

	A	B	C
1		<i>Fertilizer</i>	<i>Yield</i>
2	Fertilizer		1
3	Yield	0.981432	1

Screen 13.5

4. To calculate the coefficients of the least squares regression line, perform a hypothesis test on the slope of the regression line, and calculate a confidence interval for the slope of the regression line, select **Regression** from the list of choices within the **Data Analysis** dialog box. Enter the location of the data in the **Input Y Range** box. Click the button to identify whether the data for each sample are given in columns or rows. If your data have labels in the top row (or in the left column), click the **Labels** box. Enter the confidence level if you want something other than 95% confidence interval. Choose how you wish the output to appear. (See **Screen 13.6**.) Click **OK**.

Fertilizer	Yield
120	142
80	112
100	132
70	96
88	119
75	104
110	136

Screen 13.6

5. The output contains three tables. The first table, labeled **Regression Statistics**, contains the standard deviation of errors in the line labeled **Standard Error**. In the bottom table, the **Coefficients** column contains the values of  $a$  and  $b$ . The remaining values in the top row of this table are not used with respect to this book. The remaining descriptions correspond to the values in the bottom row of the bottom table. The value in the **Standard Error** column is the value of  $s_b$ . The next two columns contain the value of the test statistic and the two-sided  $p$ -value for a test with the slope coefficient equal to zero. The next two columns give the endpoints of the 95% confidence interval for  $B$ . If you requested a confidence level other than 95%, the endpoints will be in the last two columns. (See **Screen 13.7**.)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3		Regression Statistics							
4	Multiple R	0.9814324							
5	R Square	0.9632096							
6	Adjusted R Square	0.9558515							
7	Standard Error	3.6199022							
8	Observations	7							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	1715.338682	1715.339	130.905	8.93323E-05			
13	Residual	5	65.5184607	13.10369					
14	Total	6	1780.857143						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	37.21956	7.375682241	5.046253	0.003946	18.25976541	56.179355	18.2597654	56.17935499
18	Fertilizer	0.902742	0.078901549	11.44137	8.93E-05	0.699919069	1.1055648	0.69991907	1.105564845

Screen 13.7

## TECHNOLOGY ASSIGNMENTS

**TA13.1** In a rainy coastal town in the Pacific Northwest, the local TV weatherman is often criticized for making inaccurate forecasts for daily precipitation. On each of 30 randomly selected days last winter, his precipitation forecast ( $x$ ) for the next day was recorded along with the actual precipitation ( $y$ ) for that day. These data are shown in the following table (in inches of rain).

x	y	x	y	x	y
1.0	.6	0	0	.4	.2
0	.1	0	.1	.2	.5
.2	0	.1	.2	.1	.1
0	0	.2	.2	0	.2
.5	.3	.1	0	.1	0
1.0	1.4	2.0	2.1	.2	.1
.5	.3	.4	.2	1.4	1.2
.1	.1	.2	.1	.5	1.0
0	.1	0	0	0	.5
2.0	.3	.3	.2	0	0

- a. Construct a scatter diagram for these data.
- b. Find the correlation coefficient between the two variables.
- c. Find the regression line with actual precipitation as a dependent variable and predicted precipitation as an independent variable.
- d. Make a 95% confidence interval for  $B$ .
- e. Test at a 1% significance level whether  $B$  is positive.
- f. Using a 1% significance level, can you conclude that the linear correlation coefficient is positive?

**TA13.2** Refer to Data Set III on NFL players. Select a random sample of 30 players from that population. Do the following for the data on heights and weights of these 30 players.

- a. Construct a scatter diagram for these data.
- b. Find the correlation between these two variables.
- c. Find the regression line with weight as a dependent variable and height as an independent variable.
- d. Make a 98% confidence interval for  $B$ .
- e. Test at a 2.5% significance level whether  $B$  is positive.
- f. Make a 95% confidence interval for the mean weight of all NFL players who are 75 inches tall. Construct a 95% prediction interval for the weight of a randomly selected NFL player with a height of 75 inches.

**TA13.3** Refer to the data on the ages and the numbers of breakdowns for a sample of seven machines given in Exercise 13.95. Answer the following questions.

- a. Construct a scatter diagram for these data.
- b. Find the least squares regression line with age as an independent variable and the number of breakdowns as a dependent variable.
- c. Compute the correlation coefficient.
- d. Construct a 99% confidence interval for  $B$ .
- e. Test at a 2.5% significance level whether  $B$  is positive.

**TA13.4** Refer to the Data Set IV on the 2011 Beach to Beacon 10K Road Race that is on the Web site of the text. Take a random sample of 45 participants. Do the following for the ages and times (in seconds) of these 45 runners.

- a. Construct a scatter diagram for these data, using age as the independent variable. Discuss whether it is appropriate to fit a linear regression model to these data.
- b. Find the correlation coefficient for these two variables.
- c. Find the equation of the predictive regression line with age as the independent variable and time as the dependent variable.
- d. Make a 98% confidence interval for  $B$ . Explain what this interval means with regard to a person's time for each additional year of age.
- e. Test at a 1% significance level whether  $B$  is positive.
- f. Test at a 1% significance level whether  $B$  is greater than 30.

**TA13.5** Refer to Data Set IX on contributions and spending by candidates in the 2009–2010 elections for the U.S. Senate and House of Representatives. Take a random sample of 50 candidates for the House of Representatives and 20 for the Senate. Do the following for the net contributions and the net operating expenditures for each of the two groups of candidates.

- a. Construct a scatter diagram for these data, using net contributions as the independent variable. Discuss whether it is appropriate to fit a linear regression model to these data.
- b. Find the correlation coefficient for these two variables.
- c. Find the equation of the predictive regression line with net contributions as the independent variable and net operating expenditures as the dependent variable.
- d. Make a 99% confidence interval for  $B$ . Explain what this interval means with regard to the relationship between net operating expenditures and net contributions.
- e. Test at a 1% significance level whether  $B$  is positive.
- f. Test at a 2.5% significance level whether  $B$  is different from 1.
- g. Discuss whether it appears that the slope for the House candidates is different from the slope for the Senate candidates.

# 14



Jose Luis Pelaez Inc/Blend Images/Getty Images, Inc.

## Multiple Regression

- 14.1 Multiple Regression Analysis
- 14.2 Assumptions of the Multiple Regression Model
- 14.3 Standard Deviation of Errors
- 14.4 Coefficient of Multiple Determination
- 14.5 Computer Solution of Multiple Regression

In Chapter 13, we discussed simple linear regression and linear correlation. A simple regression model includes one independent and one dependent variable, and it presents a very simplified scenario of real-world situations. In the real world, a dependent variable is usually influenced by a number of independent variables. For example, the sales of a company's product may be determined by the price of that product, the quality of the product, and advertising expenditure incurred by the company to promote that product. Therefore, it makes more sense to use a regression model that includes more than one independent variable. Such a model is called a **multiple regression model**. In this chapter we will discuss multiple regression models.

### 14.1 Multiple Regression Analysis

The simple linear regression model discussed in Chapter 13 was written as

$$y = A + Bx + \epsilon$$

This model includes one independent variable, which is denoted by  $x$ , and one dependent variable, which is denoted by  $y$ . As we know from Chapter 13, the term represented by  $\epsilon$  in the above model is called the random error.

Usually a dependent variable is affected by more than one independent variable. When we include two or more independent variables in a regression model, it is called a **multiple regression model**. Remember, whether it is a simple or a multiple regression model, it always includes one and only one dependent variable.

A multiple regression model with  $y$  as a dependent variable and  $x_1, x_2, x_3, \dots, x_k$  as independent variables is written as

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_kx_k + \epsilon \quad (1)$$

where  $A$  represents the constant term,  $B_1, B_2, B_3, \dots, B_k$  are the regression coefficients of independent variables  $x_1, x_2, x_3, \dots, x_k$ , respectively, and  $\epsilon$  represents the random error term. This model contains  $k$  independent variables  $x_1, x_2, x_3, \dots, x_k$ . From model (1), it would seem that multiple regression models can only be used when the relationship between the dependent variable and each independent variable is linear. Furthermore, it also appears as if there can be no interaction between two or more of the independent variables. This is far from the truth. In the real world, a multiple regression model can be much more complex. Discussion of such models is outside the scope of this book. When each term contains a single independent variable raised to the first power as in model (1), we call it a **first-order multiple regression model**. This is the only type of multiple regression model we will discuss in this chapter.

In regression model (1),  $A$  represents the constant term, which gives the value of  $y$  when all independent variables assume zero values. The coefficients  $B_1, B_2, B_3, \dots, B_k$  are called the **partial regression coefficients**. For example,  $B_1$  is a partial regression coefficient of  $x_1$ . It gives the change in  $y$  due to a one-unit change in  $x_1$  when all other independent variables included in the model are held constant. In other words, if we change  $x_1$  by one unit but keep  $x_2, x_3, \dots, x_k$  unchanged, then the resulting change in  $y$  is measured by  $B_1$ . Similarly the value of  $B_2$  gives the change in  $y$  due to a one-unit change in  $x_2$  when all other independent variables are held constant. In model (1) above,  $A, B_1, B_2, B_3, \dots, B_k$  are called the *true regression coefficients or population parameters*.

A positive value for a particular  $B_i$  in model (1) will indicate a positive relationship between  $y$  and the corresponding  $x_i$  variable. A negative value for a particular  $B_i$  in that model will indicate a negative relationship between  $y$  and the corresponding  $x_i$  variable.

Remember that in a first-order regression model such as model (1), the relationship between each  $x_i$  and  $y$  is a straight-line relationship. In model (1),  $A + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_kx_k$  is called the *deterministic portion* and  $\epsilon$  is the *stochastic portion* of the model.

When we use the  $t$  distribution to make inferences about a single parameter of a multiple regression model, the **degrees of freedom** are calculated as

$$df = n - k - 1$$

where  $n$  represents the sample size and  $k$  is the number of independent variables in the model.

### Definition

**Multiple Regression Model** A regression model that includes two or more independent variables is called a multiple regression model. It is written as

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_kx_k + \epsilon$$

where  $y$  is the dependent variable,  $x_1, x_2, x_3, \dots, x_k$  are the  $k$  independent variables, and  $\epsilon$  is the random error term. When each of the  $x_i$  variables represents a single variable raised to the first power as in the above model, this model is referred to as a **first-order multiple regression model**. For such a model with a sample size of  $n$  and  $k$  independent variables, the degrees of freedom are:

$$df = n - k - 1$$

When a multiple regression model includes only two independent variables (with  $k = 2$ ), model (1) reduces to

$$y = A + B_1x_1 + B_2x_2 + \epsilon$$

A multiple regression model with three independent variables (with  $k = 3$ ) is written as

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \epsilon$$

If model (1) is estimated using sample data, which is usually the case, the estimated regression equation is written as

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k \quad (2)$$

In equation (2),  $a, b_1, b_2, b_3, \dots$ , and  $b_k$  are the sample statistics, which are the point estimators of the population parameters  $A, B_1, B_2, B_3, \dots$ , and  $B_k$ , respectively.

In model (1),  $y$  denotes the actual values of the dependent variable for members of the sample. In the estimated model (2),  $\hat{y}$  denotes the predicted or estimated values of the dependent variable. The difference between any pair of  $y$  and  $\hat{y}$  values gives the error of prediction. For a multiple regression model,

$$SSE = \sum (y - \hat{y})^2$$

where SSE stands for the error sum of squares.

As in Chapter 13, the estimated regression equation (2) is obtained by minimizing the sum of squared errors, that is,

$$\text{Minimize } \sum (y - \hat{y})^2$$

The estimated equation (2) obtained by minimizing the sum of squared errors is called the **least squares regression equation**.

Usually the calculations in a multiple regression analysis are made by using statistical software packages for computers, such as MINITAB, instead of using the formulas manually. Even for a multiple regression equation with two independent variables, the formulas are complex and manual calculations are time consuming. In this chapter we will perform the multiple regression analysis using MINITAB. The solutions obtained by using other statistical software packages such as JMP, SAS, S-Plus, or SPSS can be interpreted the same way. The TI-84 and Excel do not have built-in procedures for the multiple regression model.

## 14.2 Assumptions of the Multiple Regression Model

Like a simple linear regression model, a multiple (linear) regression model is based on certain assumptions. The following are the major assumptions for the multiple regression model (1).

**Assumption 1:** The mean of the probability distribution of  $\epsilon$  is zero, that is,

$$E(\epsilon) = 0$$

If we calculate errors for all measurements for a given set of values of independent variables for a population data set, the mean of these errors will be zero. In other words, while individual predictions will have some amount of errors, on average our predictions will be correct. Under this assumption, the mean value of  $y$  is given by the deterministic part of regression model (1). Thus,

$$E(y) = A + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_kx_k$$

where  $E(y)$  is the expected or mean value of  $y$  for the population. This mean value of  $y$  is also denoted by  $\mu_{y|x_1, x_2, \dots, x_k}$ .

**Assumption 2:** The errors associated with different sets of values of independent variables are independent. Furthermore, these errors are normally distributed and have a constant standard deviation, which is denoted by  $\sigma_\epsilon$ .

**Assumption 3:** The independent variables are not linearly related. However, they can have a nonlinear relationship. When independent variables are highly linearly correlated, it is referred to as **multicollinearity**. This assumption is about the nonexistence of the multicollinearity problem. For example, consider the following multiple regression model:

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \epsilon$$

All of the following linear relationships (and other such linear relationships) between  $x_1$ ,  $x_2$ , and  $x_3$  should be invalid for this model.

$$\begin{aligned}x_1 &= x_2 + 4x_3 \\x_2 &= 5x_1 - 2x_3 \\x_1 &= 3.5x_2\end{aligned}$$

If any linear relationship exists, we can substitute one variable for another, which will reduce the number of independent variables to two. However, nonlinear relationships, such as  $x_1 = 4x_2^2$  and  $x_2 = 2x_1 + 6x_3^2$  between  $x_1$ ,  $x_2$ , and  $x_3$  are permissible.

In practice, multicollinearity is a major issue. Examining the correlation for each pair of independent variables is a good way to determine if multicollinearity exists.

**Assumption 4:** There is no linear association between the random error term  $\epsilon$  and each independent variable  $x_i$ .

## 14.3 Standard Deviation of Errors

The **standard deviation of errors** (also called the standard error of the estimate) for the multiple regression model (1) is denoted by  $\sigma_e$ , and it is a measure of variation among errors. However, when sample data are used to estimate multiple regression model (1), the standard deviation of errors is denoted by  $s_e$ . The formula to calculate  $s_e$  is as follows.

$$s_e = \sqrt{\frac{\text{SSE}}{n - k - 1}} \quad \text{where} \quad \text{SSE} = \sum (y - \hat{y})^2$$

Note that here SSE is the error sum of squares. We will not use this formula to calculate  $s_e$  manually. Rather we will obtain it from the computer solution. Note that many software packages label  $s_e$  as Root MSE, where MSE stands for mean square error.

## 14.4 Coefficient of Multiple Determination

In Chapter 13, we denoted the coefficient of determination for a simple linear regression model by  $r^2$  and defined it as the proportion of the total sum of squares SST that is explained by the regression model. The coefficient of determination for the multiple regression model, usually called the **coefficient of multiple determination**, is denoted by  $R^2$  and is defined as the proportion of the total sum of squares SST that is explained by the multiple regression model. It tells us how good the multiple regression model is and how well the independent variables included in the model explain the dependent variable.

Like  $r^2$ , the value of the coefficient of multiple determination  $R^2$  always lies in the range 0 to 1, that is,

$$0 \leq R^2 \leq 1$$

Just as in the case of the simple linear regression model, SST is the total sum of squares, SSR is the regression sum of squares, and SSE is the error sum of squares. SST is always equal to the sum of SSE and SSR. They are calculated as follows.

$$\begin{aligned}\text{SSE} &= \sum e^2 = \sum (y - \hat{y})^2 \\ \text{SST} &= SS_{yy} = \sum (y - \bar{y})^2 \\ \text{SSR} &= \sum (\hat{y} - \bar{y})^2\end{aligned}$$

SSR is the portion of SST that is explained by the use of the regression model, and SSE is the portion of SST that is not explained by the use of the regression model. The coefficient of multiple determination is given by the ratio of SSR and SST as follows.

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

The coefficient of multiple determination  $R^2$  has one major shortcoming. The value of  $R^2$  generally increases as we add more and more explanatory variables to the regression model (even if they do not belong in the model). Just because we can increase the value of  $R^2$  does not imply that the regression equation with a higher value of  $R^2$  does a better job of predicting the dependent variable. Such a value of  $R^2$  will be misleading, and it will not represent the true explanatory power of the regression model. To eliminate this shortcoming of  $R^2$ , it is preferable to use the **adjusted coefficient of multiple determination**, which is denoted by  $\bar{R}^2$ . Note that  $\bar{R}^2$  is the coefficient of multiple determination adjusted for degrees of freedom. The value of  $\bar{R}^2$  may increase, decrease, or stay the same as we add more explanatory variables to our regression model. If a new variable added to the regression model contributes significantly to explain the variation in  $y$ , then  $\bar{R}^2$  increases; otherwise it decreases. The value of  $\bar{R}^2$  is calculated as follows.

$$\bar{R}^2 = 1 - (1 - R^2) \left( \frac{n - 1}{n - k - 1} \right) \quad \text{or} \quad 1 - \frac{\text{SSE}/(n - k - 1)}{\text{SST}/(n - 1)}$$

Thus, if we know  $R^2$ , we can find the value of  $\bar{R}^2$ . Almost all statistical software packages give the values of both  $R^2$  and  $\bar{R}^2$  for a regression model.

Another property of  $\bar{R}^2$  to remember is that whereas  $R^2$  can never be negative,  $\bar{R}^2$  can be negative.

While a general rule of thumb is that a higher value of  $\bar{R}^2$  implies that a specific set of independent variables does a better job of predicting a specific dependent variable, it is important to recognize that some dependent variables have a great deal more variability than others. Therefore,  $\bar{R}^2 = .30$  could imply that a specific model is not a very strong model, but it could be the best possible model in a certain scenario. Many *good* financial models have values of  $\bar{R}^2$  below .50.

## 14.5 Computer Solution of Multiple Regression

In this section, we take an example of a multiple regression model, solve it using MINITAB, interpret the solution, and make inferences about the population parameters of the regression model.

### ■ EXAMPLE 14-1

*Using MINITAB to find a multiple regression equation.*

A researcher wanted to find the effect of driving experience and the number of driving violations on auto insurance premiums. A random sample of 12 drivers insured with the same company and having similar auto insurance policies was selected from a large city. Table 14.1 lists

Table 14.1

Monthly Premium (dollars)	Driving Experience (years)	Number of Driving Violations (past 3 years)
148	5	2
76	14	0
100	6	1
126	10	3
194	4	6
110	8	2
114	11	3
86	16	1
198	3	5
92	9	1
70	19	0
120	13	3

the monthly auto insurance premiums (in dollars) paid by these drivers, their driving experiences (in years), and the numbers of driving violations committed by them during the past three years.

Using MINITAB, find the regression equation of monthly premiums paid by drivers on the driving experiences and the numbers of driving violations.

**Solution** Let

$y$  = the monthly auto insurance premium (in dollars) paid by a driver

$x_1$  = the driving experience (in years) of a driver

$x_2$  = the number of driving violations committed by a driver during the past three years

We are to estimate the regression model

$$y = A + B_1x_1 + B_2x_2 + \epsilon \quad (3)$$

The first step is to enter the data of Table 14.1 into MINITAB spreadsheet as shown in Screen 14.1. Here we have entered the given data in columns C1, C2, and C3 and named them Monthly Premium, Driving Experience and Driving Violations, respectively.

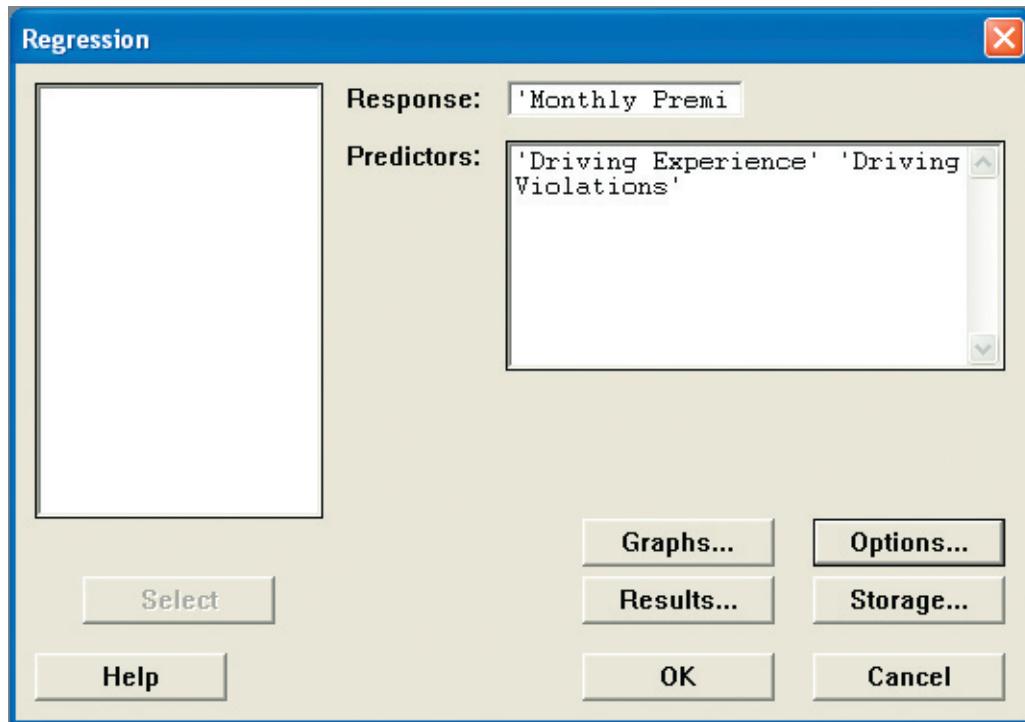
	C1	C2	C3
1	148	5	2
2	76	14	0
3	100	6	1
4	126	10	3
5	194	4	6
6	110	8	2
7	114	11	3
8	86	16	1
9	198	3	5
10	92	9	1
11	70	19	0
12	120	13	3

Screen 14.1

To obtain the estimated regression equation, select **Stat>Regression>Regression**. In the dialog box you obtain, enter *Monthly Premium* in the **Response** box, and *Driving Experience* and *Driving Violations* in the **Predictors** box as shown in Screen 14.2. Note that you can enter the column names C1, C2, and C3 instead of variable names in these boxes. Click **OK** to obtain the output, which is shown in Screen 14.3.

From the output given in Screen 14.3, the estimated regression equation is:

$$\hat{y} = 110 - 2.75x_1 + 16.1x_2$$



Screen 14.2

Session

### Regression Analysis: Monthly Prem versus Driving Experience, Driving Violations

The regression equation is  
 $\text{Monthly Premium} = 110 - 2.75 \text{ Driving Experience} + 16.1 \text{ Driving Violations}$

Predictor              Coef    SE Coef      T      P Constant              110.28    14.62      7.54    0.000 Driving Experience   -2.7473    0.9770     -2.81    0.020 Driving Violations   16.106    2.613      6.16    0.000	}	I
--	---	---

$S = 12.1459 \quad R-Sq = 93.1\% \quad R-Sq(adj) = 91.6\%$

Analysis of Variance Source              DF      SS      MS      F      P Regression        2    17961.3    8980.6    60.88    0.000 Residual Error   9    1327.7    147.5 Total              11    19289.0	}	III
---	---	-----

Source              DF      Seq SS Driving Experience   1    12357.8 Driving Violations   1    5603.5	}	IV
---	---	----

Source              DF      Seq SS Driving Experience   1    12357.8 Driving Violations   1    5603.5	}	V
---	---	---

Screen 14.3

### 14.5.1 Estimated Multiple Regression Model

Example 14–2 describes, among other things, how the coefficients of the multiple regression model are interpreted.

#### ■ EXAMPLE 14–2

Refer to Example 14–1 and the MINITAB solution given in Screen 14.3.

- Explain the meaning of the estimated regression coefficients.
- What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
- What is the predicted auto insurance premium paid per month by a driver with seven years of driving experience and three driving violations committed in the past three years?
- What is the point estimate of the expected (or mean) auto insurance premium paid per month by all drivers with 12 years of driving experience and 4 driving violations committed in the past three years?

*Interpreting parts of the  
MINITAB solution of multiple  
regression.*

#### Solution

- From the portion of the MINITAB solution that is marked **I** in Screen 14.3, the estimated regression equation is

$$\hat{y} = 110 - 2.75x_1 + 16.1x_2 \quad (4)$$

From this equation,

$$a = 110, \quad b_1 = -2.75, \quad \text{and} \quad b_2 = 16.1$$

We can also read the values of these coefficients from the column labeled **Coef** in the portion of the output marked **II** in the MINITAB solution of Screen 14.3. From this column we obtain

$$a = 110.28, \quad b_1 = -2.7473, \quad \text{and} \quad b_2 = 16.106$$

Notice that in this column the coefficients of the regression equation appear with more digits after the decimal point. With these coefficient values, we can write the estimated regression equation as

$$\hat{y} = 110.28 - 2.7473x_1 + 16.106x_2 \quad (5)$$

The value of  $a = 110.28$  in the estimated regression equation (5) gives the value of  $\hat{y}$  for  $x_1 = 0$  and  $x_2 = 0$ . Thus, a driver with no driving experience and no driving violations committed in the past three years is expected to pay an auto insurance premium of \$110.28 per month. Again, this is the technical interpretation of  $a$ . In reality, that may not be true because none of the drivers in our sample has both zero experience and zero driving violations. As all of us know, some of the highest premiums are paid by teenagers just after obtaining their drivers licenses.

The value of  $b_1 = -2.7473$  in the estimated regression model gives the change in  $\hat{y}$  for a one-unit change in  $x_1$  when  $x_2$  is held constant. Thus, we can state that a driver with one extra year of experience but the same number of driving violations is expected to pay \$2.7473 (or \$2.75) less per month for the auto insurance premium. Note that because  $b_1$  is negative, an increase in driving experience decreases the premium paid. In other words,  $y$  and  $x_1$  have a negative relationship.

The value of  $b_2 = 16.106$  in the estimated regression model gives the change in  $\hat{y}$  for a one-unit change in  $x_2$  when  $x_1$  is held constant. Thus, a driver with one extra driving violation during the past three years but with the same years of driving experience is expected to pay \$16.106 (or \$16.11) more per month for the auto insurance premium.

- (b) The values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination are given in part III of the MINITAB solution of Screen 14.3. From this part of the solution,

$$s_e = 12.1459, \quad R^2 = 93.1\%, \quad \text{and} \quad \bar{R}^2 = 91.6\%$$

Thus, the standard deviation of errors is 12.1459. The value of  $R^2 = 93.1\%$  tells us that the two independent variables, years of driving experiences and the numbers of driving violations, explain 93.1% of the variation in the auto insurance premiums. The value of  $\bar{R}^2 = 91.6\%$  is the value of the coefficient of multiple determination adjusted for degrees of freedom. It states that when adjusted for degrees of freedom, the two independent variables explain 91.6% of the variation in the dependent variable.

- (c) To Find the predicted auto insurance premium paid per month by a driver with seven years of driving experience and three driving violations during the past three years, we substitute  $x_1 = 7$  and  $x_2 = 3$  in the estimated regression model (5). Thus,

$$\begin{aligned}\hat{y} &= 110.28 - 2.7473x_1 + 16.106x_2 \\ &= 110.28 - 2.7473(7) + 16.106(3) = \$139.37\end{aligned}$$

Note that this value of  $\hat{y}$  is a point estimate of the predicted value of  $y$ , which is denoted by  $y_p$ . The concept of the predicted value of  $y$  is the same as that for a simple linear regression model discussed in Section 13.8.2 of Chapter 13.

- (d) To obtain the point estimate of the expected (mean) auto insurance premium paid per month by all drivers with 12 years of driving experience and four driving violations during the past three years, we substitute  $x_1 = 12$  and  $x_2 = 4$  in the estimated regression equation (5). Thus,

$$\begin{aligned}\hat{y} &= 110.28 - 2.7473x_1 + 16.106x_2 \\ &= 110.28 - 2.7473(12) + 16.106(4) = \$141.74\end{aligned}$$

This value of  $\hat{y}$  is a point estimate of the mean value of  $y$ , which is denoted by  $E(y)$  or  $\mu_{y|x_1, x_2}$ . The concept of the mean value of  $y$  is the same as that for a simple linear regression model discussed in Section 13.8.1 of Chapter 13. ■

### 14.5.2 Confidence Interval for an Individual Coefficient

The values of  $a$ ,  $b_1$ ,  $b_2$ ,  $b_3$ , ..., and  $b_k$  obtained by estimating model (1) using sample data give the point estimates of  $A$ ,  $B_1$ ,  $B_2$ ,  $B_3$ , ..., and  $B_k$ , respectively, which are the population parameters. Using the values of sample statistics  $a$ ,  $b_1$ ,  $b_2$ ,  $b_3$ , ..., and  $b_k$ , we can make confidence intervals for the corresponding population parameters  $A$ ,  $B_1$ ,  $B_2$ ,  $B_3$ , ..., and  $B_k$ , respectively.

Because of the assumption that the errors are normally distributed, the sampling distribution of each  $b_i$  is normal with its mean equal to  $B_i$  and standard deviation equal to  $\sigma_{b_i}$ . For example, the sampling distribution of  $b_1$  is normal with its mean equal to  $B_1$  and standard deviation equal to  $\sigma_{b_1}$ . However, usually  $\sigma_e$  is not known and, hence, we cannot find  $\sigma_{b_i}$ . Consequently, we use  $s_{b_i}$  as an estimator of  $\sigma_{b_i}$  and use the  $t$  distribution to determine a confidence interval for  $B_i$ .

The formula to obtain a confidence interval for a population parameter  $B_i$  is given below. This is the same formula we used to make a confidence interval for  $B$  in Section 13.5.2 of Chapter 13. The only difference is that to make a confidence interval for a particular  $B_i$  for a multiple regression model, the degrees of freedom are  $n - k - 1$ .

**Confidence Interval for  $B_i$**  The  $(1 - \alpha) \times 100\%$  confidence interval for  $B_i$  is given by

$$b_i \pm ts_{b_i}$$

The value of  $t$  that is used in this formula is obtained from the  $t$  distribution table for  $\alpha/2$  area in the right tail of the  $t$  distribution curve and  $(n - k - 1)$  degrees of freedom. The values of  $b_i$  and  $s_{b_i}$  are obtained from the computer solution.

Example 14–3 describes the procedure to make a confidence interval for an individual regression coefficient  $B_i$ .

### ■ EXAMPLE 14–3

Determine a 95% confidence interval for  $B_1$  (the coefficient of experience) for the multiple regression of auto insurance premium on driving experience and the number of driving violations. Use the MINITAB solution of Screen 14.3.

*Making a confidence interval for an individual coefficient of a multiple regression model.*

**Solution** To make a confidence interval for  $B_1$ , we use the portion marked **II** in the MINITAB solution of Screen 14.3. From that portion of the MINITAB solution,

$$b_1 = -2.7473 \quad \text{and} \quad s_{b_1} = .9770$$

Note that the value of the standard deviation of  $b_1$ ,  $s_{b_1} = .9770$ , is given in the column labeled **SE Coef** in part **II** of the MINITAB solution.

The confidence level is 95%. The area in each tail of the  $t$  distribution curve is obtained as follows.

$$\text{Area in each tail of the } t \text{ distribution} = (1 - .95)/2 = .025$$

The sample size is 12, which gives  $n = 12$ . Because there are two independent variables,  $k = 2$ . Therefore,

$$\text{Degrees of freedom} = n - k - 1 = 12 - 2 - 1 = 9$$

From the  $t$  distribution table (Table V of Appendix C), the value of  $t$  for .025 area in the right tail of the  $t$  distribution curve and 9 degrees of freedom is 2.262. Then, the 95% confidence interval for  $B_1$  is

$$\begin{aligned} b_1 \pm ts_{b_1} &= -2.7473 \pm 2.262(.9770) \\ &= -2.7473 \pm 2.2100 = \mathbf{-4.9573 \text{ to } -.5373} \end{aligned}$$

Thus, the 95% confidence interval for  $b_1$  is  $-.496$  to  $-.54$ . That is, we can state with 95% confidence that for one extra year of driving experience, the monthly auto insurance premium changes by an amount between  $-\$4.96$  and  $-\$.54$ . Note that since both limits of the confidence interval have negative signs, we can also state that for each extra year of driving experience, the monthly auto insurance premium decreases by an amount between  $.\$54$  and  $.\$4.96$ . ■

By applying the procedure used in Example 14–3, we can make a confidence interval for any of the coefficients (including the constant term) of a multiple regression model, such as  $A$  and  $B_2$  in model (3). For example, the 95% confidence intervals for  $A$  and  $B_2$ , respectively, are

$$a \pm ts_a = 110.28 \pm 2.262(14.62) = 77.21 \text{ to } 143.35$$

$$b_2 \pm ts_{b_2} = 16.106 \pm 2.262(2.613) = 10.20 \text{ to } 22.02$$

#### 14.5.3 Testing a Hypothesis about an Individual Coefficient

We can perform a test of hypothesis about any of the  $B_i$  coefficients of the regression model (1) using the same procedure that we used to make a test of hypothesis about  $B$  for a simple regression model in Section 13.5.3 of Chapter 13. The only difference is that the degrees of freedom are equal to  $n - k - 1$  for a multiple regression model.

Again, because of the assumption that the errors are normally distributed, the sampling distribution of each  $b_i$  is normal with its mean equal to  $B_i$  and standard deviation equal to  $\sigma_{b_i}$ . However, usually  $\sigma_e$  is not known and, hence, we cannot find  $\sigma_{b_i}$ . Consequently, we use  $s_{b_i}$  as an estimator of  $\sigma_{b_i}$ , and use the  $t$  distribution to perform the test.

**Test Statistic for  $b_i$**  The value of the test statistic  $t$  for  $b_i$  is calculated as

$$t = \frac{b_i - B_i}{s_{b_i}}$$

The value of  $B_i$  is substituted from the null hypothesis. Usually, but not always, the null hypothesis is  $H_0: B_i = 0$ . The MINITAB solution contains this value of the  $t$  statistic.

Example 14–4 illustrates the procedure for testing a hypothesis about a single coefficient.

### ■ EXAMPLE 14–4

*Testing a hypothesis about a coefficient of a multiple regression model.*

Using the 2.5% significance level, can you conclude that the coefficient of the number of years of driving experience in regression model (3) is negative? Use the MINITAB output obtained in Example 14–1 and shown in Screen 14.3 to perform this test.

**Solution** From Example 14–1, our multiple regression model (3) is

$$y = A + B_1x_1 + B_2x_2 + \epsilon$$

where  $y$  is the monthly auto insurance premium (in dollars) paid by a driver,  $x_1$  is the driving experience (in years), and  $x_2$  is the number of driving violations committed during the past three years. From the MINITAB solution, the estimated regression equation is

$$\hat{y} = 110.28 - 2.7473x_1 + 16.106x_2$$

To conduct a test of hypothesis about  $B_1$ , we use the portion marked **II** in the MINITAB solution given in Screen 14.3. From that portion of the MINITAB solution,

$$b_1 = -2.7473 \quad \text{and} \quad s_{b_1} = .9770$$

Note that the value of the standard deviation of  $b_1$ ,  $s_{b_1} = .9770$ , is given in the column labeled **SE Coef** in part **II** of the MINITAB solution.

To make a test of hypothesis about  $B_1$ , we perform the following five steps.

**Step 1.** *State the null and alternative hypotheses.*

We are to test whether or not the coefficient of the number of years of driving experience in regression model (3) is negative, that is, whether or not  $B_1$  is negative. The two hypotheses are

$$H_0: B_1 = 0$$

$$H_1: B_1 < 0$$

Note that we can also write the null hypothesis as  $H_0: B_1 \geq 0$ , which states that the coefficient of the number of years of driving experience in the regression model (3) is either zero or positive.

**Step 2.** *Select the distribution to use.*

The sample size is small ( $n < 30$ ) and  $\sigma_\epsilon$  is not known. The sampling distribution of  $b_1$  is normal because the errors are assumed to be normally distributed. Hence, we use the  $t$  distribution to make a test of hypothesis about  $B_1$ .

**Step 3.** *Determine the rejection and nonrejection regions.*

The significance level is .025. The  $<$  sign in the alternative hypothesis indicates that the test is left-tailed. Therefore, area in the left tail of the  $t$  distribution curve is  $\alpha = .025$ . The degrees of freedom are:

$$df = n - k - 1 = 12 - 2 - 1 = 9$$

From the  $t$  distribution table (Table V in Appendix C), the critical value of  $t$  for 9 degrees of freedom and .025 area in the left tail of the  $t$  distribution curve is  $-2.262$ , as shown in Figure 14.1.

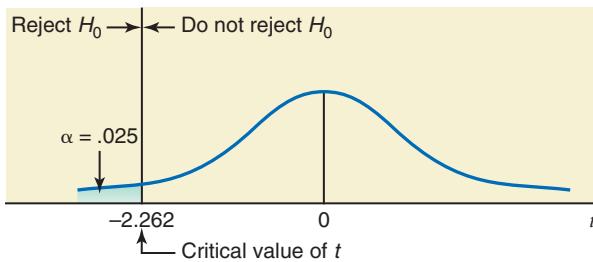


Figure 14.1

**Step 4.** Calculate the value of the test statistic and *p*-value.

The value of the test statistic  $t$  for  $b_1$  can be obtained from the MINITAB solution given in Screen 14.3. This value is given in the column labeled **T** and the row named Driving Experience in the portion marked **II** in that MINITAB solution. Thus, the observed value of  $t$  is

$$t = \frac{b_1 - B_1}{s_{b_1}} = -2.81$$

Also, in the same portion of the MINITAB solution, the *p*-value for this test is given in the column labeled **P** and the row named Driving Experience. This *p*-value is .020. However, MINITAB always gives the *p*-value for a two-tailed test. Because our test is one-tailed, the *p*-value for our test is

$$p\text{-value} = .020/2 = .010$$

**Step 5.** Make a decision.

The value of the test statistic,  $t = -2.81$ , is less than the critical value of  $t = -2.262$  and it falls in the rejection region. Consequently, we reject the null hypothesis and conclude that the coefficient of  $x_1$  in regression model (3) is negative. That is, an increase in the driving experience decreases the auto insurance premium.

Also the *p*-value for the test is .010, which is less than the significance level of  $\alpha = .025$ . Hence, based on this *p*-value also, we reject the null hypothesis and conclude that  $B_1$  is negative. ■

Note that the observed value of  $t$  in Step 4 of Example 14–4 is obtained from the MINITAB solution only if the null hypothesis is  $H_0 : B_1 = 0$ . However, if the null hypothesis is that  $B_1$  is equal to a number other than zero, then the  $t$  value obtained from the MINITAB solution is no longer valid. For example, suppose the null hypothesis in Example 14–4 is

$$H_0 : B_1 = -2$$

and the alternative hypothesis is

$$H_1 : B_1 < -2$$

In this case the observed value of  $t$  will be calculated as

$$t = \frac{b_1 - B_1}{s_{b_1}} = \frac{-2.7473 - (-2)}{.9770} = -.765$$

To calculate this value of  $t$ , the values of  $b_1$  and  $s_{b_1}$  are obtained from the MINITAB solution of Screen 14.3. The value of  $B_1$  is substituted from  $H_0$ .

## EXERCISES

### CONCEPTS AND PROCEDURES

- 14.1** How are the coefficients of independent variables in a multiple regression model interpreted? Explain.  
**14.2** What are the degrees of freedom for a multiple regression model to make inferences about individual parameters?

**14.3** What kinds of relationships among independent variables are permissible and which ones are not permissible in a linear multiple regression model?

**14.4** Explain the meaning of the coefficient of multiple determination and the adjusted coefficient of multiple determination for a multiple regression model. What is the difference between the two?

**14.5** What are the assumptions of a multiple regression model?

**14.6** The following table gives data on variables  $y$ ,  $x_1$ ,  $x_2$ , and  $x_3$ .

$y$	$x_1$	$x_2$	$x_3$
8	18	38	74
11	26	25	64
19	34	24	47
21	38	44	31
7	13	12	79
23	49	48	35
16	28	38	42
27	59	52	18
9	14	17	71
13	21	39	57

Using MINITAB, estimate the regression model

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \epsilon$$

Using the solution obtained, answer the following questions.

- Write the estimated regression equation.
- Explain the meaning of  $a$ ,  $b_1$ ,  $b_2$ , and  $b_3$  obtained by estimating the given regression model.
- What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
- What is the predicted value of  $y$  for  $x_1 = 35$ ,  $x_2 = 40$ , and  $x_3 = 65$ ?
- What is the point estimate of the expected (mean) value of  $y$  for all elements given that  $x_1 = 40$ ,  $x_2 = 30$ , and  $x_3 = 55$ ?
- Construct a 95% confidence interval for the coefficient of  $x_3$ .
- Using the 2.5% significance level, test whether or not the coefficient of  $x_1$  is positive.

**14.7** The following table gives data on variables  $y$ ,  $x_1$ , and  $x_2$ .

$y$	$x_1$	$x_2$
24	98	52
14	51	69
18	74	63
31	108	35
10	33	88
29	119	54
26	99	51
33	141	31
13	47	67
27	103	41
26	111	46

Using MINITAB, find the regression of  $y$  on  $x_1$  and  $x_2$ . Using the solution obtained, answer the following questions.

- Write the estimated regression equation.
- Explain the meaning of the estimated regression coefficients of the independent variables.

- c. What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
- d. What is the predicted value of  $y$  for  $x_1 = 87$  and  $x_2 = 54$ ?
- e. What is the point estimate of the expected (mean) value of  $y$  for all elements given that  $x_1 = 95$  and  $x_2 = 49$ ?
- f. Construct a 99% confidence interval for the coefficient of  $x_1$ .
- g. Using the 1% significance level, test if the coefficient of  $x_2$  in the population regression model is negative.

## ■ APPLICATIONS

**14.8** The salaries of workers are expected to be dependent, among other factors, on the number of years they have spent in school and their work experiences. The following table gives information on the annual salaries (in thousands of dollars) for 12 persons, the number of years each of them spent in school, and the total number of years of work experiences.

Salary	52	44	48	77	68	48	59	83	28	61	27	69
Schooling	16	12	13	20	18	16	14	18	12	16	12	16
Experience	6	10	15	8	11	2	12	4	6	9	2	18

Using MINITAB, find the regression of salary on schooling and experience. Using the solution obtained, answer the following questions.

- a. Write the estimated regression equation.
- b. Explain the meaning of the estimates of the constant term and the regression coefficients of independent variables.
- c. What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
- d. How much salary is a person with 18 years of schooling and 7 years of work experience expected to earn?
- e. What is the point estimate of the expected (mean) salary for all people with 16 years of schooling and 10 years of work experience?
- f. Determine a 99% confidence interval for the coefficient of schooling.
- g. Using the 1% significance level, test whether or not the coefficient of experience is positive.

**14.9** The CTO Corporation has a large number of chain restaurants throughout the United States. The research department at the company wanted to find if the sales of restaurants depend on the size of the population within a certain area surrounding the restaurants and the mean income of households in those areas. The company collected information on these variables for 11 restaurants. The following table gives information on the weekly sales (in thousands of dollars) of these restaurants, the population (in thousands) within five miles of the restaurants, and the mean annual income (in thousands of dollars) of the households for those areas.

Sales	19	29	17	21	14	30	33	22	18	27	24
Population	21	15	32	18	47	69	29	43	75	39	53
Income	58	69	49	52	67	76	81	46	39	64	28

Using MINITAB, find the regression of sales on population and income. Using the solution obtained, answer the following questions.

- a. Write the estimated regression equation.
- b. Explain the meaning of the estimates of the constant term and the regression coefficients of population and income.
- c. What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
- d. What are the predicted sales for a restaurant with 50 thousand people living within a five-mile area surrounding it and \$55 thousand mean annual income of households in that area.
- e. What is the point estimate of the expected (mean) sales for all restaurants with 45 thousand people living within a five-mile area surrounding them and \$46 thousand mean annual income of households living in those areas?
- f. Determine a 95% confidence interval for the coefficient of *income*.
- g. Using the 1% significance level, test whether or not the coefficient of *population* is different from zero.

## USES AND MISUSES... ADDITIVE VERSUS MULTIPLICATIVE EFFECT

A first-order multiple regression model with (quantitative) independent variables is one of the simpler types of multiple regression models. However, there are many limitations of this model. A major limitation is that the independent variables have an *additive* effect on the dependent variable. What does *additive* mean here? Suppose we have the following estimated regression equation:

$$\hat{y} = 4 + 6x_1 + 3x_2$$

From this estimated regression equation, if  $x_1$  increases by 1 unit (with  $x_2$  held constant), our predicted value of  $y$  increases by 6 units. If  $x_2$  increases by 1 unit (with  $x_1$  held constant), our predicted value of  $y$  increases by 3 units. But what happens if  $x_1$  and  $x_2$  both increase by 1 unit each? From this equation, our predicted value of  $y$  will increase by  $6 + 3 = 9$  units. The total increase in  $\hat{y}$  is simply the sum of the two increases. This change in  $\hat{y}$  does not depend on the values of  $x_1$  and  $x_2$  prior to the increase. Since the total increase in the dependent variable is equal to the sum of the increases from the two individual parts (independent variables), we say that the effect is *additive*.

Now suppose we have the following equation:

$$\hat{y} = 4 + 6x_1 + 3x_2 + 5x_1^2x_2$$

The important difference in this case is that the increase in the value of  $\hat{y}$  is no longer constant when  $x_1$  and  $x_2$  both increase by 1 unit each. Instead, it depends on the original values of  $x_1$  and  $x_2$ . For example, consider the values of  $x_1$  and  $x_2$ , and the changes in the value of  $\hat{y}$  shown in the following table.

$x_1$	$x_2$	$\hat{y}$	Change in $\hat{y}$ (versus $x_1 = 2$ and $x_2 = 3$ )
2	3	85	
3	3	166	81
2	4	108	23
3	4	214	129

Unlike the previous example, here the total increase in  $\hat{y}$  is not equal to the sum of the increases from the individual parts. In this

case, the effect is said to be *multiplicative*. It is important to recognize that the effect is multiplicative when the total increase does not equal the sum of the increases of the independent variables.

Pharmaceutical companies are always looking for multiplicative effects when creating new drugs. In many cases, a combination of two drugs might have a multiplicative effect on a certain condition. Simply stated, the two drugs provide *greater relief* when you take them together than if you take them separately so that only one drug is in your system at any time. Of course, the companies also have to look for multiplicative effects when it comes to side effects. Individual drugs may not have major side effects when taken separately, but could cause greater harm when taken together. One of the most noteworthy examples of this was the drug Fen-Phen, which was a combination of two drugs—Fenfluramine and Phentermine. Each of these two drugs had been approved for short-term (individual) control of obesity. However, the drugs used in combination became popular for long-term weight loss. Unfortunately, the combination, when associated with longtime use, resulted in severe side effects that were detailed in the following statement from the Food and Drug Administration in 1997:

Thanks to the reporting of health care professionals, as of August 22, FDA has received reports of 82 cases (including Mayo's 24 cases) of cardiac valvular disease in patients—two of whom were men—on combination fenfluramine and phentermine. These reports have been from 23 different states. Severity of the cardiac valvular disease was graded as moderate or severe in over three-fourths of the cases, and two of the reports described deterioration from no detectable heart murmur to need for a valve replacement within one-and-a-half years. Sixteen of these 82 patients required surgery to repair their heart valves. At least one of these patients died following surgery to repair the valves. (The agency's findings, as of July 31, are described in more detail in the current issue of The New England Journal of Medicine, which also carries the Mayo study.) Source: <http://www.fda.gov/cder/news/phen/fenphenupdate.htm>

## Glossary

**Adjusted coefficient of multiple determination** Denoted by  $\bar{R}^2$ , it gives the proportion of SST that is explained by the multiple regression model and is adjusted for the degrees of freedom.

**Coefficient of multiple determination** Denoted by  $R^2$ , it gives the proportion of SST that is explained by the multiple regression model.

**First-order multiple regression model** When each term in a regression model contains a single independent variable raised to the first power.

**Least squares regression model** The estimated regression model obtained by minimizing the sum of squared errors.

**Multicollinearity** When two or more independent variables in a regression model are highly correlated.

**Multiple regression model** A regression model that contains two or more independent variables.

**Partial regression coefficients** The coefficients of independent variables in a multiple regression model are called the partial regression coefficients because each of them gives the effect of the corresponding independent variable on the dependent variable when all other independent variables are held constant.

**Standard deviation of errors** Also called the *standard deviation of estimate*, it is a measure of the variation among errors.

**SSE (error sum of squares)** The sum of the squared differences between the actual and predicted values of  $y$ . It is the portion of SST that is not explained by the regression model.

**SSR (regression sum of squares)** The portion of SST that is explained by the regression model.

**SST (total sum of squares)** The sum of squared differences between actual  $y$  values and  $\bar{y}$ .

## Self-Review Test

1. When using the  $t$  distribution to make inferences about a single parameter, the degrees of freedom for a multiple regression model with  $k$  independent variables and a sample size of  $n$  are equal to
  - a.  $n + k - 1$
  - b.  $n - k + 1$
  - c.  $n - k - 1$
2. The value of  $R^2$  is always in the range
  - a. zero to 1
  - b.  $-1$  to 1
  - c.  $-1$  to zero
3. The value of  $\bar{R}^2$  is
  - a. always positive
  - b. always nonnegative
  - c. can be positive, zero, or negative
4. What is the difference between the population multiple regression model and the estimated multiple regression model?
5. Why are the regression coefficients in a multiple regression model called the partial regression coefficients?
6. What is the difference between  $R^2$  and  $\bar{R}^2$ ? Explain.
7. A real estate expert wanted to find the relationship between the sale price of houses and various characteristics of the houses. She collected data on four variables, recorded in the table, for 13 houses that were sold recently. The four variables are

Price = Sale price of a house in thousands of dollars

Lot size = Size of the lot in acres

Living area = Living area in square feet

Age = Age of a house in years

Price	Lot Size	Living Area	Age
455	1.4	2500	8
278	.9	2250	12
463	1.8	2900	5
327	.7	1800	9
505	2.6	3200	4
264	1.2	2400	28
445	2.1	2700	9
346	1.1	2050	13
487	2.8	2850	7
289	1.6	2400	16
434	3.2	2600	5
411	1.7	2300	8
223	.5	1700	19

Using MINITAB, find the regression of price on lot size, living area, and age. Using the solution obtained, answer the following questions.

- a. Indicate whether you expect a positive or a negative relationship between the dependent variable and each of the independent variables.
- b. Write the estimated regression equation. Are the signs of the coefficients of independent variables obtained in the solution consistent with your expectations of part a?
- c. Explain the meaning of the estimated regression coefficients of all independent variables.
- d. What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?

- e. What is the predicted sale price of a house that has a lot size of 2.5 acres, a living area of 3000 square feet, and is 14 years old?
- f. What is the point estimate of the mean sale price of all houses that have a lot size of 2.2 acres, a living area of 2500 square feet, and are 7 years old?
- g. Determine a 99% confidence interval for each of the coefficients of the independent variables.
- h. Construct a 98% confidence interval for the constant term in the population regression model.
- i. Using the 1% significance level, test whether or not the coefficient of *lot size* is positive.
- j. At the 2.5% significance level, test if the coefficient of *living area* is positive.
- k. At the 5% significance level, test if the coefficient of *age* is negative.

## ■ Mini-Project 14-1

Refer to the McDonald's data set explained in Appendix B and given on the Web site of this text. Use MINITAB to estimate the following regression model for that data set.

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + \epsilon$$

where

$y$  = calories

$x_1$  = fat (measured in grams)

$x_2$  = carbohydrate (measured in grams)

and

$x_3$  = protein (measured in grams)

Now research on the Internet or in a book to find the number of calories in one gram of fat, one gram of carbohydrate, and one gram of protein.

- a. Based on the information you obtain, write what the estimated regression equation should be.
- b. Are the differences between your expectation in part a and the regression equation that you obtained from MINTAB small or large?
- c. Since each gram of fat is worth a specific number of calories, and the same is true for a gram of carbohydrate, and for a gram of protein, one would expect that the predicted and observed values of  $y$  would be the same for each food item, but that is not the case. The quantities of fat, carbohydrates, and protein are reported in whole numbers. Explain why this causes the differences discussed in part b.

## DECIDE FOR YOURSELF

### Dummy Variables

In *Sanford & Son*, a very popular TV show of the 1970s, Fred Sanford would often refer to other people as *big dummies*. So, if a statistics professor questions your work and mentions a *dummy* in the process, should you be offended? Obviously, context will help you to answer that question, but if the professor is referring to a *dummy variable*, then do not take it personally.

A dummy variable is the name given to a categorical independent variable used in a multiple regression model. The simplest version occurs when there are only two categories. In this case, we assign a value of 0 to one category and 1 to the other category of the variable.

Suppose you have the following first-order regression equation to predict the amount of tar inhaled ( $y$ ) by smoking a cigarette based

on the amount of tar in the cigarette ( $x_1$ ) and the presence of a filter ( $x_2$ ). Note that here  $x_2 = 0$  implies that a cigarette does not have a filter and  $x_2 = 1$  means that a filter exists.

$$\hat{y} = .94x_1 - .45x_2$$

Answer the following questions.

1. Does the presence of a filter increase or decrease the tar consumption? What part of the regression equation tells you this?
2. On average, what percentage of the tar in a cigarette is consumed if the cigarette is unfiltered? What if the cigarette is filtered?
3. Draw a graph of the above regression equation. (Hint: The graph consists of two different regression lines with two variables, not a plane.)



© Kitch/Age Fotostock America, Inc.

## Nonparametric Methods<sup>1</sup>

What kind of soda is that on your desk? Regardless of the particular brand or flavor, it is more likely than last year that it is a diet soda. According to a Dow Jones report (August 19, 2002), diet sodas represent 30% of the soft drink market, a 6.6% sales increase over the prior year (compared to a 3.1% increase in sales of regular soft drinks). Even so, diet soft drinks represent only 18.2% of the total U.S. carbonated soft drink market, according to John Sicher, editor and publisher of *Beverage Digest*. We can conduct hypothesis tests to determine people's preferences for one type of soft drink over another.

The hypothesis tests discussed so far in this text are called **parametric tests**. In those tests, we used the normal, *t*, chi-square, and *F* distributions to make tests about population parameters such as means, proportions, variances, and standard deviations. In doing so, we made some assumptions, such as the assumption that the population from which the sample was drawn was normally distributed. This chapter discusses a few **nonparametric tests**. These tests do not require the same kinds of assumptions, and hence, they are also called **distribution-free tests**.

Nonparametric tests have several advantages over parametric tests: They are easier to use and understand; they can be applied to situations in which parametric tests cannot be used; and they do not require that the population being sampled is normally distributed. However, a major problem with nonparametric tests is that they are less efficient than parametric tests. The sample size must be larger for a nonparametric test to have the same probability of committing the two types of errors.

Although there are a large number of nonparametric tests that can be applied to conduct tests of hypothesis, this chapter discusses only six of them: the sign test, the Wilcoxon signed-rank test, the Wilcoxon rank sum test, the Kruskal-Wallis test, the Spearman rho rank correlation coefficient test, and the runs test for randomness.

<sup>1</sup>Tables VIII to XII that are needed for this chapter are given at the end of this chapter. Tables IV and VI are in Appendix C of the book.

- 15.1 The Sign Test
- 15.2 The Wilcoxon Signed-Rank Test for Two Dependent Samples
- 15.3 The Wilcoxon Rank Sum Test for Two Independent Samples
- 15.4 The Kruskal-Wallis Test
- 15.5 The Spearman Rho Rank Correlation Coefficient Test
- 15.6 The Runs Test for Randomness

## 15.1 The Sign Test

The **sign test** is one of the easiest tests to apply to test hypotheses. It uses only plus and minus signs. The sign test can be used to perform the following types of tests:

1. To determine the preference for one product or item over another, or to determine whether one outcome occurs more often than another outcome in categorical data. For example, we may test whether or not people prefer one kind of soft drink over another kind.
2. To conduct a test for the median of a single population. For example, we may use this procedure to test whether the median rent paid by all tenants in a city is different from \$1250.
3. To perform a test for the median of paired differences using data from two dependent samples. For example, we may use this procedure to test whether the median score on a standardized test increases after a preparatory course is taken.

### Definition

**Sign Test** The *sign test* is used to make hypothesis tests about preferences, a single median, and the median of paired differences for two dependent populations. We use only plus and minus signs to perform these tests.

In the following subsections we discuss these tests for small and large samples.

### 15.1.1 Tests About Categorical Data

Data that are divided into different categories for identification purposes are called **categorical data**. For example, people's opinions about a certain issue—in favor, against, or no opinion—produce categorical data. This subsection discusses how to perform tests about such data using the sign test procedure. We discuss two situations in which such tests can be performed: the small-sample case and the large-sample case.

#### The Small-Sample Case

When we apply the sign test for categorical data, if the *sample size is 25 or less* (i.e.,  $n \leq 25$ ), we consider it a small sample. Table VIII: Critical Values of  $X$  for the Sign Test (that appears at the end of this chapter), is based on the binomial probability distribution. This table gives the critical values of the test statistic for the sign test when  $n \leq 25$  using the binomial probability distribution.

The sign test can be used to test whether or not customers prefer one brand of a product over another brand of the same type of product. For example, we can test whether customers have a higher preference for Coke or for Pepsi. This procedure can also be used to test whether people prefer one of two alternatives over the other. For example, given a choice, do people prefer to live in New York City or in Los Angeles?

### ■ EXAMPLE 15–1

*Performing sign test with categorical data: small sample.*

The Top Taste Water Company produces and distributes Top Taste bottled water. The company wants to determine whether customers have a higher preference for its bottled water than for its main competitor, Spring Hill bottled water. The Top Taste Water Company hired a statistician to conduct this study. The statistician selected a random sample of 10 people and asked each of them to taste one sample of each of the two brands of water. The customers did not know the brand of each water sample. Also, the order in which each person tasted the two brands of water was determined randomly. Each person was asked to indicate which of the two samples of water he or she preferred. The following table shows the preferences of these 10 individuals.

Person	Brand Preferred
1	Spring Hill
2	Top Taste
3	Top Taste
4	Neither
5	Top Taste
6	Spring Hill
7	Spring Hill
8	Top Taste
9	Top Taste
10	Top Taste

Based on these results, can the statistician conclude that people prefer one brand of bottled water over the other? Use the significance level of 5%.

**Solution** We use the same five steps to perform this test of hypothesis that we used in earlier chapters.

**Step 1.** *State the null and alternative hypotheses.*

If we assume that people do not prefer either brand of water over the other, we would expect about 50% of the people (among those who show a preference) to indicate a preference for Top Taste water and the other 50% to indicate a preference for Spring Hill water. Let  $p$  be the proportion of all people who prefer Top Taste bottled water. The two hypotheses are as follows:

$$H_0: p = .50 \quad (\text{People do not prefer either of the two brands of water})$$

$$H_1: p \neq .50 \quad (\text{People prefer one brand of water over the other})$$

The null hypothesis states that 50% of the people prefer Top Taste water over Spring Hill water (and, hence, the other 50% prefer Spring Hill water). Note that we do not consider people who have no preference, and we drop them from the sample. If we fail to reject  $H_0$ , we will conclude that people do not prefer one brand over the other. However, if we reject  $H_0$ , we will conclude that the percentage of people who prefer Top Taste water over Spring Hill water is different from 50%. Thus, the conclusion will be that people prefer one brand over the other.

**Step 2.** *Select the distribution to use.*

We use the binomial probability distribution to make the test. Note that here there is only one sample and each member of the sample is asked to indicate a preference if he or she has one. We drop the members who do not indicate a preference and then compare the preferences of the remaining members. Also note that there are three outcomes for each person: (1) prefers Top Taste water, (2) prefers Spring Hill water, or (3) has no preference. We are to compare the two outcomes with preferences and determine whether more people belong to one of these two outcomes than to the other. All such tests of preferences are conducted by using the binomial probability distribution. If we assume that  $H_0$  is true, then the number of people who indicate a preference for Top Taste bottled water (the number of successes) follows the binomial distribution, with  $p = .50$ .

**Step 3.** *Determine the rejection and nonrejection regions.*

Note that 10 people were selected to taste the two brands of water and indicate their preferences. However, one of these individuals stated no preference. Hence, only 9 of the 10 people have indicated a preference for one or the other of the two brands of bottled water. To conduct the test, the person who has shown no preference is dropped from the sample. Thus, the *true sample size* is 9; that is,  $n = 9$ .

The significance level for the test is .05. Let  $X$  be the number of people in the sample of 9 who prefer Top Taste bottled water. Here  $X$  is called the *test statistic*. To establish a decision rule, we find the critical values of  $X$  from Table VIII for  $n = 9$ . From that table, for  $n = 9$  and  $\alpha = .05$  for a two-tailed test, the critical values of  $X$  are 1 and 8. Note that in a two-tailed test we read both the lower and the upper critical values, which are shown in Figure 15.1.

Thus, we will reject the null hypothesis if either fewer than two or more than seven people in nine indicate a preference for Top Taste bottled water.

**Figure 15.1**

0 or 1	2 to 7	8 or 9
Rejection region	Nonrejection region	Rejection region

**Critical Value(s) of  $X$**  In a sign test for a small sample, the *critical value of  $X$*  is obtained from Table VIII. If the test is two-tailed, we read both the lower and the upper critical values from that table. However, we read only the lower critical value if the test is left-tailed, and only the upper critical value if the test is right-tailed. Also note that which column we use to obtain this critical value depends on the given significance level and on whether the test is two-tailed or one-tailed.

**Step 4. Calculate the value of the test statistic.**

To record the results of the experiment, we mark a plus sign for each person who prefers Top Taste bottled water, a minus sign for each person who prefers Spring Hill bottled water, and a zero for the one who indicates no preference. This listing is shown in Table 15.1.

**Table 15.1**

Person	Brand Preferred	Sign
1	Spring Hill	–
2	Top Taste	+
3	Top Taste	+
4	Neither	0
5	Top Taste	+
6	Spring Hill	–
7	Spring Hill	–
8	Top Taste	+
9	Top Taste	+
10	Top Taste	+

Now, we count the number of plus signs (the sign that belongs to Top Taste bottled water because  $p$  in  $H_0$  refers to Top Taste water). There are six plus signs, indicating that six of the nine people in the sample stated a preference for Top Taste bottled water. Note that the sample size is 9, not 10, because we drop the person with zero sign. Thus,

$$\text{Observed value of } X = 6$$

**Observed Value of  $X$**  The *observed value of  $X$*  is given by the number of signs that belong to the category whose proportion we are testing for.

**Step 5. Make a decision.**

Because the observed value of  $X = 6$  falls in the nonrejection region (see Figure 15.1), we fail to reject  $H_0$ . Hence, we conclude that our sample does not indicate that people prefer either of these two brands of bottled water over the other.

Note that it does not matter which outcome  $p$  refers to. If we assume that  $p$  is the proportion of people who prefer Spring Hill water, then  $X$  will denote the number of people in a sample of  $n$  who prefer Spring Hill water. The observed value of  $X$  this time would be 3, which is the number of minus signs in Table 15.1. From Figure 15.1,  $X = 3$  also falls in the nonrejection region. Hence, we again fail to reject the null hypothesis. ■

## The Large-Sample Case

If we are testing a hypothesis about preference for categorical data and  $n > 25$ , we can use the normal probability distribution as an approximation to the binomial probability distribution.

**The Large-Sample Case** If  $n > 25$ , the normal distribution can be used as an approximation to the binomial probability distribution to perform a test of hypothesis about the preference for categorical data. The *observed value* of the test statistic  $z$ , in this case, is calculated as

$$z = \frac{(X \pm .5) - \mu}{\sigma}.$$

where  $X$  is the number of units in the sample that belong to the outcome referring to  $p$ . We either add .5 to  $X$  or subtract .5 from  $X$  to correct for continuity (see Section 6.7 of Chapter 6). We will add .5 to  $X$  if the value of  $X$  is less than or equal to  $n/2$ , and we will subtract .5 from  $X$  if the value of  $X$  is greater than  $n/2$ . The values of the mean and standard deviation are calculated as

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

Example 15–2 illustrates the procedure for the large-sample case.

### ■ EXAMPLE 15–2

A developer is interested in building a shopping mall adjacent to a residential area. Before granting or denying permission to build such a mall, the town council took a random sample of 75 adults from adjacent areas and asked them whether they favor or oppose construction of this mall. Of these 75 adults, 40 opposed construction of the mall, 30 favored it, and 5 had no opinion. Can you conclude that the number of adults in this area who oppose construction of the mall is higher than the number who favor it? Use  $\alpha = .01$ .

*Performing sign test with categorical data: large sample.*

**Solution** Again, each adult in the sample has to pick one of three choices: oppose, favor, or have no opinion. And we are to compare two outcomes—oppose and favor—to find out whether more adults belong to the outcome indicated by *oppose*. We can use the sign test here. To do so, we drop the subjects who have no opinion—that is, the adults who belong to the outcome that is not being compared. In our example, five adults have no opinion. Thus, we drop these adults from our sample and use the sample size of  $n = 70$  for the purposes of this test. Let  $p$  be the proportion of adults who oppose construction of the mall and  $q$  be the proportion who favor it. We apply the five steps to make this test.

**Step 1.** State the null and alternative hypotheses.

$$H_0: p = .50 \quad \text{and} \quad q = .50 \quad (\text{The two proportions are equal})$$

$$H_1: p > .50 \quad \text{or} \quad p > q \quad (\text{The proportion of adults who oppose the mall is greater than the proportion who favor it})$$

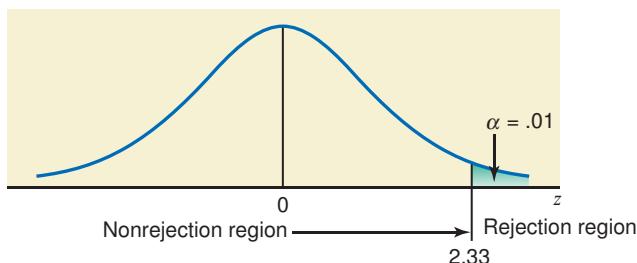
The null hypothesis here states that the proportion of adults who oppose and the proportion who favor construction of the mall are both .50, which means that  $p = .50$  and  $q = .50$ . The alternative hypothesis is that  $p > q$ , which means that more adults oppose construction of the mall than favor it. Note that  $H_1$  states that  $p > .50$  and  $q < .50$ . In other words, of those who have an opinion, more than 50% oppose and less than 50% favor construction of the shopping mall.

**Step 2.** Select the distribution to use.

As explained earlier, we will use the sign test to perform this test. Although 75 adults were asked their opinion, only 70 of them offered it and 5 did not. Hence, our sample size is 70; that is,  $n = 70$ . Because it is a large sample ( $n > 25$ ), we can use the normal approximation to perform the test.

**Step 3.** Determine the rejection and nonrejection regions.

Because  $H_1$  states that  $p > .50$ , our test is right-tailed. Also,  $\alpha = .01$ . From Table IV (the standard normal distribution table) in Appendix C, the  $z$  value for  $1.0 - .01 = .9900$  area to the left is approximately 2.33. Thus, the decision rule is that we will not reject  $H_0$  if  $z < 2.33$  and we will reject  $H_0$  if  $z \geq 2.33$ . Thus, the nonrejection region lies to the left of  $z = 2.33$  and the rejection region to the right of  $z = 2.33$ , as shown in Figure 15.2.

**Figure 15.2****Step 4.** Calculate the value of the test statistic.

Assuming that the null hypothesis is true, we expect (about) half of the adults in the population to oppose construction of the mall and the other half to favor it. Thus, we expect  $p = .50$  and  $q = .50$ . Note that we do not count the people who have no opinion. The mean and standard deviation of the binomial distribution are

$$\mu = np = 70(.50) = 35$$

$$\sigma = \sqrt{npq} = \sqrt{70(.50)(.50)} = \sqrt{17.5} = 4.18330013$$

In this example,  $p$  refers to the proportion of adults who oppose construction of the mall. Hence,  $X$  refers to the number in 70 who oppose the mall. Thus,

$$X = 40 \quad \text{and} \quad \frac{n}{2} = \frac{70}{2} = 35$$

Because  $X$  is greater than  $n/2$ , the observed value of the test statistic  $z$  is

$$z = \frac{(X - \mu) - \mu}{\sigma} = \frac{(40 - 35) - 35}{4.18330013} = 1.08$$

**Step 5.** Make a decision.

Because the observed value of  $z = 1.08$  is less than the critical value of  $z = 2.33$ , it falls in the nonrejection region. Hence, we do not reject  $H_0$ . Consequently, we conclude that the number of adults who oppose construction of the mall is not higher than the number who favor its construction. ■

### 15.1.2 Tests About the Median of a Single Population

The sign test can be used to test a hypothesis about a population median. Recall from Chapter 3 that the median is the value that divides a ranked data set into two equal parts. For example, if the median age of students in a class is 24, half of the students are younger than 24 and half are older. This section discusses how to make a test of hypothesis about the median of a population.

#### The Small-Sample Case

If  $n \leq 25$ , we use the binomial probability distribution to test a hypothesis about the median of a population. The procedure used to conduct such a test is very similar to the one explained in Example 15–1.

## ■ EXAMPLE 15–3

A real estate agent claims that the median price of homes in a small midwest city is \$137,000. A sample of 10 houses selected by a statistician produced the following data on their prices.

Home	1	2	3	4	5	6	7	8	9	10
Price (\$)	147,500	123,600	139,000	168,200	129,450	132,400	156,400	188,210	198,425	215,300

Performing sign test about a population median: small sample.

Using the 5% significance level, can you conclude that the median price of homes in this city is different from \$137,000?

**Solution** Using the given data, we prepare Table 15.2, which contains a sign row. In this row, we assign a positive sign to each price that is above the claimed median price of \$137,000 and a negative sign to each price that is below the claimed median price.

In Table 15.2, there are seven plus signs, indicating that the prices of seven houses are higher than the claimed median price of \$137,000, and there are three minus signs, showing that the prices of three homes are lower than the claimed median price. Note that if one or more values in a data set are equal to the median, then each of them is assigned a zero value and dropped from the sample. Next, we perform the following five steps to perform the test of hypothesis.

**Table 15.2**

Home	1	2	3	4	5	6	7	8	9	10
Sign	+	-	+	+	-	-	+	+	+	+

**Step 1.** State the null and alternative hypotheses.

$$H_0: \text{Median price} = \$137,000 \quad (\text{Real estate agent's claim is true})$$

$$H_1: \text{Median price} \neq \$137,000 \quad (\text{Real estate agent's claim is false})$$

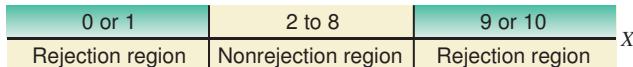
**Step 2.** Select the distribution to use.

For a test of the median of a population, we employ the sign test procedure by using the binomial probability distribution if  $n \leq 25$ . Since in our example  $n = 10$ , which is less than 25, we use the binomial probability distribution to conduct the test.

**Step 3.** Determine the rejection and nonrejection regions.

In our example,  $n = 10$  and  $\alpha = .05$ . The test is two-tailed. Let  $X$  be the test statistic that represents the number of plus signs in Table 15.2. From Table VIII, the (lower and upper) critical values of  $X$  are 1 and 9. Using these critical values, Figure 15.3 shows the rejection and nonrejection regions. Thus, we will reject the null hypothesis if the observed value of  $X$  is either 1 or 0, or 9 or 10. Note that because  $X$  represents the number of plus signs in the sample, its lowest possible value is 0 and its highest possible value is 10.

**Figure 15.3**



**Step 4.** Calculate the value of the test statistic.

The observed value of  $X$  is given by the number of plus signs in Table 15.2. Thus,

$$\text{Observed value of } X = 7$$

**Observed Value of  $X$**  When using the sign test to perform a test about a median, we can use either the number of positive signs or the number of negative signs as the observed value of  $X$  if the test is two-tailed. However, the *observed value of  $X$*  is equal to the larger of these two numbers (the number of positive and negative signs) if the test is right-tailed, and equal to the smaller of these two numbers if the test is left-tailed.

**Step 5.** Make a decision.

The observed value of  $X = 7$  falls in the nonrejection region in Figure 15.3. Hence, we do not reject  $H_0$  and conclude that the median price of homes in this city is not different from \$137,000. ■

**The Large-Sample Case**

For a test of the median of a single population, we can use the normal approximation to the binomial probability distribution when  $n > 25$ . The observed value of  $z$ , in this case, is calculated as in a test of hypothesis about the preference for categorical data (see the rule described on page 635 and Example 15–2). Example 15–4 explains the procedure for such a test.

**■ EXAMPLE 15–4**

*Performing sign test about a population median: large sample.*

A long-distance phone company believes that the median phone bill (for long-distance calls) is at least \$70 for all the families in New Haven, Connecticut. A random sample of 90 families selected from New Haven showed that the phone bills of 51 of them were less than \$70 and those of 38 of them were more than \$70, and 1 family had a phone bill of exactly \$70. Using the 1% significance level, can you conclude that the company's claim is true?

**Solution** We use the usual five steps to test this hypothesis.

**Step 1.** State the null and alternative hypotheses.

The company's claim is that the median phone bill is at least \$70. Hence, the two hypotheses are as follows:

$$H_0: \text{Median} \geq \$70 \quad (\text{Company's claim is true})$$

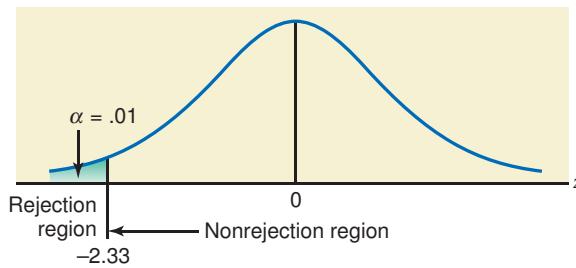
$$H_1: \text{Median} < \$70 \quad (\text{Company's claim is false})$$

**Step 2.** Select the distribution to use.

This is a test about the median and  $n > 25$ . Hence, to conduct this test, we can use the normal distribution as an approximation to the binomial probability distribution.

**Step 3.** Determine the rejection and nonrejection regions.

The test is left-tailed and  $\alpha = .01$ . From Table IV (the standard normal distribution table), the  $z$  value for .01 area in the left tail is  $-2.33$ . Note that the  $z$  value is negative because it is a left-tailed test. Thus, we will reject  $H_0$  if the observed value of  $z$  is  $-2.33$  or lower, and we will not reject  $H_0$  if the observed value of  $z$  is greater than  $-2.33$ . Figure 15.4 shows the rejection and nonrejection regions.

**Figure 15.4****Step 4.** Calculate the value of the test statistic.

In our example, 51 phone bills out of 90 are below the hypothesized median, 38 are above it, and 1 is exactly equal to the median. When we perform this test, we drop the value or values that are exactly equal to the median. Thus, after dropping one value that is equal to the median, our sample size is  $51 + 38 = 89$ ; that is,  $n = 89$ . Let a phone bill below the median be represented by a minus sign and one above the median by a plus sign. Then, in these 89 bills there are 51 minus signs (for values less than the median) and 38 plus signs (for values

greater than the median). If the given claim is true, we would expect (about) half plus signs and half minus signs. Let  $p$  be the proportion of plus signs in 89. Then, we would expect  $p = .50$  if  $H_0$  is true. Hence, the mean and the standard deviation of the binomial distribution are calculated as follows:

$$\begin{aligned} n &= 89 \quad p = .50 \quad q = 1 - p = .50 \\ \mu &= np = 89(.50) = 44.50 \\ \sigma &= \sqrt{npq} = \sqrt{89(.50)(.50)} = 4.71699057 \end{aligned}$$

In our example, 51 phone bills are below the median and 38 are above the median. Because it is a left-tailed test,  $X = 38$ , which is the smaller of the two numbers (51 and 38). Consequently, the  $z$  value is calculated as follows. Note that we have added .5 to  $X$  because the value of  $X$  is less than  $n/2$ , which is  $89/2 = 44.5$ .

$$z = \frac{(X + .5) - \mu}{\sigma} = \frac{(38 + .5) - 44.50}{4.71699057} = -1.27$$

#### **Step 5. Make a decision.**

Because  $z = -1.27$  is greater than the critical value of  $z = -2.33$ , we do not reject  $H_0$ . Hence, we conclude that the company's claim that the median phone bill is at least \$70 seems to be true. ■

### ► Observation

Note that in Example 15–4 there were 51 minus signs and 38 plus signs. We assigned the smaller of these two numbers to  $X$  so that  $X = 38$  to calculate the observed value of  $z$ . We did so to obtain a negative value of the observed  $z$  because the test is left-tailed and the critical value of  $z$  is negative. If we assigned 51 as the value of  $X$ , we would obtain  $z = +1.27$  as the observed value of  $z$ , which does not make sense. Let  $X_1$  be the number of plus signs and  $X_2$  the number of minus signs in a test about the median. Then, we can establish the following rules to calculate the observed value of  $X$ .

1. If the test is two-tailed, it does not matter which of the two values,  $X_1$  and  $X_2$ , is assigned to  $X$  to calculate the observed value of  $z$ .
2. If the test is left-tailed,  $X$  should be assigned a value equal to the smaller of the values of  $X_1$  and  $X_2$ .
3. If the test is right-tailed,  $X$  should be assigned a value equal to the larger of the values of  $X_1$  and  $X_2$ .

Note that the rule to calculate the observed value of  $z$  here is the same as explained on page 635 for the large-sample case for a test of hypothesis about the preference for categorical data.

### 15.1.3 Tests About the Median Difference Between Paired Data

We can use the sign test to perform a test of hypothesis about the difference between the medians of two dependent populations using the data obtained from paired samples. We learned in Section 10.4 of Chapter 10 that two samples are paired samples when, for each data value collected from one sample, there is a corresponding data value collected from the second sample, and both data values are collected from the same source. In this section we discuss the small-sample and the large-sample cases to conduct such tests.

#### The Small-Sample Case

If  $n \leq 25$ , we use the binomial probability distribution to perform a test about the difference between the medians of paired data. In such a case, Table VIII is used to find the critical values of the test statistic. Example 15–5 illustrates this procedure.

*Performing sign test  
about the median of paired differences: small samples.*

## ■ EXAMPLE 15–5

A researcher wanted to find the effects of a special diet on systolic blood pressure in adults. She selected a sample of 12 adults and put them on this dietary plan for three months. The following table gives the systolic blood pressure of each adult before and after the completion of the plan.

Before	210	185	215	198	187	225	234	217	212	191	226	238
After	196	192	204	193	181	233	208	211	190	186	218	236

Using the 2.5% significance level, can we conclude that the dietary plan reduces the median systolic blood pressure of adults?

**Solution** We find the sign of the difference between the two blood pressure readings of each adult by subtracting the blood pressure after completion of the dietary plan from the blood pressure before the plan. A plus sign indicates that the plan reduced that person's blood pressure and a minus sign means that it increased blood pressure. Table 15.3 gives the signs of the differences.

**Table 15.3**

Before	210	185	215	198	187	225	234	217	212	191	226	238
After	196	192	204	193	181	233	208	211	190	186	218	236
Sign of difference (before – after)	+	–	+	+	+	–	+	+	+	+	+	+

Next we perform the five steps for testing the hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

Let  $M$  denote the difference in median blood pressure readings before and after the dietary plan. The null and alternative hypotheses are as follows:

$$H_0: M = 0 \quad (\text{The dietary plan does not reduce the median blood pressure})$$

$$H_1: M > 0 \quad (\text{The dietary plan reduces the median blood pressure})$$

The alternative hypothesis is that the dietary plan reduces the median blood pressure, which means that the median systolic blood pressure of all adults after the completion of the dietary plan is lower than the median systolic blood pressure before the completion of the dietary plan. In this case, the median of the paired differences will be greater than zero.

**Step 2.** *Select the distribution to use.*

The sample size is small (that is,  $n = 12 < 25$ ), and we do not know the shape of the distribution of the population of paired differences. Hence, we use the sign test with the binomial probability distribution.

**Step 3.** *Determine the rejection and nonrejection regions.*

Because the test is right-tailed,  $n = 12$ , and  $\alpha = .025$ , the critical value of  $X$  from Table VIII is 10. Note that we use the upper critical value of  $X$  in Table VIII because, as indicated by the sign in  $H_1$ , the test is right-tailed. Thus, we will reject the null hypothesis if the observed value of  $X$  is greater than or equal to 10, and we will not reject  $H_0$  otherwise. The rejection and nonrejection regions are shown in Figure 15.5.

**Figure 15.5**

0 to 9	10 to 12
Nonrejection region	Rejection region

**Step 4.** Calculate the value of the test statistic.

In our sample data, the blood pressure of 10 adults decreases and that of 2 adults increases after the dietary plan. Note that there are 10 plus signs and 2 minus signs in Table 15.3. Whenever the test is right-tailed, the observed value of  $X$  is equal to the larger of these two numbers. Thus, for our example, observed value of  $X = 10$ .

**Step 5.** Make a decision.

Because the observed value of  $X = 10$  falls in the rejection region, we reject  $H_0$ . Hence, we conclude that the dietary plan reduces the median blood pressure of adults. ■

### The Large-Sample Case

In Example 15–5, we used Table VIII to find the critical value of the test statistic  $X$ . However, Table VIII goes up to  $n = 25$  only. If  $n > 25$ , we can use the normal distribution as an approximation to the binomial distribution to conduct a test about the difference between the medians of paired data. The following example illustrates such a case.

#### ■ EXAMPLE 15–6

Many students suffer from math anxiety. A statistics professor offered a two-hour lecture on math anxiety and ways to overcome it. A total of 42 students attended this lecture. The students were given similar statistics tests before and after the lecture. Thirty-three of the 42 students scored higher on the test after the lecture, 7 scored lower after the lecture, and 2 scored the same on both tests. Using the 1% significance level, can you conclude that the median score of students increases as a result of attending this lecture? Assume that these 42 students constitute a random sample of all students who suffer from math anxiety.

*Performing sign test  
about the median of paired differences: large samples.*

**Solution** Let  $M$  be the median of the paired differences between scores of students before and after the test, where a paired difference is obtained by subtracting the score after the lecture from the score before the lecture. In other words,

$$\text{Paired difference} = \text{Score before} - \text{Score after}$$

Thus, a positive paired difference means that the score before the lecture is higher than the score after the lecture for that student, and a negative paired difference indicates that the score before the lecture is lower than the score after the lecture for that student. Thus, there are 33 minus signs, 7 plus signs, and 2 zeros.

**Step 1.** State the null and alternative hypotheses.

$$H_0: M = 0 \quad (\text{The lecture does not increase the median score})$$

$$H_1: M < 0 \quad (\text{The lecture increases the median score})$$

The alternative hypothesis is that the lecture increases the median score, which means that the median score after the lecture is higher than the median score before the lecture. In this case, the median of the paired differences will be less than zero.

**Step 2.** Select the distribution to use.

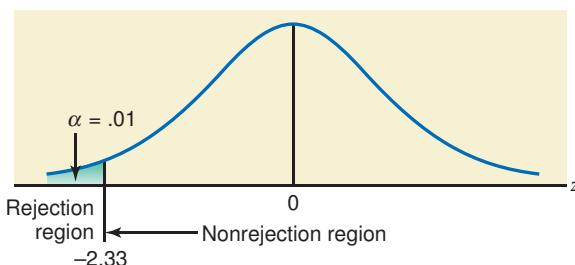
Here,  $n = 40$ . Note that to find the sample size, we drop the students whose score did not change. Because  $n > 25$ , we can use the normal distribution to test this hypothesis about the median of paired differences.

**Step 3.** Determine the rejection and nonrejection regions.

The test is left-tailed and  $\alpha = .01$ . From Table IV, the critical value of  $z$  for .01 area in the left tail is  $-2.33$ . Thus, the decision rule is that we will reject the null hypothesis if the observed

value of  $z$  is  $-2.33$  or smaller, and we will not reject  $H_0$  otherwise. The rejection and nonrejection regions are shown in Figure 15.6.

**Figure 15.6**



**Step 4. Calculate the value of the test statistic.**

If the null hypothesis is true (that is, the lecture does not increase the median score), then we would expect (about) half of the students to score higher and the other half to score lower after the lecture than before. Thus, we would expect (about) half plus signs and half minus signs in the population. In other words, if  $p$  is the proportion of plus signs, we would expect  $p = .50$  when  $H_0$  is true. Hence, the mean and standard deviation of the binomial distribution are

$$\mu = np = 40(.50) = 20$$

$$\sigma = \sqrt{npq} = \sqrt{40(.50)(.50)} = 3.16227766$$

In our example, 33 of the students scored higher after the lecture and 7 scored lower after the lecture. Thus, there are 33 minus signs and 7 plus signs. We assign the smaller of these two values to  $X$  when the test is left-tailed. Here  $X = 7$ , and the observed value of  $z$  is calculated as follows. Here, because the value of  $X$  is less than  $n/2$ , we add .5 to  $X$ .

$$z = \frac{(X + .5) - \mu}{\sigma} = \frac{(7 + .5) - 20}{3.16227766} = -3.95$$

**Step 5. Make a decision.**

Because the observed value of  $z = -3.95$  is less than the critical value of  $z = -2.33$ , it falls in the rejection region. Consequently, we reject  $H_0$  and conclude that attending the math anxiety lecture increases the median test score. ■

**Remember ►** Again, remember that if the test is left-tailed,  $X$  is assigned the value equal to the smaller number of plus or minus signs. On the other hand, if the test is right-tailed,  $X$  is assigned the value equal to the larger number of plus or minus signs. Note that the rule to calculate the observed value of  $z$  here is the same as explained on page 635 for the large-sample case for a test of hypothesis about the preference for categorical data.

## EXERCISES

### CONCEPTS AND PROCEDURES

- 15.1 Briefly explain the meaning of categorical data and give two examples.
- 15.2 When we use the sign test for categorical data, how large a sample size is required to permit the use of the normal distribution for determining the rejection region?
- 15.3 When we use the sign test for the median of a single population, how small must the sample size be to require the use of Table VIII?
- 15.4 When we use the sign test for the difference between the medians of two dependent populations, how large must  $n$  be for the large-sample case?

**15.5** Determine the rejection region for each of the following sign tests for categorical data.

- a.  $H_0: p = .50$ ,  $H_1: p > .50$ ,  $n = 15$ ,  $\alpha = .05$
- b.  $H_0: p = .50$ ,  $H_1: p \neq .50$ ,  $n = 20$ ,  $\alpha = .01$
- c.  $H_0: p = .50$ ,  $H_1: p < .50$ ,  $n = 30$ ,  $\alpha = .05$

**15.6** In each case below,  $n$  is the sample size,  $p$  is the proportion of the population possessing a certain characteristic, and  $X$  is the number of items in the sample that possess that characteristic. In each case, perform the appropriate sign test using  $\alpha = .05$ .

- a.  $n = 14$ ,  $X = 10$ ,  $H_0: p = .50$ ,  $H_1: p > .50$
- b.  $n = 10$ ,  $X = 1$ ,  $H_0: p = .50$ ,  $H_1: p \neq .50$
- c.  $n = 30$ ,  $X = 12$ ,  $H_0: p = .50$ ,  $H_1: p < .50$
- d.  $n = 27$ ,  $X = 20$ ,  $H_0: p = .50$ ,  $H_1: p > .50$

**15.7** In each case below,  $n$  is the sample size and  $X$  is the appropriate number of plus or minus signs as defined in Section 15.1.2. In each case, perform the appropriate sign test using  $\alpha = .05$ .

- a.  $n = 10$ ,  $X = 8$ ,  $H_0: \text{Median} = 28$ ,  $H_1: \text{Median} > 28$
- b.  $n = 11$ ,  $X = 1$ ,  $H_0: \text{Median} = 100$ ,  $H_1: \text{Median} < 100$
- c.  $n = 26$ ,  $X = 3$ ,  $H_0: \text{Median} = 180$ ,  $H_1: \text{Median} \neq 180$
- d.  $n = 30$ ,  $X = 6$ ,  $H_0: \text{Median} = 55$ ,  $H_1: \text{Median} < 55$

**15.8** In each case below,  $M$  is the difference between two population medians,  $n$  is the sample size, and  $X$  is the appropriate number of plus or minus signs as defined at the end of Section 15.1.3. In each case, perform the appropriate sign test using  $\alpha = .01$ .

- a.  $n = 20$ ,  $X = 6$ ,  $H_0: M = 0$ ,  $H_1: M < 0$
- b.  $n = 8$ ,  $X = 8$ ,  $H_0: M = 0$ ,  $H_1: M > 0$
- c.  $n = 29$ ,  $X = 4$ ,  $H_0: M = 0$ ,  $H_1: M \neq 0$

## ■ APPLICATIONS

**15.9** In Pine Grove, the city water is safe to drink but some people think it has a slightly unpleasant taste due to chemical treatment. Some residents prefer to buy bottled water (B) but others drink the city water (C). A random sample of 12 residents is taken. Their preferences are shown here.

B      C      B      C      C      B      C      C      C      C      B      C

At the 5% significance level, can you conclude that the residents of Pine Grove prefer either type of drinking water over the other?

**15.10** A consumer organization wanted to compare two rival brands of infant car seats, Brand A and Brand B. Fifteen families, each with a child under 12 months of age, were selected at random. Each family tested each of the two brands of car seats for one week. The order in which each family tried the two brands was decided by a coin toss. At the end of two weeks, each family indicated which brand it preferred. Their preferences are listed here. The 0 indicates that one family had no preference.

A      A      A      B      A      A      B      A      A      A      A      A      0      A      B      A

At the 5% level of significance, can you conclude that families prefer Brand A over Brand B?

**15.11** Twenty randomly chosen loyal drinkers of JW's Beer are tested to see if they can distinguish between JW's and its chief rival. Each of the 20 drinkers is given two unmarked cups, one containing JW's and the other containing the rival brand. Thirteen of the drinkers correctly indicate which cup contains JW's, but the other seven are incorrect. At the 2.5% level of significance, can you conclude that drinkers of JW's Beer are more likely to correctly identify it than not?

**15.12** Three weeks before an election for state senator, a poll of 200 randomly selected voters shows that 95 voters favor the Republican candidate, 85 favor the Democratic candidate, and the remaining 20 have no opinion. Using the sign test, can you conclude that voters prefer one candidate over the other? Use  $\alpha = .01$ .

**15.13** One hundred randomly chosen adult residents of North Dakota are asked whether they would prefer to live in another state or to stay in North Dakota. Of these 100 adults, 55 indicate that they would like to move to another state, 41 would prefer to stay, and 4 have no preference. At the 2.5% level of significance, can you conclude that less than half of all adult residents of North Dakota would prefer to stay?

**15.14** Three hundred randomly chosen doctors were asked, Which is the most important single factor in weight control: diet or exercise? Of these 300 doctors, 162 felt that diet was more important, 117 favored exercise, and 21 thought that diet and exercise were equally important. At the 1% level of significance, can you conclude that for all doctors, the number who favor diet exceeds the number who favor exercise?

**15.15** In a Gallup Poll of adults taken December 6–9, 2001, 42% reported that they frequently experienced stress in their daily lives (*USA TODAY*, January 24, 2002). Suppose that in a recent sample of 700 adults, 370 indicated that they frequently experience such stress. Using the sign test with  $\alpha = .01$ , can you conclude that currently more than half of all adults frequently experience stress in their daily lives?

**15.16** A past study claims that adults in the United States spend a median of 18 hours a week on leisure activities. A researcher took a sample of 10 adults and asked them how many hours they spend per week on leisure activities. She obtained the following data:

14      25      22      38      16      26      19      23      41      33

Using  $\alpha = .05$ , can you conclude that the median amount of time spent per week on leisure activities by all adults is more than 18 hours?

**15.17** The manager of a soft-drink bottling plant wants to see if the median amount of soda in 12-ounce bottles differs from 12 ounces. Ten filled bottles are selected at random from the bottling machine, and the amount of soda in each is carefully measured. The results (in ounces) follow:

12.10      11.95      12.00      12.01      12.02      12.05      12.02      12.03      12.04      12.06

Using the 5% level of significance, can you conclude that the median amount of soda in all such bottles differs from 12 ounces?

**15.18** According to the annual *USA TODAY/NFL* salary survey, the median salary of offensive linemen in the National Football League (NFL) was \$589,133 in 2001 (*USA TODAY*, July 29, 2002). Suppose that a recent random sample of 10 NFL offensive linemen yielded the following salaries (in thousands of dollars).

700      615      710      805      630      575      900      730      710      695

Using the 5% level of significance, can you conclude that the current median salary of all offensive linemen in the NFL exceeds \$589,133?

**15.19** A city police department claims that its median response time to 911 calls in the inner city is four minutes or less. Shown below is a random sample of 28 response times (in minutes) to 911 calls in the inner city.

6      5      7      12      2      1.5      3.5      4      10      11      4.5      6      5      8.5  
7      15      9      8      3      10      8      4.5      9      4      6      3      6      7.5

Using  $\alpha = .01$ , can you conclude that the median response time to all 911 calls in the inner city is longer than four minutes?

**15.20** According to the American Community Survey conducted during the 2000 census, New Jersey's median household income of \$54,226 was the highest among the 50 states (*USA TODAY*, August 6, 2001). Suppose that in a recent random sample of 400 New Jersey households, 220 had incomes higher than \$54,226 and 180 had incomes lower than \$54,226. Using the sign test at the 2% level of significance, can you conclude that the current median household income in New Jersey differs from \$54,226?

**15.21** The following numbers are the times served (in months) by 35 prison inmates who were released recently.

37      6      20      5      25      30      24      10      12      20  
24      8      26      15      13      22      72      80      96      33  
84      86      70      40      92      36      28      90      36      32  
72      45      38      18      9

Using  $\alpha = .01$ , test the null hypothesis that the median time served by all such prisoners is 42 months against the alternative hypothesis that the median time served is less than 42 months.

**15.22** Twelve sixth-grade boys who are underweight are put on a special diet for one month. Each boy is weighed before and after the one-month dietary regime. The weights (in pounds) of these boys are recorded here.

Before	65	63	71	60	66	72	78	74	58	59	77	65
After	70	68	75	60	69	70	81	81	66	56	79	71

Can you conclude that this diet increases the median weight of all such boys? Use the 2.5% level of significance. Assume that these 12 boys constitute a random sample of all underweight sixth-grade boys.

- 15.23** Refer to Exercise 10.52 of Chapter 10. The following table shows the self-confidence test scores of seven employees before and after they attended a course on building self-confidence.

Before	8	5	4	9	6	9	5
After	10	8	5	11	6	7	9

At the 5% level of significance, can you conclude that attending this course increases the median self-confidence test score of all employees?

- 15.24** The manager at a large factory suspects that the night-shift workers use more hours of sick leave than the day-shift workers. The workers at this factory are rotated between shifts. Each worker works the day shift for two months, then works the night shift for two months, then goes back to the day shift for two months, and so on. The manager at the factory selected 12 workers randomly and recorded the total number of hours of sick leave each of these workers used during the two months of day shift and then during the two months of night shift. The results are given in the following table.

Day shift	20	32	12	24	16	0	22	8	10	38	16	12
Night shift	16	56	0	28	36	24	40	29	30	26	32	20

Using the 5% level of significance, can you conclude that the median number of hours of sick leave used by workers is lower for the day shift than for the night shift?

- 15.25** At a large bicycle factory, employees are paid by the hour to assemble bicycles. The plant manager decides to test a modified-piecework payment schedule, whereby each worker will be paid a lower hourly wage plus an additional amount for each bicycle assembled. The manager randomly selects 27 workers and places them on the new payment schedule. For each worker in the sample, the number of bicycles assembled during the last week under the old payment system is recorded, and then the number of bicycles assembled during the first week under the new system is recorded. Nineteen of the workers assembled more bicycles under the new system, seven assembled fewer, and one assembled the same number. Using the 2% level of significance, can you conclude that the median number of bicycles assembled by all such workers is the same under both payment systems?

- 15.26** A researcher suspects that two medical laboratories, A and B, tend to give different results when determining the cholesterol content of blood samples. The researcher obtains blood samples from 30 randomly selected adults and divides each sample into two parts. One part of each blood sample is sent to Lab A, the other part to Lab B. Each lab determines the cholesterol content of each of its 30 samples and reports the results to the researcher. The following table gives the cholesterol levels (in milligrams per hundred milliliters) reported by the two labs.

Sample	Lab A	Lab B	Sample	Lab A	Lab B
1	135	137	16	214	202
2	202	195	17	255	242
3	239	250	18	233	217
4	210	202	19	246	231
5	180	185	20	292	262
6	195	195	21	229	212
7	188	177	22	170	172
8	200	204	23	261	243
9	320	300	24	310	281
10	290	269	25	302	277
11	285	271	26	283	264
12	210	216	27	221	199
13	185	176	28	208	211
14	194	184	29	344	321
15	181	182	30	170	164

Can you conclude that the median cholesterol level for all such adults as determined by Lab A is higher than that determined by Lab B? Use a significance level of 1%.

**15.27** A dairy agency wants to test a hormone that may increase cows' milk production. Some members of the group fear that the hormone could actually decrease production, so a "matched pairs" test is arranged. Thirty randomly selected cows are given the hormone, and their milk production is recorded for four weeks. Each of these 30 cows is matched with another cow of similar size, age, and prior record of milk production. This second group of 30 cows do not receive the hormone. The milk production of these cows is recorded for the same time period. In 19 of these 30 pairs, the cow taking the hormone produced more milk; in 9 of the pairs, the cow taking the hormone produced less; and in 2 of the pairs, there was no difference. Using the 5% level of significance, can you conclude that the hormone changes the median milk production of such cows?

## 15.2

# The Wilcoxon Signed-Rank Test for Two Dependent Samples

The **Wilcoxon signed-rank test** for two dependent (paired) samples is used to test whether or not the two populations from which these samples are drawn are identical. We can also test the alternative hypothesis that one population distribution lies to the left or to the right of the other. Actually, the null hypothesis in this test states that the medians of the two population distributions are equal. The alternative hypothesis states that the medians of the two populations are not equal, or that the median of the first population is less than that of the second population, or that the median of the first population is greater than that of the second population. This test is an alternative to the paired-samples test discussed in Section 10.4.2 of Chapter 10. In that section, we assumed that the paired differences have a normal distribution. Here, in the Wilcoxon signed-rank test, we do not make that assumption. In this test, we rank the absolute differences between the pairs of data values collected from two samples and then assign them the sign based on which of the paired data values is larger. Then we compare the sums of the ranks with plus and minus signs and make a decision.

### The Small-Sample Case

If the *sample size is 15 or smaller*, we find the critical value of the test statistic, denoted by  $T$ , from Table IX (given at the end of this chapter) which gives the critical values of  $T$  for the Wilcoxon signed-rank test. We also calculate the observed value of the test statistic differently in this test. However, when  $n > 15$ , we can use the normal distribution to perform the test. Example 15–7 describes the small-sample case for the Wilcoxon signed-rank test.

### ■ EXAMPLE 15–7

Performing the Wilcoxon signed-rank test for two dependent populations: small samples.

A private agency claims that the crash course it offers significantly increases the writing speed of secretaries. The following table gives the writing speeds of eight secretaries before and after they attended this course.

Before	84	75	88	91	65	71	90	75
After	97	72	93	110	78	69	115	75

Using the 2.5% significance level, can you conclude that attending this course increases the writing speed of secretaries? Use the Wilcoxon signed-rank test.

**Solution** We use the five steps to make the hypothesis test.

**Step 1.** *State the null and alternative hypotheses.*

$H_0$ : The crash course does not increase the writing speed of secretaries

$H_1$ : The crash course does increase the writing speed of secretaries

Note that the alternative hypothesis states that the population distribution of writing speeds of secretaries moves to the right after they attend the crash course. In other words, the center of the population distribution of writing speeds after the crash course is greater than the center of the population distribution of writing speeds before the crash course. If we measure the centers of the two populations by their respective medians, with  $M_A$  the median of the population distribution after the course and  $M_B$  the median of the population distribution before the course, we can rewrite the two hypotheses as follows:

$$H_0: M_A = M_B$$

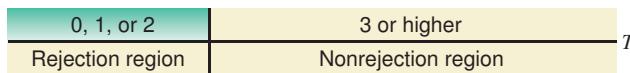
$$H_1: M_A > M_B$$

**Step 2.** *Select the distribution to use.*

We are making a test for paired samples, and the distribution of paired differences is unknown. Since  $n < 15$ , we use the Wilcoxon signed-rank test procedure for the small-sample case.

**Step 3.** *Determine the rejection and nonrejection regions.*

As mentioned earlier, we denote the test statistic in this case by  $T$ . The critical value of  $T$  is found from Table IX, which lists the critical values of  $T$  for the Wilcoxon signed-rank test for small samples ( $n \leq 15$ ). Our test is right-tailed because the alternative hypothesis is that the “after” distribution lies to the right of the “before” distribution. Also,  $\alpha = .025$  and  $n = 7$ . Note that for one pair of data, both values are the same, 75. We drop such cases when determining the sample size for the test. From Table IX, the critical value of  $T$  is 2. Thus, our decision rule will be: Reject  $H_0$  if the observed value of  $T$  is less than or equal to the critical value of  $T$ , which is 2. Note that in the Wilcoxon signed-rank test, the null hypothesis is rejected if the observed value of  $T$  is less than or equal to the critical value of  $T$ . This rule is true for a two-tailed, a right-tailed, or a left-tailed test. The observed value of  $T$  is calculated differently, depending on whether the test is two-tailed or one-tailed. This is explained in the next step. Figure 15.7 shows the rejection and nonrejection regions.



**Figure 15.7**

**Decision Rule** For the Wilcoxon signed-rank test for small samples ( $n \leq 15$ ), the critical value of  $T$  is obtained from Table IX. Note that in the Wilcoxon signed-rank test, the decision rule is to reject the null hypothesis if the observed value of  $T$  is less than or equal to the critical value of  $T$ . This rule is true for a two-tailed, a right-tailed, or a left-tailed test.

**Step 4.** *Calculate the value of the test statistic.*

The observed value of the test statistic,  $T$ , is calculated as follows. The given data on writing speeds before and after the course are reproduced in the first two columns of Table 15.4.

**Table 15.4**

Before	After	Differences (Before – After)	Absolute Differences	Ranks of Differences	Signed Ranks
84	97	-13	13	4.5	-4.5
75	72	+3	3	2	+2
88	93	-5	5	3	-3
91	110	-19	19	6	-6
65	78	-13	13	4.5	-4.5
71	69	+2	2	1	+1
90	115	-25	25	7	-7
75	75	0	0	—	—

1. We obtain the differences column by subtracting each data value after the course from the corresponding data value before the course. Thus,

$$\text{Difference} = \text{Writing speed before the course} - \text{Writing speed after the course}$$

These differences are listed in the third column of Table 15.4.

2. In the fourth column, we write the absolute values of differences. In other words, the numbers in the fourth column of the table are the same as those in the third column, but without plus and minus signs.
3. Next, we rank the absolute differences listed in the fourth column from lowest to highest. These ranks are listed in the fifth column. Note that the difference of zero is not ranked and is dropped from the sample. Among the remaining absolute differences, the smallest difference is 2, which is assigned a rank of 1. The next smallest absolute difference is 3, which is assigned a rank of 2. Next, the absolute difference of 5 is given a rank of 3. Then, two absolute differences have the same value, which is 13. We assign the average of the next two ranks,  $(4 + 5)/2 = 4.5$ , to these two values. Thus, as a rule, whenever some of the absolute differences have the same value, they are all assigned the average of their ranks.
4. In the last column of Table 15.4, we write the ranks of the fifth column with the signs of the corresponding paired differences. For example, the first difference of  $-13$  has a minus sign in the third column. Consequently, we assign a minus sign to its rank of 4.5 in the sixth column. The second difference of 3 has a positive sign. Hence, its rank of 2 is assigned a positive sign.
5. Next, we add all the positive ranks and we add the absolute values of the negative ranks separately. Thus, we obtain:

$$\text{Sum of the positive ranks} = 2 + 1 = 3$$

$$\text{Sum of the absolute values of the negative ranks} = 4.5 + 3 + 6 + 4.5 + 7 = 25$$

The observed value of the test statistic is determined as shown in the next box.

#### Observed Value of the Test Statistic $T$

- I. If the test is two-tailed with the alternative hypothesis that the two distributions are not the same, then the observed value of  $T$  is given by the smaller of the two sums, the sum of the positive ranks and the sum of the absolute values of the negative ranks. We will reject  $H_0$  if the observed value of  $T$  is less than or equal to the critical value of  $T$ .
- II. If the test is right-tailed with the alternative hypothesis that the distribution of *after* values is to the right of the distribution of *before* values, then the observed value of  $T$  is given by the sum of the values of the positive ranks. We will reject  $H_0$  if the observed value of  $T$  is less than or equal to the critical value of  $T$ .
- III. If the test is left-tailed with the alternative hypothesis that the distribution of *after* values is to the left of the distribution of *before* values, then the observed value of  $T$  is given by the sum of the absolute values of the negative ranks. We will reject  $H_0$  if the observed value of  $T$  is less than or equal to the critical value of  $T$ .

*Remember*, for the above to be true, the paired difference is defined as the *before* value minus the *after* value. In other words, the differences are obtained by subtracting the *after* values from the *before* values.

Our example is a right-tailed test. Hence,

$$\text{Observed value of } T = \text{sum of the positive ranks} = 3$$

#### Step 5. Make a decision.

Whether the test is two-tailed, left-tailed, or right-tailed, we will reject the null hypothesis if:

$$\text{Observed value of } T \leq \text{Critical value of } T$$

where the observed value of  $T$  is calculated as explained in Step 4. In this example, the observed value of  $T$  is 3 and the critical value of  $T$  is 2. Because the observed value of  $T$  is greater than the critical value of  $T$ , we do not reject  $H_0$ . Hence, we conclude that the crash course does not seem to increase the writing speed of secretaries.

### The Large-Sample Case

If  $n > 15$ , we can use the normal distribution to make a test of hypothesis about the paired differences. Example 15–8 illustrates the procedure for making such a test.

#### ■ EXAMPLE 15–8

The manufacturer of a gasoline additive claims that the use of its additive increases gasoline mileage. A random sample of 25 cars was selected, and these cars were driven for one week without the gasoline additive and then for one week with the additive. Then, the miles per gallon (mpg) were estimated for these cars without and with the additive. Next, the paired differences were calculated for these 25 cars, where a paired difference is defined as

$$\text{Paired difference} = \text{mpg without additive} - \text{mpg with additive}$$

The differences were positive for 4 cars, negative for 19 cars, and zero for 2 cars. First, the absolute values of the paired differences were ranked, and then these ranks were assigned the signs of the corresponding paired differences. The sum of the ranks of the positive paired differences was 58, and the sum of the absolute values of the ranks of the negative paired differences was 218. Can you conclude that the use of the additive increases gasoline mileage? Use the 1% significance level.

**Solution** We perform the five steps to conduct this test of hypothesis.

**Step 1.** *State the null and alternative hypotheses.*

We are to test whether or not the gasoline additive increases gasoline mileage. This will be true if the distribution of gasoline mileages with the additive lies to the right of the distribution of gasoline mileage without the additive. The median mileage with the additive will be higher than the median mileage without the additive. Let  $M_A$  and  $M_B$  be the median mileage after (with) and before (without) the gasoline additive. Then, the null and the alternative hypotheses can be written as follows:

$$H_0: M_A = M_B$$

$$H_1: M_A > M_B$$

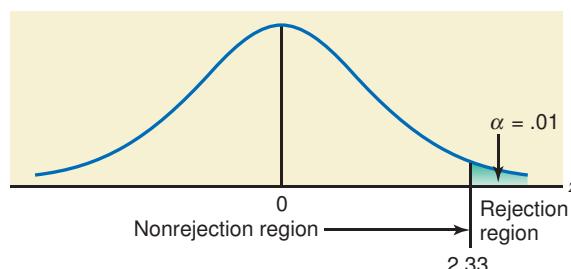
**Step 2.** *Select the distribution to use.*

We are given information about the sums of positive and negative ranks. The sample size is greater than 15. We use the Wilcoxon signed-rank test procedure with the normal distribution approximation.

**Step 3.** *Determine the rejection and nonrejection regions.*

We are using the normal distribution as an approximation to make this test. Hence, we will find the critical value of  $z$  from Table IV in Appendix C. The test is right-tailed. The significance level is .01, which gives the area to the left of critical point as  $1 - .01 = .9900$ . Therefore, the critical value of  $z$  is 2.33. The rejection and nonrejection regions are shown in Figure 15.8.

**Performing the Wilcoxon signed-rank test for paired populations: large samples.**



**Figure 15.8**

**Step 4.** Calculate the value of the test statistic.

Because the sample size is larger than 15, the test statistic  $T$  follows (an approximate) normal distribution.

**Observed Value of  $z$**  In a Wilcoxon signed-rank test for two dependent samples, when the sample size is large ( $n > 15$ ), the observed value of  $z$  for the test statistic  $T$  is calculated as

$$z = \frac{T - \mu_T}{\sigma_T}$$

$$\text{where } \mu_T = \frac{n(n + 1)}{4} \quad \text{and} \quad \sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

The value of  $T$  that is used to calculate the value of  $z$  is determined based on the alternative hypothesis, as explained next.

1. If the test is two-tailed with the alternative hypothesis that the two distributions are not the same, then the value of  $T$  may be equal to either of the two sums, the sum of the positive ranks or the sum of the absolute values of the negative ranks. We will reject  $H_0$  if the observed value of  $z$  falls in either of the rejection regions.
2. If the test is right-tailed with the alternative hypothesis that the distribution of *after* values is to the right of the distribution of *before* values, then the value of  $T$  is equal to the sum of the absolute values of the negative ranks. We will reject  $H_0$  if the observed value of  $z$  is greater than or equal to the critical value of  $z$ .
3. If the test is left-tailed with the alternative hypothesis that the distribution of *after* values is to the left of the distribution of *before* values, then the value of  $T$  is equal to the sum of the absolute values of the negative ranks. We will reject  $H_0$  if the observed value of  $z$  is less than or equal to the critical value of  $z$ .

Remember, for the above to be true, the paired difference is defined as the *before* value minus the *after* value. In other words, the differences are obtained by subtracting the *after* values from the *before* values. Also, note that whether the test is right-tailed or left-tailed, the value of  $T$  in both cases is equal to the sum of the absolute values of the negative ranks.

Using the given information, we calculate the values of  $\mu_T$  and  $\sigma_T$  and the observed value of  $z$  as follows. Note that here  $n = 23$  because two of the paired differences are zero.

$$\mu_T = \frac{n(n + 1)}{4} = \frac{23(23 + 1)}{4} = 138$$

$$\sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}} = \sqrt{\frac{23(23 + 1)(46 + 1)}{24}} = 32.87856445$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{218 - 138}{32.87856445} = 2.43$$

**Step 5.** Make a decision.

The observed value of  $z = 2.43$  falls in the rejection region. Hence, we reject the null hypothesis and conclude that the gasoline additive increases mileage. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**15.28** When would you use the Wilcoxon signed-rank test procedure instead of the paired-samples test of Chapter 10?

**15.29** Explain how the null hypothesis is usually stated in the Wilcoxon signed-rank test.

**15.30** How are ranks assigned to two or more absolute differences that have the same value in the Wilcoxon signed-rank test?

**15.31** Determine the rejection region for the Wilcoxon signed-rank test for each of the following. Indicate whether the rejection region is based on  $T$  or  $z$ .

- a.  $n = 10, H_0: M_A = M_B, H_1: M_A > M_B, \alpha = .05$
- b.  $n = 12, H_0: M_A = M_B, H_1: M_A \neq M_B, \alpha = .01$
- c.  $n = 20, H_0: M_A = M_B, H_1: M_A < M_B, \alpha = .025$
- d.  $n = 30, H_0: M_A = M_B, H_1: M_A > M_B, \alpha = .01$

**15.32** In each case, perform the Wilcoxon signed-rank test.

- a.  $n = 8, T = 5, \text{ left-tailed test using } \alpha = .05$
- b.  $n = 15, T = 20, \text{ right-tailed test using } \alpha = .01$
- c.  $n = 25, T = 51, \text{ two-tailed test using } \alpha = .02$
- d.  $n = 36, T = 238, \text{ left-tailed test using } \alpha = .01$

## ■ APPLICATIONS

**15.33** Refer to Exercise 10.96 of Chapter 10, which deals with Gamma Corporation's installation of governors on its salespersons' cars to regulate their speeds. The following table gives the number of contacts made by each of seven randomly selected sales representatives during the week before governors were installed and the number of contacts made during the week after installation.

Salesperson	A	B	C	D	E	F	G
Before	50	63	42	55	44	65	66
After	49	60	47	51	50	60	58

- a. Using the Wilcoxon signed-rank test at the 5% level of significance, can you conclude that the use of governors tends to reduce the number of contacts made per week by Gamma Corporation's sales representatives?
- b. Compare your conclusions of part a with the result of the hypothesis test that was performed (using the  $t$  distribution) in Exercise 10.96.

**15.34** Refer to Exercise 10.96 of Chapter 10. The following table gives the gas mileage (in miles per gallon) for each of seven randomly selected sales representatives' cars during the week before governors were installed and the gas mileage in the week after installation.

Salesperson	A	B	C	D	E	F	G
Before	25	21	27	23	19	18	20
After	26	24	26	25	24	22	23

- a. Using the Wilcoxon signed-rank test at the 5% level of significance, can you conclude that the use of governors tends to increase the median gas mileage for Gamma Corporation's sales representatives?
- b. Compare your conclusion of part a with the result of the hypothesis test that was performed (using the  $t$  distribution) in Exercise 10.96.

**15.35** Refer to Exercise 15.23. The following table shows the self-confidence test scores of seven employees before and after they attended a course designed to build self-confidence.

Before	8	5	4	9	6	9	5
After	10	8	5	11	6	7	9

- a. Using the Wilcoxon signed-rank test at the 5% level of significance, can you conclude that attending this course increases the median self-confidence test score of employees?
- b. Compare your conclusion of part a with the result of Exercise 15.23.

**15.36** Refer to Exercise 15.25, which compares productivity of 27 bicycle assemblers under an hourly payment system and under a modified-piecework payment scheme. The paired difference for each assembler was calculated by subtracting the number of bicycles assembled during the first week under the new payment system from the number of bicycles assembled during the last week under the hourly wage system. These paired differences are positive for 7 assemblers, negative for 19, and zero for 1 assembler. The sum of the ranks of the positive paired differences is 61, and the sum of the absolute values of the ranks of the negative paired differences is 290.

- Using the Wilcoxon signed-rank test at the 2% level of significance, can you conclude that the median number of bicycles assembled by all such assemblers is the same under both payment systems?
- Compare your conclusion of part a with the result of the sign test performed in Exercise 15.25.

**15.37** Twenty randomly selected adults who describe themselves as “couch potatoes” were given a six-week course in physical fitness. Before starting the course, each adult took a two-mile hike on the same trail. The time required to complete the hike was recorded for each adult. After finishing the course, they all took the same hike again, and their times were recorded again. The following table lists the times recorded (in minutes) before and after the course for each of the 20 adults.

Before	After	Before	After	Before	After
41	37	64	55	100	78
91	71	37	31	48	40
35	30	54	57	50	48
58	64.5	70	59	94	102
45	44	40	33	42.5	40
48.5	44	78	70.5	75	63
84	78	66	56		

Does the fitness course appear to reduce the time required to complete the two-mile hike? Use the Wilcoxon signed-rank test at the 2.5% level of significance.

**15.38** Many adults in the United States have accumulated excessive balances on their credit cards. Ninety such adults were randomly chosen to participate in a group therapy program designed to reduce their debts. Each adult’s total credit card balance was recorded twice: before the program began and three months after the program ended. The paired difference was then calculated for each adult by subtracting the balance after the program ended from the balance before the program began. These paired differences were positive for 49 adults and negative for 41 adults. The sum of the ranks of the positive paired differences was 2507, and the sum of the absolute values of the ranks of the negative paired differences was 1588. Can you conclude that this group therapy program reduces credit card debt? Use the 5% level of significance.

## 15.3 The Wilcoxon Rank Sum Test for Two Independent Samples

In Chapter 10 we discussed tests of hypotheses about the difference between the means of two independent populations using the normal and  $t$  distributions. In this section we compare two independent populations using the results obtained from samples drawn from these populations. In a **Wilcoxon rank sum test**, we assume that the two populations have identical shapes but differ only in location, which is measured by the median. Note that identical shapes do not mean that they have to have a normal distribution. To apply this test, we must be able to rank the given data. Note that the Wilcoxon rank sum test is almost identical to the **Mann-Whitney test**.

In the hypothesis tests discussed in this section, the null hypothesis is usually that the two population distributions are identical. The alternative hypothesis can be that the two population distributions are not identical or that one distribution is to the right of the other or that one distribution is to the left of the other. Assuming that the null hypothesis is true and that the two

populations are identical, we rank all the (combined) data values of the two samples as if they were drawn from the same population. Any tied data values are assigned the ranks in the same manner as in the preceding section. Then we sum the ranks for the data values of each sample separately. If the two populations are identical, the ranks should be spread randomly (and evenly) between the two samples. In this case, the sums of the ranks for the two samples should be almost equal, given that the sizes of the two samples are almost the same. However, if one of the two samples contains mostly lower ranks and the other contains mostly higher ranks, then the sums of the ranks for the two samples will be quite different. The larger the difference in the sums of the ranks of the two samples, the more convincing is the evidence that the two populations are not identical and that the null hypothesis is not true.

In this section, we discuss the Wilcoxon rank sum test first for small samples and then for large samples.

### The Small-Sample Case

If the sizes of both samples are 10 or less, we use the Wilcoxon rank sum test for small samples. Example 15–9 illustrates how the test is performed. To make this test, the population that corresponds to the smaller sample is labeled population 1 and the one that corresponds to the larger sample is called population 2. The respective samples are sample 1 and sample 2. If the sizes of the two samples are equal, either of the two populations can be labeled population 1.

#### ■ EXAMPLE 15–9

A researcher wants to determine whether the distributions of daily crimes in two cities are identical. The following data give the numbers of violent crimes on eight randomly selected days for City A and on nine days for City B.

City A	12	21	16	8	26	13	19	23
City B	18	25	14	16	23	19	28	20

Using the 5% significance level, can you conclude that the distributions of daily crimes in the two cities are different?

**Performing the Wilcoxon rank sum test for two independent populations: small samples.**

**Solution** We apply the following five steps to perform the hypothesis test.

**Step 1. State the null and alternative hypotheses.**

We are to test if the two populations are identical or different. Hence, the two hypotheses are as follows:

$H_0$ : The population distributions of daily crimes in the two cities are identical

$H_1$ : The population distributions of daily crimes in the two cities are different

**Step 2. Select the distribution to use.**

Let the distribution of daily crimes in City A be called population 1 (note that it corresponds to the smaller sample) and that in City B be called population 2. The respective samples are called sample 1 and sample 2. Because  $n_1 < 10$  and  $n_2 < 10$ , we use the Wilcoxon rank sum test for small samples.

**Step 3. Determine the rejection and nonrejection regions.**

The test statistic in Wilcoxon's rank sum test is denoted by  $T$ . The critical value or values of  $T$  are obtained from Table X that appears at the end of this chapter. In this table,  $T_L$  gives the lower critical value and  $T_U$  gives the upper critical value. If the test is two-tailed, we use both  $T_L$  and  $T_U$ . For a left-tailed test we use  $T_L$  only, and for a right-tailed test we use  $T_U$  only.

In our example, the test is two-tailed. Also,  $\alpha = .05$ ,  $n_1 = 8$ , and  $n_2 = 9$ . Hence, from Table X, the values of  $T_L$  and  $T_U$  are 51 and 93, respectively. We will reject the null hypothesis if the observed value of  $T$  is either less than or equal to  $T_L$  or greater than or equal to  $T_U$ . The rejection and nonrejection regions are shown in Figure 15.9. Thus, the decision rule is that we will reject  $H_0$  if either the observed value of  $T \leq 51$  or the observed value of  $T \geq 93$ .

**Figure 15.9**

51 or lower	52 to 92	93 or higher
Rejection region	Nonrejection region	Rejection region

**Step 4.** Calculate the value of the test statistic.

**Table 15.5**

City A		City B	
Crimes	Rank	Crimes	Rank
12	2	18	7
21	11	25	14
16	5.5	14	4
8	1	16	5.5
26	15	23	12.5
13	3	19	8.5
19	8.5	28	16
23	12.5	20	10
		31	17
Sum = 58.5		Sum = 94.5	

To find the observed value of  $T$ , first we rank all the data values of both samples as if they belonged to the same population. Then, we find the sum of the ranks for each sample separately. The observed value of the test statistic  $T$  is given by the sum of the ranks for the smaller sample. If the sizes of the samples are the same, we can use either of the rank sums as the observed value of  $T$ .

In Table 15.5, we rank all the values of both samples and find the sum of the ranks for each sample. Note that 8 is the smallest data value in both samples. Hence, it is assigned a rank of 1. The next smallest value in both samples is 12, which is assigned a rank of 2. The remaining values are assigned ranks in the same way.

Because  $n_1 = 8$  and  $n_2 = 9$ , the sample size for City A is smaller. Hence, the observed value of  $T$  is given by the sum of the ranks for city A. Thus,

$$\text{Observed value of } T = 58.5$$

**Step 5.** Make a decision.

Comparing the observed value of  $T$  with  $T_L$  and  $T_U$  (obtained from Table X in Step 3), we see that the observed value of  $T = 58.5$  is between  $T_L = 51$  and  $T_U = 93$ . Hence, we do not reject  $H_0$ , and we conclude that the two population distributions seem to be identical. ■

Below we describe the Wilcoxon rank sum test procedure for small samples for two-tailed, right-tailed, and left-tailed tests.

### Wilcoxon Rank Sum Test for Small Independent Samples

1. *A two-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the two population distributions are different. The critical values of  $T$ ,  $T_L$ , and  $T_U$  for this test are obtained from Table X for the given significance level and sample sizes. The observed value of  $T$  is given by the sum of the ranks for the smaller sample. The null hypothesis is rejected if  $T \leq T_L$  or  $T \geq T_U$ . Otherwise, the null hypothesis is not rejected.

Note that if the two sample sizes are equal, the observed value of  $T$  is given by the sum of the ranks for either sample.

2. *A right-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the distribution of population 1 (the population that corresponds to the smaller sample) lies to the right of the distribution of population 2. The critical value of  $T$  is given by  $T_U$  in Table X for the given  $\alpha$  for a one-tailed test and the given sample sizes. The observed value of  $T$  is given by the sum of the ranks for the smaller sample. The null hypothesis is rejected if  $T \geq T_U$ . Otherwise, the null hypothesis is not rejected.

Note that if the two sample sizes are equal, the observed value of  $T$  is given by the sum of the ranks for sample 1.

3. *A left-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the distribution of population 1 (the population that corresponds to the smaller sample) lies to the left of the distribution of population 2. The critical value of  $T$  in this case is given by  $T_L$  in Table X for the given  $\alpha$  for a one-tailed test and the given sample sizes. The observed value of  $T$  is given by the sum of the ranks for the smaller sample. The null hypothesis is rejected if  $T \leq T_L$ . Otherwise the null hypothesis is not rejected.

Note that if the two sample sizes are equal, the observed value of  $T$  is given by the sum of the ranks for sample 1.

### The Large-Sample Case

If either  $n_1$  or  $n_2$  or both  $n_1$  and  $n_2$  are greater than 10, we use the normal distribution as an approximation to the Wilcoxon rank sum test for two independent samples.

**Observed Value of  $z$**  In the case of a large sample, the observed value of  $z$  is calculated as

$$z = \frac{T - \mu_T}{\sigma_T}$$

Here, the sampling distribution of the test statistic  $T$  is approximately normal with mean  $\mu_T$  and standard deviation  $\sigma_T$ . The values of  $\mu_T$  and  $\sigma_T$  are calculated as

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Note that in these calculations sample 1 refers to the smaller sample and sample 2 to the larger sample. However, if the two samples are of the same size, we can label either one sample 1. The value of  $T$  used in the calculation of  $z$  is given by the sum of the ranks for sample 1.

The critical value or values of  $z$  are obtained from Table IV in Appendix C for the given significance level. We will reject the null hypothesis if the observed value of  $z$  is in the rejection region. Otherwise, we will not reject  $H_0$ . Example 15–10 illustrates the procedure for performing such a test.

## ■ EXAMPLE 15-10

*Performing the Wilcoxon rank sum test for two independent populations: large samples.*

A researcher wanted to find out whether job-related stress is lower for college and university professors than for physicians. She took random samples of 14 professors and 11 physicians and tested them for job-related stress. The following data give the stress levels for professors and physicians on a scale of 1 to 20, where 1 is the lowest level of stress and 20 is the highest.

Professors	5	9	4	12	6	15	2	8	10	4	6	11	8	3
Physicians	10	18	12	5	13	18	14	9	6	16	11			

Using the 1% significance level, can you conclude that the job-related stress level for professors is lower than that for physicians?

**Solution** Because the smaller sample should be labeled sample 1, the sample of 11 physicians will be called sample 1 and that of 14 professors will be called sample 2. The respective populations are populations 1 and 2. Thus,  $n_1 = 11$  and  $n_2 = 14$ . We perform the five steps of the hypothesis test.

**Step 1. State the null and alternative hypotheses.**

We are to test whether or not professors have lower job-related stress than physicians. Because physicians are labeled population 1 and professors population 2, professors will have a lower stress level if the distribution of population 1 is to the right of the distribution of population 2. Thus, we can state the two hypotheses as follows.

$H_0$ : The two population distributions are identical

$H_1$ : The distribution of population 1 is to the right of the distribution of population 2

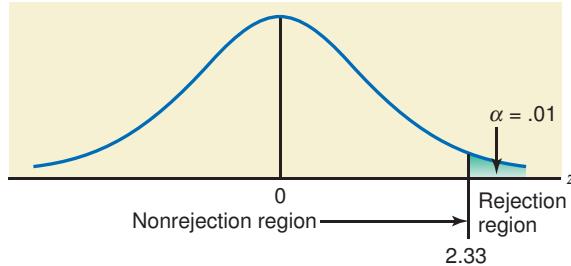
**Step 2. Select the distribution to use.**

Because  $n_1 > 10$  and  $n_2 > 10$ , we use the normal distribution to make this test as the test statistic  $T$  follows an approximately normal distribution.

**Step 3. Determine the rejection and nonrejection regions.**

The test is right-tailed and  $\alpha = .01$ . The area to the left of the critical point under the normal distribution curve is  $1 - .01 = .9900$ . From Table IV in Appendix C, the critical value of  $z$  for .9900 is 2.33. The rejection and nonrejection regions are shown in Figure 15.10. Thus, we will reject  $H_0$  if the observed value of  $z$  is 2.33 or greater. Otherwise, we will not reject  $H_0$ .

Figure 15.10



**Step 4. Calculate the value of the test statistic.**

Table 15.6 shows the rankings of all the data values for the two samples and the sums of these ranks for each sample separately.

**Table 15.6**

Physicians		Professors	
Stress Level	Rank	Stress Level	Rank
10	14.5	5	5.5
18	24.5	9	12.5
12	18.5	4	3.5
5	5.5	12	18.5
13	20	6	8
18	24.5	15	22
14	21	2	1
9	12.5	8	10.5
6	8	10	14.5
16	23	4	3.5
11	16.5	6	8
		11	16.5
		8	10.5
		3	2
Sum = 188.5		Sum = 136.5	

Hence, we calculate the value of the test statistic as follows:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{11(11 + 14 + 1)}{2} = 143$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{11(14)(11 + 14 + 1)}{12}} = 18.26654501$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{188.5 - 143}{18.26654501} = 2.49$$

Thus, the observed value of  $z$  is 2.49. Note that in the calculation of  $z$ , we used the value of  $T$  that belongs to sample 1, which should always be the case.

**Step 5. Make a decision.**

Because the observed value of  $z = 2.49$  is greater than the critical value of  $z = 2.33$ , it falls in the rejection region. Hence, we reject  $H_0$  and conclude that the distribution of population 1 is to the right of the distribution of population 2. Thus, the job-related stress level of physicians is higher than that of professors. This can also be stated as the job-related stress level of professors is lower than that of physicians. ■

Below we describe the Wilcoxon rank sum test procedure for large samples for two-tailed, right-tailed, and left-tailed tests.

**Wilcoxon Rank Sum Test for Large Independent Samples** When  $n_1 > 10$  or  $n_2 > 10$  (or both samples are greater than 10), the distribution of  $T$  (the sum of the ranks of the smaller of the two samples) is approximately normal with mean and standard deviation as follows:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

For two-tailed, right-tailed, and left-tailed tests, first calculate  $T$ ,  $\mu_T$ ,  $\sigma_T$ , and the value of the test statistic,  $z = (T - \mu_T)/\sigma_T$ . If  $n_1 = n_2$ ,  $T$  can be calculated from either sample 1 or sample 2.

1. *A two-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the two population distributions are different. At significance level  $\alpha$ , the critical values of  $z$  are obtained from Table IV in Appendix C. The null hypothesis is rejected if the observed value of  $z$  falls in the rejection region.
2. *A right-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the distribution of population 1 (the population with the smaller sample size) lies to the right of the distribution of population 2. At significance level  $\alpha$ , the critical value of  $z$  is obtained from Table IV in Appendix C. The null hypothesis is rejected if the observed value of  $z$  falls in the rejection region.
3. *A left-tailed test:* The null hypothesis is that the two population distributions are identical, and the alternative hypothesis is that the distribution of population 1 (the population with the smaller sample size) lies to the left of the distribution of population 2. At significance level  $\alpha$ , the critical value of  $z$  is found from Table IV in Appendix C. The null hypothesis is rejected if the observed value of  $z$  falls in the rejection region.

## EXERCISES

### CONCEPTS AND PROCEDURES

**15.39** Explain what determines whether to use the Wilcoxon signed-rank test or the Wilcoxon rank sum test.

**15.40** Find the rejection region for the Wilcoxon rank sum test in each of the following cases.

- $n_1 = 7$ ,  $n_2 = 8$ , right-tailed test using  $\alpha = .05$
- $n_1 = 10$ ,  $n_2 = 10$ , two-tailed test using  $\alpha = .10$
- $n_1 = 18$ ,  $n_2 = 20$ , left-tailed test using  $\alpha = .05$
- $n_1 = 25$ ,  $n_2 = 25$ , two-tailed test using  $\alpha = .01$

**15.41** In each of the following cases, perform the Wilcoxon rank sum test.

- $n_1 = 6$ ,  $n_2 = 7$ ,  $T = 22$ , two-tailed test with  $\alpha = .05$
- $n_1 = 10$ ,  $n_2 = 12$ ,  $T = 137$ , right-tailed test with  $\alpha = .025$
- $n_1 = 9$ ,  $n_2 = 11$ ,  $T = 68$ , left-tailed test with  $\alpha = .05$
- $n_1 = 22$ ,  $n_2 = 23$ ,  $T = 638$ , two-tailed test with  $\alpha = .01$

### APPLICATIONS

**15.42** A consumer agency wants to compare the caffeine content of two brands of coffee. Eight jars of each brand are analyzed, and the amount of caffeine found in each jar is recorded as shown in the table.

Brand I	82	77	85	73	84	79	81	82
Brand II	75	80	76	81	72	74	73	78

Using  $\alpha = .10$ , can you conclude that the two brands have different median caffeine contents per jar?

**15.43** In a Winter Olympics trial for women's speed skating, seven skaters use a new type of skate, while eight others use the traditional type. Each skater is timed (in seconds) in the 500-meter event. The results are given in the following table.

New skates	40.5	40.3	39.5	39.7	40.0	39.9	41.5
Traditional skates	41.0	40.8	40.9	39.8	40.6	40.7	41.1

Assuming that these 15 skaters make up a random sample of all Olympic-class 500-meter female speed skaters, can you conclude that the new skates tend to produce faster times in this event? Use the 5% level of significance.

- 15.44** During April–June 2004, the median price of homes sold in Phoenix was \$252,400, and the median price in Las Vegas was \$255,800. The following table gives the prices (in thousands of dollars) of 9 randomly selected homes in Phoenix and 10 homes in Las Vegas that were sold recently.

Phoenix	258	269	229	279	249	260	242	240	307	
Las Vegas	280	245	319	289	259	268	295	239	262	250

Using the 5% level of significance, can you conclude that the current median price of homes in Phoenix is different from the current median price of homes in Las Vegas?

- 15.45** A factory's management is concerned about the number of defective parts produced by its machinists. Management suspects that production may be improved by giving machinists frequent breaks to reduce fatigue. Twenty-four randomly chosen machinists are randomly divided into two groups (A and B) of 12 each. During the next week all 24 machinists work to manufacture similar parts. The workers in Group A get a five-minute break every hour, whereas the workers in Group B stay on the usual schedule. The number of good parts produced by each machinist during the week is recorded in the following table.

Group A	157	139	188	143	172	144	191	128	177	160	175	162
Group B	160	118	150	165	158	159	127	133	170	164	152	142

At the 1% level of significance, can you conclude that the median number of good parts produced by machinists who take a five-minute break every hour is higher than the median number of good parts produced by the machinists who do not take a break?

- 15.46** Two brands of tires are tested to compare their durability. Eleven Brand X tires and 12 Brand Y tires are tested on a machine that simulates road conditions. The mileages (in thousands of miles for each tire) are shown in the following table.

Brand X	51	55	53	49	50.5	57	54.5	48.5	51.5	52	53.5	
Brand Y	48	47	54	55.5	50	51	46	49.5	52.5	51	49	45

Using the 5% level of significance, can you conclude that the median mileage for Brand X tires is greater than the median mileage for Brand Y tires?

- 15.47** Two Midwestern towns that are 120 miles apart are served by an airline that has been plagued by delays in the past few months. Consequently, many passengers who formerly traveled by air between these towns are taking advantage of a new express bus service. Some statistics students at a local college conducted a survey to see if the bus service between these towns was faster than the air flight. The students took random samples of 15 (one-way) plane trips and 17 (one-way) bus trips between the towns, recording the times for all 32 trips. The time recorded for each trip was measured from the scheduled departure time to the actual arrival time. The sum of the ranks for the 15 plane trips was 295; the sum of the ranks for the 17 bus trips was 233. At the 5% level of significance, can you conclude that the median time for the plane trip is higher than the median time for the bus trip?

## 15.4 The Kruskal-Wallis Test

In Chapter 12 we used the one-way analysis of variance (ANOVA) procedure to test whether or not the means of three or more populations are all equal. To apply the ANOVA procedure using the  $F$  distribution, we assumed that the populations from which the samples were drawn were normally distributed with equal variance,  $\sigma^2$ . However, if the populations being sampled are not normally distributed, then we cannot apply the ANOVA procedure of Chapter 12. In such cases, we can use the **Kruskal-Wallis test**, also called the Kruskal-Wallis  $H$  test. This is a nonparametric test because to use it we do not make any assumptions about the distributions of the populations being sampled. The only assumption we make is that all

populations under consideration have identical shapes but differ only in location, which is measured by the median. Note that identical shapes do not mean that they have to have a normal distribution.

In a Kruskal-Wallis test, the null hypothesis is that the population distributions under consideration are all identical. The alternative hypothesis is that at least one of the population distributions differs and that, therefore, not all of the population distributions are identical. Note that we use the Kruskal-Wallis test to compare three or more populations. Also note that to apply the Kruskal-Wallis test, the size of each sample must be at least five.

**Kruskal-Wallis Test** To perform the Kruskal-Wallis test, we use the chi-square distribution that was discussed in Chapter 11. The test statistic in this test is denoted by  $H$ , which follows (approximately) the chi-square distribution. The critical value of  $H$  is obtained from Table VI in Appendix C for the given level of significance and  $df = k - 1$ , where  $k$  is the number of populations under consideration. *Note that the Kruskal-Wallis test is always right-tailed.*

To find the observed value of the test statistic  $H$ , we first rank the combined data from all samples in the same way as in a Wilcoxon rank sum test. The tied data values are handled the same way as in a Wilcoxon test. Then the observed value of  $H$  is calculated as explained below.

**Observed Value of the Test Statistic  $H$**  The observed value of the test statistic  $H$  is calculated using the following formula:

$$H = \frac{12}{n(n+1)} \left( \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k} \right) - 3(n+1)$$

where

$R_1$  = sum of the ranks for sample 1

$R_2$  = sum of the ranks for sample 2

$\vdots$

$R_k$  = sum of the ranks for sample  $k$

$n_1$  = sample size for sample 1

$n_2$  = sample size for sample 2

$\vdots$

$n_k$  = sample size for sample  $k$

$n = n_1 + n_2 + \cdots + n_k$

$k$  = number of samples

The test statistic  $H$  measures the extent to which the  $k$  samples differ with regard to the ranks assigned to their data values. Basically,  $H$  is a measure of the variance of ranks (or of the variance of the means of ranks) for different samples. If all  $k$  samples have exactly the same mean of ranks,  $H$  will have a value of zero. The value of  $H$  becomes larger as the difference between the means of ranks for different samples increases. Thus, a larger observed value of  $H$  indicates that the distributions of the given populations do not seem to be identical.

Example 15–11 illustrates the procedure for applying the Kruskal-Wallis test.

## ■ EXAMPLE 15-11

A researcher wanted to find out whether the population distributions of salaries of computer programmers are identical in three cities, Boston, San Francisco, and Atlanta. Three different samples—one from each city—produced the following data on the annual salaries (in thousands of dollars) of computer programmers.

*Performing the  
Kruskal-Wallis test.*

Boston	San Francisco	Atlanta
43	54	57
39	33	68
62	58	60
73	38	44
51	43	39
46	55	28
	34	49
		57

Using the 2.5% significance level, can you conclude that the population distributions of salaries for computer programmers in these three cities are all identical?

**Solution** We apply the five steps to perform this hypothesis test.

**Step 1.** *State the null and alternative hypotheses.*

$H_0$ : The population distributions of salaries of computer programmers in the three cities are all identical

$H_1$ : The population distributions of salaries of computer programmers in the three cities are not all identical

Note that the alternative hypothesis states that the population distribution of at least one city is different from those of the other two cities.

**Step 2.** *Select the distribution to use.*

The shapes of the population distributions are unknown. We are comparing three populations. Hence, we apply the Kruskal-Wallis procedure to perform this test, and we use the chi-square distribution.

**Step 3.** *Determine the rejection and nonrejection regions.*

In this example,

$$\alpha = .025 \quad \text{and} \quad df = k - 1 = 3 - 1 = 2$$

Hence, from Table VI in Appendix C, the critical value of  $\chi^2$  is 7.378, as shown in Figure 15.11.

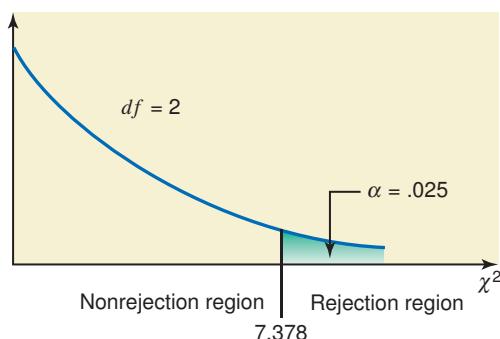


Figure 15.11

**Step 4.** Calculate the value of the test statistic.

To calculate the observed value of the test statistic  $H$ , we first rank the combined data for all three samples and find the sum of ranks for each sample separately. This is done in Table 15.7.

**Table 15.7**

Boston		San Francisco		Atlanta	
Salary	Rank	Salary	Rank	Salary	Rank
43	7.5	54	13	57	15.5
39	5.5	33	2	68	20
62	19	58	17	60	18
73	21	38	4	44	9
51	12	43	7.5	39	5.5
46	10	55	14	28	1
		34	3	49	11
				57	15.5
$n_1 = 6$	$R_1 = 75$	$n_2 = 7$	$R_2 = 60.5$	$n_3 = 8$	$R_3 = 95.5$

We have

$$n = n_1 + n_2 + n_3 = 6 + 7 + 8 = 21$$

and

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left( \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k} \right) - 3(n+1) \\ &= \frac{12}{21(21+1)} \left( \frac{(75)^2}{6} + \frac{(60.5)^2}{7} + \frac{(95.5)^2}{8} \right) - 3(21+1) \\ &= 1.543 \end{aligned}$$

**Step 5.** Make a decision.

Because the observed value of  $H = 1.543$  is less than the critical value of  $H = 7.378$  and it falls in the nonrejection region, we do not reject the null hypothesis. Consequently, we conclude that the population distributions of salaries of computer programmers in the three cities seem to be all identical. ■

**EXERCISES****CONCEPTS AND PROCEDURES**

**15.48** Briefly explain when the Kruskal-Wallis test is used to make a test of hypothesis.

**15.49** What assumption that is required for the ANOVA procedure of Chapter 12 is not necessary for the Kruskal-Wallis test?

**15.50** Describe the form of the null and alternative hypotheses for a Kruskal-Wallis test.

**15.51** Here,  $n_i$  is the size of the  $i$ th sample and  $R_i$  is the sum of ranks for the  $i$ th sample. For each of the following cases, perform the Kruskal-Wallis test using the 5% level of significance.

- a.  $n_1 = 9$ ,  $n_2 = 8$ ,  $n_3 = 5$ ,  $R_1 = 81$ ,  $R_2 = 102$ ,  $R_3 = 70$
- b.  $n_1 = n_2 = n_3 = n_4 = 5$ ,  $R_1 = 27$ ,  $R_2 = 30$ ,  $R_3 = 83$ ,  $R_4 = 70$
- c.  $n_1 = 6$ ,  $n_2 = 10$ ,  $n_3 = 6$ ,  $R_1 = 93$ ,  $R_2 = 70$ ,  $R_3 = 90$
- d.  $n_1 = 8$ ,  $n_2 = 9$ ,  $n_3 = 8$ ,  $n_4 = 10$ ,  $n_5 = 9$ ,  
 $R_1 = 210$ ,  $R_2 = 195$ ,  $R_3 = 178$ ,  $R_4 = 212$ ,  $R_5 = 195$

- 15.52** The following table gives the ranked data for three samples. Perform the Kruskal-Wallis test using the 1% level of significance.

Sample I	Sample II	Sample III
3	14	2
1	11	4.5
10	16	13
7	15	4.5
9	12	8
6		

## ■ APPLICATIONS

- 15.53** Refer to Examples 12–2 and 12–3 of Chapter 12. Fifteen randomly selected fourth-grade students were randomly assigned to three groups, and each group was taught arithmetic by a different method. At the end of the semester, all 15 students took the same arithmetic test. Their test scores are given in the following table.

Method I	Method II	Method III
48	55	84
73	85	68
51	70	95
65	69	74
87	90	67

- a. At the 1% level of significance, can you reject the null hypothesis that the median arithmetic test scores of all fourth-grade students taught by these three methods are all equal?
- b. Compare your answer to part a with the result of the hypothesis test in Example 12–3.

- 15.54** A consumer agency investigated the premiums charged by four auto insurance companies. The agency randomly selected five drivers insured by each company who had similar driving records, autos, and insurance coverages. The following table gives the monthly premiums paid by the 20 drivers.

Company A	Company B	Company C	Company D
\$65	\$48	\$57	\$62
73	69	61	53
54	88	89	45
43	75	77	51
70	72	69	44

Can you reject the null hypothesis that the distributions of auto insurance premiums paid per month by all such drivers are the same for all four companies? Use  $\alpha = .05$ .

- 15.55** Refer to Problem 10 of the Self-Review Test in Chapter 12. A small college town has four pizza parlors that make deliveries. A student doing a research paper for her business management class decides to compare how promptly the four parlors deliver. On six randomly chosen nights, she orders a large pepperoni pizza from each establishment and then records the elapsed time until the pizza is delivered to her apartment. Assume that her apartment is approximately at the same distance from the four pizza parlors. The following table shows the delivery times (in minutes) for these orders.

Tony's	Luigi's	Angelo's	Kowalski's
20.0	22.1	22.3	23.9
24.0	27.0	26.0	24.1
18.3	20.2	24.0	25.8
22.0	32.0	30.1	29.0
20.8	26.0	28.0	25.0
19.0	24.8	25.8	24.2

- Test the null hypothesis that the distributions of delivery times are identical for the four pizza parlors. Use the 5% level of significance.
- Compare your conclusion of part a here with that of part a of Problem 10 of the Self-Review Test in Chapter 12.

**15.56** Refer to Exercise 12.27 of Chapter 12. A resort area has three seafood restaurants, which employ students during the summer season. The local chamber of commerce took a random sample of five servers from each restaurant and recorded the tips they received on a recent Friday night. The results of the survey are shown in the table below. Assume that the Friday night for which the data were collected is typical of all Friday nights of the summer season.

Barzini's	Hwang's	Jack's
\$ 97	\$67	\$ 93
114	85	102
105	92	98
85	78	80
120	90	91

- Would a student seeking a server's job at one of these three restaurants conclude that the population distributions of tips on Friday nights are identical for the three restaurants? Use the 5% level of significance.
- Compare your conclusion of part a with that of part a of Exercise 12.27 of Chapter 12.
- What would your decision be if the probability of making a Type I error were zero in part a? Explain.

**15.57** A factory operates three shifts a day, five days per week, each with the same number of workers and approximately the same level of production. The following table gives the number of defective parts produced during each shift over a period of five days.

First Shift	Second Shift	Third Shift
23	25	33
36	35	44
32	41	50
40	38	52
45	50	60

At the 5% level of significance, can you conclude that the median number of defective parts is the same for all three shifts?

**15.58** A consumer group wanted to compare the service time at three fast-food restaurants, Al's, Eduardo's, and Patel's. Every Tuesday and Wednesday for four weeks, three staff members of the group were randomly assigned to these three restaurants. Each staff member went to his or her assigned restaurant and ordered a hamburger, fries, and a Coke and then recorded the time that elapsed from entering the restaurant until receiving the food. The service times (in minutes) for these eight days for the three restaurants are listed below.

Al's	Eduardo's	Patel's
7.0	3.3	1.1
8.3	11.0	2.4
6.9	5.7	1.8
1.3	8.1	3.0
6.7	6.6	4.1
7.1	13.0	12.0
5.5	2.3	1.5
6.6	5.9	3.1

Assume that these service times make up random samples of all service times at the respective restaurants. At the 10% level of significance, can you conclude that there is a difference in the median service times at these three restaurants?

## 15.5 The Spearman Rho Rank Correlation Coefficient Test

In Chapter 13 we discussed the linear correlation coefficient between two variables  $x$  and  $y$ . We also learned how to make a test of hypothesis about the population correlation coefficient  $\rho$  using the information from a sample. In that chapter we used the  $t$  distribution to perform this test about  $\rho$ . However, using the procedure of Chapter 13 and using the  $t$  distribution to make this test of hypothesis about  $\rho$  require that both variables  $x$  and  $y$  are normally distributed.

The **Spearman rho rank correlation coefficient** (Spearman's rho) is a nonparametric analog of the linear correlation coefficient of Chapter 13. It helps us decide what type of relationship, if any, exists between data from populations with unknown distributions. The Spearman rho rank correlation coefficient is denoted by  $r_s$  for sample data and by  $\rho_s$  for population data. This correlation coefficient is simply the linear correlation coefficient between the ranks of the data on variables  $x$  and  $y$ . To make a test of hypothesis about the Spearman rho rank correlation coefficient, we do not need to make any assumptions about the populations of  $x$  and  $y$  variables.

**Spearman Rho Rank Correlation Coefficient** The Spearman rho rank correlation coefficient is denoted by  $r_s$  for sample data and by  $\rho_s$  for population data. This correlation coefficient is simply the linear correlation coefficient between the ranks of the data. To calculate the value of  $r_s$ , we rank the data for each variable,  $x$  and  $y$ , separately and denote those ranks by  $u$  and  $v$ , respectively. Then we take the difference between each pair of ranks and denote it by  $d$ . Thus,

$$\text{Difference between each pair of ranks} = d = u - v$$

Next, we square each difference  $d$  and add these squared differences to find  $\sum d^2$ . Finally, we calculate the value of  $r_s$  using the formula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

In a test of hypothesis about the Spearman rho rank correlation coefficient  $\rho_s$ , the test statistic is  $r_s$  and its observed value is calculated by using the above formula.

Example 15–12 shows how to calculate the Spearman rho rank correlation coefficient  $r_s$  and how to perform a test of hypothesis about  $\rho_s$ .

### ■ EXAMPLE 15–12

Suppose we want to investigate the relationship between the per capita income (in thousands of dollars) and the infant mortality rate (in percent) for different states. The following table gives data on these two variables for a random sample of eight states.

*Performing the Spearman rho rank correlation coefficient test.*

Per capita income ( $x$ )	29.85	19.0	19.18	31.78	25.22	16.68	23.98	26.33
Infant mortality ( $y$ )	8.3	10.1	10.3	7.1	9.9	11.5	8.7	9.8

Based on these data, can you conclude that there is no significant (linear) correlation between the per capita incomes and the infant mortality rates for all states? Use  $\alpha = .05$ .

**Solution** We perform the five steps to test the null hypothesis that there is no correlation between the two variables against the alternative hypothesis that there is a significant correlation.

**Step 1.** *State the null and alternative hypotheses.*

The null and alternative hypotheses are as follows:

$H_0$ : There is no correlation between per capita incomes and infant mortality rates in all states

$H_1$ : There is a correlation between per capita incomes and infant mortality rates in all states

If we denote the Spearman correlation coefficient by  $\rho_s$ , the null hypothesis and the alternative hypothesis can be written as

$$H_0: \rho_s = 0$$

$$H_1: \rho_s \neq 0$$

Note that this is a two-tailed test.

**Step 2. Select the distribution to use.**

Because the sample is taken from a small population and the variables do not follow a normal distribution, we use the Spearman rho rank correlation coefficient test procedure to make this test.

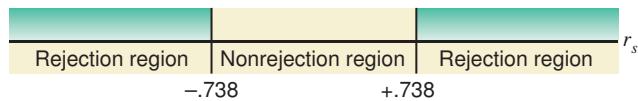
**Step 3. Determine the rejection and nonrejection regions.**

The test statistic that is used to make this test is  $r_s$ , and its critical values are given in Table XI that appears at the end of this chapter. Note that, for this example,

$$n = 8 \quad \text{and} \quad \alpha = .05$$

To read the critical value of  $r_s$  from Table XI, we locate 8 in the column labeled  $n$  and .05 in the top row of the table for a two-tailed test. The critical values of  $r_s$  are  $\pm .738$ , or  $+.738$  and  $-.738$ . Thus, we will reject the null hypothesis if the observed value of  $r_s$  is either  $-.738$  or less, or  $+.738$  or greater. The rejection and nonrejection regions for this example are shown in Figure 15.12.

**Figure 15.12**



**Critical Value of  $r_s$**  The critical value of  $r_s$  is obtained from Table XI for the given sample size and significance level. If the test is two-tailed, we use two critical values, one negative and one positive. However, we use only the negative value of  $r_s$  if the test is left-tailed, and only the positive value of  $r_s$  if the test is right-tailed.

**Step 4. Calculate the value of the test statistic.**

In the Spearman rho rank correlation coefficient test, the test statistic is denoted by  $r_s$ , which is simply the linear correlation coefficient between the ranks of the data. As explained in the beginning of this section, to calculate the observed value of  $r_s$ , we use the formula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where  $d = u - v$ , and  $u$  and  $v$  are the ranks of variables  $x$  and  $y$ , respectively.

**Table 15.8**

$u$	7	2	3	8	5	1	4	6	
$v$	2	6	7	1	5	8	3	4	
$d$	5	-4	-4	7	0	-7	1	2	
$d^2$	25	16	16	49	0	49	1	4	$\Sigma d^2 = 160$

Table 15.8 shows the ranks for  $x$  and  $y$ , which are denoted by  $u$  and  $v$ , respectively. The table also lists the values of  $d$ ,  $d^2$ , and  $\Sigma d^2$ . Note that if two or more values are equal, we use the average of their ranks for all of them. Hence, the observed value of  $r_s$  is

$$r_s = 1 - \frac{6(160)}{8(64 - 1)} = 1 - \frac{960}{504} = -.905$$

Note that Spearman's rho rank correlation coefficient has the same properties as the linear correlation coefficient (discussed in Chapter 13). Thus,  $-1 \leq r_s \leq 1$  or  $-1 \leq \rho_s \leq 1$ , depending on whether sample or population data are used to calculate the Spearman rho rank correlation coefficient. If  $\rho_s = 0$ , there is no relationship between the  $x$  and  $y$  data. If  $0 < \rho_s \leq 1$ , on average a larger value of  $x$  is associated with a larger value of  $y$ . Similarly, if  $-1 \leq \rho_s < 0$ , on average a larger value of  $x$  is associated with a smaller value of  $y$ .

#### Step 5. Make a decision.

Because  $r_s = -.905$  is less than  $-.738$  and it falls in the rejection region, we reject  $H_0$  and conclude that there is a correlation between the per capita incomes and the infant mortality rates in all states. Because the value of  $r_s$  from the sample is negative, we can also state that as per capita income increases, infant mortality tends to decrease. ■

**Decision Rule for the Spearman Rho Rank Correlation Coefficient** The null hypothesis is always  $H_0: \rho_s = 0$ . The observed value of the test statistic is always the value of  $r_s$  computed from the sample data. Let  $\alpha$  denote the significance level, and  $-c$  and  $+c$  be the critical values for the Spearman rho rank correlation coefficient test obtained from Table XI.

- For a two-tailed test, the alternative hypothesis is  $H_1: \rho_s \neq 0$ . If  $\pm c$  are the critical values corresponding to sample size  $n$  and two-tailed  $\alpha$ , we reject  $H_0$  if either  $r_s \leq -c$  or  $r_s \geq +c$ ; that is, reject  $H_0$  if  $r_s$  is "too small" or "too large."
- For a right-tailed test, the alternative hypothesis is  $H_1: \rho_s > 0$ . If  $+c$  is the critical value corresponding to sample size  $n$  and one-sided  $\alpha$ , we reject  $H_0$  if  $r_s \geq +c$ ; that is, reject  $H_0$  if  $r_s$  is "too large."
- For a left-tailed test, the alternative hypothesis is  $H_1: \rho_s < 0$ . If  $-c$  is the critical value corresponding to sample size  $n$  and one-sided  $\alpha$ , we reject  $H_0$  if  $r_s \leq -c$ ; that is, reject  $H_0$  if  $r_s$  is "too small."

## EXERCISES

### CONCEPTS AND PROCEDURES

**15.59** What assumptions that are required for hypothesis tests about the linear correlation coefficient  $\rho$  in Chapter 13 are not required for testing a hypothesis about the Spearman rho rank correlation coefficient?

**15.60** Two sets of paired data on two variables,  $x$  and  $y$ , have been ranked. In each case, the ranks for  $x$  and  $y$  are denoted by  $u$  and  $v$ , respectively, and are shown in the tables. Calculate the Spearman rho rank correlation coefficient for each case.

a.

$u$	2	1	3	4	6	5	7	8
$v$	8	6	7	4	5	2	1	3

b.

$u$	1	2	3	4	5	6	7
$v$	4	2	1	5	3	7	6

**15.61** Calculate the Spearman rho rank correlation coefficient for each of the following data sets.

a.

$x$	5	10	15	20	25	30
$y$	17	15	12	14	10	9

b.

$x$	27	15	32	21	16	40	8
$y$	95	81	102	88	75	120	62

**15.62** Perform the indicated hypothesis test in each of the following cases.

- $n = 9, r_s = .575, H_0: \rho_s = 0, H_1: \rho_s > 0, \alpha = .025$
- $n = 15, r_s = -.575, H_0: \rho_s = 0, H_1: \rho_s < 0, \alpha = .005$
- $n = 20, r_s = .554, H_0: \rho_s = 0, H_1: \rho_s \neq 0, \alpha = .01$
- $n = 20, r_s = .554, H_0: \rho_s = 0, H_1: \rho_s > 0, \alpha = .01$

## ■ APPLICATIONS

**15.63** The following data are a random sample of the heights (in inches) and weights (in pounds) of 10 NBA players selected at random.

Height	84	76	79	79	84	74	83	81	83	75
Weight	240	208	205	215	265	182	225	220	250	190

- Based on the reasonable assumption that as height increases, weight tends to increase, do you expect the value of  $r_s$  to be positive or negative? Why?
- Compute the value of  $r_s$ . Does it agree with your expectation of its value in part a?

**15.64** Let  $\rho_s$  be the Spearman rho rank correlation coefficient between heights (in inches) and weights (in pounds) for the entire population of NBA players listed in Data Set II. Using the value of  $r_s$  computed from the sample data in Exercise 15.63, test the null hypothesis  $H_0: \rho_s = 0$  against the alternative hypothesis  $H_1: \rho_s > 0$  at the significance level of  $\alpha = .01$ .

**15.65** In Example 13–1 of Chapter 13, we estimated the regression line for the data given in Table 13.2 on food expenditures (in hundred dollars) and incomes (in hundred dollars). Those data are reproduced here.

Income ( $x$ )	35	49	21	39	15	28	25
Food expenditure ( $y$ )	9	15	7	11	5	8	9

The estimated regression line in Example 13–1 had a slope of .2642.

- Do you expect the Spearman rho rank correlation coefficient for these data to be positive or negative? Why?
- Compute  $r_s$  for these data. Did it come out as expected?

**15.66** In Example 13–7 of Chapter 13, the null hypothesis  $H_0: \rho = 0$  was tested against the alternative hypothesis  $H_1: \rho > 0$ , where  $\rho$  is the linear correlation coefficient for the population. At a significance level of  $\alpha = .01$ ,  $H_0: \rho = 0$  was rejected there. Hence, it seemed that in fact  $\rho > 0$ .

- If  $\rho_s$  is the Spearman rho rank correlation coefficient for the entire population for food expenditure and income data, and you test  $H_0: \rho_s = 0$  against  $H_1: \rho_s > 0$  at the significance level of  $\alpha = .01$ , based on the results of Example 13–7, do you expect to reject or accept  $H_0$ ? Why?
- Perform the hypothesis test stated in part a.

**15.67** The following table shows the combined (math and verbal) SAT scores (denoted by  $x$ ) and college grade point averages (denoted by  $y$ ) on completion of a bachelor's degree for nine randomly chosen recent college graduates who had taken the SAT test.

$x$	1105	990	1040	1215	1405	975	1300	1010	1080
$y$	3.33	2.62	3.05	3.60	3.85	2.43	3.90	2.40	2.95

- Find the Spearman rho rank correlation coefficient for this data set.
- Test  $H_0: \rho_s = 0$  against  $H_1: \rho_s > 0$  using the 5% level of significance.
- Does your test result indicate a positive relationship between the variables  $x$  and  $y$ ?

**15.68** A day-care center operator is concerned about the aggressive behavior of young boys left in her care. She feels that prolonged watching of television tends to promote aggressive behavior. She selects seven boys at random and ranks them according to the level of aggressiveness in their behavior, with a rank of 7 indicating most aggressive and a rank of 1 denoting least aggressive. Then she asks each boy's parent(s) to estimate the average number of hours per week the boy spends watching television. The following table shows the data collected on the aggressiveness ranks of these boys and the number of hours spent watching TV per week.

Aggressiveness rank	1	2	3	4	5	6	7
Weekly TV hours	15	21	28	8	24	32	20

- Calculate  $r_s$  for these data.
- At the 5% level of significance, can you conclude that there is a positive relationship between aggressiveness and hours spent watching television?

## 15.6 The Runs Test for Randomness

The **runs test for randomness** tests the null hypothesis that a sequence of events has occurred randomly against the alternative hypothesis that this sequence of events has not occurred randomly.

### The Small-Sample Case

As an example of a runs test, in apartment complexes families with children are often assigned to units close to one another to lessen the impact of noise on childless tenants. We would like to determine whether a landlord has randomly assigned units to tenants irrespective of whether they have children, or has tried to cluster tenants with children. Here what we really mean by *random* is independence in the sense that if, say, a childless tenant lives in unit 3, this provides no additional information about whether the tenants in units 2 and 4 are childless. With non-randomness, we would have some additional information.

Suppose that a part of the apartment complex consists of a single building with 10 adjacent units numbered 1 to 10. We specify the family status of the 10 tenants via a string of 10 letters, using C for “has children” and D for “no children.” One possible string is D D C D C C D D C C, which would mean that the tenants in units 3, 5, 6, 9, and 10 have children and those in 1, 2, 4, 7, and 8 do not. (Note that in our example the numbers of tenants with and without children are equal, but they do not have to be equal.)

Two extreme arrangements that may not be random are C C C C C D D D D D and C D C D C D C D. In the first arrangement, all the tenants with children are adjacent, as are those without children. In the second case, tenants with and without children alternate perfectly. Note that in these two examples exactly half of the tenants have children and half do not.

A characteristic of a string of the letters C and D that helps determine the randomness of the string is called a **run**. A run is a sequence of the same symbol (in this case the letter C or D) appearing one or more times. The arrangement C C C C C D D D D D has two runs; the arrangement C D C D C D C D C D has 10 runs. Intuitively, we know that in the string with 2 runs the arrangement is not random because there are too few runs, whereas in the string with 10 runs the arrangement is not random due to too many runs.

If the number of tenants with children, the number without children, and their arrangement in the 10 units are all random, then the number of runs in the arrangement, which we denote by  $R$ , will also be random. Thus,  $R$  is a statistic with its own sampling distribution. Table XII (that appears at the end of this chapter) gives the critical values of  $R$  for a significance level of 5%—that is, for  $\alpha = .05$ . There are two parameters associated with the distribution of  $R$ ,  $n_1$  and  $n_2$ . Here  $n_1$  is the number of times the first symbol (in our example the letter C) appears in the string, and  $n_2$  is the number of times the second symbol (in our example the letter D) appears. Table XII provides critical values of  $R$  for values of  $n_1$  and  $n_2$  up to 15. If either  $n_1 > 15$  or  $n_2 > 15$ , we can apply the normal approximation (discussed later in the section on the large-sample case) to perform the test. For each pair of  $n_1$  and  $n_2$ , there are two critical values: a smaller value (denoted by  $c_1$ ) and a larger value (denoted by  $c_2$ ).

### Definition

**Run** A *run* is a sequence of one or more consecutive occurrences of the same outcome in a sequence of occurrences in which there are only two outcomes. The number of runs in a sequence is denoted by  $R$ . The value of  $R$  obtained for a sequence of outcomes for a sample gives the observed value of the test statistic for the runs test for randomness.

Suppose we formally set up the following hypotheses:

$H_0$ : Tenants with and without children are randomly mixed among the 10 units

$H_1$ : These tenants are not randomly mixed

We will reject  $H_0$  if either of the following occurs:

$$R \leq c_1 \text{ (too few runs)} \quad \text{or} \quad R \geq c_2 \text{ (too many runs)}$$

Let's apply these rules to the hypothetical strings listed earlier to determine whether or not to reject  $H_0$  at a significance level of  $\alpha = .05$ .

1. Let the string of letters be C C C C C D D D D. Here  $n_1 = 5$ ,  $n_2 = 5$ , and  $R = 2$ . From Table XII,  $c_1 = 2$  and  $c_2 = 10$ . Since  $R \leq c_1$ , we reject  $H_0$  on the basis that there are too few runs.
2. Let the string of letters be C D C D C D C D C D. Here  $n_1 = 5$ ,  $n_2 = 5$ , and  $R = 10$ . From Table XII,  $c_1 = 2$  and  $c_2 = 10$ . Since  $R \geq c_2$ , we reject  $H_0$  on the basis that there are too many runs.
3. Let the string of letters be D D C D C C D D C C. Here  $n_1 = 5$ ,  $n_2 = 5$ , and  $R = 6$ . From Table XII,  $c_1 = 2$  and  $c_2 = 10$ . Because the value of  $R = 6$  is between  $c_1 = 2$  and  $c_2 = 10$ , we do not reject  $H_0$ .

Example 15–13 illustrates the application of the runs test for randomness.

### ■ EXAMPLE 15–13

*Performing the runs test for randomness: small sample.*

A college admissions office is interested in knowing whether applications for admission arrive randomly with respect to gender. The genders of 25 consecutively arriving applications were found to arrive in the following order (here M denotes a male applicant and F a female applicant).

M F M M F F F M F M M M F F F F M M M F F M F M M

Can you conclude that the applications for admission arrive randomly with respect to gender? Use  $\alpha = .05$ .

**Solution** We perform the following five steps in this hypothesis test.

**Step 1.** State the null and alternative hypotheses.

$H_0$ : Applications arrive in a random order with respect to gender

$H_1$ : Applications do not arrive in a random order with respect to gender

**Step 2.** Select the distribution to use.

Let  $n_1$  and  $n_2$  be the number of male and female applicants, respectively. Then

$$n_1 = 13 \quad \text{and} \quad n_2 = 12$$

Because both  $n_1$  and  $n_2$  are less than 15, we use the runs test to check for randomness.

**Step 3.** Determine the rejection and nonrejection regions.

For  $n_1 = 13$ ,  $n_2 = 12$ , and  $\alpha = .05$ , the critical values from Table XII are  $c_1 = 8$  and  $c_2 = 19$ . Thus, we will not reject the null hypothesis if the observed value of  $R$  is in the interval 9 to 18. We will reject the null hypothesis if the observed value of  $R$  is either 8 or lower, or 19 or higher. The rejection and nonrejection regions are shown in Figure 15.13.

**Figure 15.13**

8 or lower	9 to 18	19 or higher
Rejection region	Nonrejection region	Rejection region

**Step 4.** Calculate the value of the test statistic.

As the given data show, the 25 applications included in the sample were received in the following order with respect to gender:

M F M M F F F M F M M M F F F F M M M M F F M F M M

Because this string of the letters M and F has 13 runs,

Observed value of  $R = 13$

**Step 5.** Make a decision.

Because  $R = 13$  is between 9 and 18, we do not reject  $H_0$ . Hence, we conclude that the applications for admission arrive in a random order with respect to gender. ■

## The Large-Sample Case

If either  $n_1 > 15$  or  $n_2 > 15$ , the sample is considered large for the purpose of applying the runs test for randomness and we use the normal distribution to perform the test.

**Observed Value of  $z$**  For large values of  $n_1$  and  $n_2$ , the distribution of  $R$  (the number of runs in the sample) is approximately normal with its mean and standard deviation given as

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

The observed value of  $z$  for  $R$  is calculated using the formula

$$z = \frac{R - \mu_R}{\sigma_R}$$

In this case, rather than using Table XII to find the critical values of  $R$ , we use the standard normal distribution table (Table IV in Appendix C) to find the critical values of  $z$  for the given significance level. Then, we make a decision to reject or not to reject the null hypothesis based on whether the observed value of  $z$  falls in the rejection or the nonrejection region. Example 15–14 describes the application of this procedure.

### ■ EXAMPLE 15–14

Refer to Example 15–13. Suppose that the admissions officer examines 50 consecutive applications and observes that  $n_1 = 22$ ,  $n_2 = 28$ , and  $R = 20$ , where  $n_1$  is the number of male applicants,  $n_2$  the number of female applicants, and  $R$  the number of runs. Can we conclude that the applications for admission arrive randomly with respect to gender? Use  $\alpha = .01$ .

*Performing the runs test for randomness: large sample.*

**Solution** We perform the following five steps to make this test.

**Step 1.** State the null and alternative hypotheses.

$H_0$ : Applications arrive in a random order with respect to gender

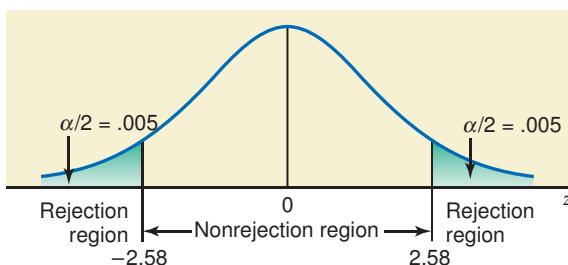
$H_1$ : Applications do not arrive in a random order with respect to gender

**Step 2.** Select the distribution to use.

Here,  $n_1 = 22$  and  $n_2 = 28$ . Since  $n_1$  and  $n_2$  are both greater than 15, we use the normal distribution to make the runs test. Note that only one of  $n_1$  and  $n_2$  has to be greater than 15 to apply the normal distribution.

**Step 3.** Determine the rejection and nonrejection regions.

The significance level is .01 and the test is two-tailed. From Table IV in Appendix C, the critical values of  $z$  for .005 and .9950 areas to the left are  $-2.58$  and  $2.58$ , respectively. The rejection and nonrejection regions are shown in Figure 15.14.

**Figure 15.14****Step 4.** Calculate the value of the test statistic.

To find the observed value of  $z$ , we first find the mean and standard deviation of  $R$  as follows:

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{(2)(22)(28)}{22 + 28} + 1 = 25.64$$

$$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(22)(28)(2 \cdot 22 \cdot 28 - 22 - 28)}{(22 + 28)^2 (22 + 28 - 1)}} = 3.44783162$$

The observed value of the test statistic  $z$  is

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{20 - 25.64}{3.44783162} = -1.64$$

Note that the value of  $R$  is given in the example to be 20.

**Step 5.** Make a decision.

Since  $z = -1.64$  is between  $-2.58$  and  $2.58$ , we do not reject  $H_0$ , and we conclude that the applications for admission arrive in a random order with respect to gender. ■

## EXERCISES

### CONCEPTS AND PROCEDURES

**15.69** Briefly explain the term *run* as used in a runs test.

**15.70** What is the usual form of the null hypothesis in a runs test for randomness?

**15.71** Under what conditions may we use the normal distribution to perform a runs test?

**15.72** Using the runs test for randomness, indicate whether the null hypothesis should be rejected in each of the following cases.

- a.  $n_1 = 10$ ,  $n_2 = 12$ ,  $R = 17$ ,  $\alpha = .05$
- b.  $n_1 = 20$ ,  $n_2 = 23$ ,  $R = 35$ ,  $\alpha = .01$
- c.  $n_1 = 15$ ,  $n_2 = 17$ ,  $R = 7$ ,  $\alpha = .05$
- d.  $n_1 = 14$ ,  $n_2 = 13$ ,  $R = 21$ ,  $\alpha = .05$

**15.73** In Example 15–13, if we use the symbol 0 for male and 1 for female instead of M and F, would this affect the test in any way? Why or why not?

**15.74** For each of the following sequences of observations, determine the values of  $n_1$ ,  $n_2$ , and  $R$ .

- a. X X Y X Y Y X Y X Y X X X Y Y
- b. F M F F F F M M F F F F F F F
- c. + + + - - - - + + - + - + + + + +
- d. 1 1 0 0 0 0 1 1 0 0 1 1 1 1

## ■ APPLICATIONS

- 15.75** A psychic claims that she can cause a nonrandom sequence of heads (H) and tails (T) to appear when a coin is tossed a number of times. A fair coin was tossed 20 times in her presence, and the following sequence of heads and tails was obtained.

H H T H T H T H T H H T T H H H T T H

Can you conclude that the psychic's claim is true? Use  $\alpha = .05$ .

- 15.76** At a small soda factory, the amount of soda put into each 12-ounce bottle by the bottling machine varies slightly. The plant manager suspects that the machine has a nonrandom pattern of overfilling and underfilling the bottles. The following are the results of filling 18 bottles, where O denotes 12 ounces or more of soda in a bottle and U denotes less than 12 ounces of soda.

U U U O O O O U U O O O O U O U U U U

Using the runs test at the 5% significance level, can you conclude that there is a nonrandom pattern of overfilling and underfilling such bottles?

- 15.77** An experimental planting of a new variety of pear trees consists of a single row of 20 trees. Several of these trees were affected by an unknown disease. The order of diseased and normal trees is shown below, where D denotes a diseased and N denotes a normal tree.

N N N D D D N N N N N N D D D D D N N N N

If the sequence of diseased and normal trees falls into a nonrandom pattern with clusters of diseased trees, that would suggest that the disease may be contagious. Perform the runs test at the 5% significance level to determine if there is evidence of a nonrandom pattern in the sequence.

- 15.78** A fourth-grade teacher asks her class orally one by one whether "potato" or "potatOE" is the correct spelling for the common vegetable. She suspects the children may engage in copycat behavior, in which an incorrect spelling by one student is likely to be followed by an incorrect spelling by the next student. If this theory is true, there should be fewer runs than expected. The responses of the 22 students in the class are as follows, where a correct answer is denoted by C and an incorrect answer by I.

C C C I I I C C C C C I I C C C C C I C C C C

At the 5% significance level, test the null hypothesis that the correct (and incorrect) responses are randomly distributed in the population of all fourth-graders against the alternative hypothesis that they are not randomly distributed in the population. Assume that this class is a random sample of all fourth-graders.

- 15.79** Do baseball players' hits come in streaks? Seventy-five consecutive at-bats of a baseball player were recorded to determine whether the hits were randomly distributed or nonrandomly distributed for him, thus possibly indicating the presence of "hitting streaks." The observation of these 75 at-bats produced the following data.

$$n_1 = \text{number of hits} = 22, \quad n_2 = \text{number of nonhits} = 53, \quad R = \text{number of runs} = 37$$

Can you conclude that hits occur randomly for this player? Use  $\alpha = .01$ .

- 15.80** A researcher wanted to determine whether the stock market moves up and down randomly. He recorded the movement of the Dow Jones Industrial Average for 40 consecutive business days. The observed data showed that the market moved up 16 times, moved down 24 times, and had 11 runs during these 40 days. Using a significance level of 5%, do you think that the movements in the stock market are random?

- 15.81** Many state lotteries offer a "daily number" game, in which a three-digit number is randomly drawn every day. Suppose that a certain state generates its daily number by a computer program, and the state lottery commissioner suspects that the process is flawed. Specifically, she feels that if today's lucky number is higher than 500, tomorrow's number is more likely to exceed 500; and if today's number is below 500, tomorrow's number is more likely to be under 500. Suppose that a sequence of the state's lucky numbers for a period of consecutive days is analyzed for runs above 500 and runs below 500. If the commissioner is right, there should be fewer runs than expected by chance. Analysis of a sequence for 50 consecutive days produced the following information.

$$n_1 = 27 \text{ numbers above } 500 \quad n_2 = 23 \text{ numbers below } 500 \quad R = \text{number of runs} = 11$$

Using the 2.5% level of significance, can you conclude that the sequence of all this state's daily numbers for this game is nonrandom?

## USES AND MISUSES...

### 1. I'M FREE!

Imagine that you turn on the news one morning. Still a bit sleepy, you hear a report: "And this just in from the state government: all road signs have been removed. That's right, folks. No more traffic lights, no more 'no parking' signs, no more speed limits! And now, the weather..." You would rightfully be scared to death. During your drive to school or work, you would be much more vigilant than the day before. You would approach an intersection more slowly, you would have to rely on published maps given the absence of exit signs, and you would keep your eye out for very fast and very dangerous drivers. Quite simply, you would not be able to rely on assumptions regarding the rules of the road.

Assumptions can be good. This text has emphasized that the parametric methods (including hypothesis testing, chi-square tests, *F*-tests, and linear regression models), require that your samples, populations, and errors be normally distributed. Because of the central limit theorem, the assumption that population data and large samples taken from them are normally distributed is typically pretty good. Standard statistical computer packages are able to compare your samples to a normal distribution and should be used to do so. The assumption that a statistical sample comes from a normal distribution is equivalent to augmenting your data set. In the example above, the assumption is roughly equivalent to the rules of the road: You do not have to check on the driver next to you all the time because you can assume that he or she will behave in a certain way.

However, for every case in which the assumption of normality applies, there is at least one in which it does not. Without the distribution parameters of the mean and variance, you are going to need to gather more data to make up for the fact that you do not have these parameters, or accept a larger confidence interval than you would have preferred. You might even find that as you collect more data, they resemble a normal distribution after all. Additionally, the nonparametric tests have their own applications and assumptions, and keeping them straight can be tricky: The Kruskal-Wallis test investigates if the distributions under investigation are identical but tells us no more.

Fortunately, the nonparametric methods do not abandon the rules of the road. Each nonparametric method is based on properties of probability distributions and the assumption of random sampling of the populations under investigation. Each has explicit null

and alternative hypotheses, and proper application requires a detailed specification. Similarly, if you were to drive your car down the road on that dangerous morning, you would probably find that everyone was still driving on the right side of the road.

### 2. MORE POWER TO YOU!

Chapters 8 through 10, and 12 through 14 discussed a variety of (parametric) procedures related to statistical inference and each of these procedures requires that certain conditions hold true in order to produce valid results. The nonparametric methods presented in Chapter 11 and 15 also require a set of conditions that must hold true, but these conditions are much less restrictive than the ones in other chapters. If the results of a parametric test are valid, then the results from an equivalent nonparametric test will also be valid. However, the inverse of this relationship is not true. Having said that, many people wonder why not just forget about the stricter conditions and instead use a nonparametric test each time? That is a good question, and it has a good answer.

As you may recall from Chapter 9, every hypothesis test has the potential to produce an incorrect result. The possible incorrect results were referred to as Type I and Type II errors in that chapter. You may remember from Chapter 9 that the significance level (denoted by  $\alpha$ ) gives the probability of a Type I error. In practice, the statistician specifies the significance level that is used in a hypothesis test. The probability of a Type II error (denoted by  $\beta$ ) in a hypothesis test problem is a function of the test being used, the sample size, and the significance level. The power of a hypothesis test is defined to be 1 minus the probability of a Type II error, i.e.,  $1 - \beta$ . So we try to use a test that has the highest power of the (hypothesis) test given a specific significance level.

Parametric tests, such as the single-sample *t* test for the mean and a one-way ANOVA test, have higher powers of the test than their nonparametric compatriots. Therefore, if it is reasonable to conclude that the stricter conditions of a parametric test have been met, using a parametric test will result in a higher power of the test. However, if it is not reasonable to assume that the conditions of a parametric test have been met, then using a parametric test will not produce valid results. So while the power of the test will be lower when using a nonparametric test, you can rely on the fact that the designated significance level will be met.

## Glossary

**Categorical data** Data divided into different categories for identification purposes only.

**Distribution-free test** A hypothesis test in which no assumptions are made about the specific population distribution from which the sample is selected.

**H** The test statistic used in the Kruskal-Wallis test.

**Kruskal-Wallis test** A distribution-free method used to test the hypothesis that three or more populations have identical distributions.

**Nonparametric test** A hypothesis test in which the sample data are not assumed to come from a specific type of population distribution, such as the normal distribution.

**$\rho_s$**  The value of the Spearman rho rank correlation coefficient between the ranks of the values of two variables for population data.

**$r_s$**  The value of the Spearman rho rank correlation coefficient between the ranks of the values of two variables for sample data.

**R** The number of runs in a runs test used to test for randomness.

**Run** A sequence of one or more consecutive occurrences of the same outcome in a sequence of occurrences in which there are only two possible outcomes.

**Runs test for randomness** A test that is used to test the null hypothesis that a sequence of events has occurred randomly.

**Sign test** A nonparametric test that is used to test a population proportion (with categorical data), a population median (with numerical data), or the difference in population medians for two dependent and paired sets of numerical data.

**Spearman rho rank correlation coefficient** The linear correlation coefficient between the ranks of paired data for two samples or populations.

**T** The test statistic used in the Wilcoxon signed-rank test and Wilcoxon rank sum test.

**$T_U, T_L$**  The upper and lower critical values for the Wilcoxon rank sum test obtained from the table.

**Wilcoxon rank sum test** A nonparametric test that is used to test whether two independent samples come from identically distributed populations by analyzing the ranks of the pooled sample data.

**Wilcoxon signed-rank test** A nonparametric test that is used to test whether two paired and dependent samples come from identically distributed populations by analyzing the ranks of the paired differences of the samples.

## Supplementary Exercises

**15.82** Fifteen cola drinkers are given two paper cups, one containing Brand A cola and the other containing Brand B cola. Each person tries both drinks and then indicates which one he or she prefers. The drinks are offered in random order (some people are given Brand A first, and others get Brand B first). Ten of the people prefer Brand A, while five prefer Brand B. Using the sign test at the .05 significance level, can you conclude that among all cola drinkers there is a preference for Brand A?

**15.83** Twenty-four randomly selected people are given samples of two brands of low-fat ice cream. Seventeen of them prefer Brand B, and 7 prefer Brand A. Using the sign test at the .05 significance level, can you conclude that among all people there is a preference for Brand B?

**15.84** Four hundred randomly selected football fans were asked whether they prefer watching college football or professional football. Of these fans, 220 said they prefer the professional games, 168 prefer college games, and 12 have no preference. At the 2% level of significance, can you conclude that among all football fans there is a preference for either professional or college football?

**15.85** A random sample of 200 customers of a large bank are asked whether they prefer using an automatic teller machine (ATM) or seeing a human teller for deposits and withdrawals. Of these customers, 122 said they prefer an ATM, 66 prefer a teller, and 12 have no opinion. At the 1% level of significance, can you conclude that more than half of all customers of this bank prefer an ATM?

**15.86** Suppose that a polling agency is conducting a telephone survey. When prospective participants answer the phone, they are told that the survey will take just five minutes of their time. Ten randomly selected calls are monitored. The lengths of time (in minutes) required for the survey in these 10 cases are shown here.

7.1      6.3      4.9      5.0      5.7      9.0      8.2      5.9      6.5      7.7

Using the sign test at the 5% level of significance, can you conclude that the median time for the survey exceeds 5 minutes?

**15.87** In 2001, the median age of buyers of Harley-Davidson motorcycles was 45 years (*USA TODAY*, June 7, 2002). Suppose that a random sample of 25 persons who bought Harley-Davidson motorcycles recently showed that 16 of them were over 45 years of age, 7 were under 45, and 2 were 45 years old. At the 5% level of significance, can you conclude that the current median age of Harley-Davidson buyers is over 45 years?

**15.88** A state motor vehicle department requires auto owners to bring their autos to state emission centers periodically for testing. State officials claim that the median waiting time between the hours of 8 A.M. and 11 A.M. on weekdays at a particular site is 25 minutes. In a check of 30 randomly selected motorists during this time period at this site, 9 motorists waited for less than 25 minutes, 2 waited exactly 25 minutes, and 19 waited longer than 25 minutes.

- Using the sign test at the 5% significance level, can you conclude that the median waiting time at this site during these hours exceeds 25 minutes?
- Perform the test of part a at the 2.5% level of significance.
- Comment on the results of parts a and b.

**15.89** The following data give the amounts (in dollars) spent on textbooks by 35 college students during the 2005–2006 academic year.

475	418	680	610	655	488	710	375	250
695	420	610	380	98	530	415	757	357
409	611	455	618	395	612	468	610	780
450	880	490	490	626	850	688	588	

Using  $\alpha = .05$ , can you conclude that the median expenditure on textbooks by all such students in 2005–2006 was different from \$650?

**15.90** Two brothers, Bob and Morris, who work the same hours at the same company in a large city, share an apartment on the outskirts of the city. When weather permits, Bob rides his bicycle to work, but Morris always drives. Although they always leave for work at exactly the same time each morning, Morris often arrives later than Bob because of the heavy traffic. Last year, on 21 randomly selected days of good weather, Bob arrived at work first 16 times, and Morris was first on 5 days. At the 5% level of significance, can you conclude that the median morning commuting time for Bob is less than that for Morris?

**15.91** Refer to Exercise 15.34 and to Exercise 10.96 of Chapter 10, which concern Gamma Corporation's installation of governors on its salespersons' cars to regulate their speeds. The following table gives the gas mileage (in miles per gallon) for each of seven sales representatives' cars during the week before governors were installed and the gas mileage in the week after installation.

Salesperson	A	B	C	D	E	F	G
Before	25	21	27	23	19	18	20
After	26	24	26	25	24	22	23

- a. Using the sign test at the 5% level of significance, can you conclude that the use of governors tends to increase the median gas mileage for Gamma Corporation's sales representatives' cars?
- b. Compare your conclusion of part a with the result of the Wilcoxon signed-rank test that was performed in part a of Exercise 15.34 and with the result of the corresponding hypothesis test (using the  $t$  distribution) of Exercise 10.96.
- c. If there is a difference in the three conclusions, how can you account for it?

**15.92** A reporter for a travel magazine wanted to compare the effectiveness of two large travel agencies (X and Y) in finding the lowest airfares to given destinations. She randomly chose 32 destinations from the many offered by both agencies. She and her assistants requested the lowest available fare for each destination from each agency. For 18 of these destinations Agency X quoted a fare lower than that of Agency Y, for 8 destinations Agency Y found a lower fare, and in 6 cases the fares were the same. At the 2% level of significance, can you conclude that there is any difference in the median fares quoted by Agency X and Agency Y for all destinations they both offer?

**15.93** Thirty-five patients with high blood pressures are given medication to lower their blood pressures. For all 35 patients, their blood pressures are measured before they begin the medication and again after they have finished taking medication for 30 days. For 25 patients the blood pressures were lower after finishing the medication, in 7 cases they were higher, and for 3 patients there was no change. Assume that these 35 patients make up a random sample of all people suffering from high blood pressures. At the 2.5% level of significance, can you conclude that the median blood pressure in all such patients is lower after the medication than before?

**15.94** The following table shows the one-week sales of six salespersons before and after they attended a course on "how to be a successful salesperson."

Before	12	18	25	9	14	16
After	18	24	24	14	19	20

- a. Using the Wilcoxon signed-rank test at the 5% significance level, can you conclude that the weekly sales for all salespersons tend to increase as a result of attending this course?
- b. Perform the test of part a using the sign test at the 5% significance level.
- c. Compare your conclusions from parts a and b.

**15.95** An official at a figure skating competition thinks that two of the judges tend to score skaters differently. Shown next are the two judges' scores for eight skaters.

Skater	1	2	3	4	5	6	7	8
Judge A	5.8	5.7	5.6	5.9	5.8	5.9	5.8	5.6
Judge B	5.4	5.5	5.7	5.4	5.6	5.3	5.4	5.6

Using the Wilcoxon signed-rank test at the 5% level of significance, can you conclude that either judge tends to give higher median scores than the other?

**15.96** Refer to Exercise 15.26. Consider the data given in that exercise on the cholesterol levels (in milligrams per hundred milliliters) for 30 randomly selected adults as determined by two laboratories, A and B.

- Using the Wilcoxon signed-rank test at the 1% level of significance, can you conclude that the median cholesterol level for all such adults as determined by Lab A is higher than that determined by Lab B?
- Compare your conclusion in part a to that of Exercise 15.26.

**15.97** A consumer agency conducts a fuel economy test on two new subcompact cars, the Mouse (M) and the Road Runner (R). Each of 18 randomly selected drivers takes both cars on an 80-mile test run. For each driver, the gas mileage (in miles per gallon) is recorded for both cars; then the gas mileage for the R car is subtracted from the gas mileage for the M car. Thus, a minus difference indicates a higher gas mileage for the R car. One of the 18 drivers obtains exactly the same gas mileage for both cars. For the other drivers, the differences are ranked. The sum of the positive ranks is 31, and the sum of the absolute values of the negative ranks is 122. Can you conclude that the R car gets better gas mileage than the M car? Use the 2.5% level of significance.

**15.98** Each of the two supermarkets, Al's and Bart's, in River City claims to offer lower-cost shopping. Fifty people who normally do the grocery shopping for their families are chosen at random. Each shopper makes up a list for a week's supply of groceries. Then these items are priced and the total cost is computed for each store. The paired differences are then calculated for each of the 50 shoppers, where a paired difference is defined as the cost of a cart of groceries at Al's minus the cost of the same cart of groceries at Bart's. These paired differences were positive for 21 shoppers and negative for 29 shoppers. The sum of ranks of the positive paired differences was 527 and the sum of the absolute values of the ranks of the negative paired differences was 748. Using the 1% level of significance, can you conclude that either store is less expensive than the other?

**15.99** A consumer advocate is comparing the prices of eggs at supermarkets in the suburbs with the prices of eggs at supermarkets in the cities. The following data give the prices (in dollars) of a dozen large eggs in 13 supermarkets, 6 of which are in cities and 7 are in suburbs.

City	1.49	1.29	1.35	1.58	1.33	1.47
Suburb	.99	1.09	1.39	1.28	1.16	1.44

Using the .05 level of significance and the Wilcoxon rank sum test, can you conclude that egg prices tend to be higher in the city?

**15.100** Many VCR owners have difficulty learning to program the VCR to record TV programs. A consumer magazine tested two new VCRs, Brands X and Y, which are claimed to be user-friendly by their manufacturers. A random sample of 13 adults (6 for Brand X, 7 for Brand Y) are observed to see how quickly they can learn to program the VCRs properly. The following table gives the times (in minutes).

Brand X	32	36	28	43	98	39
Brand Y	33	18	21	25	24	27

Using  $\alpha = .05$  and the Wilcoxon rank sum test, can you conclude that learning times tend to be longer for Brand X?

**15.101** A researcher obtains a random sample of 24 students taking elementary statistics at a large university and divides them randomly into two groups. Group A receives instruction to use Software A to do a statistics assignment, whereas Group B is taught to use Software B to do the same statistics assignment. The time (in minutes) taken by each student to complete this assignment is given in the table.

Group A	123	101	112	85	87	133	129	114	150	110	180	115
Group B	65	115	95	100	94	72	60	110	99	102	88	97

- Using the 5% level of significance and the Wilcoxon rank sum test, can you conclude that the median time required for all students taking elementary statistics at this university to complete this assignment is longer for Software A than for Software B?
- Would a paired-samples sign test be appropriate here? Why or why not?

**15.102** Refer to Exercise 15.101. The scores on the homework assignment for the 24 students are given in the table.

Group A	48	38	45	31	42	25	40	43	50	30	33	46
Group B	37	21	40	27	49	44	36	41	20	39	18	40

Using the 10% significance level and the Wilcoxon rank sum test, can you conclude that there is a difference in the median scores for all students using Software A and all students using Software B?

**15.103** Manufacturers of luxury cars are very much interested in knowing the age distribution of their customers because then they can change these models to attract younger buyers without losing the older customers who have traditionally favored such cars. According to data from CNW Marketing Research, the median ages of drivers (primary drivers using vehicles for personal use only) of Rolls-Royce, Mercedes, and Cadillac automobiles were 62.9, 58.7, and 53.4 years, respectively, at the time of the survey (*USA TODAY*, February 17, 2005). The following table gives the ages of seven randomly selected primary drivers of each of these three makes of cars.

Rolls-Royce	Mercedes	Cadillac
64	61	52
61	47	63
70	66	39
68	71	55
55	44	50
64	53	47
68	58	61

At the 5% level of significance, can you reject the null hypothesis that the median age of drivers for each of these three makes of cars is the same?

**15.104** An academic employment service compared the starting salaries of May 2005 graduates in three major fields. Random samples were taken of 8 engineering majors, 10 business majors, and 7 mathematics majors. The starting salaries of all 25 graduates were determined and then ranked, yielding the following rank sums:

Engineering: 137      Business: 126      Mathematics: 62

At the 5% level of significance, can you reject the null hypothesis that the median starting salary is the same for May 2005 graduates in these three fields?

**15.105** A sports magazine conducted a test of three brands (A, B, and C) of golf balls by having a professional golfer drive six balls of each brand. The lengths of the drives (in yards) from this test are listed in the table.

Brand A	Brand B	Brand C
275	245	267
266	256	283
301	261	259
281	270	250
288	259	263
277	262	256

At the 5% level of significance, can you reject the null hypothesis that the median distance of drives by this golfer is the same for all three brands of golf balls?

**15.106** A students' group at a state university wanted to compare textbook costs for students majoring in economics, history, and psychology. The group obtained data from random samples of 10 economics majors, 9 history majors, and 11 psychology majors, all in the second semester of their junior year. The total textbook costs of the 30 students were recorded and ranked. The rank sums for economics and history majors were 134 and 157, respectively.

- a. Find the rank sum for psychology majors. [Hint: The sum of  $n$  integers from 1 through  $n$  is given by  $n(n + 1)/2$ .]

- b.** At the 2.5% significance level, can you reject the null hypothesis that the median textbook costs are the same for students in all three majors who are in the second semester of the junior year?

**15.107** The following table shows the average verbal SAT score and the percentage of high school graduates who took the SAT in 2002 for a random sample of 10 states.

State	Average Verbal SAT Score	Percentage of Graduates Taking SAT	
Connecticut	509		83
Georgia	489		65
Illinois	578		11
Kentucky	550		12
Michigan	558		11
New Jersey	498		82
South Carolina	488		59
South Dakota	576		5
Vermont	512		69
Wisconsin	583		7

Source: The College Board, *The World Almanac and Book of Facts*, 2003.

- a.** For all 50 states, would you expect  $\rho_s$  to be positive, negative, or near zero? Why?  
**b.** Calculate  $r_s$  for the sample of 10 states and indicate whether its value is consistent with your answer to part a.  
**c.** Using the value of  $r_s$  calculated in part b, test  $H_0: \rho_s = 0$  against  $H_1: \rho_s \neq 0$  using the 5% level of significance.

**15.108** The Spearman rho rank correlation coefficient may be used in cases where data for one or both variables are given in the form of ranks. Suppose that a film critic views 10 randomly chosen new movies and ranks them, with a rank of 10 being assigned to the film that he thinks will have the highest box office receipts, a rank of 9 to the next most profitable, and so forth. Three months after each film is released, its total box office receipts (in millions of dollars) are tabulated. The following table shows the ranking and receipts for each of these 10 films.

Rank	7	3	10	1	4	5	2	6	8	9
Receipts	40	5	66	2	3	10	28	15	30	17

- a.** Calculate  $r_s$  for the sample of 10 films.  
**b.** Using the 5% level of significance, test  $H_0: \rho_s = 0$  against  $H_1: \rho_s > 0$ .  
**c.** Based on your conclusion in part b, is there sufficient evidence that this critic can predict the box office performance of a film?

**15.109** Refer to Example 13–8, which contained data on monthly auto insurance premiums and years of driving experiences. Those data are reproduced here.

Driving Experience (years)	Monthly Auto Insurance Premium
5	\$64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

The estimated regression line was found to be  $\hat{y} = 76.6605 - 1.5476x$  in Example 13–8 and the simple linear correlation coefficient for the sample data was  $- .77$ . The true regression slope  $B$  was found to be significantly less than zero at the significance level of 5%.

- Based on this information, if  $\rho_s$  is the Spearman rho rank correlation coefficient for the entire population from which this sample was taken, what would you expect from a test of  $H_0: \rho_s = 0$  against  $H_1: \rho_s < 0$  at the significance level of 5%?
- Perform the hypothesis test mentioned in part a.

**15.110** A researcher wonders if men still tend to stand aside and let women board elevators ahead of them. She observes 10 men and 10 women boarding the same elevator. The order in which they boarded is given here.

W W W M M W W M W W W M M M M W W M M M

Using the 5% level of significance, can you conclude that the order of boarding is nonrandom with respect to gender?

**15.111** A machinist is making precision cutting tools. Because of the exacting specifications for these tools, about 20% of them fail to pass inspection and are judged defective. The shop supervisor feels that the machinist tends to produce defective tools in clusters, perhaps due to fatigue or distraction. If this is true, then a sequence of tools produced by this machinist will tend to have fewer runs of defective and good tools than expected by chance. The supervisor chooses a day at random and observes the following sequence of 18 tools, where G denotes a tool that passed inspection and D indicates a defective tool.

G G G G G D D G G G G D G G G D D G

Do you think that there is evidence of nonrandomness in this sequence? Use the 5% level of significance.

**15.112** Some states require periodic testing of cars to monitor the emission of pollutants. A state official suspects that the inspection process at a particular station is faulty, that a car's test result may be affected by the tests of preceding cars. Analysis of the sequence of test results for a random sample chosen on a day yields the following information:

$$n_1 = \text{number of cars that passed the test} = 157$$

$$n_2 = \text{number of cars that failed the test} = 143$$

$$R = \text{number of runs} = 41$$

Using the 1% level of significance, can you conclude that the test results for this emissions test station are not random?

**15.113** The following data give the sequence of wins and losses in 30 consecutive games for a baseball team during a season.

L W L W L W L L W W W L L W L L L L W L L L L W L W L L

Can you conclude that the wins are randomly distributed for this baseball team? Use the 2% significance level.

### Advanced Exercises

**15.114** A medical researcher wants to study the effects of a low-calorie diet on the longevity of laboratory mice. She randomly divides 20 mice into two groups. Group A gets a standard diet, while Group B receives a diet that contains all the necessary nutrients but provides only 70% as many calories as Group A's diet. The experiment is conducted for 36 months and the length of life (in days) of each mouse is recorded. The data obtained on the lives of these mice are shown in the following table. In these data, the asterisk (\*) indicates that this mouse was still alive at the end of the 36-month experiment.

Group A	900	907	751	833	920	787	850	877	848	901
Group B	1037	905	1023	988	1078	1011	*	1063	898	1033

- Using the 2.5% level of significance and the Wilcoxon rank sum test, does the low-calorie diet appear to lengthen the longevity of laboratory mice? Should the mouse that was still alive at the end of the experiment be eliminated from your analysis or is there a way to include it?
- Would a Wilcoxon signed-rank test be appropriate in this example? Why or why not?

**15.115** The editor of an automotive magazine has asked you to compare the median gas mileages for driving in the city for three models of compact cars. The editor has made available to you one car of each of the three models, three drivers, and a budget sufficient to buy gas and pay the drivers for approximately 500 miles of city driving for each car.

- Explain how you would conduct an experiment and gather the data for a magazine article comparing gas mileages.
- Suppose you wish to test the null hypothesis that the median gas mileages for driving in the city are the same for all three models of cars. Outline the procedure for using your data to conduct this test. Do not assume that the gas mileages for all cars of each model are normally distributed.

**15.116** Refer to Exercise 10.96 in Chapter 10. Suppose Gamma Corporation decides to test the governors on seven cars. However, management is afraid that the speed limit imposed by the governors will reduce the number of contacts the salespersons can make each day. Thus, both the fuel consumption and the number of contacts made are recorded for each car/salesperson for each week of the testing period, both before and after the installation of governors.

Salesperson	Number of Contacts		Fuel Consumption (mpg)	
	Before	After	Before	After
A	50	49	25	26
B	63	60	21	24
C	42	47	27	26
D	55	51	23	25
E	44	50	19	24
F	65	60	18	22
G	66	58	20	23

Suppose that you are directed to prepare a brief report that includes statistical analysis and interpretation of the data. Management will use your report to help decide whether or not to install governors on all salespersons' cars. Use 5% significance levels for any hypothesis tests you perform to make suggestions. In contrast to Exercise 10.96, do not assume that the numbers of contacts, fuel consumption, or differences are normally distributed.

**15.117** Suppose that you are a newspaper reporter and your editor has asked you to compare the hourly wages of carpenters, plumbers, electricians, and masons in your city. Because many of these workers are not union members, the wages may vary considerably among individuals in the same trade.

- What data should you gather to make this statistical analysis and how would you collect them? What sample statistics would you present in your article and how would you calculate them? Assume that your newspaper is not intended for technical readers.
- Suppose that you must submit your findings to a technical journal that does require statistical analysis of your data. If you want to determine whether or not the median hourly wages are the same for all four trades, briefly describe how you would analyze the data. Do not assume that the hourly wages for these populations are normally distributed.

**15.118** Consider the data in the following table.

x	10	20	30	40	50	60
y	12	15	19	21	25	30

- Suppose that each value of y in the table is increased by 5 but the x values remain unchanged. What effect will this have on the rank of each value of y? Do you expect the value of  $r_s$  to increase, decrease, or remain the same? Explain why.
- Now, first calculate the value of  $r_s$  for the data in the table, and then increase each value of y by 5 and recalculate the value of  $r_s$ . Does the value of  $r_s$  increase, decrease, or remain the same? Does this result agree with your expectation in part a?

**15.119** The English department at a college has hired a new instructor to teach the composition course to first-year students. The department head is concerned that the new instructor's grading practices might not

be consistent with those of the professor (let us call him Professor A) who taught this course previously. She randomly selects 10 essays written by students for this class and makes two copies of each essay. She asks Professor A and this instructor (working independently) to assign a numerical grade to each of the 10 essays. The results are shown in the following table.

Essay	1	2	3	4	5	6	7	8	9	10
Professor A	75	62	90	48	67	82	94	76	78	84
Instructor	80	50	85	55	63	78	89	81	75	83

- a. Suppose the department head wants to determine whether the instructor tends to grade higher or lower than Professor A. Which of the statistical tests discussed in this chapter could she use? Note that more than one test may be appropriate.
- b. Using an appropriate test from your answer in part a, can you conclude that the instructor tends to grade higher or lower than Professor A? Use  $\alpha = .05$ .
- c. Suppose the department head wants to determine whether the instructor is consistent with Professor A in the sense that they tend to agree on which paper is best, which is second best, and so forth. Which test from this chapter would be appropriate to use? State the relevant null and alternative hypotheses.
- d. Using the test you chose in part c, can you conclude that Professor A and the instructor are consistent in their grading? Use the 5% level of significance.

**15.120** Three doctors are employed at a large clinic. The manager at the clinic wants to know whether these three doctors spend the same amount of time per patient. The manager randomly chooses 10 routine appointments of patients with each of the three doctors and times them. Thus, the data set consists of 10 observations on the time spent with patients by each doctor.

- a. To test the null hypothesis that the mean or median times are equal for all three doctors against the alternative hypothesis that they are not all equal, which tests from Chapters 12 and 15 are appropriate?
- b. For each test that you indicated in part a, specify whether the test is about the means or the medians.
- c. What assumptions are required for the test from Chapter 12?

**15.121** An educational researcher is studying the relationship between high school grade point averages (GPAs) and SAT scores. She obtains GPAs and SAT scores for a random sample of 25 students and wants to test the null hypothesis that there is no correlation between GPAs and SAT scores against the alternative hypothesis that these variables are positively correlated.

- a. If she wants to base her test on the linear correlation coefficient of Chapter 13, what assumptions are required about the two variables (GPAs and SAT scores)?
- b. If the assumptions required in part a are not satisfied, what other test(s) might she use?

**15.122** To test the effectiveness of a new six-week body-building course, 12 tenth-grade boys are randomly selected. Each boy is tested before and after the course to see how much weight he can lift.

- a. To test whether or not the mean or median weight lifted by all such boys tends to be greater after the course than before, which tests from Chapters 10 and 15 might be used?
- b. For each test that you indicated in part a, specify whether it involves the mean or median.
- c. If the paired differences in weights lifted before and after the test are not normally distributed, which of the tests indicated in part a could be used?

**15.123** Suppose in a sample we have 10 A's and 15 B's. What is the maximum number of runs possible in a sequence of these 25 letters?

**15.124** Refer to Exercises 12.27 and 15.56. In these two problems you were asked to perform an ANOVA and a Kruskal-Wallis test, respectively, on the data. In both cases, the results were significant at the 5% significance level. Change the values of the tips in those data so that the  $p$ -value for the Kruskal-Wallis test remains the same, but the ANOVA results are no longer significant at the 5% level. (Hint: When making the changes, the ranks of the fifteen data points should not change.)

**15.125** A student who typically does not do his homework was asked to toss a coin 20 times and write down the sequence of results. Instead of tossing the coin, the student simply wrote down the following sequence (reading from left to right) of hypothetical outcomes.

H	T	H	T	H	T	H	H	T	H
T	T	H	T	H	T	H	T	H	T

Use the appropriate test to show that the professor was justified in accusing the student of not actually tossing the coin.

## Self-Review Test

1. Nonparametric tests
  - a. are more efficient than the corresponding parametric tests
  - b. do not require that the population being sampled has a normal distribution
  - c. generally require more assumptions about the population than parametric tests do
2. For small samples ( $n \leq 25$ ), the critical value(s) for the sign test are based on the
  - a. binomial distribution
  - b. normal distribution
  - c.  $t$  distribution
3. Which of the following tests may be used to test hypotheses about one median?
  - a. The sign test
  - b. The Kruskal-Wallis test
  - c. The Wilcoxon rank sum test
4. The Wilcoxon signed-rank test may be used to test
  - a. for a difference between the medians of two independent samples
  - b. for a preference for one product over another
  - c. hypotheses involving paired samples
5. When we use the Wilcoxon signed-rank test,
  - a. all observations are ranked
  - b. the difference for each pair is calculated and then all the differences are ranked according to their absolute values
  - c. only the signs of the differences are used to calculate the value of the test statistic
6. In order to perform a Wilcoxon rank sum test, one must calculate the
  - a. standard deviation of each sample
  - b. range of the data
  - c. rank of each observation
7. Which of the following tests may be used with paired samples? Circle all that apply.
  - a. Sign test
  - b. Wilcoxon signed-rank test
  - c. Wilcoxon rank sum test
  - d. Spearman rho rank correlation coefficient test
8. The Spearman rho rank correlation coefficient is calculated as the
  - a. simple linear correlation coefficient between the two sets of observations
  - b. simple linear correlation coefficient between the ranks of the two sets of observations
  - c. square of the simple linear correlation coefficient between two sets of observations
9. In order to test a hypothesis about the Spearman rho rank correlation coefficient
  - a. both sets of data must come from normally distributed populations
  - b. one set of data must come from a normally distributed population
  - c. either set of data can have any distribution
10. The Spearman rho rank correlation coefficient is positive when
  - a. there is no relationship between the two sets of observations
  - b. the values in one set of observations increase as the values of the corresponding observations in the other set decrease
  - c. the values in one set of observations increase as the values of the corresponding observations in the other set increase
11. For the runs test for randomness, which of the following statements are true?
  - a. We notice which of the two possible outcomes has occurred at each stage in a list of consecutive outcomes.
  - b. A run is one or more consecutive occurrences of either one of the two possible outcomes.
  - c. We are testing the hypothesis that one of the two possible outcomes occurred significantly more frequently than the other.
12. In the runs test for randomness, we reject the null hypothesis
  - a. only when there is a very large number of runs
  - b. only when there is a very small number of runs
  - c. if there is either a very large or a very small number of runs
13. In the runs test for randomness, the distribution of  $R$  (the total number of runs) is approximately normal when
  - a.  $R$  is greater than 10
  - b. at least one of the two possible outcomes occurs more than 15 times
  - c. each of the two possible outcomes occurs more than 15 times

**14.** A large pool of prospective jurors is made up of an equal number of men and women. A 12-person jury selected from this pool consists of 2 women and 10 men. At the 5% level of significance, can we reject the null hypothesis that the selection process is unbiased in terms of gender?

**15.** A September 2002 *USA TODAY/Gallup* poll asked Americans whether they favored a proposal to put Social Security payroll taxes into personal retirement accounts. Fifty-two percent of the respondents were in favor of the proposal (*USA TODAY*, September 25, 2002). Suppose that the 2002 poll consisted of 1000 respondents, so that 520 favored the proposal. Using the 2.5% level of significance, can you conclude that more than half of all Americans favor putting Social Security payroll taxes into personal retirement accounts?

**16.** The past records of a supermarket show that its customers spent a median of \$65 per visit. After a promotional campaign designed to increase spending, the store took a sample of 12 customers and recorded the amounts (in dollars) they spent. The amounts are listed here.

88      69      141      28      106      45      32      51      78      54      110      83

Using  $\alpha = .05$ , can you conclude that the median amount spent by all customers at this store after the campaign exceeds \$65?

**17.** According to a U.S. Census Bureau survey of households, women living alone had a median income of \$20,264 in 2001 (*USA TODAY*, September 25, 2002). Suppose that in a recent random sample of 400 women living alone, 229 had incomes under \$20,264 and 171 had incomes over \$20,264. At the 1% level of significance, can you conclude that the median income of women living alone currently is different from \$20,264?

**18.** The following table gives the cholesterol levels for seven adults before and after they completed a special dietary plan.

Before	210	180	195	220	231	199	224
After	193	186	186	223	220	183	233

- a.** Using the sign test at the 5% significance level, can you conclude that the median cholesterol levels are the same before and after the diet?
- b.** Using the Wilcoxon signed-rank test at the 5% significance level, can you conclude that the median cholesterol levels are the same before and after the diet?
- c.** Compare your conclusions for parts a and b.

**19.** An archeologist wants to compare two methods (I and II) of radioactive dating of artifacts. He submits a random sample of 33 artifacts that are suitable for radioactive dating. Each one of these artifacts is dated by both methods. The paired differences are then calculated for each of the 33 artifacts, where a paired difference is defined as the age of an artifact dated by Method I minus the age of the same artifact dated by Method II. These paired differences were positive for 11 of the artifacts, negative for 20, and zero for 2 artifacts. Using the sign test at the 2% level of significance, can you conclude that the median estimated ages of such artifacts differ for the two methods?

**20.** A professor at a large university suspects that the grades of engineering majors tend to be lower in the spring semester than in the fall semester. He randomly selects 10 sophomore electrical engineering majors and records their grade point averages (GPAs) for the fall and the spring semesters. The data obtained are shown in the table.

Student	1	2	3	4	5	6	7	8	9	10
Fall GPA	3.20	3.56	3.05	3.78	4.00	2.85	3.33	2.67	3.00	3.67
Spring GPA	3.15	3.40	2.88	3.67	4.00	3.00	3.30	3.05	2.95	3.50

Using the Wilcoxon signed-rank test at the 5% level of significance, can you conclude that the median GPA of all sophomore electrical engineering majors at this university tends to be lower in the spring semester than in the fall semester?

**21.** A random sample of 30 students was selected to test the effectiveness of a course designed to improve memory. Each student was given a memory test before and after taking the course. Each student's score after taking the course was subtracted from his or her score before the course; then the 30 differences were ranked. Thus, a negative rank denotes an improved score after taking the course. The sum of the positive ranks was 102; the sum of the absolute values of the negative ranks was 276. Three students scored exactly the same on both tests. Using the 2.5% level of significance, can you conclude that the course tends to improve scores on memory tests?

- 22.** A commuter has two alternative routes (Route 1 and Route 2) to drive to work. Picking days at random, she drives to work using each route for eight days and records the time (in minutes) taken to commute from home to work on each day. These times are shown in the following table.

Route 1	45	43	38	56	41	43	46	44
Route 2	38	40	39	42	50	37	46	36

Using the Wilcoxon rank sum test at the 5% level of significance, can you reject the null hypothesis that the median commuting time is the same for both routes?

- 23.** An accounting firm has hired two temporary employees, A and B, to prepare individual federal income tax returns during the tax season. Clients who have relatively simple tax situations are randomly assigned to either A or B. The firm randomly selected 18 income tax returns prepared by each of these two employees and recorded the times taken to prepare these tax returns. After these times taken to prepare 36 tax returns were ranked, the sum of the ranks for A was found to be 298 and the sum of the ranks for B was equal to 368. Using the Wilcoxon rank sum test at the 2.5% level of significance, can you conclude that there is a difference in the median times taken to prepare such income tax returns by A and B?

- 24.** The following table lists the numbers of cases of telemarketing fraud reported to law-enforcement officials during several randomly chosen weeks in 2002 for three large cities of approximately equal populations.

City A	City B	City C
53	29	75
46	35	49
59	44	62
33	31	68
60	50	52
	48	

- a. At the 2.5% level of significance, can you reject the null hypothesis that the distributions of the numbers of such reported cases are identical for all three cities?
- b. Can you reject the null hypothesis of part a at the 1% level of significance?
- c. Comment on the results of parts a and b.

- 25.** The following is a list of home runs (denoted by  $x$ ) and runs batted in (denoted by  $y$ ) as of July 1, 2005, by 10 players selected at random from a minor league baseball team.

Player	1	2	3	4	5	6	7	8	9	10
$x$	10	7	13	2	8	4	16	11	5	4
$y$	49	38	54	20	41	27	62	40	22	19

- a. As home runs increase, runs batted in tend to increase. From this, do you expect the value of the Spearman rho rank correlation coefficient to be positive or negative?
- b. Compute  $r_s$  for the data.
- c. Suppose  $\rho_s$  is the value of the Spearman rho rank correlation coefficient for all players in this league. Using the 2.5% significance level, test the null hypothesis  $H_0: \rho_s = 0$  against the alternative hypothesis  $H_1: \rho_s > 0$ .

- 26.** Ramon fishes in a lake where the minimum size for bass to be kept is 12 inches long; all smaller bass must be returned to the water. He thinks that most of the “keepers” (bass 12 inches or longer) are caught early in the morning. If he is right, there should be a few long runs of keepers caught in the early morning followed by a few long runs of smaller bass caught later on during the day. Thus, if the fish are recorded in sequence, there should be fewer runs than expected by chance. Last Saturday Ramon fished from 6 A.M. to 11 A.M. and caught 14 bass in the following order, where K denotes a keeper and S denotes a bass shorter than 12 inches.

K    K    K    K    S    K    K    S    K    S    S    S    S

Using the 5% significance level, does this sequence support Ramon’s theory?

- 27.** As of June 1, 2005, a minor league baseball team had played 54 games, winning 30 and losing 24. In these 54 consecutive games, there were 15 runs (in the statistical sense). Using the runs test for randomness, can we conclude that the 30 wins and 24 losses are randomly spread out among the 54 games? Use a significance level of 5%.

## Mini-Projects

### MINI-PROJECT 15-1

For a period of 30 business days, record the daily price of crude oil and the price of a stock that you think might be affected by the oil price (for example, an oil company or an alternative energy company).

- Compute the Spearman rho rank correlation coefficient for these data.
- Can you conclude that there is a relationship between the two sets of prices? Use  $\alpha = .05$ .

### MINI-PROJECT 15-2

For a period of 30 business days, record whether the Dow Jones Industrial Average moves up or down. Use your data to perform an appropriate test at the 1% level of significance to see if the sequence of upward and downward movements of the Dow appears to be random over this period. (As an alternative, you might use the NASDAQ index, the price of an individual stock, or the price of gold.)

### MINI-PROJECT 15-3

On December 8, 2005, the U.S. national debt was approximately \$8.13 trillion. Take random samples of 10 or more students from each of three different majors and ask each student to estimate the size of the national debt.

- At the 5% level of significance and using the Kruskal-Wallis test, can you conclude that the median perceived values of the national debt are the same for all three majors?
- Find the current value of the national debt. On the whole, did the students in your sample tend to overestimate its value, or did they tend to underestimate?

## DECIDE FOR YOURSELF

### Using Nonparametric Methods

Prior to this chapter, you used inferential methods to work with quantitative data that are classified as scale data or with categorical data that are classified as nominal data.<sup>2</sup> A third type of data, called ordinal data, often requires nonparametric methods for analysis. Ordinal data are data that can be ranked. For example, insurance companies will classify policy holders by their age groups, as opposed to their specific ages, for assessing risk. Since nobody can be in two different age groups at the same time, the data identifying age groups can be ranked.

Suppose you are interested in studying the relationship between the successes of Division I men's and women's basketball teams at colleges that have both of these sports. We can use Spearman's Rho to calculate the rank correlation between the power index rankings (RPIs) of the men's and women's teams. We have provided three scatterplots here. Specifically we have the scatterplot of the ranks of the men's teams versus the ranks of the women's teams for various colleges (Figure 15.15), the scatterplot of the RPIs of men's teams versus the RPIs of women's teams (Figure 15.16), and the scatterplot of the RPIs of men's teams versus their ranks (Figure 15.17). These data are for all Division I programs that had men's and women's basketball programs during the 2004–2005 season.

<sup>2</sup>Based on what are called the scales or levels of measurement, data can be classified into four scales or levels—nominal, ordinal, interval, and ratio scales. Data that can be divided into different categories only for identification purposes are said to have a nominal scale. An example of such data is the names given to different makes of cars, such as Town Car, Toyota Camry, and so forth. Data that can be divided into different categories so that categories can be ranked are said to have an ordinal scale. An example

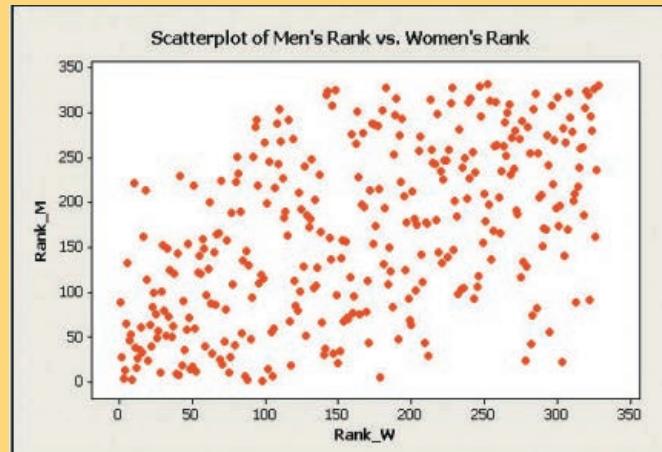
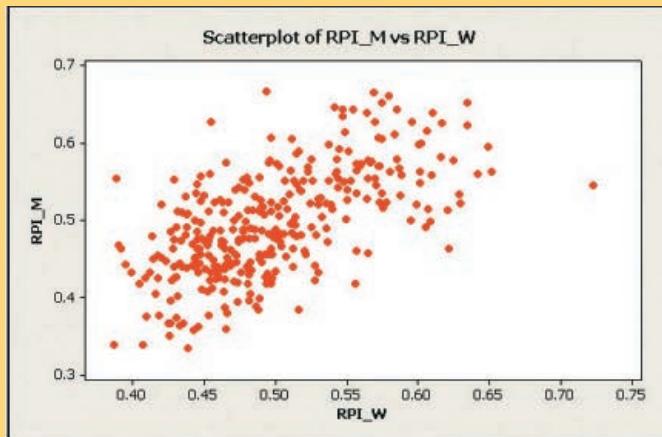


Figure 15.15

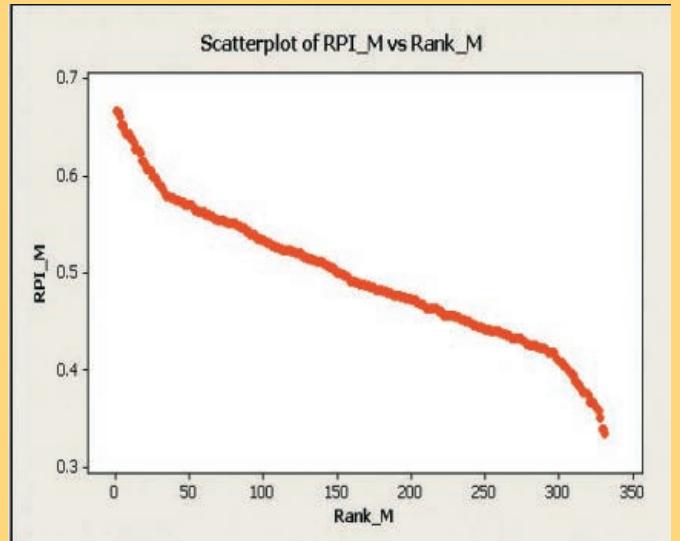
Answer the following questions.

- Which of the three relationships should have its strength measured by Spearman's Rho rank correlation instead of the Pearson correlation coefficient?

of such data is an evaluation of a product as excellent, good, or poor. Data that can be ranked and for which the difference between any two values can be calculated (and is meaningful) are said to have an interval scale. An example of such data is temperatures in two cities. Data that can be ranked and for which all arithmetic operations (such as addition, subtraction, multiplication and division) can be done are said to have ratio scale. An example of such data is gross sales of two companies.

**Figure 15.16**

2. Which of the three relationships is inappropriate to be discussed in terms of its correlation?
3. There is one point that stands out on the scatterplot of RPI values (Figure 15.16). Identify the location of this point on the scatterplot of ranks (Figure 15.15).

**Figure 15.17**

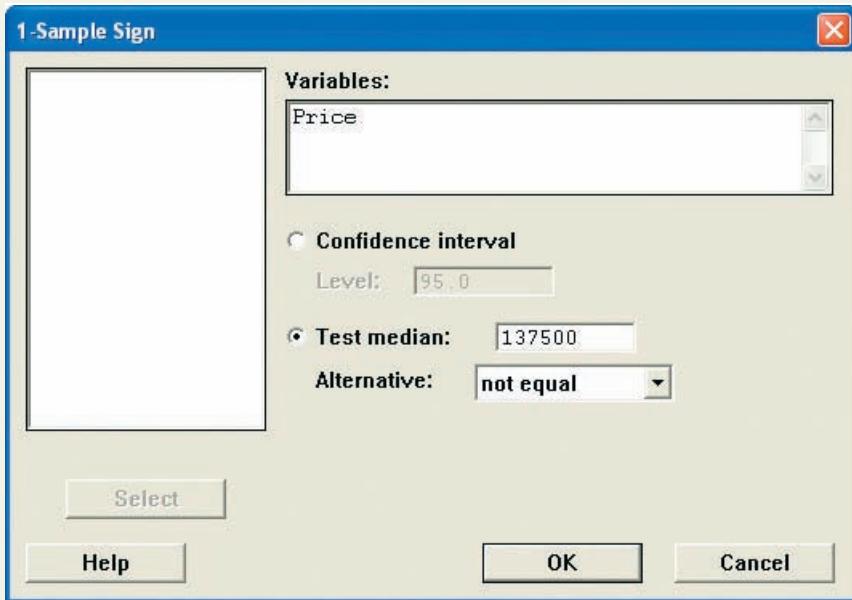
## TECHNOLOGY INSTRUCTION

### Nonparametric Methods

#### TI-84

1. The TI-84 does not contain any built-in nonparametric methods.

#### MINITAB

**Screen 15.1**

1. To perform a sign test about a population median, select **Stat>Nonparametrics>1-Sample Sign**. Enter the name of the column containing your sample data in the box below **Variables**, select **Test median**, enter the hypothesized value of the median, and select your alternative hypothesis. Click **OK** to see the results. (See Screens 15.1 and 15.2.) Then use the *p*-value obtained in this output to make a decision.
2. To run a Wilcoxon rank sum test to determine if the populations from which two independent samples are drawn are identical, select **Stat>Nonparametrics>Mann-Whitney**. (Note that MINITAB does not have procedures for the Wilcoxon rank sum test. The Mann-Whitney test is very

**Sign Test for Median: Price**

```
Sign test of median = 137500 versus not = 137500
N Below Equal Above P Median
Price 10 3 0 7 0.3438 151950
```

Screen 15.2

similar to the Wilcoxon rank sum test and we can use it here.) Enter the names of the two columns containing your data in their respective boxes. Select the alternative hypothesis. Click **OK** to see the results. (See Screens 15.3 and 15.4.) Then use the *p*-value obtained in this output to make a decision.

- To run a Kruskal-Wallis test to determine if three or more populations have identical distributions, enter the data on the response variable in one column and the factor in another column. Select **Stat>Nonparametrics>Kruskal-Wallis** and enter the columns

**Mann-Whitney**

C1	Price
C2	City A
C3	City B

First Sample: 'City A'

Second Sample: 'City B'

Confidence level: 95.0

Alternative: not equal

Select

Help

OK

Cancel

Screen 15.3

**Mann-Whitney Test and CI: City A, City B**

```
N Median
City A 8 17.50
City B 9 20.00

Point estimate for ETA1-ETA2 is -4.00
95.1 Percent CI for ETA1-ETA2 is (-11.00,3.00)
W = 58.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.2110
The test is significant at 0.2101 (adjusted for ties)
```

Screen 15.4

containing the response variable and factor in their respective boxes. Click **OK** to see the results. (See Screens 15.5 and 15.6.) Then use the *p*-value obtained in this output to make a decision.

- To perform a Wilcoxon signed-rank test, enter the pairwise differences in a single column, making sure that you calculate each difference in the same order. Select **Stat>Nonparametrics>1-Sample Wilcoxon**. Enter the name of the column containing the differences in the box below **Variables**, select **Test median**, enter the hypothesized value of the median, and select your alternative hypothesis. Click **OK** to

**Kruskal-Wallis**

Response: Salary

Factor: City

Select

Help OK Cancel

Screen 15.5

**Kruskal-Wallis Test: Salary versus City**

Kruskal-Wallis Test on Salary

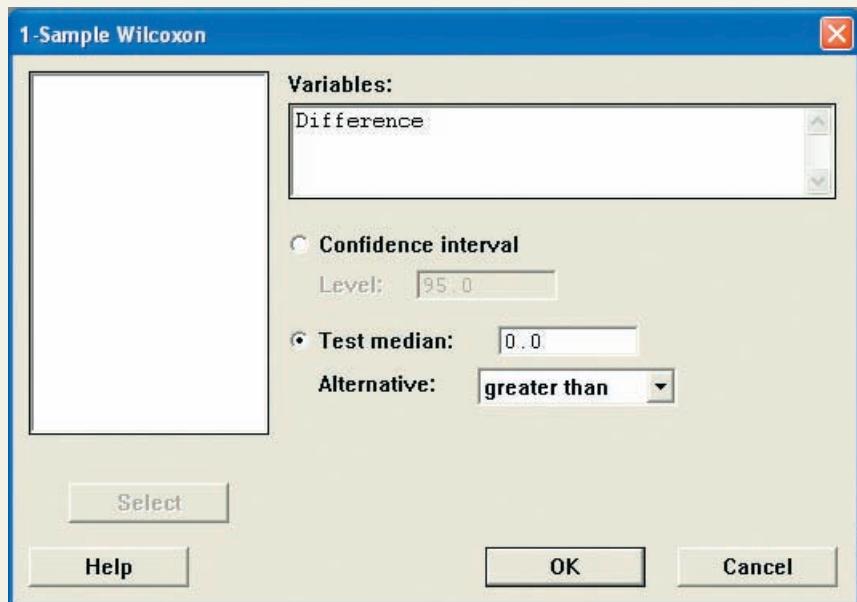
City	N	Median	Ave Rank	Z
Atlanta	8	53.00	11.9	0.54
Boston	6	48.50	12.5	0.70
San Francisco	7	43.00	8.6	-1.23
Overall	21		11.0	

H = 1.54 DF = 2 P = 0.462  
H = 1.55 DF = 2 P = 0.462 (adjusted for ties)

Screen 15.6

see the results. (See Screens 15.7 and 15.8.) Then use the *p*-value obtained in this output to make a decision.

5. To perform a runs test to see if a set of data is random, enter the given data into a column. Note that categorical data must be given numeric values. Select **Stat>Nonparametrics>Runs Test** and enter the column containing data in the box below **Variables**. Click **OK** to see the results. (See Screens 15.9 and 15.10.) Then use the *p*-value obtained in this output to make a decision.



Screen 15.7

#### Wilcoxon Signed Rank Test: Difference

```
Test of median = 0.000000 versus median > 0.000000

      N
      for   Wilcoxon      Estimated
      N  Test Statistic   P   Median
Difference 8      7       25.0  0.038     8.750
```

Screen 15.8

#### Runs Test: Gender

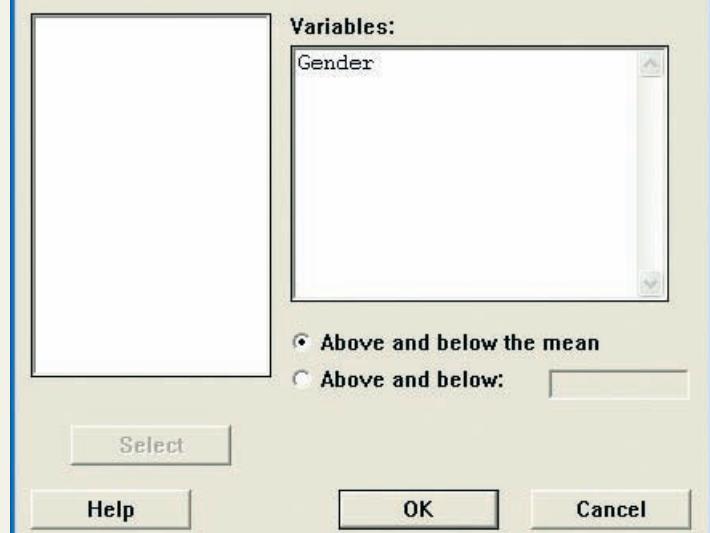
```
Runs test for Gender

Runs above and below K = 0.48

The observed number of runs = 13
The expected number of runs = 13.48
12 observations above K, 13 below
P-value = 0.844
```

Screen 15.10

#### Runs Test



Screen 15.9

## Excel

Excel does not contain any built-in nonparametric methods.

## TECHNOLOGY ASSIGNMENTS

**TA15.1** Fifteen coffee drinkers are selected at random and asked to test and state their preferences for Brand X coffee, Brand Y coffee, or neither (N). The results are as follows:

X    X    Y    X    N    Y    Y    X    Y    X    X    Y    Y    X    X

Let  $p$  be the proportion of coffee drinkers in the population who prefer Brand X. Using the sign test, perform the test  $H_0: p = .50$  against  $H_1: p > .50$ . Use a significance level of 2.5%.

**TA15.2** Twelve sixth-graders were selected at random and asked how many hours per week they spend watching television. The data obtained are shown here.

23      30      22.5      28      29      24.5      25      32      31      26      27      21

Using the sign test, can you conclude that the median number of hours spent per week watching television by all sixth-graders is less than 28? Use a significance level of 5%.

**TA15.3** The manufacturer of an engine oil additive, Hyper-Slick, claims that this product reduces the engine friction and, consequently, increases the miles per gallon (mpg). To test this claim, 10 cars are driven on a fixed 300-mile course without the oil additive, and each car's mpg is calculated and recorded. Then the engine oil additive is added to each car and the process is repeated. The data obtained are shown in the table.

Car	MPG without Additive	MPG with Additive
1	20.00	19.90
2	23.60	27.85
3	29.40	28.70
4	25.70	28.20
5	35.80	37.30
6	32.20	31.30
7	26.30	26.10
8	31.80	36.80
9	29.00	32.75
10	24.70	29.20

Using the sign test, can you conclude that the manufacturer's claim is true? Use a significance level of 5%.

**TA15.4** Do Technology Assignment TA15.3 using the Wilcoxon signed-rank test and a significance level of 5%. Compare your conclusion with that of Technology Assignment TA15.3 and comment.

**TA15.5** Refer to Exercise 15.43. In a winter Olympics trial for women's speed skating, seven skaters use a new type of skate, while eight others use the traditional type. Each skater is timed (in seconds) in the 500-meter event. The results are given in the table.

New skates	40.5	40.3	39.5	39.7	40.0	39.9	41.5
Traditional skates	41.0	40.8	40.9	39.8	40.6	40.7	41.1

Assume that these 15 skaters make up a random sample of all Olympic-class 500-meter female speed skaters. Using the Wilcoxon rank sum test (Mann-Whitney test), can you conclude that the new skates tend to produce faster times in this event? Use the 5% significance level.

**TA15.6** Refer to Exercise 15.46. Two brands of tires are tested to compare their durability. Eleven Brand X tires and 12 Brand Y tires are tested on a machine that simulates road conditions. The mileages (in thousands of miles for each tire) are shown in the following table.

Brand X	51	55	53	49	50.5	57	54.5	48.5	51.5	52	53.5
Brand Y	48	47	54	55.5	50	51	46	49.5	52.5	51	49

Using the Wilcoxon rank sum (Mann-Whitney) test, can you conclude that the median mileage for Brand X tires is greater than the median mileage for Brand Y tires? Use the 5% level of significance.

**TA15.7** Three brands of 60-watt lightbulbs—Brand A, Brand B, and a generic brand—are tested for their lives. The following table shows the lives (in hours) of these bulbs.

	Brand A	Brand B	Generic
	975	1001	899
	1050	1099	789
	890	915	824
	933	959	1011

962	986	907
925	957	923
1007	987	937
855	881	865
	1025	1024

Using the Kruskal-Wallis test with a significance level of .05, can you conclude that the distributions of the lives of lightbulbs are the same for all three brands?

**TA15.8** Refer to Exercise 15.54. A consumer agency investigated the premiums charged by four auto insurance companies. The agency randomly selected five drivers insured by each company who had similar driving records, autos, and insurance coverages. The following table gives the monthly premiums paid by these 20 drivers.

Company A	Company B	Company C	Company D
\$65	\$48	\$57	\$62
73	69	61	53
54	88	89	45
43	75	77	51
70	72	69	44

Using the Kruskal-Wallis test at the 5% significance level, can you reject the null hypothesis that the distributions of auto insurance premiums paid per month by all such drivers are the same for all four companies?

**TA15.9** Refer to Exercise 15.75. A fair coin is tossed 20 times in the presence of a psychic who claims that she can cause a nonrandom sequence of heads (H) and tails (T) to appear. The following sequence of heads and tails is obtained in these 20 tosses.

H H T H T H T H T H H T T H H H T T H

Using the runs test, can you conclude that the psychic's claim is true? Use a significance level of 5%.

**TA15.10** At a small soda factory, the amount of soda put into each 12-ounce bottle by the bottling machine varies slightly for each filling. The plant manager suspects that the machine has a random pattern of overfilling and underfilling the bottles. The following are the results of filling 18 bottles, where O denotes 12 ounces or more of soda in a bottle and U denotes less than 12 ounces of soda.

U U U O O O U U O O O U O U U U U

Using the runs test at the 5% significance level, can you conclude that there is a nonrandom pattern of overfilling and underfilling such bottles?

**Table VIII Critical Values of  $X$  for the Sign Test**

$n$	One tail $\alpha = .005$ Two tail $\alpha = .01$		One tail $\alpha = .01$ Two tail $\alpha = .02$		One tail $\alpha = .025$ Two tail $\alpha = .05$		One tail $\alpha = .05$ Two tail $\alpha = .10$	
	Lower critical value	Upper critical value	Lower critical value	Upper critical value	Lower critical value	Upper critical value	Lower critical value	Upper critical value
1	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	0	5
6	—	—	—	—	0	6	0	6
7	—	—	0	7	0	7	0	7
8	0	8	0	8	0	8	1	7
9	0	9	0	9	1	8	1	8
10	0	10	0	10	1	9	1	9
11	0	11	1	10	1	10	2	9
12	1	11	1	11	2	10	2	10
13	1	12	1	12	2	11	3	10
14	1	13	2	12	2	12	3	11
15	2	13	2	13	3	12	3	12
16	2	14	2	14	3	13	4	12
17	2	15	3	14	4	13	4	13
18	3	15	3	15	4	14	5	13
19	3	16	4	15	4	15	5	14
20	3	17	4	16	5	15	5	15
21	4	17	4	17	5	16	6	15
22	4	18	5	17	5	17	6	16
23	4	19	5	18	6	17	7	16
24	5	19	5	19	6	18	7	17
25	5	20	6	19	7	18	7	18

Source: D. B. Owen, *Handbook of Statistical Tables*. © 1962 by Addison-Wesley Publishing Company, Inc. Reprinted by permission of Addison Wesley Longman.

**Table IX Critical Values of  $T$  for the Wilcoxon Signed-Rank Test**

$n$	One-tailed $\alpha = .005$ Two-tailed $\alpha = .01$	One-tailed $\alpha = .01$ Two-tailed $\alpha = .02$	One-tailed $\alpha = .025$ Two-tailed $\alpha = .05$	One-tailed $\alpha = .05$ Two-tailed $\alpha = .10$
1	—	—	—	—
2	—	—	—	—
3	—	—	—	—
4	—	—	—	—
5	—	—	—	1
6	—	—	1	2
7	—	0	2	4
8	0	2	4	6
9	2	3	6	8
10	3	5	8	11
11	5	7	11	14
12	7	10	14	17
13	10	13	17	21
14	13	16	21	26
15	16	20	25	30

Source: *Some Rapid Approximate Statistical Procedures*, 1964. Reprinted with permission of Lederle Pharmaceutical Division of American Cyanamid Company, Philadelphia, PA.

**Table X Critical Values of  $T$  for the Wilcoxon Rank Sum Test****a. One-tailed  $\alpha = .025$ ; Two-tailed  $\alpha = .05$** 

$n_1 \backslash n_2$	3		4		5		6		7		8		9		10	
	$T_L$	$T_U$														
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

**b. One-tailed  $\alpha = .05$ ; Two-tailed  $\alpha = .10$** 

$n_1 \backslash n_2$	3		4		5		6		7		8		9		10	
	$T_L$	$T_U$														
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Source: *Some Rapid Approximate Statistical Procedures*, 1964. Reprinted with the permission of Lederle Pharmaceutical Division of American Cyanamid Company, Philadelphia, PA.

**Table XI Critical Values for the Spearman Rho Rank Correlation Coefficient Test**

<i>n</i>	One-tailed $\alpha$			
	.05	.025	.01	.005
	Two-tailed $\alpha$			
.10	.05	.02	.01	
5	±.900	—	—	—
6	±.829	±.886	±.943	—
7	±.714	±.786	±.893	±.929
8	±.643	±.738	±.833	±.881
9	±.600	±.700	±.783	±.833
10	±.564	±.648	±.745	±.794
11	±.536	±.618	±.709	±.755
12	±.503	±.587	±.678	±.727
13	±.475	±.566	±.672	±.744
14	±.456	±.544	±.645	±.714
15	±.440	±.524	±.622	±.688
16	±.425	±.506	±.601	±.665
17	±.411	±.490	±.582	±.644
18	±.399	±.475	±.564	±.625
19	±.388	±.462	±.548	±.607
20	±.377	±.450	±.534	±.591
21	±.368	±.438	±.520	±.576
22	±.359	±.428	±.508	±.562
23	±.351	±.418	±.496	±.549
24	±.343	±.409	±.485	±.537
25	±.336	±.400	±.475	±.526
26	±.329	±.392	±.465	±.515
27	±.323	±.384	±.456	±.505
28	±.317	±.377	±.448	±.496
29	±.311	±.370	±.440	±.487
30	±.305	±.364	±.432	±.478

**Table XII Critical Values for a Two-Tailed Runs Test with  $\alpha = .05$** 

$n_1 \backslash n_2$	5	6	7	8	9	10	11	12	13	14	15
2	—	—	—	—	—	—	—	2	2	2	2
								6	6	6	6
3	—	2	2	2	2	2	2	2	2	2	3
		8	8	8	8	8	8	8	8	8	8
4	2	2	2	3	3	3	3	3	3	3	3
	9	9	10	10	10	10	10	10	10	10	10
5	2	3	3	3	3	3	4	4	4	4	4
	10	10	11	11	12	12	12	12	12	12	12
6	3	3	3	3	4	4	4	4	5	5	5
	10	11	12	12	13	13	13	13	14	14	14
7	3	3	3	4	4	5	5	5	5	5	6
	11	12	13	13	14	14	14	14	15	15	15
8	3	3	4	4	5	5	5	6	6	6	6
	11	12	13	14	14	15	15	16	16	16	16
9	3	4	4	5	5	5	6	6	6	7	7
	12	13	14	14	15	16	16	16	17	17	18
10	3	4	5	5	5	6	6	7	7	7	7
	12	13	14	15	16	16	17	17	18	18	18
11	4	4	5	5	6	6	7	7	7	8	8
	12	13	14	15	16	17	17	18	19	19	19
12	4	4	5	6	6	7	7	7	8	8	8
	12	13	14	16	16	17	18	19	19	20	20
13	4	5	5	6	6	7	7	8	8	9	9
	12	14	15	16	17	18	19	19	20	20	21
14	4	5	5	6	7	7	8	8	9	9	9
	12	14	15	16	17	18	19	20	20	21	22
15	4	5	6	6	7	7	8	8	9	9	10
	12	14	15	16	18	18	19	20	21	22	22

Source: Frieda S. Swed and C. Eisenhart, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *The Annals of Statistics* 14(1943). Reprinted with permission of the Institute of Mathematical Statistics.



# Sample Surveys, Sampling Techniques, and Design of Experiments

## A.1 Sources of Data

The availability of accurate data is essential for deriving reliable results and making accurate decisions. As the truism “garbage in, garbage out” (GIGO) indicates, policy decisions based on the results of poor data may prove to be disastrous.

Data sources can be divided into three categories: internal sources, external sources, and surveys and experiments.

### A.1.1 Internal Sources

Often data come from **internal sources**, such as a company’s personnel files or accounting records. A company that wants to forecast the future sales of its products might use data from its records for previous periods. A police department might use data that exist in its records to analyze changes in the nature of crimes over a period of time.

### A.1.2 External Sources

All needed data may not be available from internal sources. Hence, to obtain data we may have to depend on sources outside the company, called **external sources**. Data obtained from external sources may be primary or secondary data. Data obtained from the organization that originally collected them are called **primary data**. If we obtain data from the Bureau of Labor Statistics that were collected by this organization, then these are primary data. Data obtained from a source that did not originally collect them are called **secondary data**. For example, data originally collected by the Bureau of Labor Statistics and published in the *Statistical Abstract of the United States* are secondary data.

### A.1 Sources of Data

### A.2 Sample Surveys and Sampling Techniques

### A.3 Design of Experiments

### A.1.3 Surveys and Experiments

Sometimes the data we need may not be available from internal or external sources. In such cases, we may have to obtain data by conducting our own survey or experiment.

#### Surveys

In a **survey**, we do not exercise any control over the factors when we collect information.

#### Definition

**Survey** In a *survey*, data are collected from the members of a population or sample in such a way that we have no particular control over the factors that may affect the characteristic of interest or the results of the survey.

For example, if we want to collect data on the money various families spent last month on clothes, we will ask each of the families included in the survey how much it spent last month on clothes. Then we will record this information.

A survey may be a census or a sample survey.

#### (i) Census

A **census** includes every member of the population of interest, which is called the **target population**.

#### Definition

**Census** A survey that includes every member of the population is called a *census*.

In practice, a census is rarely taken because it is very expensive and time consuming. Furthermore, in many cases it is impossible to identify each member of the target population. We discuss these reasons in more detail in Section A.2.1.

#### (ii) Sample Survey

Usually, to conduct research, we select a portion of the target population. This portion of the population is called a **sample**. Then we collect the required information from the elements included in the sample.

#### Definition

**Sample Survey** The technique of collecting information from a portion of the population is called a *sample survey*.

A survey can be conducted by personal interviews, by telephone, or by mail. The personal interview technique has the advantages of a high response rate and a high quality of answers obtained. However, it is the most expensive and time-consuming technique. The telephone survey also gives a high response rate. It is less expensive and less time-consuming than personal interviews. Nonetheless, a problem with telephone surveys is that many people do not like to be called at home, and those who do not have a phone are left out of the survey. A survey conducted by mail is the least expensive method, but the response rate is usually very low. Many people included in such a survey do not return the questionnaires.

Conducting a survey that gives accurate and reliable results is not an easy task. To quote Warren Mitofsky, director of Elections and Surveys for CBS News, “Any damn fool with 10 phones and a typewriter thinks he can conduct a poll.”<sup>1</sup> Preparing a questionnaire is probably

<sup>1</sup>“The Numbers Racket: How Polls and Statistics Lie,” *U.S. News & World Report*, July 11, 1988.

the most difficult part of a survey. The way a question is phrased can affect the results of the survey.

Section A.2 discusses sample surveys and sampling techniques in detail.

## Experiments

In an **experiment**, we exercise control over some factors when we collect information.

### Definition

**Experiment** In an *experiment*, data are collected from members of a population or sample in such a way that we have some control over the factors that may affect the characteristic of interest or the results of the experiment.

For example, how is a new drug to be tested to find out whether or not it cures a disease? This is done by designing an experiment in which the patients under study are divided into two groups as follows:

1. The **treatment group**—the members of this group receive the actual drug.
2. The **control group**—the members of this group do not receive the actual drug but are given a substitute (called a placebo) that appears to be the actual drug.

The two groups are formed in such a way that the patients in one group are similar to the patients in the other group. This is done by making random assignments of patients to the two groups. Neither the doctors nor the patients know to which group a patient belongs. Such an experiment is called a **double-blind experiment**. Then, after a comparison of the percentage of patients cured in each of the two groups, a decision is made about the effectiveness or noneffectiveness of the new drug. For more on experiments, refer to Section A.3 on experimental design.

## A.2 Sample Surveys and Sampling Techniques

In this section first we discuss the reasons sample surveys are preferred over a census, and then we discuss a representative sample, random and nonrandom samples, sampling and nonsampling errors, and random sampling techniques.

### A.2.1 Why Sample?

As mentioned in the previous section, most of the time surveys are conducted by using samples and not a census of the population. Three of the main reasons for conducting a sample survey instead of a census are listed next.

#### Time

In most cases, the size of the population is quite large. Consequently, conducting a census takes a long time, whereas a sample survey can be conducted very quickly. It is time-consuming to interview or contact hundreds of thousands or even millions of members of a population. On the other hand, a survey of a sample of a few hundred elements may be completed in little time. In fact, because of the amount of time needed to conduct a census, by the time the census is completed, the results may be obsolete.

#### Cost

The cost of collecting information from all members of a population may easily fall outside the limited budget of most, if not all, surveys. Consequently, to stay within the available resources, conducting a sample survey may be the best approach.

## Impossibility of Conducting a Census

Sometimes it is impossible to conduct a census. First, it may not be possible to identify and access each member of the population. For example, if a researcher wants to conduct a survey about homeless people, it is not possible to locate each member of the population and include him or her in the survey. Second, sometimes conducting a survey means destroying the items included in the survey. For example, to estimate the mean life of lightbulbs would necessitate burning out all the bulbs included in the survey. The same is true about finding the average life of batteries. In such cases, only a portion of the population can be selected for the survey.

### A.2.2 Random and Nonrandom Samples

Depending on how a sample is drawn, it may be a **random sample** or a **nonrandom sample**.

#### Definition

**Random and Nonrandom Samples** A *random sample* is a sample drawn in such a way that each member of the population has some chance of being selected in the sample. In a *nonrandom sample*, some members of the population may not have any chance of being selected in the sample.

Suppose we have a list of 100 students and we want to select 10 of them. If we write the names of all 100 students on pieces of paper, put them in a hat, mix them, and then draw 10 names, the result will be a random sample of 10 students. However, if we arrange the names of these 100 students alphabetically and pick the first 10 names, it will be a nonrandom sample because the students who are not among the first 10 have no chance of being selected in the sample.

A random sample is usually a representative sample. Note that for a random sample, each member of the population may or may not have the same chance of being included in the sample. Four types of random samples are discussed in Section A.2.4.

Two types of nonrandom samples are a *convenience sample* and a *judgment sample*. In a **convenience sample**, the most accessible members of the population are selected to obtain the results quickly. For example, an opinion poll may be conducted in a few hours by collecting information from certain shoppers at a single shopping mall. In a **judgment sample**, the members are selected from the population based on the judgment and prior knowledge of an expert. Although such a sample may happen to be a representative sample, the chances of it being so are small. If the population is large, it is not an easy task to select a representative sample based on judgment.

The so-called *pseudo polls* are examples of nonrepresentative samples. For instance, a survey conducted by a magazine that includes only its own readers does not usually involve a representative sample. Similarly, a poll conducted by a television station giving two separate telephone numbers for *yes* and *no* votes is not based on a representative sample. In these two examples, respondents will be only those people who read that magazine or watch that television station, who do not mind paying the postage or telephone charges, or who feel compelled to respond.

Another kind of sample is the **quota sample**. To select such a sample, we divide the target population into different subpopulations based on certain characteristics. Then we select a sub-sample from each subpopulation in such a way that each subpopulation is represented in the sample in exactly the same proportion as in the target population. As an example of a quota sample, suppose we want to select a sample of 1000 persons from a city whose population has 48% men and 52% women. To select a quota sample, we choose 480 men from the male population and 520 women from the female population. The sample selected in this way will contain exactly 48% men and 52% women. Another way to select a quota sample is to select from the population one person at a time until we have exactly 480 men and 520 women.

Until the 1948 presidential election in the United States, quota sampling was the most commonly used sampling procedure to conduct opinion polls. The voters included in the samples

were selected in such a way that they represented the population proportions of voters based on age, sex, education, income, race, and so on. However, this procedure was abandoned after the 1948 presidential election, in which the underdog, Harry Truman, defeated Thomas E. Dewey, who was heavily favored based on the opinion polls. First, the quota samples failed to be representative because the interviewers were allowed to fill their quotas by choosing voters based on their own judgments. This caused the selection of more upper-income and highly educated people, who happened to be Republicans. Thus, the quota samples were unrepresentative of the population because Republicans were overrepresented in these samples. Second, the results of the opinion polls based on quota sampling happened to be false because a large number of factors differentiate voters, but the pollsters considered only a few of those factors. A quota sample based on a few factors will skew the results. A random sample (one that is not based on quotas) has a much better chance of being representative of the population of all voters than a quota sample based on a few factors.

### A.2.3 Sampling and Nonsampling Errors

The results obtained from a sample survey may contain two types of errors: sampling and nonsampling errors. The sampling error is also called the chance error, and nonsampling errors are also called the systematic errors.

#### Sampling or Chance Error

Usually, all samples selected from the same population will give different results because they contain different elements of the population. Moreover, the results obtained from any one sample will not be exactly the same as the ones obtained from a census. The difference between a sample result and the result we would have obtained by conducting a census is called the **sampling error**, assuming that the sample is random and no nonsampling error has been made.

#### Definition

**Sampling Error** The *sampling error* is the difference between the result obtained from a sample survey and the result that would have been obtained if the whole population had been included in the survey.

The sampling error occurs because of chance, and it cannot be avoided. A sampling error can occur only in a sample survey. It does not occur in a census. Sampling error is discussed in detail in Section 7.1 of Chapter 7, and an example of it is given there.

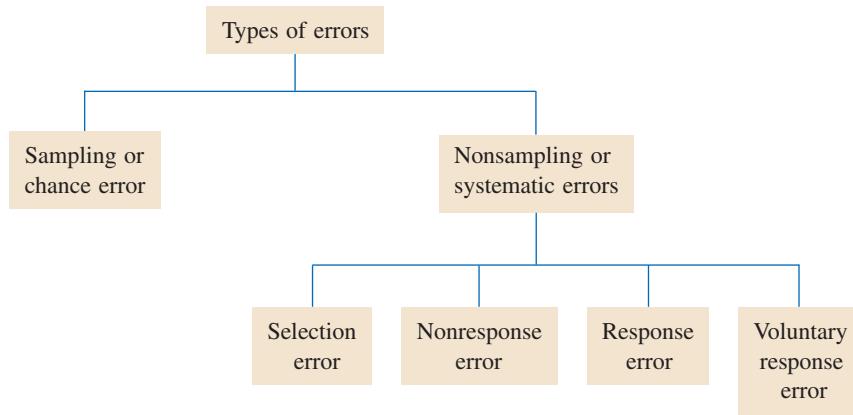
#### Nonsampling or Systematic Errors

**Nonsampling errors** can occur both in a sample survey and in a census. Such errors occur because of human mistakes and not chance.

#### Definition

**Nonsampling Errors** The errors that occur in the collection, recording, and tabulation of data are called *nonsampling errors*.

Nonsampling errors occur because of human mistakes and not chance. Nonsampling errors can be minimized if questions are prepared carefully and data are handled cautiously. Many types of systematic errors or biases can occur in a survey, including selection error, nonresponse error, response error, and voluntary response error. The following chart shows the types of errors.



### (i) Selection Error

When we need to select a sample, we use a list of elements from which we draw a sample, and this list usually does not include many members of the target population. Most of the time it is not feasible to include every member of the target population in this list. This list of members of the population that is used to select a sample is called the **sampling frame**. For example, if we use a telephone directory to select a sample, the list of names that appears in this directory makes the sampling frame. In this case we will miss the people who are not listed in the telephone directory. The people we miss, for example, will be poor people (including homeless people) who do not have telephones and people who do not want to be listed in the directory. Thus, the sampling frame that is used to select a sample may not be representative of the population. This may cause the sample results to be different from the population results. The error that occurs because the sampling frame is not representative of the population is called the **selection error**.

#### Definition

**Selection Error** The list of members of the target population that is used to select a sample is called the sampling frame. The error that occurs because the sampling frame is not representative of the population is called the *selection error*.

If a sample is nonrandom (and, hence, nonrepresentative), the sample results may be quite different from the census results.

### (ii) Nonresponse Error

Even if our sampling frame and, consequently, the sample are representative of the population, **nonresponse error** may occur because many of the people included in the sample did not respond to the survey.

#### Definition

**Nonresponse Error** The error that occurs because many of the people included in the sample do not respond to a survey is called the *nonresponse error*.

This type of error occurs especially when a survey is conducted by mail. A lot of people do not return the questionnaires. It has been observed that families with low and high incomes do not respond to surveys by mail. Consequently, such surveys overrepresent middle-income families. This kind of error occurs in other types of surveys, too. For instance, in a face-to-face survey where the interviewer interviews people in their homes, many people may not be home when the interviewer visits their homes. The people who are home at the time the interviewer

visits and the ones who are not home at that time may differ in many respects, causing a bias in the survey results. This kind of error may also occur in a telephone survey. Many people may not be home when the interviewer calls. This may distort the results. To avoid the nonresponse error, every effort should be made to contact all people included in the sample.

### (iii) Response Error

The **response error** occurs when the answer given by a person included in the survey is not correct. This may happen for many reasons. One reason is that the respondent may not have understood the question. Thus, the wording of the question may have caused the respondent to answer incorrectly. It has been observed that when the same question is worded differently, many people do not respond the same way. Usually such an error on the part of respondents is not intentional.

#### Definition

**Response Error** The *response error* occurs when people included in the survey do not provide correct answers.

Sometimes the respondents do not want to give correct information when answering a question. For example, many respondents will not disclose their true incomes on questionnaires or in interviews. When information on income is provided, it is almost always biased in the upward direction.

Sometimes the race of the interviewer may affect the answers of respondents. This is especially true if the questions asked are about race relations. The answers given by respondents may differ depending on the race of the interviewer.

### (iv) Voluntary Response Error

Another source of systematic error is a survey based on a voluntary response sample.

#### Definition

**Voluntary Response Error** *Voluntary response error* occurs when a survey is not conducted on a randomly selected sample but a questionnaire is published in a magazine or newspaper and people are invited to respond to that questionnaire.

The polls conducted based on samples of readers of magazines and newspapers suffer from **voluntary response error** or **bias**. Usually only those readers who have very strong opinions about the issues involved respond to such surveys. Surveys in which the respondents are required to call some telephone numbers also suffer from this type of error. Here, to participate, many times a respondent has to pay for the call, and many people do not want to bear this cost. Consequently, the sample is usually neither random nor representative of the target population because participation is voluntary.

## A.2.4 Random Sampling Techniques

There are many ways to select a random sample. Four of these techniques are discussed next.

### Simple Random Sampling

Under this sampling technique, each sample of the same size selected from the same population has the same probability of being selected.

#### Definition

**Simple Random Sampling** In this sampling technique, each sample of the same size has the same probability of being selected. Such a sample is called a *simple random sample*.

One way to select a simple random sample is by a lottery or drawing. For example, if we need to select 5 students from a class of 50, we write each of the 50 names on a separate piece of paper. Then, we place all 50 names in a hat and mix them thoroughly. Next, we draw 1 name randomly from the hat. We repeat this experiment four more times. The 5 drawn names make up a simple random sample.

The second procedure to select a simple random sample is to use a table of random numbers, which has become an outdated procedure. In this age of technology, it is much easier to use a statistical package, such as Minitab, to select a simple random sample.

### Systematic Random Sampling

The simple random sampling procedure becomes very tedious if the size of the population is large. For example, if we need to select 150 households from a list of 45,000, it is very time-consuming either to write the 45,000 names on pieces of paper and then select 150 households or to use a table of random numbers. In such cases, it is more convenient to use **systematic random sampling**.

The procedure to select a systematic random sample is as follows. In the example just mentioned, we would arrange all 45,000 households alphabetically (or based on some other characteristic). Since the sample size should equal 150, the ratio of population to sample size is  $45,000/150 = 300$ . Using this ratio, we randomly select one household from the first 300 households in the arranged list using either method. Suppose by using either of the methods, we select the 210th household. We then select every 210th household from every 300 households in the list. In other words, our sample includes the households with numbers 210, 510, 810, 1110, 1410, 1710, and so on.

#### Definition

**Systematic Random Sample** In *systematic random sampling*, we first randomly select one member from the first  $k$  units. Then every  $k$ th member, starting with the first selected member, is included in the sample.

### Stratified Random Sampling

Suppose we need to select a sample from the population of a city, and we want households with different income levels to be proportionately represented in the sample. In this case, instead of selecting a simple random sample or a systematic random sample, we may prefer to apply a different technique. First, we divide the whole population into different groups based on income levels. For example, we may form three groups of low-, medium-, and high-income households. We will now have three *subpopulations*, which are usually called **strata**. We then select one sample from each subpopulation or stratum. The collection of all three samples selected from three strata gives the required sample, called the **stratified random sample**. Usually, the sizes of the samples selected from different strata are proportionate to the sizes of the subpopulations in these strata. Note that the elements of each stratum are identical with regard to the possession of a characteristic.

#### Definition

**Stratified Random Sample** In a *stratified random sample*, we first divide the population into subpopulations, which are called strata. Then, one sample is selected from each of these strata. The collection of all samples from all strata gives the stratified random sample.

Thus, whenever we observe that a population differs widely in the possession of a characteristic, we may prefer to divide it into different strata and then select one sample from each stratum. We can divide the population on the basis of any characteristic, such as income, expenditure, sex, education, race, employment, or family size.

## Cluster Sampling

Sometimes the target population is scattered over a wide geographical area. Consequently, if a simple random sample is selected, it may be costly to contact each member of the sample. In such a case, we divide the population into different geographical groups or clusters and as a first step select a random sample of certain clusters from all clusters. We then take a random sample of certain elements from each selected cluster. For example, suppose we are to conduct a survey of households in the state of New York. First, we divide the whole state of New York into, say, 40 regions, which are called **clusters** or **primary units**. We make sure that all clusters are similar and, hence, representative of the population. We then select at random, say, 5 clusters from 40. Next, we randomly select certain households from each of these 5 clusters and conduct a survey of these selected households. This is called **cluster sampling**. Note that all clusters must be representative of the population.

### Definition

**Cluster Sampling** In *cluster sampling*, the whole population is first divided into (geographical) groups called clusters. Each cluster is representative of the population. Then a random sample of clusters is selected. Finally, a random sample of elements from each of the selected clusters is selected.

## A.3 Design of Experiments

As mentioned earlier, to use statistical methods to make decisions, we need access to data. Consider the following examples about decision making.

1. A government agency wants to find the average income of households in the United States.
2. A company wants to find the percentage of defective items produced on a machine.
3. A researcher wants to know if there is an association between eating unhealthy food and cholesterol level.
4. A pharmaceutical company has developed a new medicine for a disease and it wants to check if this medicine cures the disease.

All of these cases relate to decision making. We cannot reach a conclusion in these examples unless we have access to data. Data can be obtained from observational studies, experiments, or surveys. This section is devoted mainly to controlled experiments. However, it also explains observational studies and how they differ from surveys.

Suppose two diets, Diet 1 and Diet 2, are being promoted by two different companies, and each of these companies claims that its diet is successful in reducing weight. A research nutritionist wants to compare these diets with regard to their effectiveness for losing weight. Following are the two alternatives for the researcher to conduct this research.

1. The researcher contacts the persons who are using these diets and collects information on their weight loss. The researcher may contact as many persons as she has the time and financial resources for. Based on this information, the researcher makes a decision about the comparative effectiveness of these diets.
2. The researcher selects a sample of persons who want to lose weight, divides them randomly into two groups, and assigns each group to one of the two diets. Then she compares these two groups with regard to the effectiveness of these diets.

The first alternative is an example of an **observational study**, and the second is an example of a **controlled experiment**.

### Definition

**Treatment** A condition (or a set of conditions) that is imposed on a group of elements by the experimenter is called a *treatment*.

In an observational study the investigator does not impose a *treatment* on subjects or elements included in the study. For instance, in the first alternative mentioned above, the researcher simply collects information from the persons who are currently using these diets. In this case, the persons were not assigned to the two diets at random; instead, they chose the diets voluntarily. In this situation the researcher's conclusion about the comparative effectiveness of the two diets may not be valid because the effects of the diets will be **confounded** with many other factors or variables. When the effects of one factor cannot be separated from the effects of some other factors, the effects are said to be confounded. The persons who chose Diet 1 may be completely different with regard to age, gender, and eating and exercise habits from the persons who chose Diet 2. Thus, the weight loss may not be due entirely to the diet but to other factors or variables as well. Persons in one group may aggressively manage both diet and exercise, for example, whereas persons in the second group may depend entirely on diet. Thus, the effects of these other variables will get mixed up (confounded) with the effect of the diets.

Under the second alternative, the researcher selects a group of people, say 100, and randomly assigns them to two diets. One way to make random assignments is to write the name of each of these persons on a piece of paper, put them in a hat, and then randomly draw 50 names from this hat. These 50 persons will be assigned to one of the two diets, say Diet 1. The remaining 50 persons will be assigned to the second diet, Diet 2. This procedure is called **randomization**. Note that random assignments can also be made by using other methods such as a table of random numbers or technology.

### Definition

**Randomization** The procedure in which elements are assigned to different groups at random is called *randomization*.

When people are assigned to one or the other of two diets at random, the other differences among people in the two groups almost disappear. In this case these groups will not differ very much with regard to such factors as age, gender, and eating and exercise habits. The two groups will be very similar to each other. By using the random process to assign people to one or the other of two diets, we have *controlled* the other factors that can affect the weights of people. Consequently, this is an example of a **designed experiment**.

As mentioned earlier, a condition (or a set of conditions) that is imposed on a group of elements by the experimenter is called a treatment. In the example on diets, each of the two diet types is called a treatment. The experimenter randomly assigns the elements to these two treatments. Again, in such cases the study is called a designed experiment.

### Definition

**Designed Experiment and Observational Study** When the experimenter controls the (random) assignment of elements to different treatment groups, the study is said to be a *designed experiment*. In contrast, in an *observational study* the assignment of elements to different treatments is voluntary, and the experimenter simply observes the results of the study.

The group of people who receive a treatment is called the **treatment group**, and the group of people who do not receive a treatment is called the **control group**. In our example on diets, both groups are treatment groups because each group is assigned to one of the two types of diet. That example does not contain a control group.

### Definition

**Treatment and Control Groups** The group of elements that receives a treatment is called the *treatment group*, and the group of elements that does not receive a treatment is called the *control group*.

## ■ EXAMPLE A-1

Suppose a pharmaceutical company has developed a new medicine to cure a disease. To see whether or not this medicine is effective in curing this disease, it will have to be tested on a group of humans. Suppose there are 100 persons who have this disease; 50 of them voluntarily decide to take this medicine, and the remaining 50 decide not to take it. The researcher then compares the cure rates for the two groups of patients. Is this an example of a designed experiment or an observational study?

*An example of an observational study.*

**Solution** This is an example of an observational study because 50 patients voluntarily joined the treatment group; they were not randomly selected. In this case, the results of the study may not be valid because the effects of the medicine will be confounded with other variables. All of the patients who decided to take the medicine may not be similar to the ones who decided not to take it. It is possible that the persons who decided to take the medicine are in the advanced stages of the disease. Consequently, they do not have much to lose by being in the treatment group. The patients in the two groups may also differ with regard to other factors such as age, gender, and so on. ■

## ■ EXAMPLE A-2

Reconsider Example A-1. Now, suppose that out of the 100 people who have this disease, 50 are selected at random. These 50 people make up one group, and the remaining 50 belong to the second group. One of these groups is the treatment group, and the second is the control group. The researcher then compares the cure rates for the two groups of patients. Is this an example of a designed experiment or an observational study?

*An example of a designed experiment.*

**Solution** In this case, the two groups will be very similar to each other. Note that we do not expect the two groups to be exactly identical. However, when randomization is used, the two groups will be very similar. After these two groups have been formed, one group will be given the actual medicine. This group is called the treatment group. The other group will be administered a placebo (a dummy medicine that looks exactly like the actual medicine). This group is called the control group. This is an example of a designed experiment because the patients are assigned to one of two groups—the treatment or the control group—randomly. ■

Usually in an experiment like the one in Example A-2, patients do not know which group they belong to. Most of the time the experimenters do not know which group a patient belongs to. This is done to avoid any bias or distortion in the results of the experiment. When neither patients nor experimenters know who is taking the real medicine and who is taking the placebo, it is called a **double-blind experiment**. For the results of the study to be unbiased and valid, an experiment must be a double-blind designed experiment. Note that if either experimenters or patients or both have access to information regarding which patients belong to treatment or control groups, it will no longer be a double-blind experiment.

The use of placebos in medical experiments is very important. A placebo is just a dummy pill that looks exactly like the real medicine. Often, patients respond to any kind of medicine. Many studies have shown that even when the patients were given sugar pills (and did not know it), many of them indicated a decrease in pain. Patients respond to placebos because they have confidence in their physicians and medicines. This is called the **placebo effect**.

Note that there can be more than two groups of elements in an experiment. For example, an investigator may need to compare three diets for chickens with regard to weight gain. Here, in a designed experiment, the chickens will be randomly assigned to one of the three diets, which are the three treatments.

In some instances we have to base our research on observational studies because it is not feasible to conduct a designed experiment. For example, suppose a researcher wants to compare the starting salaries of business and psychology majors. The researcher will have to depend on an observational study. She will select two samples, one of recent business majors and another of recent psychology majors. Based on the starting salaries of these two groups, the researcher will make a decision. Note that, here, the effects of the majors on the starting salaries of the two

groups of graduates will be confounded with other variables. One of these other factors is that the business and psychology majors may be different in regard to intelligence level, which may affect their salaries. However, the researcher cannot conduct a designed experiment in this case. She cannot select a group of persons randomly and ask them to major in business and select another group and ask them to major in psychology. Instead, persons voluntarily choose their majors.

In a survey we do not exercise any control over the factors when we collect information. This characteristic of a survey makes it very close to an observational study. However, a survey may be based on a probability sample, which differentiates it from an observational study.

If an observational study or a survey indicates that two variables are related, it does not mean that there is a cause-and-effect relationship between them. For example, if an economist takes a sample of families, collects data on the incomes and rents paid by these families, and establishes an association between these two variables, it does not necessarily mean that families with higher incomes pay higher rents. Here the effects of many variables on rents are confounded. A family may pay a higher rent not because of higher income but because of various other factors, such as family size, preferences, or place of residence. We cannot make a statement about the cause-and-effect relationship between incomes and rents paid by families unless we control for these other variables. The association between incomes and rents paid by families may fit any of the following scenarios.

1. These two variables have a cause-and-effect relationship. Families that have higher incomes do pay higher rents. A change in incomes of families causes a change in rents paid.
2. The incomes and rents paid by families do not have a cause-and-effect relationship. Both of these variables have a cause-and-effect relationship with a third variable. Whenever that third variable changes, these two variables change.
3. The effect of income on rent is confounded with other variables, and this indicates that income affects rent paid by families.

If our purpose in a study is to establish a cause-and-effect relationship between two variables, we must control for the effects of other variables. In other words, we must conduct a designed study.

## EXERCISES

**A.1** Briefly describe the various sources of data.

**A.2** What is the difference between internal and external sources of data? Explain.

**A.3** Explain the difference between a sample survey and a census. Why is a sample survey usually preferred over a census?

**A.4** What is the difference between a survey and an experiment? Explain.

**A.5** Explain the following.

- |                    |                     |                       |
|--------------------|---------------------|-----------------------|
| a. Random sample   | b. Nonrandom sample | c. Convenience sample |
| d. Judgment sample | e. Quota sample     |                       |

**A.6** Explain briefly the following four sampling techniques.

- |                               |                               |
|-------------------------------|-------------------------------|
| a. Simple random sampling     | b. Systematic random sampling |
| c. Stratified random sampling | d. Cluster sampling           |

**A.7** In which sampling technique do all samples of the same size selected from a population have the same chance of being selected?

**A.8** A statistics professor wanted to find out the average GPA (grade point average) for all students at her university. She used all students enrolled in her statistics class as a sample and collected information on their GPAs to find the average GPA.

- a. Is this sample a random or a nonrandom sample? Explain.
- b. What kind of sample is it? In other words, is it a simple random sample, a systematic sample, a stratified sample, a cluster sample, a convenience sample, a judgment sample, or a quota sample? Explain.
- c. What kind of systematic error, if any, will be made with this kind of sample? Explain.

**A.9** A professor wanted to select 20 students from his class of 300 students to collect detailed information on the profiles of his students. He used his knowledge and expertise to select these 20 students.

- a. Is this sample a random or a nonrandom sample? Explain.
- b. What kind of sample is it? In other words, is it a simple random sample, a systematic sample, a stratified sample, a cluster sample, a convenience sample, a judgment sample, or a quota sample? Explain.
- c. What kind of systematic error, if any, will be made with this kind of sample? Explain.

**A.10** Refer to Exercise A.8. Suppose the professor obtains a list of all students enrolled at the university from the registrar's office and then selects 150 students at random from this list using a statistical software package such as Minitab.

- a. Is this sample a random or a nonrandom sample? Explain.
- b. What kind of sample is it? In other words, is it a simple random sample, a systematic sample, a stratified sample, a cluster sample, a convenience sample, a judgment sample, or a quota sample? Explain.
- c. Do you think any systematic error will be made in this case? Explain.

**A.11** Refer to Exercise A.9. Suppose the professor enters the names of all students enrolled in his class on a computer. He then selects a sample of 20 students at random using a statistical software package such as Minitab.

- a. Is this sample a random or a nonrandom sample? Explain.
- b. What kind of sample is it? In other words, is it a simple random sample, a systematic sample, a stratified sample, a cluster sample, a convenience sample, a judgment sample, or a quota sample? Explain.
- c. Do you think any systematic error will be made in this case? Explain.

**A.12** A company has 1000 employees, of whom 58% are men and 42% are women. The research department at the company wanted to conduct a quick survey by selecting a sample of 50 employees and asking them about their opinions on an issue. They divided the population of employees into two groups, men and women, and then selected 29 men and 21 women from these respective groups. The interviewers were free to choose any 29 men and 21 women they wanted. What kind of sample is it? Explain.

**A.13** A magazine published a questionnaire for its readers to fill out and mail to the magazine's office. In the questionnaire, cell phone owners were asked how much they would have to be paid to do without their cell phones for one month. The magazine received responses from 5439 cell phone owners.

- a. Based on the discussion of types of samples in Section A.2.2, what type of sample is this? Explain.
- b. To what kind(s) of systematic error, if any, would this survey be subject?

**A.14** A researcher wanted to conduct a survey of major companies to find out what benefits are offered to their employees. She mailed questionnaires to 2500 companies and received questionnaires back from 493 companies. What kind of systematic error does this survey suffer from? Explain.

**A.15** An opinion poll agency conducted a survey based on a random sample in which the interviewers called the parents included in the sample and asked them the following questions:

- i. Do you believe in spanking children?
- ii. Have you ever spanked your children?
- iii. If the answer to the second question is yes, how often?

What kind of systematic error, if any, does this survey suffer from? Explain.

**A.16** A survey based on a random sample taken from a borough of New York City showed that 65% of the people living there would prefer to live somewhere other than New York City if they had the opportunity to do so. Based on this result, can the researcher say that 65% of people living in New York City would prefer to live somewhere else if they had the opportunity to do so? Explain.

**A.17** In March 2005, the *New England Journal of Medicine* published the results of a 10-year clinical trial of low-dose aspirin therapy for the cardiovascular health of women (*Time*, March 21, 2005). The study was based on 40,000 healthy women, most of whom were in their 40s and 50s when the trial began. Half of these women were administered 100 mg of aspirin every other day, and the others were given a placebo. Assume that the women were assigned randomly to these two groups.

- a. Is this an observational study or a designed experiment? Explain.
- b. From the information given above, can you determine whether or not this is a double-blind study? Explain. If not, what additional information would you need?

**A.18** Refer to Exercise A.17. That study also looked at the incidences of heart attacks in the two groups of women. Overall the study did not find a statistically significant difference in heart attacks between the two groups of women. However, the study noted that among women who were at least 65 years old when

the study began, there was a lower incidence of heart attack for those who took aspirin than for those who took a placebo. Suppose that some medical researchers want to study this phenomenon more closely. They recruit 2000 healthy women aged 65 years and older, and randomly divide them into two groups. One group takes 100 mg of aspirin every other day, and the other group takes a placebo. The women did not know to which group they belonged, but the doctors who conducted the study had access to this information.

- a. Is this an observational study or a designed experiment? Explain.
- b. Is this a double-blind study? Explain.

**A.19** Refer to Exercise A.18. Now suppose that neither patients nor doctors knew what group patients belonged to.

- a. Is this an observational study or a designed experiment? Explain.
- b. Is this study a double-blind study? Explain.

**A.20** A federal government think tank wanted to investigate whether a job training program helps the families who are on welfare to get off the welfare program. The researchers at this agency selected 5000 volunteer families who were on welfare and offered the adults in those families free job training. The researchers selected another group of 5000 volunteer families who were on welfare and did not offer them such job training. After 3 years the two groups were compared in regard to the percentage of families who got off welfare. Is this an observational study or a designed experiment? Explain.

**A.21** Refer to Exercise A.20. Now suppose the agency selected 10,000 families at random from the list of all families that were on welfare. Of these 10,000 families, the agency randomly selected 5000 families and offered them free job training. The remaining 5000 families were not offered such job training. After 3 years the two groups were compared in regard to the percentage of families who got off welfare. Is this an observational study or a designed experiment? Explain.

**A.22** Refer to Exercise A.20. Based on that study, the researchers concluded that the job training program causes (helps) families who are on welfare to get off the welfare program. Do you agree with this conclusion? Explain.

**A.23** Refer to Exercise A.21. Based on that study, the researchers concluded that the job training program causes (helps) families who are on welfare to get off the welfare program. Do you agree with this conclusion? Explain.

**A.24** A researcher advertised for volunteers to study the relationship between the amount of meat consumed and cholesterol level. In response to this advertisement, 3476 persons volunteered. The researcher collected information on the meat consumption and cholesterol level of each of these persons. Based on these data, the researcher concluded that there is a very strong positive association between these two variables.

- a. Is this an observational study or a designed experiment? Explain.
- b. Based on this study, can the researcher conclude that consumption of meat increases cholesterol level? Explain why or why not.

**A.25** A pharmaceutical company developed a new medicine for compulsive behavior. To test this medicine on humans, the company advertised for volunteers who were suffering from this disease and wanted to participate in the study. As a result, 1820 persons responded. Using their own judgment, the group of physicians who were conducting this study assigned 910 of these patients to the treatment group and the remaining 910 to the control group. The patients in the treatment group were administered the actual medicine, and the patients in the control group were given a placebo. Six months later the conditions of the patients in the two groups were examined and compared. Based on this comparison, the physicians concluded that this medicine improves the condition of patients suffering from compulsive behavior.

- a. Comment on this study and its conclusion.
- b. Is this an observational study or a designed experiment? Explain.
- c. Is this a double-blind study? Explain.

**A.26** Refer to Exercise A.25. Suppose the physicians conducting this study obtained a list of all patients suffering from compulsive behavior who were being treated by doctors in all hospitals in the country. Further assume that this list is representative of the population of all such patients. The physicians then randomly selected 1820 patients from this list. Of these 1820, a randomly selected group of 910 patients were assigned to the treatment group, and the remaining 910 patients were assigned to the control group. The patients did not know which group they belonged to, but the doctors had access to such information. Six months later the conditions of the patients in the two groups were examined and compared. Based on this comparison, the physicians concluded that this medicine improves the condition of patients suffering from compulsive behavior.

- a. Comment on this study and its conclusion.
- b. Is this an observational study or a designed experiment? Explain.
- c. Is this a double-blind study? Explain.

**A.27** Refer to Exercise A.26. Now suppose that neither patients nor doctors knew what group the patients belonged to.

- Is this an observational study or a designed experiment? Explain.
- Is this a double-blind study? Explain.

**A.28** The Centre for Nutrition and Food Research at Queen Margaret University College in Edinburgh studied the relationship between sugar consumption and weight gain (*Fitness*, May 2002). All the people who participated in the study were divided into two groups, and both of these groups were put on low-calorie, low-fat diets. The diet of the people in the first group was low in sugar, but the people in the second group received as much as 10% of their calories from sucrose. Both groups stayed on their respective diets for 8 weeks. During these 8 weeks, participants in both groups lost 1/2 to 3/4 pound per week.

- Was this a designed experiment or an observational study?
- Was there a control group in this study?
- Was this a double-blind experiment?

**A.29** A psychologist needs 10 pigs for a study of the intelligence of pigs. She goes to a pig farm where there are 40 young pigs in a large pen. Assume that these pigs are representative of the population of all pigs. She selects the first 10 pigs she can catch and uses them for her study.

- Do these 10 pigs make a random sample?
- Are these 10 pigs likely to be representative of the entire population? Why or why not?
- If these 10 pigs do not form a random sample, what type of sample is it?
- Can you suggest a better procedure for selecting a sample of 10 from the 40 pigs in the pen?

**A.30** A newspaper wants to conduct a poll to estimate the percentage of its readers who favor a gambling casino in their city. People register their opinions by placing a phone call that costs them \$1.

- Is this method likely to produce a random sample?
- Which, if any, of the types of biases listed in this appendix are likely to be present and why?

## Advanced Exercises

**A.31** A researcher sent out questionnaires to 5000 randomly chosen members of HMOs (health maintenance organizations). Only 1200 of these members completed their questionnaires and returned them. Seventy-eight percent of the respondents reported that they had experienced denial of claims by their HMOs. Of those who experienced such denials, 25% had been unable to resolve the problem to their satisfaction in at least one such instance. Write an article for a business magazine summarizing the results of the survey and cautioning the readers about possible bias in the results. Indicate which types of biases are likely to be present, how they could arise, and whether the percentages given above are likely to overestimate the true percentages of all HMO members who have experienced the denial of claims by HMOs.

**A.32** A college is planning to finance an expansion of its student center through a special \$20 annual fee to be levied on each student for the next 4 years. Because the project will take 2 years to complete, the students who are currently juniors or seniors will not benefit from the expansion. The campus newspaper wants to conduct a poll to seek the opinions of students on this expansion. Such opinions of students are likely to depend on their current class status, so the newspaper decides to use a stratified random sample with four class levels (freshmen, sophomores, juniors, seniors) as strata. The current student body consists of 4000 freshmen, 3200 sophomores, 2800 juniors, and 2000 seniors. The sample will contain a total of 300 students, and the size of the sample from each stratum is to be proportional to the size of the subpopulation in each stratum.

- How many freshmen should be in the sample?
- How many students should be chosen from each of the other three class levels?

**A.33** A college mailed a questionnaire to all 5432 of its alumni who graduated in the last 5 years. One of the questions was about the current annual incomes of these alumni. Only 1620 of these alumni returned the completed questionnaires, and 1240 of them answered that question. The current mean annual income of these 1240 respondents was \$61,200.

- Do you think \$61,200 is likely to be an unbiased estimate of the current mean annual income of all 5432 alumni? If so, explain why.
- If you think that \$61,200 is probably a biased estimate of the current mean annual income of all 5432 alumni, what sources of systematic errors discussed in Section A.2.3 do you think are present here?
- Do you expect the estimate of \$61,200 to be above or below the current mean annual income of all 5432 alumni? Explain.

**A.34** A group of veterinarians wants to test a new canine vaccine for Lyme disease. (Lyme disease is transmitted by the bite of an infected deer tick.) One hundred dogs are randomly selected to receive the vaccine (with their owners' permission) from an area that has a high incidence of Lyme disease. These dogs are examined by veterinarians for symptoms of Lyme disease once a month for a period of 12 months. During this 12-month period, 10 of these 100 dogs are diagnosed with Lyme disease. During the same 12-month period, 18% of the unvaccinated dogs in the area are found to have contracted Lyme disease.

- a. Does this experiment have a control group?
- b. Is this a double-blind experiment?
- c. Identify any potential sources of bias in this experiment.
- d. Explain how this experiment could have been designed to reduce or eliminate the bias pointed out in part c.

## Glossary

**Census** A survey conducted by including every element of the population.

**Cluster** A subgroup (usually geographical) of the population that is representative of the population.

**Cluster sampling** A sampling technique in which the population is divided into clusters and a sample is chosen from one or a few clusters.

**Control group** The group on which no condition is imposed.

**Convenience sample** A sample that includes the most accessible members of the population.

**Designed experiment** A study in which the experimenter controls the assignment of elements to different treatment groups.

**Double-blind experiment** An experiment in which neither the doctors (or researchers) nor the patients (or members) know to which group a patient (or member) belongs.

**Experiment** A method of collecting data by controlling some or all factors.

**Judgment sample** A sample that includes the elements of the population selected based on the judgment and prior knowledge of an expert.

**Nonresponse error** The error that occurs because many of the people included in the sample do not respond.

**Nonsampling or systematic errors** The errors that occur in the collection, recording, and tabulation of data.

**Observational study** A study in which the assignment of elements to different treatments is voluntary, and the researcher simply observes the results of the study.

**Quota sample** A sample selected in such a way that each group or subpopulation is represented in the sample in exactly the same proportion as in the target population.

**Random sample** A sample that assigns some chance of being selected in the sample to each member of the population.

**Randomization** The procedure in which elements are assigned to different (treatment and control) groups at random.

**Representative sample** A sample that contains the characteristics of the population as closely as possible.

**Response error** The error that occurs because people included in the survey do not provide correct answers.

**Sample** A portion of the population of interest.

**Sample survey** A survey that includes elements of a sample.

**Sampling frame** The list of elements of the target population that is used to select a sample.

**Sampling or chance error** The difference between the result obtained from a sample survey and the result that would be obtained from the census.

**Selection error** The error that occurs because the sampling frame is not representative of the population.

**Simple random sampling** If all samples of the same size selected from a population have the same chance of being selected, it is called simple random sampling. Such a sample is called a simple random sample.

**Stratified random sampling** A sampling technique in which the population is divided into different strata and a sample is chosen from each stratum.

**Stratum** A subgroup of the population whose members are identical with regard to the possession of a characteristic.

**Survey** Collecting data from the elements of a population or sample.

**Systematic random sampling** A sampling method used to choose a sample by selecting every  $k$ th unit from the list.

**Target population** The collection of all subjects of interest.

**Treatment** A condition (or a set of conditions) that is imposed on a group of elements by the experimenter. This group is called the **treatment group**.

**Voluntary response error** The error that occurs because a survey is not conducted on a randomly selected sample, but people are invited to respond voluntarily to the survey.

# Explanation of Data Sets

This textbook is accompanied by 13 large data sets that can be used for statistical analysis using technology. These data sets are:

Data Set I	City Data
Data Set II	Data on States
Data Set III	NFL Data
Data Set IV	Beach to Beacon 10k Road Race Data
Data Set V	Sample of 500 Observations Selected From Beach to Beacon 10k Road Race Data
Data Set VI	Data on Movies
Data Set VII	Standard & Poor's 100 Index Data
Data Set VIII	McDonald's Data
Data Set IX	Candidate Data
Data Set X	Kickers2010 Data
Data Set XI	Billboard Data
Data Set XII	Motorcycle Data
Data Set XIII	Simulated Data

These data sets are available in Minitab, Excel, and a few other formats on the Web site for this text, [www.wiley.com/college/mann](http://www.wiley.com/college/mann). Once you are on this Web site, click on companion sites next to the cover of the book. Choose one of the two sites. Click on Data Sets on the left side. These data sets can be downloaded from this Web site. If you need more information on these data sets, you may either contact John Wiley's area representative or send an email to the author (see Preface). The Web site contains the following files:

1. CITYDATA (This file contains Data Set I)
2. STATEDATA (This file contains Data Set II)
3. NFL (This file contains Data Set III)
4. ROADRACE (This file contains the population data for Data Set IV)
5. RRSAMPLE (This file contains Data Set V)
6. MOVIEDATA (This file contains Data Set VI)
7. S&PDATA (This file contains Data Set VII)
8. MCDONALDDATA (This file contains Data Set VIII)
9. Candidate Data (This file contains Data Set IX)
10. KICKERS2010 (This file contains Data Set X)
11. BILLBOARD (This file contains Data Set XI)
12. MOTORCYCLE (This file contains Data Set XII)
13. SIMULATED (This file contains Data Set XIII)

The following are the explanations of these data sets.

## Data Set I: City Data<sup>1</sup>

This data set contains prices (in dollars) of selected products for selected cities across the country. This data set is reproduced from the ACCRA Cost of Living Index Survey for the second quarter 2011. It is reproduced with the permission of American Chamber of Commerce Researchers Association. This data set has 44 columns that contain the following variables.

- C1** Name of the city
- C2** Price of T-bone steak per pound
- C3** Price of 1 pound of ground beef (minimum 80% lean)
- C4** Price of 1 pound of sausage, Jimmy Dean or Owens brand, 100% pork
- C5** Price per pound of a whole fryer chicken
- C6** Price of a 6-ounce can of chunk light tuna, Starkist or Chicken of the Sea
- C7** Price of one half-gallon carton of whole milk
- C8** Price of one dozen large eggs, grade A/AA
- C9** Price of 1 pound of stick margarine, Blue Bonnet or Parkay
- C10** Price of Kraft parmesan cheese, grated, 8-ounce canister
- C11** Price of potatoes, 5 pounds, white or red
- C12** Price per pound of bananas
- C13** Price of one head of iceberg lettuce
- C14** Price of a loaf of white bread
- C15** Price of fresh orange juice, 64 ounces, Tropicana or Florida Natural brand
- C16** Price of an 11.5-ounce can or brick of coffee
- C17** Price of a 1-pound box of granulated sugar
- C18** Price of 18 ounces of corn flakes, Kellogg's or Post Toasties
- C19** Price of a 15-ounce can of sweet peas, Del Monte or Green Giant
- C20** Price of a 29-ounce can of peaches, Hunts, Del Monte, or Libby's, halves or slices
- C21** Price of facial tissue, 200 count, Kleenex
- C22** Price of 75-ounce Cascade dishwashing powder
- C23** Price of canola oil, 48 ounces, store brand.
- C24** Price of frozen prepared food, 8 to 10 ounces, frozen chicken entrée, Healthy Choice or Lean Cuisine.
- C25** Price of 16-ounce whole-kernel frozen corn, lowest priced
- C26** Price of potato chips, 13.75 or 20 ounces, Lay's plain
- C27** Price of 2-liter Coca-Cola, excluding any deposit
- C28** Monthly rent of an unfurnished two-bedroom apartment (excluding all utilities except water), 1½ or 2 baths, 950 square feet
- C29** Price of 1 gallon of regular unleaded gas, natural brand, including all taxes; cash price at self-service pump, if available
- C30** Cost of a visit to the optometrist's office for a full vision eye examination, established patient
- C31** Cost of a visit to the doctor's office for a routine examination for a problem with low to moderate severity
- C32** Cost of a tooth cleaning visit to the dentist's office (established patients only)
- C33** Price of Advil, 200-mg tablets, 100 count
- C34** Price of a 1/4-pound patty with cheese, pickle, onion, mustard, and ketchup (McDonald's Quarter-Pounder with cheese where available)

<sup>1</sup>We are thankful to Mr. Sean McNamara (COO and Chief Administrative Director) and Dean Frutiger (Project Manager, COLI) of the Council for Community and Economic Research for providing us these ACCRA data.

- C35** Price of an 11 to 12-inch thin-crust regular cheese pizza (no extra cheese) at Pizza Hut and/or Pizza Inn
- C36** Price of fried chicken, thigh and drumstick, with or without extras, whichever is lower, Kentucky Fried Chicken or Church's where available
- C37** Price of a man's barbershop haircut, no styling
- C38** Price of a woman's haircut with shampoo and blow-dry at beauty salons that make appointments and allow customer to select stylist
- C39** Price of toothpaste, 6 to 6.4-ounce-tube, Crest or Colgate
- C40** Price of a movie ticket; first run (new release), indoor 6 to 10 PM, Saturday evening rates, no discounts
- C41** Bowling, average price per line (game), Saturday evening, with rates in effect from 6 to 10 PM
- C42** Price of a can of three extra-duty tennis balls, Wilson or Penn brand
- C43** Price of Heineken's beer; six-pack of 12-ounce containers, excluding any deposit
- C44** Price of wine, Livingston Cellars or Gallo brand Chablis or Chenin Blanc, 1.5-liter bottle

## Data Set II: Data on States

This data set contains information on different variables for all 50 states of the United States and the district of Columbia. This data set has eight columns that contain the following variables:

- C1** Name of the state
- C2** Per capita personal income (in current dollars), 2010 (Source: U.S. Bureau of Economic Analysis)
- C3** Traffic fatalities, 2009 (Source: U.S. National Highway Traffic Safety Administration)
- C4** Percentage decrease in traffic fatalities, 2000–2009
- C5** Labor force participation rate (in percent), August 2011 (Source: U.S. Bureau of Labor Statistics)
- C6** Average salaries of teachers (in dollars), 2008–09 (Source: Current NEA Estimates Data Base)
- C7** Percent of the population (25 years and older) with a bachelor's degree or higher, 2005–09 (U.S. Census Bureau)
- C8** Location (East/West of the Mississippi River)

## Data Set III: NFL Data

This data set contains information on players who were on the rosters of National Football League (NFL) teams as of October 31, 2011. This data set has 11 columns that contain the following variables:

- C1** Uniform number
- C2** Name
- C3** Position
- C4** Position group
- C5** Status (active, physically unable to play, reserve, suspended)
- C6** Weight (pounds)
- C7** Years of experience
- C8** College
- C9** NFL Team
- C10** Height (inches)
- C11** Age (as of October 31, 2011)

## Data Set IV: Beach to Beacon 10k Road Race Data

This data set contains information on the runners who completed the 14th Annual Beach To Beacon 10K Road Race held on August 6, 2011, in Cape Elizabeth, Maine. The total distance of this race is 10 kilometers (6.2137 miles), and it is held every year on the first Saturday in August. A total of 5875 individuals completed the race in August 2011. The data set contains twelve columns that contain the following variables:

- C1** Overall place
- C2** Place within gender/age division
- C3** Number of entrants in gender/age division
- C4** Gender/age division
- C5** Age (years)
- C6** Gender (M/F)
- C7** State of residence (country for non-U.S. resident)
- C8** Time to complete the race (in seconds)
- C9** Pace per mile (in seconds)
- C10** U.S. resident (Yes/No)
- C11** Maine resident (Maine/Away)
- C12** Time to complete the race (in minutes)

## Data Set V: Sample of 500 Observations Selected From Beach to Beacon 10k Road Race Data

This data set contains a random sample of 500 observations selected from Data Set IV. It has 12 columns containing the same variables as listed in Data Set IV.

## Data Set VI: Data on Movies

This data set contains information on the top 150 films from 2010 in terms of gross revenue in the United States. This data set contains 8 columns that contain the following variables (source: <http://www.boxofficemojo.com>):

- C1** Rank
- C2** Movie title
- C3** Name of studio that produced the film
- C4** Gross revenue during entire theater release period
- C5** Number of theaters that showed the film during release period
- C6** Gross revenue during first week of theater release
- C7** Number of theaters that showed the film during first week of theater release
- C8** Length of release period (in days)

## Data Set VII: Standard & Poor's 100 Index Data

This data set contains trading and value information on the 100 stocks in the Standard & Poor's 100 Index as of Friday, February 17, 2012. This data set has 10 columns that contain the following variables (source: <http://finance.yahoo.com>):

- C1** Company's stock exchange symbol
- C2** Company name

- C3** Company's economic sector (e.g., manufacturing)
- C4** Stock price at close of business on Thursday, February 16, 2012
- C5** Stock price at close of business on Friday, February 17, 2012
- C6** Change in stock price from close of business on 2/16/2012 to 2/17/2012
- C7** Opening bid for stock price on 2/17/2012
- C8** Highest stock price attained on 2/17/2012
- C9** Lowest stock price attained on 2/17/2012
- C10** Number of shares traded on 2/17/2012

## Data Set VIII: McDonald's Data

This data set contains information on the nutritional aspects of McDonald's food. This data set is reproduced from McDonald's Web site ([http://www.mcdonalds.com/usa/eat/nutrition\\_info.html](http://www.mcdonalds.com/usa/eat/nutrition_info.html)). The only alteration involves the approximation of the dietary fiber content of four food items listed as having less than 1 gram of dietary fiber each, which were all changed to .5 gram. Condiments (ketchup, salad dressing, dipping sauces, and so forth) are not included. This data set has 25 columns that contain the following variables:

- C1** Menu item
- C2** Serving size (in ounces)
- C3** Serving size (in grams)
- C4** Calories
- C5** Calories from fat
- C6** Total fat (in grams)
- C7** Percent daily value of fat
- C8** Saturated fat (in grams)
- C9** Percent daily value of saturated fat
- C10** Trans fat (in grams)
- C11** Cholesterol (in milligrams)
- C12** Percent daily value of cholesterol
- C13** Sodium (in milligrams)
- C14** Percent daily value of sodium
- C15** Carbohydrates (in milligrams)
- C16** Percent daily value of carbohydrates
- C17** Dietary fiber (in grams)
- C18** Percent daily value of dietary fiber
- C19** Sugars (in grams)
- C20** Protein (in grams)
- C21** Percent daily value of Vitamin A
- C22** Percent daily value of Vitamin C
- C23** Percent daily value of Calcium
- C24** Percent daily value of Iron
- C25** Menu category (e.g., sandwich, non-sandwich chicken, breakfast, and so forth)

## Data Set IX: Candidate Data

This data set contains information on all of the candidates who ran for office for the U.S. Senate or House of Representatives in the 2010 election (<http://explore.data.gov/Elections/>

2009–2010-Candidate-Summary-File/38zs-22s9). This data set has 11 columns that contain the following variables:

- C1** Name
- C2** Outcome (win/loss)
- C3** Office (House/Senate)
- C4** State (postal abbreviation)
- C5** Party affiliation
- C6** Candidate status (challenger, incumbent, open)
- C7** Contributions received from individuals
- C8** Gross contributions received
- C9** Net contributions received
- C10** Net operating expenditures
- C11** Debt owed

## Data Set X: Kickers2010 Data

This data set contains information on the place kickers in the National Football League (NFL) and Canadian Football League (CFL) for the 2010 season. It contains 57 observations in ten columns with the following information:

- C1** Player's name
- C2** Player's team
- C3** Number of field goals made
- C4** Number of field goals attempted
- C5** Field goal completion percentage
- C6** Longest field goal made (in yards)
- C7** Number of extra points made
- C8** Number of extra points attempted
- C9** Extra point completion percentage
- C10** Conference/League (AFC = American Football Conference, CFL = Canadian Football League, NFC = National Football Conference) Both the AFC and the NFC are in the NFL

## Data Set XI: Billboard Data

This data set contains 100 observations in two columns, and was collected from the Billboard Hot 100 Popular Music charts for the week of July 9, 2011. The two columns contain the following information:

- C1** Number of weeks (for each song) in the Hot 100 chart
- C2** Ranking group for the week of July 9, 2011 (1-50 or 51-100)

## Data Set XII: Motorcycle Data

This data set contains information on the number of fatal motorcycle accidents during 2009 that occurred in each county of South Carolina. It contains 46 observations in two columns with the following information:

- C1** County name
- C2** Number of fatal motorcycle accidents

## Data Set XIII: Simulated Data

This data set contains four columns of simulated data from four different probability distributions. There are 1000 observations in each of the four columns, and these columns contains the following information:

- C1** Simulated data from probability distribution 1
- C2** Simulated data from probability distribution 2
- C3** Simulated data from probability distribution 3
- C4** Simulated data from probability distribution 4

This page is intentionally left blank



# Statistical Tables

Table I Table of Binomial Probabilities

Table II Values of  $e^{-\lambda}$

Table III Table of Poisson Probabilities

Table IV Standard Normal Distribution Table

Table V The  $t$  Distribution Table

Table VI Chi-Square Distribution Table

Table VII The  $F$  Distribution Table

Note: The following tables are on the Web site of the text along with Chapters 14 and 15.

Table VIII Critical Values of  $X$  for the Sign Test

Table IX Critical Values of  $T$  for the Wilcoxon Signed-Rank Test

Table X Critical Values of  $T$  for the Wilcoxon Rank Sum Test

Table XI Critical Values for the Spearman Rho Rank Correlation Coefficient Test

Table XII Critical Values for a Two-Tailed Runs Test with  $\alpha = .05$

**Table I Table of Binomial Probabilities**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
1	0	.9500	.9000	.8000	.7000	.6000	.5000	.4000	.3000	.2000	.1000	.0500
	1	.0500	.1000	.2000	.3000	.4000	.5000	.6000	.7000	.8000	.9000	.9500
2	0	.9025	.8100	.6400	.4900	.3600	.2500	.1600	.0900	.0400	.0100	.0025
	1	.0950	.1800	.3200	.4200	.4800	.5000	.4800	.4200	.3200	.1800	.0950
	2	.0025	.0100	.0400	.0900	.1600	.2500	.3600	.4900	.6400	.8100	.9025
3	0	.8574	.7290	.5120	.3430	.2160	.1250	.0640	.0270	.0080	.0010	.0001
	1	.1354	.2430	.3840	.4410	.4320	.3750	.2880	.1890	.0960	.0270	.0071
	2	.0071	.0270	.0960	.1890	.2880	.3750	.4320	.4410	.3840	.2430	.1354
	3	.0001	.0010	.0080	.0270	.0640	.1250	.2160	.3430	.5120	.7290	.8574
4	0	.8145	.6561	.4096	.2401	.1296	.0625	.0256	.0081	.0016	.0001	.0000
	1	.1715	.2916	.4096	.4116	.3456	.2500	.1536	.0756	.0256	.0036	.0005
	2	.0135	.0486	.1536	.2646	.3456	.3750	.3456	.2646	.1536	.0486	.0135
	3	.0005	.0036	.0256	.0756	.1536	.2500	.3456	.4116	.4096	.2916	.1715
	4	.0000	.0001	.0016	.0081	.0256	.0625	.1296	.2401	.4096	.6561	.8145
5	0	.7738	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	.0000
	1	.2036	.3280	.4096	.3602	.2592	.1562	.0768	.0284	.0064	.0005	.0000
	2	.0214	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	.0011
	3	.0011	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	.0214
	4	.0000	.0004	.0064	.0283	.0768	.1562	.2592	.3601	.4096	.3281	.2036
	5	.0000	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	.7738
6	0	.7351	.5314	.2621	.1176	.0467	.0156	.0041	.0007	.0001	.0000	.0000
	1	.2321	.3543	.3932	.3025	.1866	.0937	.0369	.0102	.0015	.0001	.0000
	2	.0305	.0984	.2458	.3241	.3110	.2344	.1382	.0595	.0154	.0012	.0001
	3	.0021	.0146	.0819	.1852	.2765	.3125	.2765	.1852	.0819	.0146	.0021
	4	.0001	.0012	.0154	.0595	.1382	.2344	.3110	.3241	.2458	.0984	.0305
	5	.0000	.0001	.0015	.0102	.0369	.0937	.1866	.3025	.3932	.3543	.2321
	6	.0000	.0000	.0001	.0007	.0041	.0156	.0467	.1176	.2621	.5314	.7351
7	0	.6983	.4783	.2097	.0824	.0280	.0078	.0016	.0002	.0000	.0000	.0000
	1	.2573	.3720	.3670	.2471	.1306	.0547	.0172	.0036	.0004	.0000	.0000
	2	.0406	.1240	.2753	.3177	.2613	.1641	.0774	.0250	.0043	.0002	.0000
	3	.0036	.0230	.1147	.2269	.2903	.2734	.1935	.0972	.0287	.0026	.0002
	4	.0002	.0026	.0287	.0972	.1935	.2734	.2903	.2269	.1147	.0230	.0036
	5	.0000	.0002	.0043	.0250	.0774	.1641	.2613	.3177	.2753	.1240	.0406
	6	.0000	.0000	.0004	.0036	.0172	.0547	.1306	.2471	.3670	.3720	.2573
	7	.0000	.0000	.0000	.0002	.0016	.0078	.0280	.0824	.2097	.4783	.6983
8	0	.6634	.4305	.1678	.0576	.0168	.0039	.0007	.0001	.0000	.0000	.0000
	1	.2793	.3826	.3355	.1977	.0896	.0312	.0079	.0012	.0001	.0000	.0000

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
2	2	.0515	.1488	.2936	.2965	.2090	.1094	.0413	.0100	.0011	.0000	.0000
3	3	.0054	.0331	.1468	.2541	.2787	.2187	.1239	.0467	.0092	.0004	.0000
4	4	.0004	.0046	.0459	.1361	.2322	.2734	.2322	.1361	.0459	.0046	.0004
5	5	.0000	.0004	.0092	.0467	.1239	.2187	.2787	.2541	.1468	.0331	.0054
6	6	.0000	.0000	.0011	.0100	.0413	.1094	.2090	.2965	.2936	.1488	.0515
7	7	.0000	.0000	.0001	.0012	.0079	.0312	.0896	.1977	.3355	.3826	.2793
8	8	.0000	.0000	.0000	.0001	.0007	.0039	.0168	.0576	.1678	.4305	.6634
9	0	.6302	.3874	.1342	.0404	.0101	.0020	.0003	.0000	.0000	.0000	.0000
	1	.2985	.3874	.3020	.1556	.0605	.0176	.0035	.0004	.0000	.0000	.0000
	2	.0629	.1722	.3020	.2668	.1612	.0703	.0212	.0039	.0003	.0000	.0000
	3	.0077	.0446	.1762	.2668	.2508	.1641	.0743	.0210	.0028	.0001	.0000
	4	.0006	.0074	.0661	.1715	.2508	.2461	.1672	.0735	.0165	.0008	.0000
	5	.0000	.0008	.0165	.0735	.1672	.2461	.2508	.1715	.0661	.0074	.0006
	6	.0000	.0001	.0028	.0210	.0743	.1641	.2508	.2668	.1762	.0446	.0077
	7	.0000	.0000	.0003	.0039	.0212	.0703	.1612	.2668	.3020	.1722	.0629
	8	.0000	.0000	.0000	.0004	.0035	.0176	.0605	.1556	.3020	.3874	.2985
	9	.0000	.0000	.0000	.0000	.0003	.0020	.0101	.0404	.1342	.3874	.6302
10	0	.5987	.3487	.1074	.0282	.0060	.0010	.0001	.0000	.0000	.0000	.0000
	1	.3151	.3874	.2684	.1211	.0403	.0098	.0016	.0001	.0000	.0000	.0000
	2	.0746	.1937	.3020	.2335	.1209	.0439	.0106	.0014	.0001	.0000	.0000
	3	.0105	.0574	.2013	.2668	.2150	.1172	.0425	.0090	.0008	.0000	.0000
	4	.0010	.0112	.0881	.2001	.2508	.2051	.1115	.0368	.0055	.0001	.0000
	5	.0001	.0015	.0264	.1029	.2007	.2461	.2007	.1029	.0264	.0015	.0001
	6	.0000	.0001	.0055	.0368	.1115	.2051	.2508	.2001	.0881	.0112	.0010
	7	.0000	.0000	.0008	.0090	.0425	.1172	.2150	.2668	.2013	.0574	.0105
	8	.0000	.0000	.0001	.0014	.0106	.0439	.1209	.2335	.3020	.1937	.0746
	9	.0000	.0000	.0000	.0001	.0016	.0098	.0403	.1211	.2684	.3874	.3151
	10	.0000	.0000	.0000	.0000	.0001	.0010	.0060	.0282	.1074	.3487	.5987
11	0	.5688	.3138	.0859	.0198	.0036	.0005	.0000	.0000	.0000	.0000	.0000
	1	.3293	.3835	.2362	.0932	.0266	.0054	.0007	.0000	.0000	.0000	.0000
	2	.0867	.2131	.2953	.1998	.0887	.0269	.0052	.0005	.0000	.0000	.0000
	3	.0137	.0710	.2215	.2568	.1774	.0806	.0234	.0037	.0002	.0000	.0000
	4	.0014	.0158	.1107	.2201	.2365	.1611	.0701	.0173	.0017	.0000	.0000
	5	.0001	.0025	.0388	.1321	.2207	.2256	.1471	.0566	.0097	.0003	.0000
	6	.0000	.0003	.0097	.0566	.1471	.2256	.2207	.1321	.0388	.0025	.0001
	7	.0000	.0000	.0017	.0173	.0701	.1611	.2365	.2201	.1107	.0158	.0014
	8	.0000	.0000	.0002	.0037	.0234	.0806	.1774	.2568	.2215	.0710	.0137
	9	.0000	.0000	.0000	.0005	.0052	.0269	.0887	.1998	.2953	.2131	.0867

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
10	0	.0000	.0000	.0000	.0000	.0007	.0054	.0266	.0932	.2362	.3835	.3293
10	1	.0000	.0000	.0000	.0000	.0000	.0005	.0036	.0198	.0859	.3138	.5688
12	0	.5404	.2824	.0687	.0138	.0022	.0002	.0000	.0000	.0000	.0000	.0000
12	1	.3413	.3766	.2062	.0712	.0174	.0029	.0003	.0000	.0000	.0000	.0000
12	2	.0988	.2301	.2835	.1678	.0639	.0161	.0025	.0002	.0000	.0000	.0000
12	3	.0173	.0852	.2362	.2397	.1419	.0537	.0125	.0015	.0001	.0000	.0000
12	4	.0021	.0213	.1329	.2311	.2128	.1208	.0420	.0078	.0005	.0000	.0000
12	5	.0002	.0038	.0532	.1585	.2270	.1934	.1009	.0291	.0033	.0000	.0000
12	6	.0000	.0005	.0155	.0792	.1766	.2256	.1766	.0792	.0155	.0005	.0000
12	7	.0000	.0000	.0033	.0291	.1009	.1934	.2270	.1585	.0532	.0038	.0002
12	8	.0000	.0000	.0005	.0078	.0420	.1208	.2128	.2311	.1329	.0213	.0021
12	9	.0000	.0000	.0001	.0015	.0125	.0537	.1419	.2397	.2362	.0852	.0173
12	10	.0000	.0000	.0000	.0002	.0025	.0161	.0639	.1678	.2835	.2301	.0988
12	11	.0000	.0000	.0000	.0000	.0003	.0029	.0174	.0712	.2062	.3766	.3413
12	12	.0000	.0000	.0000	.0000	.0000	.0002	.0022	.0138	.0687	.2824	.5404
13	0	.5133	.2542	.0550	.0097	.0013	.0001	.0000	.0000	.0000	.0000	.0000
13	1	.3512	.3672	.1787	.0540	.0113	.0016	.0001	.0000	.0000	.0000	.0000
13	2	.1109	.2448	.2680	.1388	.0453	.0095	.0012	.0001	.0000	.0000	.0000
13	3	.0214	.0997	.2457	.2181	.1107	.0349	.0065	.0006	.0000	.0000	.0000
13	4	.0028	.0277	.1535	.2337	.1845	.0873	.0243	.0034	.0001	.0000	.0000
13	5	.0003	.0055	.0691	.1803	.2214	.1571	.0656	.0142	.0011	.0000	.0000
13	6	.0000	.0008	.0230	.1030	.1968	.2095	.1312	.0442	.0058	.0001	.0000
13	7	.0000	.0001	.0058	.0442	.1312	.2095	.1968	.1030	.0230	.0008	.0000
13	8	.0000	.0000	.0011	.0142	.0656	.1571	.2214	.1803	.0691	.0055	.0003
13	9	.0000	.0000	.0001	.0034	.0243	.0873	.1845	.2337	.1535	.0277	.0028
13	10	.0000	.0000	.0000	.0006	.0065	.0349	.1107	.2181	.2457	.0997	.0214
13	11	.0000	.0000	.0000	.0001	.0012	.0095	.0453	.1388	.2680	.2448	.1109
13	12	.0000	.0000	.0000	.0000	.0001	.0016	.0113	.0540	.1787	.3672	.3512
13	13	.0000	.0000	.0000	.0000	.0000	.0001	.0013	.0097	.0550	.2542	.5133
14	0	.4877	.2288	.0440	.0068	.0008	.0001	.0000	.0000	.0000	.0000	.0000
14	1	.3593	.3559	.1539	.0407	.0073	.0009	.0001	.0000	.0000	.0000	.0000
14	2	.1229	.2570	.2501	.1134	.0317	.0056	.0005	.0000	.0000	.0000	.0000
14	3	.0259	.1142	.2501	.1943	.0845	.0222	.0033	.0002	.0000	.0000	.0000
14	4	.0037	.0349	.1720	.2290	.1549	.0611	.0136	.0014	.0000	.0000	.0000
14	5	.0004	.0078	.0860	.1963	.2066	.1222	.0408	.0066	.0003	.0000	.0000
14	6	.0000	.0013	.0322	.1262	.2066	.1833	.0918	.0232	.0020	.0000	.0000
14	7	.0000	.0002	.0092	.0618	.1574	.2095	.1574	.0618	.0092	.0002	.0000
14	8	.0000	.0000	.0020	.0232	.0918	.1833	.2066	.1262	.0322	.0013	.0000
14	9	.0000	.0000	.0003	.0066	.0408	.1222	.2066	.1963	.0860	.0078	.0004
14	10	.0000	.0000	.0000	.0014	.0136	.0611	.1549	.2290	.1720	.0349	.0037

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
11	0	.0000	.0000	.0000	.0002	.0033	.0222	.0845	.1943	.2501	.1142	.0259
11	1	.0000	.0000	.0000	.0000	.0005	.0056	.0317	.1134	.2501	.2570	.1229
11	2	.0000	.0000	.0000	.0000	.0001	.0009	.0073	.0407	.1539	.3559	.3593
11	3	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0068	.0440	.2288	.4877
12	0	.4633	.2059	.0352	.0047	.0005	.0000	.0000	.0000	.0000	.0000	.0000
12	1	.3658	.3432	.1319	.0305	.0047	.0005	.0000	.0000	.0000	.0000	.0000
12	2	.1348	.2669	.2309	.0916	.0219	.0032	.0003	.0000	.0000	.0000	.0000
12	3	.0307	.1285	.2501	.1700	.0634	.0139	.0016	.0001	.0000	.0000	.0000
12	4	.0049	.0428	.1876	.2186	.1268	.0417	.0074	.0006	.0000	.0000	.0000
12	5	.0006	.0105	.1032	.2061	.1859	.0916	.0245	.0030	.0001	.0000	.0000
12	6	.0000	.0019	.0430	.1472	.2066	.1527	.0612	.0116	.0007	.0000	.0000
12	7	.0000	.0003	.0138	.0811	.1771	.1964	.1181	.0348	.0035	.0000	.0000
12	8	.0000	.0000	.0035	.0348	.1181	.1964	.1771	.0811	.0138	.0003	.0000
12	9	.0000	.0000	.0007	.0116	.0612	.1527	.2066	.1472	.0430	.0019	.0000
12	10	.0000	.0000	.0001	.0030	.0245	.0916	.1859	.2061	.1032	.0105	.0006
12	11	.0000	.0000	.0000	.0006	.0074	.0417	.1268	.2186	.1876	.0428	.0049
12	12	.0000	.0000	.0000	.0001	.0016	.0139	.0634	.1700	.2501	.1285	.0307
12	13	.0000	.0000	.0000	.0000	.0003	.0032	.0219	.0916	.2309	.2669	.1348
12	14	.0000	.0000	.0000	.0000	.0000	.0005	.0047	.0305	.1319	.3432	.3658
12	15	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0047	.0352	.2059	.4633
13	0	.4401	.1853	.0281	.0033	.0003	.0000	.0000	.0000	.0000	.0000	.0000
13	1	.3706	.3294	.1126	.0228	.0030	.0002	.0000	.0000	.0000	.0000	.0000
13	2	.1463	.2745	.2111	.0732	.0150	.0018	.0001	.0000	.0000	.0000	.0000
13	3	.0359	.1423	.2463	.1465	.0468	.0085	.0008	.0000	.0000	.0000	.0000
13	4	.0061	.0514	.2001	.2040	.1014	.0278	.0040	.0002	.0000	.0000	.0000
13	5	.0008	.0137	.1201	.2099	.1623	.0667	.0142	.0013	.0000	.0000	.0000
13	6	.0001	.0028	.0550	.1649	.1983	.1222	.0392	.0056	.0002	.0000	.0000
13	7	.0000	.0004	.0197	.1010	.1889	.1746	.0840	.0185	.0012	.0000	.0000
13	8	.0000	.0001	.0055	.0487	.1417	.1964	.1417	.0487	.0055	.0001	.0000
13	9	.0000	.0000	.0012	.0185	.0840	.1746	.1889	.1010	.0197	.0004	.0000
13	10	.0000	.0000	.0002	.0056	.0392	.1222	.1983	.1649	.0550	.0028	.0001
13	11	.0000	.0000	.0000	.0013	.0142	.0666	.1623	.2099	.1201	.0137	.0008
13	12	.0000	.0000	.0000	.0002	.0040	.0278	.1014	.2040	.2001	.0514	.0061
13	13	.0000	.0000	.0000	.0000	.0008	.0085	.0468	.1465	.2463	.1423	.0359
13	14	.0000	.0000	.0000	.0000	.0001	.0018	.0150	.0732	.2111	.2745	.1463
13	15	.0000	.0000	.0000	.0000	.0000	.0002	.0030	.0228	.1126	.3294	.3706
13	16	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0033	.0281	.1853	.4401
14	0	.4181	.1668	.0225	.0023	.0002	.0000	.0000	.0000	.0000	.0000	.0000
14	1	.3741	.3150	.0957	.0169	.0019	.0001	.0000	.0000	.0000	.0000	.0000
14	2	.1575	.2800	.1914	.0581	.0102	.0010	.0001	.0000	.0000	.0000	.0000

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	3	.0415	.1556	.2393	.1245	.0341	.0052	.0004	.0000	.0000	.0000	.0000
	4	.0076	.0605	.2093	.1868	.0796	.0182	.0021	.0001	.0000	.0000	.0000
	5	.0010	.0175	.1361	.2081	.1379	.0472	.0081	.0006	.0000	.0000	.0000
	6	.0001	.0039	.0680	.1784	.1839	.0944	.0242	.0026	.0001	.0000	.0000
	7	.0000	.0007	.0267	.1201	.1927	.1484	.0571	.0095	.0004	.0000	.0000
	8	.0000	.0001	.0084	.0644	.1606	.1855	.1070	.0276	.0021	.0000	.0000
	9	.0000	.0000	.0021	.0276	.1070	.1855	.1606	.0644	.0084	.0001	.0000
	10	.0000	.0000	.0004	.0095	.0571	.1484	.1927	.1201	.0267	.0007	.0000
	11	.0000	.0000	.0001	.0026	.0242	.0944	.1839	.1784	.0680	.0039	.0001
	12	.0000	.0000	.0000	.0006	.0081	.0472	.1379	.2081	.1361	.0175	.0010
	13	.0000	.0000	.0000	.0001	.0021	.0182	.0796	.1868	.2093	.0605	.0076
	14	.0000	.0000	.0000	.0000	.0004	.0052	.0341	.1245	.2393	.1556	.0415
	15	.0000	.0000	.0000	.0000	.0001	.0010	.0102	.0581	.1914	.2800	.1575
	16	.0000	.0000	.0000	.0000	.0000	.0001	.0019	.0169	.0957	.3150	.3741
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0023	.0225	.1668	.4181
18	0	.3972	.1501	.0180	.0016	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3763	.3002	.0811	.0126	.0012	.0001	.0000	.0000	.0000	.0000	.0000
	2	.1683	.2835	.1723	.0458	.0069	.0006	.0000	.0000	.0000	.0000	.0000
	3	.0473	.1680	.2297	.1046	.0246	.0031	.0002	.0000	.0000	.0000	.0000
	4	.0093	.0700	.2153	.1681	.0614	.0117	.0011	.0000	.0000	.0000	.0000
	5	.0014	.0218	.1507	.2017	.1146	.0327	.0045	.0002	.0000	.0000	.0000
	6	.0002	.0052	.0816	.1873	.1655	.0708	.0145	.0012	.0000	.0000	.0000
	7	.0000	.0010	.0350	.1376	.1892	.1214	.0374	.0046	.0001	.0000	.0000
	8	.0000	.0002	.0120	.0811	.1734	.1669	.0771	.0149	.0008	.0000	.0000
	9	.0000	.0000	.0033	.0386	.1284	.1855	.1284	.0386	.0033	.0000	.0000
	10	.0000	.0000	.0008	.0149	.0771	.1669	.1734	.0811	.0120	.0002	.0000
	11	.0000	.0000	.0001	.0046	.0374	.1214	.1892	.1376	.0350	.0010	.0000
	12	.0000	.0000	.0000	.0012	.0145	.0708	.1655	.1873	.0816	.0052	.0002
	13	.0000	.0000	.0000	.0002	.0045	.0327	.1146	.2017	.1507	.0218	.0014
	14	.0000	.0000	.0000	.0000	.0011	.0117	.0614	.1681	.2153	.0700	.0093
	15	.0000	.0000	.0000	.0000	.0002	.0031	.0246	.1046	.2297	.1680	.0473
	16	.0000	.0000	.0000	.0000	.0000	.0006	.0069	.0458	.1723	.2835	.1683
	17	.0000	.0000	.0000	.0000	.0000	.0001	.0012	.0126	.0811	.3002	.3763
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0180	.1501	.3972
19	0	.3774	.1351	.0144	.0011	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3774	.2852	.0685	.0093	.0008	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1787	.2852	.1540	.0358	.0046	.0003	.0000	.0000	.0000	.0000	.0000
	3	.0533	.1796	.2182	.0869	.0175	.0018	.0001	.0000	.0000	.0000	.0000
	4	.0112	.0798	.2182	.1491	.0467	.0074	.0005	.0000	.0000	.0000	.0000

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
5	5	.0018	.0266	.1636	.1916	.0933	.0222	.0024	.0001	.0000	.0000	.0000
6	6	.0002	.0069	.0955	.1916	.1451	.0518	.0085	.0005	.0000	.0000	.0000
7	7	.0000	.0014	.0443	.1525	.1797	.0961	.0237	.0022	.0000	.0000	.0000
8	8	.0000	.0002	.0166	.0981	.1797	.1442	.0532	.0077	.0003	.0000	.0000
9	9	.0000	.0000	.0051	.0514	.1464	.1762	.0976	.0220	.0013	.0000	.0000
10	10	.0000	.0000	.0013	.0220	.0976	.1762	.1464	.0514	.0051	.0000	.0000
11	11	.0000	.0000	.0003	.0077	.0532	.1442	.1797	.0981	.0166	.0002	.0000
12	12	.0000	.0000	.0000	.0022	.0237	.0961	.1797	.1525	.0443	.0014	.0000
13	13	.0000	.0000	.0000	.0005	.0085	.0518	.1451	.1916	.0955	.0069	.0002
14	14	.0000	.0000	.0000	.0001	.0024	.0222	.0933	.1916	.1636	.0266	.0018
15	15	.0000	.0000	.0000	.0000	.0005	.0074	.0467	.1491	.2182	.0798	.0112
16	16	.0000	.0000	.0000	.0000	.0001	.0018	.0175	.0869	.2182	.1796	.0533
17	17	.0000	.0000	.0000	.0000	.0000	.0003	.0046	.0358	.1540	.2852	.1787
18	18	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0093	.0685	.2852	.3774
19	19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0011	.0144	.1351	.3774
20	0	.3585	.1216	.0115	.0008	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3774	.2702	.0576	.0068	.0005	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1887	.2852	.1369	.0278	.0031	.0002	.0000	.0000	.0000	.0000	.0000
	3	.0596	.1901	.2054	.0716	.0123	.0011	.0000	.0000	.0000	.0000	.0000
	4	.0133	.0898	.2182	.1304	.0350	.0046	.0003	.0000	.0000	.0000	.0000
	5	.0022	.0319	.1746	.1789	.0746	.0148	.0013	.0000	.0000	.0000	.0000
	6	.0003	.0089	.1091	.1916	.1244	.0370	.0049	.0002	.0000	.0000	.0000
	7	.0000	.0020	.0545	.1643	.1659	.0739	.0146	.0010	.0000	.0000	.0000
	8	.0000	.0004	.0222	.1144	.1797	.1201	.0355	.0039	.0001	.0000	.0000
	9	.0000	.0001	.0074	.0654	.1597	.1602	.0710	.0120	.0005	.0000	.0000
	10	.0000	.0000	.0020	.0308	.1171	.1762	.1171	.0308	.0020	.0000	.0000
	11	.0000	.0000	.0005	.0120	.0710	.1602	.1597	.0654	.0074	.0001	.0000
	12	.0000	.0000	.0001	.0039	.0355	.1201	.1797	.1144	.0222	.0004	.0000
	13	.0000	.0000	.0000	.0010	.0146	.0739	.1659	.1643	.0545	.0020	.0000
	14	.0000	.0000	.0000	.0002	.0049	.0370	.1244	.1916	.1091	.0089	.0003
	15	.0000	.0000	.0000	.0000	.0013	.0148	.0746	.1789	.1746	.0319	.0022
	16	.0000	.0000	.0000	.0000	.0003	.0046	.0350	.1304	.2182	.0898	.0133
	17	.0000	.0000	.0000	.0000	.0000	.0011	.0123	.0716	.2054	.1901	.0596
	18	.0000	.0000	.0000	.0000	.0000	.0002	.0031	.0278	.1369	.2852	.1887
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0068	.0576	.2702	.3774
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0115	.1216	.3585
21	0	.3406	.1094	.0092	.0006	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3764	.2553	.0484	.0050	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1981	.2837	.1211	.0215	.0020	.0001	.0000	.0000	.0000	.0000	.0000

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	3	.0660	.1996	.1917	.0585	.0086	.0006	.0000	.0000	.0000	.0000	.0000
	4	.0156	.0998	.2156	.1128	.0259	.0029	.0001	.0000	.0000	.0000	.0000
	5	.0028	.0377	.1833	.1643	.0588	.0097	.0007	.0000	.0000	.0000	.0000
	6	.0004	.0112	.1222	.1878	.1045	.0259	.0027	.0001	.0000	.0000	.0000
	7	.0000	.0027	.0655	.1725	.1493	.0554	.0087	.0005	.0000	.0000	.0000
	8	.0000	.0005	.0286	.1294	.1742	.0970	.0229	.0019	.0000	.0000	.0000
	9	.0000	.0001	.0103	.0801	.1677	.1402	.0497	.0063	.0002	.0000	.0000
	10	.0000	.0000	.0031	.0412	.1342	.1682	.0895	.0176	.0008	.0000	.0000
	11	.0000	.0000	.0008	.0176	.0895	.1682	.1342	.0412	.0031	.0000	.0000
	12	.0000	.0000	.0002	.0063	.0497	.1402	.1677	.0801	.0103	.0001	.0000
	13	.0000	.0000	.0000	.0019	.0229	.0970	.1742	.1294	.0286	.0005	.0000
	14	.0000	.0000	.0000	.0005	.0087	.0554	.1493	.1725	.0655	.0027	.0000
	15	.0000	.0000	.0000	.0001	.0027	.0259	.1045	.1878	.1222	.0112	.0004
	16	.0000	.0000	.0000	.0000	.0007	.0097	.0588	.1643	.1833	.0377	.0028
	17	.0000	.0000	.0000	.0000	.0001	.0029	.0259	.1128	.2156	.0998	.0156
	18	.0000	.0000	.0000	.0000	.0000	.0006	.0086	.0585	.1917	.1996	.0660
	19	.0000	.0000	.0000	.0000	.0000	.0001	.0020	.0215	.1211	.2837	.1981
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0050	.0484	.2553	.3764
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0092	.1094	.3406
22	0	.3235	.0985	.0074	.0004	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3746	.2407	.0406	.0037	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2070	.2808	.1065	.0166	.0014	.0001	.0000	.0000	.0000	.0000	.0000
	3	.0726	.2080	.1775	.0474	.0060	.0004	.0000	.0000	.0000	.0000	.0000
	4	.0182	.1098	.2108	.0965	.0190	.0017	.0001	.0000	.0000	.0000	.0000
	5	.0034	.0439	.1898	.1489	.0456	.0063	.0004	.0000	.0000	.0000	.0000
	6	.0005	.0138	.1344	.1808	.0862	.0178	.0015	.0000	.0000	.0000	.0000
	7	.0001	.0035	.0768	.1771	.1314	.0407	.0051	.0002	.0000	.0000	.0000
	8	.0000	.0007	.0360	.1423	.1642	.0762	.0144	.0009	.0000	.0000	.0000
	9	.0000	.0001	.0140	.0949	.1703	.1186	.0336	.0032	.0001	.0000	.0000
	10	.0000	.0000	.0046	.0529	.1476	.1542	.0656	.0097	.0003	.0000	.0000
	11	.0000	.0000	.0012	.0247	.1073	.1682	.1073	.0247	.0012	.0000	.0000
	12	.0000	.0000	.0003	.0097	.0656	.1542	.1476	.0529	.0046	.0000	.0000
	13	.0000	.0000	.0001	.0032	.0336	.1186	.1703	.0949	.0140	.0001	.0000
	14	.0000	.0000	.0000	.0009	.0144	.0762	.1642	.1423	.0360	.0007	.0000
	15	.0000	.0000	.0000	.0002	.0051	.0407	.1314	.1771	.0768	.0035	.0001
	16	.0000	.0000	.0000	.0000	.0015	.0178	.0862	.1808	.1344	.0138	.0005
	17	.0000	.0000	.0000	.0000	.0004	.0063	.0456	.1489	.1898	.0439	.0034
	18	.0000	.0000	.0000	.0000	.0001	.0017	.0190	.0965	.2108	.1094	.0182
	19	.0000	.0000	.0000	.0000	.0000	.0004	.0060	.0474	.1775	.2080	.0726

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
20	0	.0000	.0000	.0000	.0000	.0000	.0001	.0014	.0166	.1065	.2808	.2070
21	0	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0037	.0406	.2407	.3746
22	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0074	.0985	.3235
23	0	.3074	.0886	.0059	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3721	.2265	.0339	.0027	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2154	.2768	.0933	.0127	.0009	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0794	.2153	.1633	.0382	.0041	.0002	.0000	.0000	.0000	.0000	.0000
	4	.0209	.1196	.2042	.0818	.0138	.0011	.0000	.0000	.0000	.0000	.0000
	5	.0042	.0505	.1940	.1332	.0350	.0040	.0002	.0000	.0000	.0000	.0000
	6	.0007	.0168	.1455	.1712	.0700	.0120	.0008	.0000	.0000	.0000	.0000
	7	.0001	.0045	.0883	.1782	.1133	.0292	.0029	.0001	.0000	.0000	.0000
	8	.0000	.0010	.0442	.1527	.1511	.0584	.0088	.0004	.0000	.0000	.0000
	9	.0000	.0002	.0184	.1091	.1679	.0974	.0221	.0016	.0000	.0000	.0000
	10	.0000	.0000	.0064	.0655	.1567	.1364	.0464	.0052	.0001	.0000	.0000
	11	.0000	.0000	.0019	.0332	.1234	.1612	.0823	.0142	.0005	.0000	.0000
	12	.0000	.0000	.0005	.0142	.0823	.1612	.1234	.0332	.0019	.0000	.0000
	13	.0000	.0000	.0001	.0052	.0464	.1364	.1567	.0655	.0064	.0000	.0000
	14	.0000	.0000	.0000	.0016	.0221	.0974	.1679	.1091	.0184	.0002	.0000
	15	.0000	.0000	.0000	.0004	.0088	.0584	.1511	.1527	.0442	.0010	.0000
	16	.0000	.0000	.0000	.0001	.0029	.0292	.1133	.1782	.0883	.0045	.0001
	17	.0000	.0000	.0000	.0000	.0008	.0120	.0700	.1712	.1455	.0168	.0007
	18	.0000	.0000	.0000	.0000	.0002	.0040	.0350	.1332	.1940	.0505	.0042
	19	.0000	.0000	.0000	.0000	.0000	.0011	.0138	.0818	.2042	.1196	.0209
	20	.0000	.0000	.0000	.0000	.0000	.0002	.0041	.0382	.1633	.2153	.0794
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0009	.0127	.0933	.2768	.2154
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0027	.0339	.2265	.3721
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0059	.0886	.3074
24	0	.2920	.0798	.0047	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3688	.2127	.0283	.0020	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2232	.2718	.0815	.0097	.0006	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0862	.2215	.1493	.0305	.0028	.0001	.0000	.0000	.0000	.0000	.0000
	4	.0238	.1292	.1960	.0687	.0099	.0006	.0000	.0000	.0000	.0000	.0000
	5	.0050	.0574	.1960	.1177	.0265	.0025	.0001	.0000	.0000	.0000	.0000
	6	.0008	.0202	.1552	.1598	.0560	.0080	.0004	.0000	.0000	.0000	.0000
	7	.0001	.0058	.0998	.1761	.0960	.0206	.0017	.0000	.0000	.0000	.0000
	8	.0000	.0014	.0530	.1604	.1360	.0438	.0053	.0002	.0000	.0000	.0000
	9	.0000	.0003	.0236	.1222	.1612	.0779	.0141	.0008	.0000	.0000	.0000
	10	.0000	.0000	.0088	.0785	.1612	.1169	.0318	.0026	.0000	.0000	.0000
	11	.0000	.0000	.0028	.0428	.1367	.1488	.0608	.0079	.0002	.0000	.0000

**Table I Table of Binomial Probabilities (continued)**

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
12	0	.0000	.0000	.0008	.0199	.0988	.1612	.0988	.0199	.0008	.0000	.0000
12	1	.0000	.0000	.0002	.0079	.0608	.1488	.1367	.0428	.0028	.0000	.0000
12	2	.0000	.0000	.0000	.0026	.0318	.1169	.1612	.0785	.0088	.0000	.0000
12	3	.0000	.0000	.0000	.0008	.0141	.0779	.1612	.1222	.0236	.0003	.0000
12	4	.0000	.0000	.0000	.0002	.0053	.0438	.1360	.1604	.0530	.0014	.0000
12	5	.0000	.0000	.0000	.0000	.0017	.0206	.0960	.1761	.0998	.0058	.0001
12	6	.0000	.0000	.0000	.0000	.0004	.0080	.0560	.1598	.1552	.0202	.0008
12	7	.0000	.0000	.0000	.0000	.0001	.0025	.0265	.1177	.1960	.0574	.0050
12	8	.0000	.0000	.0000	.0000	.0000	.0006	.0099	.0687	.1960	.1292	.0238
12	9	.0000	.0000	.0000	.0000	.0000	.0001	.0028	.0305	.1493	.2215	.0862
12	10	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0097	.0815	.2718	.2232
12	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0020	.0283	.2127	.3688
12	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0047	.0798	.2920
25	0	.2774	.0718	.0038	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
25	1	.3650	.1994	.0236	.0014	.0000	.0000	.0000	.0000	.0000	.0000	.0000
25	2	.2305	.2659	.0708	.0074	.0004	.0000	.0000	.0000	.0000	.0000	.0000
25	3	.0930	.2265	.1358	.0243	.0019	.0001	.0000	.0000	.0000	.0000	.0000
25	4	.0269	.1384	.1867	.0572	.0071	.0004	.0000	.0000	.0000	.0000	.0000
25	5	.0060	.0646	.1960	.1030	.0199	.0016	.0000	.0000	.0000	.0000	.0000
25	6	.0010	.0239	.1633	.1472	.0442	.0053	.0002	.0000	.0000	.0000	.0000
25	7	.0001	.0072	.1108	.1712	.0800	.0143	.0009	.0000	.0000	.0000	.0000
25	8	.0000	.0018	.0623	.1651	.1200	.0322	.0031	.0001	.0000	.0000	.0000
25	9	.0000	.0004	.0294	.1336	.1511	.0609	.0088	.0004	.0000	.0000	.0000
25	10	.0000	.0001	.0118	.0916	.1612	.0974	.0212	.0013	.0000	.0000	.0000
25	11	.0000	.0000	.0040	.0536	.1465	.1328	.0434	.0042	.0001	.0000	.0000
25	12	.0000	.0000	.0012	.0268	.1140	.1550	.0760	.0115	.0003	.0000	.0000
25	13	.0000	.0000	.0003	.0115	.0760	.1550	.1140	.0268	.0012	.0000	.0000
25	14	.0000	.0000	.0001	.0042	.0434	.1328	.1465	.0536	.0040	.0000	.0000
25	15	.0000	.0000	.0000	.0013	.0212	.0974	.1612	.0916	.0118	.0001	.0000
25	16	.0000	.0000	.0000	.0004	.0088	.0609	.1511	.1336	.0294	.0004	.0000
25	17	.0000	.0000	.0000	.0001	.0031	.0322	.1200	.1651	.0623	.0018	.0000
25	18	.0000	.0000	.0000	.0000	.0009	.0143	.0800	.1712	.1108	.0072	.0001
25	19	.0000	.0000	.0000	.0000	.0002	.0053	.0442	.1472	.1633	.0239	.0010
25	20	.0000	.0000	.0000	.0000	.0000	.0016	.0199	.1030	.1960	.0646	.0060
25	21	.0000	.0000	.0000	.0000	.0000	.0004	.0071	.0572	.1867	.1384	.0269
25	22	.0000	.0000	.0000	.0000	.0000	.0001	.0019	.0243	.1358	.2265	.0930
25	23	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0074	.0708	.2659	.2305
25	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0014	.0236	.1994	.3650
25	25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0038	.0718	.2774

**Table II Values of  $e^{-\lambda}$** 

$\lambda$	$e^{-\lambda}$	$\lambda$	$e^{-\lambda}$
0.0	1.00000000	3.9	.02024191
0.1	.90483742	4.0	.01831564
0.2	.81873075	4.1	.01657268
0.3	.74081822	4.2	.01499558
0.4	.67032005	4.3	.01356856
0.5	.60653066	4.4	.01227734
0.6	.54881164	4.5	.01110900
0.7	.49658530	4.6	.01005184
0.8	.44932896	4.7	.00909528
0.9	.40656966	4.8	.00822975
1.0	.36787944	4.9	.00744658
1.1	.33287108	5.0	.00673795
1.2	.30119421	5.1	.00609675
1.3	.27253179	5.2	.00551656
1.4	.24659696	5.3	.00499159
1.5	.22313016	5.4	.00451658
1.6	.20189652	5.5	.00408677
1.7	.18268352	5.6	.00369786
1.8	.16529889	5.7	.00334597
1.9	.14956862	5.8	.00302755
2.0	.13533528	5.9	.00273944
2.1	.12245643	6.0	.00247875
2.2	.11080316	6.1	.00224287
2.3	.10025884	6.2	.00202943
2.4	.09071795	6.3	.00183630
2.5	.08208500	6.4	.00166156
2.6	.07427358	6.5	.00150344
2.7	.06720551	6.6	.00136037
2.8	.06081006	6.7	.00123091
2.9	.05502322	6.8	.00111378
3.0	.04978707	6.9	.00100779
3.1	.04504920	7.0	.00091188
3.2	.04076220	7.1	.00082510
3.3	.03688317	7.2	.00074659
3.4	.03337327	7.3	.00067554
3.5	.03019738	7.4	.00061125
3.6	.02732372	7.5	.00055308
3.7	.02472353	7.6	.00050045
3.8	.02237077	7.7	.00045283

**Table II Values of  $e^{-\lambda}$  (continued)**

$\lambda$	$e^{-\lambda}$	$\lambda$	$e^{-\lambda}$
7.8	.00040973	9.5	.00007485
7.9	.00037074	9.6	.00006773
8.0	.00033546	9.7	.00006128
8.1	.00030354	9.8	.00005545
8.2	.00027465	9.9	.00005017
8.3	.00024852	10.0	.00004540
8.4	.00022487	11.0	.00001670
8.5	.00020347	12.0	.00000614
8.6	.00018411	13.0	.00000226
8.7	.00016659	14.0	.00000083
8.8	.00015073	15.0	.00000031
8.9	.00013639	16.0	.00000011
9.0	.00012341	17.0	.00000004
9.1	.00011167	18.0	.000000015
9.2	.00010104	19.0	.000000006
9.3	.00009142	20.0	.000000002
9.4	.00008272		

**Table III Table of Poisson Probabilities**

**Table III Table of Poisson Probabilities (continued)**

$x$	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

$x$	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

$x$	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149

**Table III Table of Poisson Probabilities (continued)**

$x$	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001

$x$	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0244	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0263
13	.0058	.0065	.0073	.0081	.0089	.0099	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

**Table III Table of Poisson Probabilities (continued)**

$x$	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1381	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

$x$	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1255	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0775	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728

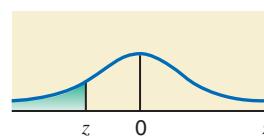
**Table III Table of Poisson Probabilities (continued)**

$x$	$\lambda$									
	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001
$x$	$\lambda$									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1209	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

**Table III Table of Poisson Probabilities (continued)**

**Table IV Standard Normal Distribution Table**

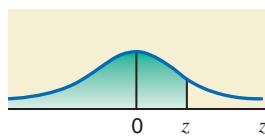
The entries in the table on this page give the cumulative area under the standard normal curve to the left of  $z$  with the values of  $z$  equal to 0 or negative.



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

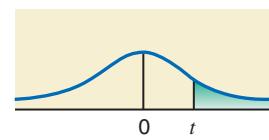
**Table IV Standard Normal Distribution Table (continued)**

The entries in the table on this page give the cumulative area under the standard normal curve to the left of  $z$  with the values of  $z$  equal to 0 or positive.



**Table V The *t* Distribution Table**

The entries in this table give the critical values of *t* for the specified number of degrees of freedom and areas in the right tail.



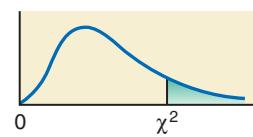
<i>df</i>	Area in the Right Tail Under the <i>t</i> Distribution Curve					
	.10	.05	.025	.01	.005	.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340

**Table V The *t* Distribution Table (continued)**

<i>df</i>	Area in the Right Tail Under the <i>t</i> Distribution Curve					
	.10	.05	.025	.01	.005	.001
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
41	1.303	1.683	2.020	2.421	2.701	3.301
42	1.302	1.682	2.018	2.418	2.698	3.296
43	1.302	1.681	2.017	2.416	2.695	3.291
44	1.301	1.680	2.015	2.414	2.692	3.286
45	1.301	1.679	2.014	2.412	2.690	3.281
46	1.300	1.679	2.013	2.410	2.687	3.277
47	1.300	1.678	2.012	2.408	2.685	3.273
48	1.299	1.677	2.011	2.407	2.682	3.269
49	1.299	1.677	2.010	2.405	2.680	3.265
50	1.299	1.676	2.009	2.403	2.678	3.261
51	1.298	1.675	2.008	2.402	2.676	3.258
52	1.298	1.675	2.007	2.400	2.674	3.255
53	1.298	1.674	2.006	2.399	2.672	3.251
54	1.297	1.674	2.005	2.397	2.670	3.248
55	1.297	1.673	2.004	2.396	2.668	3.245
56	1.297	1.673	2.003	2.395	2.667	3.242
57	1.297	1.672	2.002	2.394	2.665	3.239
58	1.296	1.672	2.002	2.392	2.663	3.237
59	1.296	1.671	2.001	2.391	2.662	3.234
60	1.296	1.671	2.000	2.390	2.660	3.232
61	1.296	1.670	2.000	2.389	2.659	3.229
62	1.295	1.670	1.999	2.388	2.657	3.227
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.220
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.650	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211
71	1.294	1.667	1.994	2.380	2.647	3.209
72	1.293	1.666	1.993	2.379	2.646	3.207
73	1.293	1.666	1.993	2.379	2.645	3.206
74	1.293	1.666	1.993	2.378	2.644	3.204
75	1.293	1.665	1.992	2.377	2.643	3.202
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090

**Table VI Chi-Square Distribution Table**

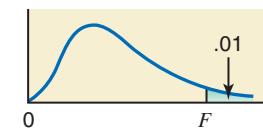
The entries in this table give the critical values of  $\chi^2$  for the specified number of degrees of freedom and areas in the right tail.



df	Area in the Right Tail Under the Chi-square Distribution Curve									
	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

**Table VII The F Distribution Table**

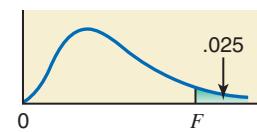
The entries in the table on this page give the critical values of  $F$  for .01 area in the right tail under the  $F$  distribution curve and specified degrees of freedom for the numerator and denominator.



	Degrees of Freedom for the Numerator																		
	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	50	100
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6157	6209	6240	6261	6287	6303	6334
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.24
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.58
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.13
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56	7.40	7.30	7.23	7.14	7.09	6.99
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.75
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52	5.36	5.26	5.20	5.12	5.07	4.96
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.41
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.01
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.71
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.47
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.27
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.11
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.52	3.37	3.28	3.21	3.13	3.08	2.98
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.86
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.76
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.68
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.60
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.54
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.48
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.42
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.37
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.33
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.29
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.13
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52	2.37	2.27	2.20	2.11	2.06	1.94
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.82
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.60

**Table VII The F Distribution Table (continued)**

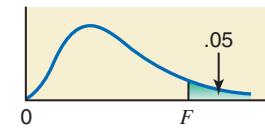
The entries in the table on this page give the critical values of  $F$  for .025 area in the right tail under the  $F$  distribution curve and specified degrees of freedom for the numerator and denominator.



Degrees of Freedom for the Denominator	Degrees of Freedom for the Numerator																		
	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	50	100
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	973.0	976.7	984.9	993.1	998.1	1001	1006	1008	1013
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.96
4	12.22	10.65	9.98	9.61	6.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.32
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.08
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.92
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.21
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.74
9	7.21	5.72	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.40
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.15
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.33	3.23	3.16	3.12	3.06	3.03	2.96
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.80
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.67
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.56
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.47
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.40
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.33
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.27
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.22
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.17
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.13
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.09
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.06
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.02
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.00
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.88
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.74
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.66
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.48

**Table VII The F Distribution Table (continued)**

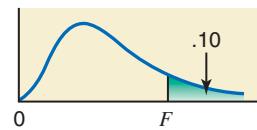
The entries in the table on this page give the critical values of  $F$  for .05 area in the right tail under the  $F$  distribution curve and specified degrees of freedom for the numerator and denominator.



	Degrees of Freedom for the Numerator																		
	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	50	100
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	246.0	248.0	249.3	250.1	251.1	251.8	253.0
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.11	3.08	3.04	3.02	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.76
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.59
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.46
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.35
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.26
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.19
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.07
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.14	2.11	2.06	2.04	1.98
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.94
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.91
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.88
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.02	1.97	1.94	1.91	1.85
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.82
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.16	2.03	1.97	1.94	1.89	1.86	1.80
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.78
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.70
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.59
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.52
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.39

**Table VII The F Distribution Table (continued)**

The entries in the table on this page give the critical values of  $F$  for .10 area in the right tail under the  $F$  distribution curve and specified degrees of freedom for the numerator and denominator.



Degrees of Freedom for the Denominator	Degrees of Freedom for the Numerator																		
	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	50	100
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	61.22	61.74	62.05	62.26	62.53	62.69	63.01
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.14
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.78
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.13
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.75
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.50
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.32
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.19
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.09
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.01
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.94
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.88
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.83
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.79
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.76
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.73
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.70
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.93	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.67
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.65
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.90	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.63
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.61
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.87	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.59
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.58
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.84	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.56
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.51
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.43
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.70	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.39
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.64	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.29



## Statistical Tables on the Web Site

Note: The following tables are on the Web site of the text along with Chapters 14 and 15.

Table VIII Critical Values of  $X$  for the Sign Test

Table IX Critical Values of  $T$  for the Wilcoxon Signed-Rank Test

Table X Critical Values of  $T$  for the Wilcoxon Rank Sum Test

Table XI Critical Values for the Spearman Rho Rank Correlation Coefficient Test

Table XII Critical Values for a Two-Tailed Runs Test with  $\alpha = .05$

# Answers To Selected Odd-Numbered Exercises and Self-Review Tests

(Note: Due to differences in rounding, the answers obtained by readers may differ slightly from the ones given in this Appendix.)

## Chapter 1

- 1.7 a. population    b. sample    c. population  
d. population    e. sample
- 1.11 a. number of dog bites reported last year  
b. six observations    c. six elements
- 1.15 a. quantitative    b. quantitative    c. qualitative  
d. qualitative    e. quantitative
- 1.17 a. continuous    b. continuous  
e. continuous
- 1.21 a. cross-section data    b. cross-section data  
c. time-series data    d. time-series data
- 1.23 a.  $\Sigma f = 69$     b.  $\Sigma m^2 = 1363$     c.  $\Sigma mf = 922$   
d.  $\Sigma m^2 f = 17,128$
- 1.25 a.  $\Sigma x = 120$     b.  $\Sigma y = 45$     c.  $\Sigma xy = 237$   
d.  $\Sigma y^2 = 285$     e.  $(\Sigma y)^2 = 2025$
- 1.27 a.  $\Sigma x = 856$     b.  $(\Sigma x)^2 = 732,736$   
c.  $\Sigma x^2 = 157,574$
- 1.29 a.  $\Sigma x = 2847$     b.  $(\Sigma x)^2 = 8,105,409$   
c.  $\Sigma x^2 = 1,158,777$
- 1.33 a. sample    b. population for the year  
c. sample    d. population
- 1.35 a. sampling without replacement  
b. sampling with replacement
- 1.37 a.  $\Sigma x = 47$     b.  $(\Sigma x)^2 = 2209$     c.  $\Sigma x^2 = 443$
- 1.39 a.  $\Sigma m = 59$     b.  $\Sigma f^2 = 2662$     c.  $\Sigma mf = 1508$   
d.  $\Sigma m^2 f = 24,884$     e.  $\Sigma m^2 = 867$
- 1.41 draft round: quantitative, discrete  
40-yard-dash speed: quantitative, continuous  
position: qualitative  
drafting team's current payroll: quantitative, continuous  
power ratio: quantitative, continuous  
quality starter: qualitative  
standing high jump: quantitative, continuous

## Self-Review Test

1. b    2. c
3. a. sampling without replacement  
b. sampling with replacement
4. a. qualitative    b. quantitative (continuous)  
c. quantitative (discrete)    d. qualitative
6. a.  $\Sigma x = 33$     b.  $(\Sigma x)^2 = 1089$     c.  $\Sigma x^2 = 231$
7. a.  $\Sigma m = 35$     b.  $\Sigma f = 429$     c.  $\Sigma m^2 = 203$   
d.  $\Sigma mf = 1315$     e.  $\Sigma m^2 f = 4345$   
f.  $(\Sigma f)^2 = 184,041$

## Chapter 2

- 2.3 c. 26.7%    d. 73.3%
- 2.5 c. 42.2%    2.7 c. 50%    2.15 d. 82.2%
- 2.17 a. class limits: \$1–\$25, \$26–\$50, \$51–\$75, \$76–\$100, \$101–\$125, \$126–\$150    b. class boundaries: \$5, \$25.5, \$50.5, \$75.5, \$100.5, \$125.5, \$150.5; width = \$25    c. class midpoints: \$13, \$38, \$63, \$88, \$113, \$138
- 2.19 d. 60.7%    2.29 c. .792
- 2.35 c. 43.1%    e. about 86.2%    2.43 12 teams
- 2.47 218, 245, 256, 329, 367, 383, 397, 404, 427, 433, 471, 523, 537, 551, 563, 581, 592, 622, 636, 647, 655, 678, 689, 810, 841
- 2.67 d. 50%    2.69 c. 16.7%
- 2.71 c. 56.7%
- 2.73 d. Boundaries of the fourth class are \$4200.5 and \$5600.5; width = \$1400.
- 2.87 No. The older group may drive more miles per week than the younger group.

## Self-Review Test

2. a. 5    b. 7    c. 17    d. 6.5    e. 13  
f. 90    g. .30
4. c. 35%    5. c. 70.8%
8. 30, 33, 37, 42, 44, 46, 47, 49, 51, 53, 53, 56, 60, 67, 67, 71, 79

## Chapter 3

- 3.5 mode    3.9 mean = 3.00; median = 3.50; no mode
- 3.11 mean = \$3881.67; median = \$3250
- 3.13 a. mean = \$289.04 billion; median = \$173.5 billion  
b. mode = \$49 billion
- 3.15 mean = \$1919.71 million; median = \$485 million
- 3.17 mean = \$9.42 million;  
median = \$7.60 million; no mode
- 3.19 mean = 2.92 power outages; median = 2.5 power outages; mode = 2 power outages
- 3.21 mean = 29.4; median = 28.5; mode = 23
- 3.23 a. mean = 1803; median = 1270    b. outlier = 5490; when the outlier is dropped: mean = 1467.8; median = 1166; mean changes by a larger amount  
c. median
- 3.25 combined mean = \$148.89    3.27 total = \$1055
- 3.29 age of the sixth person = 48 years

## AN2 Answers to Selected Odd-Numbered Exercises and Self-Review Tests

- 3.31** mean for data set I = 24.60; mean for data set II = 31.60  
The mean of the second data set is equal to the mean of the first data set plus 7.
- 3.33** 10% trimmed mean = 38.25 years
- 3.35** weighted mean = 77.5
- 3.41** range = 25;  $\sigma^2 = 61.5$ ;  $\sigma = 7.84$
- 3.43** a.  $\bar{x} = 9$ ; deviations from the mean: -2, 1, -1, -6, 6, 3, -3, 2. The sum of these deviations is zero.  
b. range = 12;  $s^2 = 14.2857$ ;  $s = 3.78$
- 3.45** range = 4;  $s^2 = 1.6319$ ;  $s = 1.28$
- 3.47** range = 22 indictments;  $s^2 = 43.2$ ;  $s = 6.57$  indictments
- 3.49** range = 17 women;  $s^2 = 27.9697$ ;  $s = 5.29$  women
- 3.51** range = 30;  $s^2 = 107.4286$ ;  $s = 10.36$
- 3.53** range = 38;  $s^2 = 135.9015$ ;  
 $s = 11.66$   
 $s = 0$
- 3.57** CV for salaries = 10.94%; CV for years of experience = 13.33%; The relative variation in salaries is lower.  
 $s = 14.64$  for both data sets
- 3.63**  $\bar{x} = 9.40$ ;  $s^2 = 37.7114$ ;  $s = 6.14$
- 3.65**  $\mu = 11.24$  hours;  $\sigma^2 = 36.3824$ ;  $\sigma = 6.03$  hours
- 3.67**  $\bar{x} = 19.67$ ;  $s^2 = 67.6979$ ;  $s = 8.23$
- 3.69**  $\bar{x} = 36.80$  minutes;  $s^2 = 597.7143$ ;  $s = 24.45$  minutes
- 3.71**  $\bar{x} = 13.03$  hours;  $s^2 = 78.2648$ ;  $s = 8.85$  hours
- 3.75** at least 75%; at least 84%; at least 89%
- 3.77** 68%; 95%; 99.7%
- 3.79** a. at least 75%    b. at least 84%  
c. at least 89%
- 3.81** a. i. at least 75%    ii. at least 89%  
b. \$1515 to \$3215
- 3.83** a. 99.7%    b. 68%    c. 95%
- 3.85** a. i. 99.7%    ii. 68%    b. 66 to 78 mph
- 3.91** a.  $Q_1 = 69$ ;  $Q_2 = 73$ ;  $Q_3 = 76.5$ ;  $IQR = 7.5$   
b.  $P_{35} = 71$     c. 30.77%
- 3.93** a.  $Q_1 = 300$ ;  $Q_2 = 322.5$ ;  $Q_3 = 347$ ;  $IQR = 47$   
b.  $P_{57} = 330$     c. 40%
- 3.95** a.  $Q_1 = 25$ ;  $Q_2 = 28.5$ ;  $Q_3 = 33$ ;  $IQR = 8$   
b.  $P_{65} = 31$     c. 33.33%
- 3.97** a.  $Q_1 = 533$ ;  $Q_2 = 626.5$ ;  $Q_3 = 728$ ;  $IQR = 195$   
b.  $P_{30} = 572$     c. 22.73%
- 3.99** no outlier
- 3.109** a. mean = \$106.5 thousand; median = \$76 thousand  
b. outlier = 382; when the outlier is dropped: mean = \$75.9 thousand; median = \$74 thousand; mean changes by a larger amount    c. median
- 3.111** a. mean = 1889.4 points; median = 1902.5 points;  
mode none    b. range = 539 points;  
 $s^2 = 26,219.98$ ;  $s = 161.93$  points
- 3.113**  $\bar{x} = 5.08$  inches;  $s^2 = 6.8506$ ;  $s = 2.62$  inches
- 3.115** a. i. at least 75%    ii. at least 89%  
b. 160 to 240 minutes
- 3.117** a. i. 68%    ii. 95%    b. 140 to 260 minutes
- 3.119** a.  $Q_1 = 60$ ;  $Q_2 = 76$ ;  $Q_3 = 97$ ;  $IQR = 37$   
b.  $P_{70} = 84$     c. 70%
- 3.121** The data set is skewed slightly to the right; 135 is an outlier.
- 3.123** The minimum score is 169.
- 3.125** a. new mean = 76.4 inches; new median = 78 inches; new range = 13 inches    b. new mean = 75.2 inches
- 3.127** mean = \$94.85 per barrel
- 3.129** a. trimmed mean = 9.5    b. 14.3%

- 3.131** a. age 30 and under: rate for A = 25; rate for B = 20    b. age 31 and over: rate for A = 100; rate for B = 85.7    c. overall: rate for A = 50; rate for B = 58.3    d. Country A has the lower overall average because 66.67% of its population is under 30.
- 3.133** a.  $k = 1.41$     b.  $k = 2.24$     **3.135** b. median
- 3.137** b. For men: mean = 82, median = 79, modes = 75, 79, and 92,  $s = 12.08$ ,  $Q_1 = 73.5$ ,  $Q_3 = 89.5$ , and  $IQR = 16$ . For women: mean = 97.53, median = 98, modes = 94 and 100,  $s = 8.44$ ,  $Q_1 = 94$ ,  $Q_3 = 101$ , and  $IQR = 7$
- 3.139** a. mean = 30    b. mean = 50
- 3.141** a. at least 55.56%    b. 1 to 11 inches  
c. 2.66 to 9.34 inches
- 3.143** a. For men: mean = 174.91 lbs = 76,189.05 grams = 12.49 stone, median = 179 lbs = 77,970.61 grams = 12.79 stone, and st. dev. = 19.12 lbs = 8328.48 grams = 1.37 stone. For women: mean = 124.95 lbs = 54,426.97 grams = 8.93 stone, median = 123 lbs = 53,577.57 grams = 8.79 stone, st. dev. = 17.48 lbs = 7614.11 grams = 1.25 stone.    b. See answer to a, as answers are identical.  
c. yes    d & e. Smaller unit has more variability.
- 3.145** 108 to 111

## Self-Review Test

1. b    2. a and d    3. c    4. c    5. b  
6. b    7. a    8. a    9. b    10. a    11. b  
12. c    13. a    14. a  
15. mean = 14.1; median = 13.5; modes = 13,22; range = 21;  $\sigma^2 = 39.2653$ ;  $\sigma = 6.27$   
19. b.  $\bar{x} = 19.46$ ;  $s^2 = 44.0400$ ;  $s = 6.64$   
20. a. i. at least 75%    ii. at least 84%  
b. 43.2 to 140.4 minutes  
21. a. i. 68%    ii. 99.7%    b. 2.9 to 11.7 years  
22. a.  $Q_1 = 3$ ;  $Q_2 = 8$ ;  $Q_3 = 13$ ;  $IQR = 10$   
b.  $P_{60} = 10$     c. 66.67%  
23. Data are skewed slightly to the right.  
24. combined mean = \$1066.43  
25. GPA of fifth student = 3.17  
26. 10% trimmed mean = 376.625; trimmed mean is a better measure  
27. a. mean for data set I = 19.75; mean for data set II = 16.75. The mean of the second data set is equal to the mean of the first data set minus 3.    b.  $s = 11.32$  for both data sets.

## Chapter 4

- 4.3  $S = \{\text{AB, AC, BA, BC, CA, CB}\}$   
4.5 four possible outcomes;  $S = \{\text{NN, NI, IN, II}\}$   
4.7 four possible outcomes;  $S = \{\text{DD, DG, GD, GG}\}$   
4.9  $S = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$   
4.11 a. {NI and IN}; a compound event  
b. {II, NI, and IN}; a compound event  
c. {NN, IN, and NI}; a compound event  
d. {IN}; a simple event  
4.13 a. {DG, GD, and GG}; a compound event  
b. {DG and GD}; a compound event  
c. {GD}; a simple event  
d. {DD, DG, and GD}; a compound event  
4.19 2.4, - .63,  $9/4$ ,  $-2/9$   
4.21 not equally likely outcomes; use relative frequency approach

- 4.23** subjective probability  
**4.25** a. .450 b. .550  
**4.27** .660 **4.29** .160  
**4.31** a. .250 b. .750  
**4.33** .9094; .0906  
**4.35** .325; .675  
**4.37** a. .4285 b. .4986 c. .0728  
**4.39** use relative frequency approach  
**4.45** a. no; no b. no; yes; no  
c.  $\bar{A} = \{a, c, f, g, h, i, k\}; P(\bar{A}) = .636$   
 $\bar{B} = \{b, d, e, g, h, i, k\}; P(\bar{B}) = .636$   
 $\bar{C} = \{a, b, d, e, f, h, i, j\}; P(\bar{C}) = .727$   
**4.47** a. i. .600 ii. .600 iii. .375 iv. .583  
b. Events “male” and “female” are mutually exclusive.  
Events “have shopped” and “male” are not mutually exclusive.  
c. Events “female” and “have shopped” are dependent.  
**4.49** a. i. .3475 ii. .5425 iii. .2727  
iv. .4545 b. Events “male” and “in favor” are not mutually exclusive. Events “in favor” and “against” are mutually exclusive.  
c. Events “female” and “no opinion” are dependent.  
**4.51** a. i. .1012 ii. .4835 iii. .5524  
iv. .1014 b. Events “Airline A” and “more than 1 hour late” are not mutually exclusive. Events “less than 30 minutes late” and “more than one hour late” are mutually exclusive.  
c. Events “Airline B” and “30 minutes to 1 hour late” are dependent.  
**4.53** Events “female” and “pediatrician” are dependent but not mutually exclusive.  
**4.55** Events “female” and “first 5K race” are dependent but not mutually exclusive.  
**4.57**  $P(A) = .3333; P(\bar{A}) = .6667$  **4.59** .88  
**4.65** a. .6006 b. .0084  
**4.67** a. .1885 b. .0084  
**4.69** a. .1050 b. .1200  
**4.71** .8276 **4.73** .500  
**4.75** a. i. .3844 ii. .1590 b. .0000  
**4.77** a. i. .350 ii. .150  
**4.79** a. i. .3147 ii. .2071 b. .0000  
**4.81** .3529 **4.83** .1110; .4302 **4.85** .1600  
**4.87** a. .0025 b. .9025 **4.89** .5120  
**4.91** .5278 **4.93** .40  
**4.99** a. .59 b. .49 **4.101** a. .74 b. .82  
**4.103** a. .6358 b. .9075  
**4.105** a. .750 b. .750 c. 1.0  
**4.107** a. .780 b. .550 c. .790  
**4.109** .344 **4.111** .77  
**4.113** .700 **4.115** .80 **4.117** .9744 **4.119** 1024  
**4.121**  $6! = 720; 11! = 39,916,800; (7 - 2)! = 120;$   
 $(15 - 5)! = 3,628,800; {}_8C_2 = 28; {}_5C_0 = 1; {}_5C_5 = 1;$   
 ${}_6C_4 = 15; {}_{11}C_7 = 330; {}_9P_6 = 60,480; {}_{12}P_8 = 19,958,400$   
**4.123** 384  
**4.125** 240  
**4.127**  ${}_{25}C_2 = 300; {}_{25}P_2 = 600$   
**4.129**  ${}_{25}C_4 = 12,650; {}_{25}P_4 = 303,600$   
**4.131**  ${}_{16}C_2 = 120; {}_{16}P_2 = 240$   
**4.133**  ${}_{15}C_5 = 3003$   
**4.135** a. .2571 b. .1429
- 4.137** a. i. .4360 ii. .4800 iii. .3462  
iv. .6809 v. .3400 vi. .6600  
b. Events “female” and “prefers watching sports” are dependent but not mutually exclusive.  
**4.139** a. i. .750 ii. .700 iii. .225 iv. .775  
b. Events “student athlete” and “should be paid” are dependent but not mutually exclusive.  
**4.141** a. .5118 b. .4882  
**4.143** .0605  
**4.145** .0048 **4.147** a. 17,576,000 b. 5200  
**4.149** a.  $1/195,249,054 = .0000000051$   
b.  $1/5,138,133 = .00000019$   
**4.151** a. .5000 b. .3333 c. No; the sixth toss is independent of the first five tosses. Equivalent to part a.  
**4.153** a. .030 b. .150  
**4.155** a. .50 b. .50 **4.157** a. .8333 b. .1667  
**4.159** a. .0001 b. i. .0024 ii. .0012  
iii. .0006 iv. .0004 **4.161** a. .8851  
b. .0035  
**4.163** a. 1,099,511,627,776 b. 466,560,000  
c. .999957

### Self-Review Test

1. a 2. b 3. c 4. a 5. a 6. b
7. c 8. b 9. b 10. c 11. b
12. 120 **13.** a. .3333 b. .6667
14. a. Events “female” and “out of state” are dependent but not mutually exclusive. b. i. .4500 ii. .6364
15. .825 **16.** .3894 **17.** .4225 **18.** .40; .60
19. a. .279 b. .829
20. a. i. .358 ii. .405 iii. .235 iv. .5593  
b. Events “woman” and “yes” are dependent but not mutually exclusive.

### Chapter 5

- 5.3** a. continuous random variable b. discrete random variable c. continuous random variable  
d. discrete random variable e. continuous random variable f. continuous random variable  
**5.5** discrete random variable  
**5.9** a. not a valid probability distribution b. a valid probability distribution c. not a valid probability distribution  
**5.11** a. .17 b. .20 c. .58 d. .42  
e. .42 f. .27 g. .68  
**5.13** b. i. .51 ii. .235 iii. .285 iv. .305  
**5.15** a. 

$x$	1	2	3	4	5
$P(x)$	.10	.25	.30	.20	.15

b. approximate c. i. .30 ii. .65  
iii. .75 iv. .65  
**5.17**

$x$	0	1	2
$P(x)$	.7039	.2702	.0259

  
**5.19**

$x$	0	1	2
$P(x)$	.9274	.0712	.0014

  
**5.21**

$x$	0	1	2
$P(x)$	.4789	.4422	.0789

  
**5.23** a.  $\mu = 1.590; \sigma = .960$  b.  $\mu = 7.070; \sigma = 1.061$

## AN4 Answers to Selected Odd-Numbered Exercises and Self-Review Tests

- 5.25**  $\mu = .440$  error;  $\sigma = .852$  error  
**5.27**  $\mu = 2.94$  camcorders;  $\sigma = 1.441$  camcorders  
**5.29**  $\mu = 1.00$  head;  $\sigma = .707$  head  
**5.31**  $\mu = 2.561$  tires;  $\sigma = 1.322$  tires  
**5.33**  $\mu = .100$  lemon;  $\sigma = .308$  lemon  
**5.35**  $\mu = \$3.9$  million;  $\sigma = \$3.015$  million  
**5.37**  $\mu = .500$  person;  $\sigma = .584$  person  
**5.41** a. not a binomial experiment  
 b. a binomial experiment  
 c. a binomial experiment  
**5.43** a. .2541    b. .1536    c. .3241  
**5.45** b.  $\mu = 2.100$ ;  $\sigma = 1.212$   
**5.49** a. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17  
 b. .1540  
**5.51** a. .7095    b. .7332    c. .5000  
**5.53** a. .0750    b. .0000    c. .1836  
**5.55** a. .2725    b. .0839  
**5.57** a.  $\mu = 5.6$  customers;  $\sigma = 1.058$  customers    b. .1147  
**5.59** a.  $\mu = 5.600$  customers;  $\sigma = 1.296$  customers  
 b. .0467  
**5.61** a. .4286    b. .0714    c. .5  
**5.63** a. .3818    b. .0030    c. .5303  
**5.65** a. .4747    b. .0440    c. .3407  
**5.67** a. .1078    b. .5147    c. .8628  
**5.71** a. .0404    b. .2565  
**5.73** a.  $\mu = 1.3$ ;  $\sigma^2 = 1.3$ ;  $\sigma = 1.140$     b.  $\mu = 2.1$ ;  
 $\sigma^2 = 2.1$ ;  $\sigma = 1.449$   
**5.75** .1496    **5.77** .1185  
**5.79** a. .1162    b. i. .6625    ii. .1699  
 iii. .4941  
**5.81** a. .3033    b. i. .0900    ii. .0018  
 iii. .9098  
**5.83** a. i. .0629    ii. .0722  
 b. i. .9719    ii. .6400    iii. .5718  
**5.85** a. .2466    c.  $\mu = 1.4$   $\sigma^2 = 1.4$   
 $\sigma = 1.183$   
**5.87** a. .0446    b. i. .0390    ii. .2580    iii. .0218  
**5.89**  $\mu = 4.11$ ;  $\sigma = 1.019$ ; This mechanic repairs, on average,  
 4.11 cars per day.  
**5.91** b.  $\mu = \$557,000$ ;  $\sigma = \$1,288,274$ ;  $\mu$  gives the  
 company's expected profit.  
**5.93** a. .0000    b. .0351    c. .7214  
**5.95** a. .9246    b. .0754  
**5.97** a. .3692    b. .1429    c. .0923  
**5.99** a. .8643    b. .1357  
**5.101** a. .0912    b. i. .5502    ii. .0817    iii. .2933  
**5.103** a. .2466  
**5.105**  $\Sigma x P(x) = -2.22$ . This game is not fair to you, and you  
 should not play, as you expect to lose an average of \$2.22  
 per play.  
**5.107** a. .0625    b. .125    c. .3125  
**5.109** c. .7149    d. 3 nights  
**5.111** 8 cheesecakes  
**5.113** a. 35    b. 10    c. .2857    **5.117** \$6  
**5.119** a. .0211    b. .0475    c. .4226
8. b    9. a    10. c    12. a  
**14.**  $\mu = 2.040$  homes;  $\sigma = 1.449$  homes  
**15.** a. i. .2128    ii. .8418    iii. .0153  
 b.  $\mu = 7.2$  adults;  $\sigma = 1.697$  adults  
**16.** a. .4525    b. .0646    c. .0666  
**17.** a. i. .0521    ii. .2203    iii. .2013

## Chapter 6

- 6.11** .8664    **6.13** .9876  
**6.15** a. .4744    b. .4798    c. .1162    d. .0610  
 e. .9400  
**6.17** a. .0869    b. .0244    c. .9798    d. .9608  
**6.19** a. .5 approximately    b. .5 approximately  
 c. .00 approximately    d. .00 approximately  
**6.21** a. .9613    b. .4783    c. .4767    d. .0694  
**6.23** a. .0096    b. .2466    c. .1570    d. .9625  
**6.25** a. .8365    b. .8947    c. approximately .5  
 d. approximately .5    e. approximately .00  
 f. approximately .00  
**6.27** a. 1.80    b. -2.20    c. -1.20    d. 2.80  
**6.29** a. .4599    b. .1598    c. .2223  
**6.31** a. .3336    b. .9564    c. .9686  
 d. approximately .00  
**6.33** a. .2178    b. .6440  
**6.35** a. .8212    b. .2810    c. .0401    d. .7190  
**6.37** a. .0764    b. .1126  
**6.39** a. .0838    b. .7026  
**6.41** a. 93.32%    b. 15.57%  
**6.43** a. .0359    b. .1515  
**6.45** a. .8264    b. 12.83%  
**6.47** a. 15.62%    b. 7.64%  
**6.49** a. .11%    b. 49.06%    c. .69%    d. 47.78%  
**6.51** 2.64%  
**6.53** a. 2.00    b. -2.02 approximately  
 c. -.37 approximately    d. 1.02 approximately  
**6.55** a. approximately 1.65    b. -1.96    c. -2.33  
 approximately    d. 2.58 approximately  
**6.57** a. 208.50    b. 241.25    c. 178.50  
 d. 145.75    e. 158.25    f. 251.25  
**6.59** 19 minutes approximately  
**6.61** 2060 kilowatt-hours  
**6.63** \$153.99 approximately  
**6.65**  $np > 5$  and  $nq > 5$   
**6.67** a. .7688    b. .7697; difference is .0009  
**6.69** a.  $\mu = 72$ ;  $\sigma = 5.36656315$     b. .3192  
 c. .4564  
**6.71** a. .0764    b. .6793    c. .8413    d. .8238  
**6.73** .0901    **6.75** a. .0568    b. .9671    c. .8903  
**6.77** a. .0454    b. .0516    c. .8646  
**6.79** a. .7549    b. .2451  
**6.81** a. .1093    b. 9.31%    c. 57.33%  
 d. It is possible, but its probability is close to zero.  
**6.83** .0124 or 1.24%  
**6.85** a. 8304 hours    b. 8132 hours approximately  
**6.87** \$121,660  
**6.89** a. .0151    b. .0465    c. .8340    d. .2540  
**6.91** 16.23 oz  
**6.93** .0637  
**6.95** Plant B  
**6.97** 8:10 am

## Self-Review Test

2. probability distribution table  
 3. a    4. b    6. b    7. a

- 6.99** a. 106.32    b. .0808  
**6.101** a. single-number bet    b. single-number bet: .4866  
     color bet: .3974  
**6.103** a. .0005    b. .7714  
**6.105** a. .0375    b. .1952    c. .6624  
     d. .1679

**Self-Review Test**

1. a    2. a    3. d    4. b    5. a    6. c  
 7. b    8. b  
 9. a. .1878    b. .9304    c. .0985    d. .7704  
**10.** a. -1.28 approximately    b. .61    c. 1.65  
     approximately    d. -1.07 approximately  
**11.** a. .5608    b. .0015    c. .0170    d. .1165  
**12.** a. 48669.8    b. 40162  
**13.** a. i. .0318    ii. .9453    iii. .9099  
     iv. .0268    v. .4632    b. .7054    c. .3986

**Chapter 7**

- 7.5** a. 16.60    b. sampling error = -.27  
     c. sampling error = -.27; nonsampling error = 1.11  
     d.  $\bar{x}_1 = 16.22$ ;  $\bar{x}_2 = 15.67$ ;  $\bar{x}_3 = 17.00$ ;  $\bar{x}_4 = 16.33$ ;  
      $\bar{x}_5 = 17.44$ ;  $\bar{x}_6 = 16.78$ ;  $\bar{x}_7 = 17.22$ ;  
      $\bar{x}_8 = 17.67$ ;  $\bar{x}_9 = 16.56$ ;  $\bar{x}_{10} = 15.11$   
**7.7** b.  $\bar{x}_1 = 22.25$ ;  $\bar{x}_2 = 28.50$ ;  $\bar{x}_3 = 29.00$ ;  $\bar{x}_4 = 29.25$ ;  
      $\bar{x}_5 = 30.25$ ;  $\bar{x}_6 = 29.75$ ;  $\bar{x}_7 = 30.75$ ;  $\bar{x}_8 = 32.25$ ;  
      $\bar{x}_9 = 31.75$ ;  $\bar{x}_{10} = 36.00$ ;  $\bar{x}_{11} = 37.00$ ;  
      $\bar{x}_{12} = 37.75$ ;  $\bar{x}_{13} = 38.50$ ;  $\bar{x}_{14} = 39.25$ ;  
      $\bar{x}_{15} = 40.25$ ;    c.  $\mu = 32.83$   
**7.13** a.  $\mu_{\bar{x}} = 60$ ;  $\sigma_{\bar{x}} = 2.357$     b.  $\mu_{\bar{x}} = 60$ ;  $\sigma_{\bar{x}} = 1.054$   
**7.15** a.  $\sigma_{\bar{x}} = 1.400$     b.  $\sigma_{\bar{x}} = 2.500$   
**7.17** a.  $n = 100$     b.  $n = 256$   
**7.19**  $\mu_{\bar{x}} = \$25,000$ ;  $\sigma_{\bar{x}} = \$314$   
**7.21**  $\mu_{\bar{x}} = \$25,510$ ;  $\sigma_{\bar{x}} = \$321.73$   
**7.23**  $n = 256$   
**7.25** a.  $\mu_{\bar{x}} = 80.60$     b.  $\sigma_{\bar{x}} = 3.302$     d.  $\sigma_{\bar{x}} = 3.302$   
**7.33**  $\mu_{\bar{x}} = 20.20$  hours;  $\sigma_{\bar{x}} = .613$  hours; the normal  
     distribution  
**7.35**  $\mu_{\bar{x}} = 3.020$ ;  $\sigma_{\bar{x}} = .042$ ; approximately normal  
     distribution  
**7.37**  $n = 25$ :  $\mu_{\bar{x}} = 28.2$  years;  $\sigma_{\bar{x}} = 1.2$  years; skewed to  
     the right  
 $n = 100$ :  $\mu_{\bar{x}} = 28.2$  years;  $\sigma_{\bar{x}} = .6$  year;  
     approximately normal distribution  
**7.39**  $\mu_{\bar{x}} = 151$  min;  $\sigma_{\bar{x}} = 1.414$  min; approximately normal  
     distribution; no, sample size  $\geq 30$   
**7.41** 86.64%  
**7.43** a.  $z = 2.44$     b.  $z = -7.25$     c.  $z = -3.65$   
     d.  $z = 5.82$   
**7.45** a. .1940    b. .8749  
**7.47** a. .0003    b. .9292  
**7.49** a. .1093    b. .0322    c. .7776  
**7.51** a. .0559    b. .0222    c. .7812  
**7.53** a. .8203    b. .9750  
**7.55** a. .1147    b. .9164    c. .1251  
**7.57** a. .1032    b. .3172    c. .0016    d. .9049  
**7.59** .0124    **7.61**  $p = .12$ ;  $\hat{p} = .15$   
**7.63** 7125 subjects in the population; 312 subjects in the  
     sample

- 7.65** sampling error = -.05  
**7.71** a.  $\mu_{\hat{p}} = .21$ ;  $\sigma_{\hat{p}} = .020$   
     b.  $\mu_{\hat{p}} = .21$ ;  $\sigma_{\hat{p}} = .015$   
**7.73** a.  $\sigma_{\hat{p}} = .051$     b.  $\sigma_{\hat{p}} = .071$   
**7.77** a.  $p = .667$     b. 6    d. -0.67, -0.67, .133, .133,  
     -.067, -.067  
**7.79**  $\mu_{\hat{p}} = .86$ ;  $\sigma_{\hat{p}} = .017$ ; approximately normal distribution  
**7.81**  $\mu_{\hat{p}} = .65$ ;  $\sigma_{\hat{p}} = .019$ ; approximately normal distribution  
**7.83** 95.44%  
**7.85** a.  $z = -.61$     b.  $z = 1.83$     c.  $z = -1.22$   
     d.  $z = 1.22$   
**7.87** a. .1251    b. .1147  
**7.89** a. .9649    b. .8789  
**7.91** .1515  
**7.93**  $\mu_{\bar{x}} = 8000$  hours;  $\sigma_{\bar{x}} = 80$  hours; the normal  
     distribution  
**7.95** a. .0838    b. .0991    c. .8968    d. .0301  
**7.97** a. .0582    b. .8325    c. .9991    d. .0045  
**7.99**  $\mu_{\hat{p}} = .88$ ;  $\sigma_{\hat{p}} = .036$ ; approximately normal distribution  
**7.101** a. i. .9788    ii. .8903    b. .9090    c. .0212  
**7.103** .6778  
**7.105** 10 approximately  
**7.107** a. .8023    b. 754 approximately  
**7.109** .0035

**Self-Review Test**

1. b    2. b    3. a    4. a    5. b  
 6. b    7. c    8. a    9. a  
**10.** a. **11.** c    **12.** a  
**14.** a.  $\mu_{\bar{x}} = 145$  pounds;  $\sigma_{\bar{x}} = 3.600$  pounds; approximately  
     normal distribution  
     b.  $\mu_{\bar{x}} = 145$  pounds;  $\sigma_{\bar{x}} = 1.800$  pounds; approximately  
     normal distribution  
**15.** a.  $\mu_{\bar{x}} = \$650,000$ ;  $\sigma_{\bar{x}} = \$31,305$ ; unknown distribution  
     b.  $\mu_{\bar{x}} = \$650,000$ ;  $\sigma_{\bar{x}} = \$14,000$ ; approximately normal  
     distribution  
     c.  $\mu_{\bar{x}} = \$650,000$ ;  $\sigma_{\bar{x}} = \$7000$ ; approximately normal  
     distribution  
**16.** a. .1261    b. .9128    c. .9236    d. .1528  
     e. .2389    f. .7611    g. .0764    h. .6188  
**17.** a. i. .1203    ii. .1335    iii. .7486  
     b. .9736    c. .0013  
**18.** a.  $\mu_{\hat{p}} = .15$   $\sigma_{\hat{p}} = .065$ ; unknown distribution  
     b.  $\mu_{\hat{p}} = .15$   $\sigma_{\hat{p}} = .021$ ; approximately normal  
     distribution  
     c.  $\mu_{\hat{p}} = .15$   $\sigma_{\hat{p}} = .007$ ; approximately normal  
     distribution  
**19.** a. i. .0869    ii. .8919    iii. .0212  
     iv. .1450    v. .7517    vi. .7517  
     b. .9090    c. .0424    d. .0869

**Chapter 8**

- 8.11** a. 24.5    b. 22.71 to 26.29    c. 1.79  
**8.13** a. 70.59 to 79.01    b. 69.80 to 79.80  
     c. 68.22 to 81.38    d. yes  
**8.15** a. 77.84 to 85.96    b. 78.27 to 85.53  
     c. 78.65 to 85.15    d. yes  
**8.17** a. 38.34    b. 37.30 to 39.38    c. 1.04  
**8.19** a.  $n = 167$     b.  $n = 65$

## AN6 Answers to Selected Odd-Numbered Exercises and Self-Review Tests

- 8.21** a.  $n = 299$    b.  $n = 126$    c.  $n = 61$
- 8.23** \$191.37 to \$225.33
- 8.25** a. 48,903.27 to 58,196.73 labor-hours
- 8.27** 31.86 to 32.02 ounces; no adjustment needed
- 8.29** a. \$1532.41 to \$1617.59
- 8.31**  $n = 167$    **8.33**  $n = 72$
- 8.41** a.  $t = -1.325$    b.  $t = 2.160$    c.  $t = 3.281$   
d.  $t = -2.715$
- 8.43** a.  $\alpha \approx .10$ , left tail   b.  $\alpha = .005$ , right tail  
c.  $\alpha = .10$ , right tail   d.  $\alpha \approx .01$ , left tail
- 8.45** a.  $t = 2.080$    b.  $t = 1.671$    c.  $t = 2.807$
- 8.47** a. 1.41   b. -3.40 to 6.22   c. 4.81
- 8.49** a. 24.06 to 26.94   b. 23.58 to 27.42  
c. 23.73 to 27.27
- 8.51** a. 91.03 to 93.87   b. 90.06 to 93.44  
c. 88.06 to 91.20   d. confidence intervals of parts b and c cover  $\mu$ , that of part a does not
- 8.53** 40.04 to 42.36 bushels
- 8.55** 162.42 to 181.58 minutes
- 8.57** 18.64 to 25.36 minutes
- 8.59** a. 21.56 to 24.44 hours
- 8.61** 4.88 to 11.12 hours
- 8.63** a.  $\bar{x} = \$24.14$    b. \$17.11 to \$31.17
- 8.65** a. yes  
b.  $\bar{x} = 284.3$   
c. 265 to 304
- 8.71** a. yes, sample size is large   b. no, sample size is not large   c. yes, sample size is large   d. yes, sample size is large
- 8.73** a. .297 to .343   b. .336 to .384  
c. .277 to .323   d. confidence intervals of parts a and b cover  $p$ , but that of part c does not
- 8.75** a. .189 to .351   b. .202 to .338  
c. .218 to .322   d. yes
- 8.77** a. .284 to .336   b. .269 to .351  
c. .209 to .411   d. yes
- 8.79** a.  $n = 668$    b.  $n = 671$
- 8.81** a.  $n = 1432$    b.  $n = 196$    c.  $n = 353$
- 8.83** a. .29 to .45
- 8.85** a. 40%   b. 33.1% to 46.9%; margin of error = 6.9%
- 8.87** a. 20.3% to 55.7%   **8.89** a. .627 to .673
- 8.91** a. .333   b. 8.5% to 58.1%
- 8.93**  $n = 1084$
- 8.95**  $n = 1849$
- 8.99** a. \$2640   b. \$2514.57 to \$2765.43
- 8.101** 3.969 to 4.011 inches; the machine needs to be adjusted
- 8.103** 12.5 to 16.5 gallons
- 8.105** 21.76 to 26.24 minutes
- 8.107** 4.4 to 4.6 hours
- 8.109** 144.33 to 158.47 calories
- 8.111** a. .033   b. .016 to .050
- 8.113** 6.1% to 56.4%
- 8.115**  $n = 221$    **8.117**  $n = 359$
- 8.121**  $n = 74$
- 8.123** a.  $n = 20$  days   b. 90%   c. 75 cars
- 2.** b   **3.** a   **4.** a   **5.** c   **6.** b
- 7.** a. \$159,000  
b. \$147,390 to \$170,610; margin of error = \$11,610
- 8.** \$571,283.30 to \$649,566.70   **9.** a. .83  
b. .799 to .861
- 10.**  $n = 83$    **11.**  $n = 273$    **12.**  $n = 229$

## Chapter 9

- 9.5** a. a left-tailed test   b. a right-tailed test  
c. a two-tailed test
- 9.7** a. Type II error   b. Type I error
- 9.9** a.  $H_0: \mu = 20$  hours;  $H_1: \mu \neq 20$  hours;  
a two-tailed test   b.  $H_0: \mu = 10$  hours;  
 $H_1: \mu > 10$  hours; a right-tailed test  
c.  $H_0: \mu = 3$  years;  $H_1: \mu \neq 3$  years; a two-tailed test  
d.  $H_0: \mu = \$1000$ ;  $H_1: \mu < \$1000$ ;  
a left-tailed test   e.  $H_0: \mu = 12$  minutes;  
 $H_1: \mu > 12$  minutes; a right-tailed test
- 9.17** a.  $p$ -value = .0188   b.  $p$ -value = .0116  
c.  $p$ -value = .0087
- 9.19** a.  $p$ -value = .0166   b. no, do not reject  $H_0$   
c. yes, reject  $H_0$
- 9.21** a. rejection region is at and to the left of -2.58 and at and to the right of 2.58; nonrejection region is between -2.58 and 2.58   b. rejection region is at and to the left of -2.58; nonrejection region is to the right of -2.58  
c. rejection region is at and to the right of 1.96; nonrejection region is to the left of 1.96
- 9.23** Statistically not significant
- 9.25** a. .10   b. .02   c. .005
- 9.27** a. observed value of  $z$  is .58; critical values of  $z$  are  $\pm 1.96$   
b. observed value of  $z$  is .58; critical value of  $z$  is 1.65
- 9.29** a. reject  $H_0$  if  $z \geq 1.65$    b. reject  $H_0$  if  $z \leq -1.65$   
c. reject  $H_0$  if  $z \leq -1.96$  or  $z \geq 1.96$
- 9.31** a. critical value:  $z = -1.96$ ; test statistic:  $z = -2.67$ ;  
reject  $H_0$    b. critical value:  $z = -1.96$ ; test statistic:  $z = -1.00$ ; do not reject  $H_0$
- 9.33** a. critical values:  $z = -1.65$  and 1.65; test statistic:  $z = -1.34$ ; do not reject  $H_0$    b. critical value:  $z = -2.33$ ;  
test statistic:  $z = -6.44$ ; reject  $H_0$   
c. critical value:  $z = 1.65$ ; test statistic:  $z = 8.70$ ;  
reject  $H_0$
- 9.35** a.  $H_0: \mu = 45$  months;  $H_1: \mu < 45$  months;  $p$ -value = .0170; if  $\alpha = .025$  reject  $H_0$    b. test statistic:  $z = -2.12$ ; Critical value:  $z = -1.96$ ; reject  $H_0$
- 9.37** a.  $H_0: \mu = \$1038$ ;  $H_1: \mu > \$1038$ ;  $p$ -value = .0030; if  $\alpha = .025$ , reject  $H_0$   
b. Critical value:  $z = 1.96$ ; observed value:  $z = 2.75$ ;  
reject  $H_0$
- 9.39** a.  $H_0: \mu = 10$  minutes;  $H_1: \mu \neq 10$  minutes; test statistic:  $z = -2.11$ ;  $p$ -value = .0348. If  $\alpha = .02$ , do not reject  $H_0$ .  
If  $\alpha = .05$ , reject  $H_0$ .   b. Observed value  $z = -2.11$ ;  
If  $\alpha = .02$ , critical values:  $z = -2.33$  and 2.33; do not  
reject  $H_0$ . If  $\alpha = .05$ , critical values:  $z = -1.96$  and 1.96;  
reject  $H_0$ .
- 9.41** a. test statistic:  $z = -2.33$ ;  $p$ -value = .0198;  
If  $\alpha = .01$ , do not reject  $H_0$ ; If  $\alpha = .05$ , reject  $H_0$ .  
b. Observed value  $z = -2.33$ ; If  $\alpha = .01$ , critical values:  
 $z = -2.58$  and 2.58, do not reject  $H_0$ ; If  $\alpha = .05$ , critical  
values:  $z = -1.96$  and 1.96; reject  $H_0$ .

## Self-Review Test

1. a. population parameter; sample statistic  
b. sample statistic; population parameter  
c. sample statistic; population parameter

- 9.43** a.  $H_0: \mu \geq 47.93$  boxes;  $H_1: \mu < 47.93$  boxes; critical value:  $z = -1.28$ ; test statistic:  $z = -1.16$ ; do not reject  $H_0$   
b. do not reject  $H_0$ .
- 9.45** a.  $H_0: \mu \geq 8$  hours;  $H_1: \mu < 8$  hours; critical value:  $z = -2.33$ ;  $\alpha = .01$ ; test statistic:  $z = -.68$ ; p-value = .2483 do not reject  $H_0$ .    b. critical value:  $z = -1.96$ ; test statistic:  $z = -.68$ ; do not reject  $H_0$ .
- 9.49** a. reject  $H_0$  if  $t \leq -2.977$  or  $t \geq 2.977$     b. reject  $H_0$  if  $t \leq -2.797$     c. reject  $H_0$  if  $t \geq 2.080$
- 9.51** a. critical values:  $t = -2.365$  and  $2.365$ ; observed value:  $t = -2.097$ ;  $.05 < p\text{-value} < .10$     b. critical value:  $t = -1.895$ ; observed value:  $t = -2.097$ ;  $.025 < p\text{-value} < .05$
- 9.53** a. reject  $H_0$  if  $t \geq 1.672$     b. reject  $H_0$  if  $t \leq -1.672$   
c. reject  $H_0$  if  $t \leq -2.002$  or  $t \geq 2.002$
- 9.55** a. critical value:  $t = 1.998$ ; test statistic:  $t = 4.800$ ; reject  $H_0$     b. critical value:  $t = 1.998$ ; test statistic:  $t = 1.143$ ; do not reject  $H_0$
- 9.57** a. critical value:  $t = -1.363$ ; test statistic:  $t = -1.252$ ; do not reject  $H_0$     b. critical values:  $t = -2.064$  and  $2.064$ ; test statistic:  $t = 2.258$ ; reject  $H_0$     c. critical value:  $t = 3.143$ ; test statistic:  $t = 2.658$ ; do not reject  $H_0$
- 9.59**  $H_0: \mu = 26.1$  years;  $H_1: \mu > 26.1$  years; critical value:  $t = 2.001$ ; test statistic:  $t = 2.434$ ; reject  $H_0$ ;  $.005 < p\text{-value} < .01$ ; reject  $H_0$
- 9.61**  $H_0: \mu = \$850$ ;  $H_1: \mu < \$850$ ; critical value:  $t = -2.397$ ; test statistic:  $t = -2.257$ ; do not reject  $H_0$ ; if  $\alpha = .025$ , critical value:  $-2.005$ ; reject  $H_0$
- 9.63**  $H_0: \mu = \$650,000$ ;  $H_1: \mu \neq \$650,000$ ; test statistic:  $t = 2.125$ ;  $p\text{-value} > .02$ ; do not reject  $H_0$
- 9.65** a.  $H_0: \mu \geq \$150$ ;  $H_1: \mu < \$150$ ; test statistic:  $t = -1.964$ ;  $.025 < p\text{-value} < .050$ ; do not reject  $H_0$ ; for  $\alpha = .01$ , critical value:  $t = -2.492$ ; test statistic:  $t = -1.964$ ; do not reject  $H_0$     b.  $\alpha = .01$
- 9.67** a.  $H_0: \mu = 58$  years;  $H_1: \mu \neq 58$  years; if  $\alpha = 0$ , do not reject  $H_0$     b. test statistic:  $t = -4.183$ ;  $p\text{-value} < .002$ ; for  $\alpha = .01$ , reject  $H_0$ ; critical values:  $t = -2.649$  and  $2.649$ ; test statistic:  $t = -4.183$ ; reject  $H_0$
- 9.69**  $H_0: \mu = \$95$ ;  $H_1: \mu > \$95$ ; critical value:  $t = 1.771$ ; test statistic:  $t = 2.130$ ; reject  $H_0$
- 9.71**  $H_0: \mu = \$34,400$ ;  $H_1: \mu > \$34,400$ ; test statistic:  $t = 16.2$ ;  $0 < p\text{-value} < .01$ ; reject  $H_0$ ; critical value:  $t = 2.326$ ; reject  $H_0$
- 9.75** a. not large enough    b. large enough  
c. not large enough    d. large enough
- 9.77** a. reject  $H_0$  if  $z \leq -1.65$  or  $z \geq 1.65$     b. reject  $H_0$  if  $z \leq -2.33$     c. reject  $H_0$  if  $z \geq 1.65$
- 9.79** a. critical value:  $z = 1.65$ ; observed value:  $z = 3.90$   
b. critical values:  $z = -1.96$  and  $1.96$ ; observed value:  $z = 3.90$
- 9.81** a. reject  $H_0$  if  $z \leq -1.65$     b. reject  $H_0$  if  $z \leq -1.96$  or  $z \geq 1.96$     c. reject  $H_0$  if  $z \geq 1.65$
- 9.83** a. critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = -1.07$ ; do not reject  $H_0$     b. critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = 3.21$ ; reject  $H_0$
- 9.85** a. critical values:  $z = -1.65$  and  $1.65$ ; test statistic:  $z = .80$ ; do not reject  $H_0$     b. critical value:  $z = -1.65$ ; test statistic:  $z = -4.71$ ; reject  $H_0$     c. critical value:  $z = 2.33$ ; test statistic:  $z = .93$ ; do not reject  $H_0$
- 9.87**  $H_0: p \leq .11$ ;  $H_1: p > .11$ ; test statistic:  $z = 3.84$ ; p-value = 0; for  $\alpha = .05$ , reject  $H_0$ ; critical value for  $\alpha = .05$ :  $z = 1.65$ ; test statistic:  $z = 3.84$ ; reject  $H_0$
- 9.89**  $H_0: p = .55$ ;  $H_1: p > .55$ ; critical value:  $z = 2.05$ ; test statistic:  $z = 5.12$ ; reject  $H_0$ ; p-value = 0; for  $\alpha = .02$ , reject  $H_0$
- 9.91**  $H_0: p \geq .75$ ;  $H_1: p < .75$ ; critical value:  $z = -2.33$ ; test statistic:  $z = -2.31$ ; do not reject  $H_0$ ; p-value: = .0104; for  $\alpha = .01$ , do not reject  $H_0$
- 9.93** a.  $H_0: p \geq .35$ ;  $H_1: p < .35$ ; critical value:  $z = -1.96$ ; test statistic:  $z = -2.94$ ; reject  $H_0$     b. do not reject  $H_0$   
c.  $\alpha = .025$ ; p-value = .0016; reject  $H_0$
- 9.95** a. critical value:  $z = 1.96$ ; test statistic:  $z = 2.27$ ; reject  $H_0$ ; adjust machine    b. critical value:  $z = 2.33$ ; test statistic:  $z = 2.27$ ; do not reject  $H_0$ ; do not adjust the machine
- 9.99** a. critical value:  $z = 1.96$ ; test statistic:  $z = 2.10$ ; reject  $H_0$     b.  $P(\text{Type I error}) = .025$   
c. p-value = .0179; do not reject  $H_0$  if  $\alpha = .01$ ; reject  $H_0$  if  $\alpha = .05$
- 9.101** a. critical values:  $z = -2.33$  and  $2.33$ ; test statistic:  $z = 2.55$ ; reject  $H_0$     b.  $P(\text{Type I error}) = .02$   
c. p-value = .0108; reject  $H_0$  if  $\alpha = .025$ ; do not reject  $H_0$  if  $\alpha = .01$
- 9.103** a.  $H_0: \mu = 151$  minutes;  $H_1: \mu > 151$  minutes; test statistic:  $z = 4.02$ ; p-value = .000; if  $\alpha = .05$ , reject  $H_0$   
b. critical value:  $z = 2.33$ ; test statistic:  $z = 4.02$ ; reject  $H_0$
- 9.105** a.  $H_0: \mu \geq 50$ ;  $H_1: \mu < 50$ ; critical value of  $z = -1.96$ ; test statistic:  $z = -3.00$ ; reject  $H_0$   
b.  $P(\text{Type I error}) = .025$     c. do not reject  $H_0$   
d. p-value = .0013; for  $\alpha = .025$ , reject  $H_0$
- 9.107** a.  $H_0: \mu \leq 2400$  square feet;  $H_1: \mu > 2400$  square feet; critical value:  $t = 1.677$ ; test statistic:  $t = 2.097$ ; reject  $H_0$   
b. for  $\alpha = .01$ , critical value:  $t = 2.405$ ; test statistic:  $t = 2.097$ ; do not reject  $H_0$
- 9.109**  $H_0: \mu \leq 15$  minutes;  $H_1: \mu > 15$  minutes; critical value:  $t = 2.438$ ; test statistic:  $t = 1.875$ ; do not reject  $H_0$
- 9.111**  $H_0: \mu = 25$  minutes;  $H_1: \mu \neq 25$  minutes; critical values:  $t = -2.947$  and  $2.947$ ; test statistic:  $t = 2.083$ ; do not reject  $H_0$
- 9.113** a.  $H_0: \mu \leq 2$  hours;  $H_1: \mu > 2$  hours; critical value:  $t = 2.718$ ; test statistic:  $t = 1.682$ ; do not reject  $H_0$
- 9.115** a.  $H_0: p = .69$ ;  $H_1: p \neq .69$ ; critical values:  $z = -1.96$  and  $1.96$ ; test statistic:  $z = -3.96$ ; reject  $H_0$   
b.  $P(\text{Type I error}) = .05$     c.  $\alpha = .05$ ; p-value = 0; reject  $H_0$
- 9.117**  $H_0: p = .40$ ;  $H_1: p \neq .40$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = -1.62$ ; do not reject  $H_0$ ; p-value = .1052; do not reject  $H_0$
- 9.119** a.  $H_0: p = .80$ ;  $H_1: p < .80$ ; critical value:  $z = -2.33$ ; test statistic:  $z = -.79$ ; do not reject  $H_0$   
b. do not reject  $H_0$
- 9.121** a. .0238    b.  $\alpha = .0238$
- 9.123**  $\alpha = .2776$
- 9.125**  $H_0: \mu = 8000$  hours;  $H_1: \mu < 8000$  hours; reject  $H_0$  if  $\bar{x} < 7890$ ;  $\alpha = .0239$ ; reject  $H_0$  if  $\bar{x} < 7857$ ;  $\alpha = .0049$
- 9.129** a. 29 or more, or 11 or less    b. 226 or more, or 174 or less    c. 2081 or more, or 1919 or less

### Self-Review Test

1. a    2. b    3. a    4. b    5. a    6. a
7. a    8. b    9. c    10. a    11. c    12. b
13. c    14. a    15. b
16. a.  $H_0: \mu = \$921$ ;  $H_1: \mu \neq \$921$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = 2.27$ ; do not reject  $H_0$   
 b.  $H_0: \mu = \$921$ ;  $H_1: \mu > \$921$ ; critical value:  $z = 1.96$ ; test statistic:  $z = 2.27$ ; reject  $H_0$   
 c. in part a,  $\alpha = .01$ ; in part b,  $\alpha = .025$   
 d.  $p$ -value =  $.0232$ ; do not reject  $H_0$   
 e.  $p$ -value =  $.0116$ ; reject  $H_0$
17. a.  $H_0: \mu = 185$  minutes;  $H_1: \mu < 185$  minutes; critical value:  $t = -2.438$  test statistic:  $t = -3.000$ ; reject  $H_0$   
 b.  $P(\text{Type I error}) = .01$   
 c. do not reject  $H_0$   
 d.  $.001 < p\text{-value} < .005$ ; for  $\alpha = .01$ , reject  $H_0$
18. a.  $H_0: \mu \geq 31$  months;  $H_1: \mu < 31$  months; critical value:  $t = -2.131$ ; test statistic:  $t = -3.333$ ; reject  $H_0$   
 b.  $P(\text{Type I error}) = .025$   
 c. critical value:  $t = -3.733$ ; do not reject  $H_0$
19. a.  $H_0: p = .5$ ;  $H_1: p < .5$ ; critical value:  $z = -1.65$ ; test statistic:  $z = -3.16$ ; reject  $H_0$   
 b.  $P(\text{Type I error}) = .05$   
 c. do not reject  $H_0$   
 d.  $p$ -value =  $.0008$ ; reject  $H_0$  if  $\alpha = .05$ ; reject  $H_0$  if  $\alpha = .01$

### Chapter 10

- 10.3 a. 1.83; b.  $-.72$  to  $4.38$ ; margin of error =  $2.55$
- 10.5  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -1.96$  and  $1.96$ ; test statistic:  $z = 1.85$ ; do not reject  $H_0$
- 10.7  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $z = -1.65$ ; test statistic:  $z = -1.47$ ; do not reject  $H_0$
- 10.9 a. 9 hours  
 b.  $1.65$  to  $16.35$  hours;  
 c.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.33$  and  $2.33$ ; test statistic:  $z = 2.66$ ; reject  $H_0$ ;  $p$ -value =  $.0078$ ; for  $\alpha = .02$ , reject  $H_0$
- 10.11 a. .74  
 b.  $.373$  to  $1.11$   
 c.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $2.33$ ; test statistic:  $z = 3.95$ ; reject  $H_0$ ;  $p$ -value =  $.0000$ ; for  $\alpha = .01$ , reject  $H_0$
- 10.13 a.  $-\$1024.54$  to  $-\$75.46$   
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = -2.99$ ; reject  $H_0$   
 c. do not reject  $H_0$
- 10.15 a.  $-6.87$  to  $.87$  calories  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $z = -2.33$ ; test statistic:  $z = -1.81$ ; do not reject  $H_0$   
 c.  $p$ -value =  $.0351$ ; reject  $H_0$  for  $\alpha = .05$ ; do not reject  $H_0$  for  $\alpha = .025$
- 10.17 a.  $-1.58$   
 b.  $-3.82$  to  $.66$
- 10.19  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $t = -2.023$  and  $2.023$ ; test statistic  $t = -1.430$ ; do not reject  $H_0$
- 10.21  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.426$ ; test statistic:  $t = -1.430$ ; do not reject  $H_0$
- 10.23 a. 2.61  
 b.  $-5.86$  to  $11.08$   
 c.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.500$ ; test statistic:  $t = .77$ ; do not reject  $H_0$
- 10.25 a.  $-46.80$  to  $-7.20$  miles;  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.326$ ; test statistic:  $t = -2.67$ ; reject  $H_0$
- 10.27 a.  $2.29$  to  $5.71$  mph  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.416$ ; test statistic:  $t = 5.658$ ; reject  $H_0$

- 10.29 a.  $-12.95$  to  $2.95$  minutes  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.412$ ; test statistic:  $t = -1.691$ ; do not reject  $H_0$
- 10.31 a.  $-.61$  to  $-.39$   
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical value:  $t = -2.576$  and  $2.576$ ; test statistic:  $t = -10.130$ ; reject  $H_0$
- 10.33  $-7.86$  to  $-1.04$
- 10.35  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $t = -2.101$  and  $2.101$ ; test statistic:  $t = -2.740$ ; reject  $H_0$
- 10.37  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.552$ ; test statistic:  $t = -2.740$ ; reject  $H_0$
- 10.39 a.  $-47.01$  to  $-6.99$  miles;  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.326$ ; test statistic:  $t = -2.64$ ; reject  $H_0$   
 c.  $-48.30$  to  $-5.70$ ; critical value:  $t = -2.397$ ; test statistic:  $t = -2.54$ ; reject  $H_0$
- 10.41 a.  $2.23$  to  $5.77$  mph  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.445$ ; test statistic:  $t = 5.513$ ; reject  $H_0$   
 c.  $1.81$  to  $6.20$  mph; critical value:  $t = 2.492$ ; test statistic:  $t = 4.541$ ; reject  $H_0$
- 10.43 a.  $-12.86$  to  $2.86$  minutes  
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.414$ ; test statistic:  $t = -1.713$ ; do not reject  $H_0$   
 c.  $-13.34$  to  $3.34$  minutes; critical value:  $t = -2.431$ ; test statistic:  $t = -1.63$ ; do not reject  $H_0$
- 10.45 a.  $-.61$  to  $-.39$   
 b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $t = -2.576$  and  $2.576$ ; test statistic:  $t = -10.162$ ; reject  $H_0$   
 c.  $-.62$  to  $-.38$ ; critical values:  $t = -2.576$  and  $2.576$ ; test statistic:  $t = -10.10$ ; reject  $H_0$ ; confidence interval is slightly wider
- 10.49 a.  $11.85$  to  $23.15$   
 b.  $50.08$  to  $61.72$ ; c.  $25.66$  to  $32.94$
- 10.51 a. critical values:  $t = -2.060$  and  $2.060$ ; test statistic:  $t = 12.551$ ; reject  $H_0$   
 b. critical value:  $t = 2.624$ ; test statistic:  $t = 7.252$ ; reject  $H_0$   
 c. critical value:  $t = -1.328$ ; test statistic:  $t = -14.389$ ; reject  $H_0$
- 10.53 a.  $-2.98$  to  $9.84$  minutes  
 b.  $H_0: \mu_d = 0$ ;  $H_1: \mu_d > 0$ ; critical value:  $t = 2.447$ ; test statistic:  $t = 1.983$ ; do not reject  $H_0$
- 10.55 a.  $13.22$  to  $30.01$  seconds  
 b.  $H_0: \mu_d = 15$ ;  $H_1: \mu_d > 15$ ; critical value:  $t = 1.356$ ; test statistic:  $t = 1.72$ , reject  $H_0$
- 10.57 a.  $-1.02$  to  $1.52$   
 b.  $H_0: \mu_d = 0$ ;  $H_1: \mu_d \neq 0$ ; critical values:  $t = -2.093$  and  $2.093$ ; test statistic:  $t = .4122$ ; do not reject  $H_0$
- 10.61 a.  $-.062$  to  $.142$
- 10.63  $H_0: p_1 - p_2 = 0$ ;  $H_1: p_1 - p_2 \neq 0$ ; critical values:  $z = -1.96$  and  $1.96$ ; test statistic:  $z = .76$ ; do not reject  $H_0$
- 10.65  $H_0: p_1 - p_2 = 0$ ;  $H_1: p_1 - p_2 > 0$ ; critical value:  $z = 2.05$ ; test statistic:  $z = .76$ ; do not reject  $H_0$
- 10.67 a.  $-.04$   
 b.  $-.086$  to  $.006$   
 c. rejection region at and to the left of  $z = -2.33$ ; non-rejection region to the right of  $z = -2.33$   
 d. test statistic:  $z = -2.02$   
 e. do not reject  $H_0$
- 10.69 a.  $-.019$  to  $.059$   
 b.  $H_0: p_1 - p_2 = 0$ ;  $H_1: p_1 - p_2 \neq 0$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = 1.11$ ; do not reject  $H_0$ ;  $p$ -value =  $.2670$ ; for  $\alpha = .01$ , do not reject  $H_0$
- 10.71 a.  $.024$   
 b.  $-.020$  to  $.068$   
 c.  $H_0: p_1 - p_2 = 0$ ;  $H_1: p_1 - p_2 \neq 0$ ; critical values:  $z = -1.96$  and  $1.96$ ; test

- statistic:  $z = 1.09$ ; do not reject  $H_0$ ;  $p$ -value = .2758; for  $\alpha = .05$ , do not reject  $H_0$
- 10.73** **a.** .10    **b.** .018 to .182    **c.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = 3.04$ ; reject  $H_0$
- 10.75** **a.** -.013 to .093    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $z = 2.33$ ; test statistic:  $z = 1.75$ ; do not reject  $H_0$
- 10.77** **a.** -\$131.30 to -\$120.70    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $z = -1.96$ ; test statistic:  $z = -46.58$ ; reject  $H_0$
- 10.79** **a.** \$1061.95 to \$3278.05    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.326$ ; test statistic:  $t = 4.56$ ; reject  $H_0$
- 10.81** **a.** -8.42 to -1.82 cards    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $t = -1.645$  and  $1.645$ ; test statistic:  $t = -3.04$ ; reject  $H_0$
- 10.83** **a.** \$1056.40 to \$3283.60;  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_0: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.326$ ; test statistic:  $t = 4.54$ ; reject  $H_0$     **b.** \$1118.41 to \$3221.59;  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $t = 2.326$ ; test statistic:  $t = 4.81$ ; reject  $H_0$
- 10.85** **a.** -8.35 to -1.89 cards;  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical value:  $t = -1.645$  and  $1.645$ ; test statistic:  $t = -3.11$ ; reject  $H_0$     **b.** -8.55 to -1.69 cards;  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $t = -1.645$  and  $1.645$ ; test statistic:  $t = -2.93$ ; reject  $H_0$
- 10.87** **a.** -9.54 to -2.4    **b.**  $H_0: \mu_d = 0$ ;  $H_1: \mu_d < 0$ ; critical value:  $t = -2.896$ ; test statistic:  $t = -2.425$ ; do not reject  $H_0$
- 10.89** **a.** -.063 to .023    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = -.91$ ; do not reject  $H_0$ ;  $p$ -value: .3628; for  $\alpha = .01$ , do not reject  $H_0$
- 10.91** **a.** .053 to .127    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.33$  and  $z = 2.33$ ; test statistic:  $z = 4.79$ ; reject  $H_0$ ;  $p$ -value: 0; for  $\alpha = .02$ , reject  $H_0$
- 10.93** .2611
- 10.95**  $n = 9$
- 10.97** **a.**  $n = 545$     **b.** .8708
- 10.101** **a.** .3564    **b.** .0793    **c.** .0013

## Self-Review Test

1. **a**
3. **a.** 1.62 to 2.78    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 > 0$ ; critical value:  $z = 1.96$ ; test statistic:  $z = 9.86$ ; reject  $H_0$
4. **a.** -2.72 to -1.88 hours    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.416$ ; test statistic:  $t = -10.997$ ; reject  $H_0$
5. **a.** -2.70 to -1.90 hours    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 < 0$ ; critical value:  $t = -2.421$ ; test statistic:  $t = -11.474$ ; reject  $H_0$
6. **a.** -\$53.60 to \$186.18    **b.**  $H_0: \mu_d = 0$ ;  $H_1: \mu_d \neq 0$ ; critical values:  $t = -2.447$  and  $2.447$ ; test statistic:  $t = 2.050$ ; do not reject  $H_0$
7. **a.** -.052 to .092    **b.**  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_1: \mu_1 - \mu_2 \neq 0$ ; critical values:  $z = -2.58$  and  $2.58$ ; test statistic:  $z = .60$ ; do not reject  $H_0$

## Chapter 11

- 11.3**  $\chi^2 = 41.337$     **11.5**  $\chi^2 = 41.638$
- 11.7** **a.**  $\chi^2 = 5.009$     **b.**  $\chi^2 = 3.565$
- 11.13** critical value:  $\chi^2 = 11.070$ ; test statistic:  $\chi^2 = 5.200$ ; do not reject  $H_0$
- 11.15** critical value:  $\chi^2 = 9.348$ ; test statistic:  $\chi^2 = 6.994$ ; do not reject  $H_0$
- 11.17** critical value:  $\chi^2 = 13.277$ ; test statistic:  $\chi^2 = 19.328$ ; reject  $H_0$
- 11.19** critical value:  $\chi^2 = 9.488$ ; test statistic:  $\chi^2 = 6.752$ ; do not reject  $H_0$
- 11.21** critical value:  $\chi^2 = 9.348$ ; test statistic:  $\chi^2 = 65.087$ ; reject  $H_0$
- 11.27** **a.**  $H_0$ : the proportion in each row is the same for all four populations;  
**b.** the proportion in each row is not the same for all four populations  
**c.** critical value:  $\chi^2 = 14.449$     **d.** test statistic:  $\chi^2 = 52.451$     **e.** reject  $H_0$
- 11.29** critical value:  $\chi^2 = 5.024$ ; test statistic:  $\chi^2 = 1.980$ ; do not reject  $H_0$
- 11.31** **a.** critical value:  $\chi^2 = 6.635$ ; test statistic:  $\chi^2 = 24.834$ ; reject  $H_0$     **b.** critical value:  $\chi^2 = 6.635$ ; test statistic:  $\chi^2 = 22.588$ ; reject  $H_0$
- 11.33** critical value:  $\chi^2 = 7.815$ ; test statistic:  $\chi^2 = 2.587$ ; do not reject  $H_0$
- 11.35** critical value:  $\chi^2 = 6.635$ ; test statistic:  $\chi^2 = 8.178$ ; reject  $H_0$
- 11.37** critical value:  $\chi^2 = 12.592$ ; test statistic:  $\chi^2 = 30.663$ ; reject  $H_0$
- 11.39** critical value:  $\chi^2 = 7.378$ ; test statistic:  $\chi^2 = 2.404$ ; do not reject  $H_0$
- 11.41** **a.** 18.4376 to 84.9686    **b.** 21.3393 to 67.7365  
**c.** 23.0674 to 60.6586
- 11.43** **a.**  $H_0: \sigma^2 = 1.75$ ;  $H_1: \sigma^2 > 1.75$     **b.** reject  $H_0$  if  $\chi^2 > 34.170$     **c.** test statistic:  $\chi^2 = 22.514$   
**d.** do not reject  $H_0$
- 11.45** **a.**  $H_0: \sigma^2 = 2.2$ ;  $H_1: \sigma^2 \neq 2.2$     **b.** reject  $H_0$  if  $\chi^2 < 7.564$  or  $\chi^2 > 30.191$     **c.** test statistic:  $\chi^2 = 35.545$     **d.** reject  $H_0$
- 11.47** **a.** .8120 to 3.3160; .9011 to 1.8210    **b.**  $H_0: \sigma^2 \leq 1.0$ ;  $H_1: \sigma^2 > 1.0$ ; critical value:  $\chi^2 = 41.638$ ; test statistic:  $\chi^2 = 33.810$ ; do not reject  $H_0$
- 11.49** **a.** 2739.3051 to 12,623.9126; 52.338 to 112.356  
**b.**  $H_0: \sigma^2 = 4200$ ;  $H_1: \sigma^2 \neq 4200$ ; critical values:  $\chi^2 = 12.401$  and  $39.364$ ; test statistic:  $\chi^2 = 29.714$ ; do not reject  $H_0$
- 11.51** critical value:  $\chi^2 = 7.815$ ; test statistic:  $\chi^2 = 10.464$ ; reject  $H_0$
- 11.53** critical value:  $\chi^2 = 13.277$ ; test statistic:  $\chi^2 = 73.25$ ; reject  $H_0$
- 11.55** critical value:  $\chi^2 = 11.345$ ; test statistic:  $\chi^2 = 15.920$ ; reject  $H_0$
- 11.57** critical value:  $\chi^2 = 9.488$ ; test statistic:  $\chi^2 = 29.622$ ; reject  $H_0$
- 11.59** critical value:  $\chi^2 = 9.210$ ; test statistic:  $\chi^2 = 13.593$ ; reject  $H_0$
- 11.61** critical value:  $\chi^2 = 16.812$ ; test statistic:  $\chi^2 = 10.181$ ; do not reject  $H_0$
- 11.63** **a.** 3.4064 to 24.0000; 1.846 to 4.899    **b.** 8.3336 to 33.2628; 2.887 to 5.767

## AN10 Answers to Selected Odd-Numbered Exercises and Self-Review Tests

- 11.65**  $H_0: \sigma^2 = 1.1$ ;  $H_1: \sigma^2 > 1.1$ ; critical value:  $\chi^2 = 28.845$ ; test statistic:  $\chi^2 = 24.727$ ; do not reject  $H_0$
- 11.67**  $H_0: \sigma^2 = 10.4$ ;  $H_1: \sigma^2 \neq 10.4$ ; critical values:  $\chi^2 = 7.564$  and  $30.191$ ; test statistic:  $\chi^2 = 24.192$ ; do not reject  $H_0$
- 11.69** **a.**  $H_0: \sigma^2 = 5000$ ;  $H_1: \sigma^2 < 5000$ ; critical value:  $\chi^2 = 8.907$ ; test statistic:  $\chi^2 = 12.065$ ; do not reject  $H_0$
- b.** 1666.8509 to 7903.1835; 40.827 to 88.900
- 11.71** **a.** .1001 to .4613; .316 to .679    **b.**  $H_0: \sigma^2 = .13$ ;  $H_1: \sigma^2 \neq .13$ ; critical values:  $\chi^2 = 9.886$  and  $45.559$ ; test statistic:  $\chi^2 = 35.077$ ; do not reject  $H_0$
- 11.73** **a.**  $s^2 = 1840.6964$     **b.** 804.6509 to 7624.1864; 28.366 to 87.317    **c.**  $H_0: \sigma^2 = 750$ ;  $H_1: \sigma^2 \neq 750$ ; critical values:  $\chi^2 = 1.690$  and  $16.013$ ; test statistic:  $\chi^2 = 17.180$ ; reject  $H_0$
- 11.75** critical value:  $\chi^2 = 5.991$ ; test statistic:  $\chi^2 = 12.931$ ; reject  $H_0$
- 11.77** critical value:  $\chi^2 = 9.488$ ; test statistic:  $\chi^2 = 11.823$ ; reject  $H_0$
- 11.79** critical value:  $\chi^2 = 16.919$ ; test statistic:  $\chi^2 = 215.568$ ; reject  $H_0$
- 11.81** **a.** test statistic:  $\chi^2 = 2.480$     **b.** no;  $p$ -value  $> .10$

### Self-Review Test

- 1.** b    **2.** a    **3.** c    **4.** a    **5.** b    **6.** b  
**7.** c    **8.** b    **9.** a
- 10.** critical value:  $\chi^2 = 11.070$ ; test statistic:  $\chi^2 = 8.641$ ; do not reject  $H_0$
- 11.** critical value:  $\chi^2 = 11.345$ ; test statistic:  $\chi^2 = 31.188$ ; reject  $H_0$
- 12.** critical value:  $\chi^2 = 9.488$ ; test statistic:  $\chi^2 = 82.450$ ; reject  $H_0$
- 13.** **a.** .2364 to 1.3326; .486 to 1.154    **b.**  $H_0: \sigma^2 = .25$ ;  $H_1: \sigma^2 > .25$ ; critical value:  $\chi^2 = 36.191$ ; test statistic:  $\chi^2 = 36.480$ ; reject  $H_0$

## Chapter 12

- 12.3** **a.** 7.26    **b.** 5.82    **c.** 5.27
- 12.5** **a.** 9.00    **b.** 2.59    **c.** 1.79
- 12.7** **a.** 9.96    **b.** 6.57    **12.9** **a.** 4.85    **b.** 3.22
- 12.13** **a.**  $\bar{x}_1 = 15$ ;  $\bar{x}_2 = 11$ ;  $s_1 = 4.50924975$ ;  $s_2 = 4.39696865$   
**b.**  $H_0: \mu_1 = \mu_2$ ;  $H_1: \mu_1 \neq \mu_2$ ; critical values:  $t = -2.179$  and  $2.179$ ; test statistic:  $t = 1.680$ ; do not reject  $H_0$   
**c.** critical value:  $F = 4.75$ ; test statistic:  $F = 2.82$ ; do not reject  $H_0$     **d.** conclusions are the same
- 12.15** **b.** critical value:  $F = 3.29$ ; test statistic:  $F = 4.07$ ; reject  $H_0$
- 12.17** **a.** numerator:  $df = 2$ ; denominator:  $df = 27$ ;  $SSB = 51,423.2$ ;  $MSW = 829.1944$ ;  $F = 31.01$   
**b.**  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_1$ : all three population means are not equal; critical value:  $F = 3.35$ ; reject  $H_0$
- 12.19** **a.**  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ ;  $H_1$ : all four population means are not equal    **b.** numerator:  $df = 3$ ; denominator:  $df = 28$     **c.**  $SSB = .0105$ ;  $SSW = 1.1449$ ;  $SST = 1.1554$     **d.** reject  $H_0$  if  $F > 4.58$   
**e.**  $MSB = .0035$ ;  $MSW = .0409$     **f.** critical value:  $F = 4.58$     **g.** test statistic:  $F = .0856$   
**i.** do not reject  $H_0$
- 12.21** critical value:  $F = 3.55$ ; test statistic:  $F = 1.30$ ; do not reject  $H_0$

- 12.23** critical value:  $F = 3.72$ ; test statistic:  $F = 5.44$ ; reject  $H_0$
- 12.25** **a.** critical value:  $F = 2.05$ ; test statistic:  $F = 2.12$ ; reject  $H_0$     **b.** .10
- 12.27** **a.** critical value:  $F = 6.93$ ; test statistic:  $F = 1.24$ ; do not reject  $H_0$
- 12.29** **a.** critical value:  $F = 3.89$ ; test statistic:  $F = 4.89$ ; reject  $H_0$     **b.** do not reject  $H_0$
- 12.31** critical value:  $F$  is 5.29; test statistic:  $F = .57$ ; do not reject  $H_0$
- 12.35** **a.** 5 groups with 10 members each.    **b.** 36 members each.

### Self-Review Test

- 1.** a    **2.** b    **3.** c    **4.** a    **5.** a  
**6.** a    **7.** b    **8.** a  
**10.** **a.** critical value:  $F = 3.10$ ; test statistic:  $F = 4.46$ ; reject  $H_0$   
**b.** Type I error

## Chapter 13

- 13.15** **a.**  $y$ -intercept = 100; slope = 5; positive relationship  
**b.**  $y$ -intercept = 400; slope =  $-4$ ; negative relationship
- 13.17**  $\mu_{y|x} = -5.5815 + .2886x$
- 13.19**  $\hat{y} = -83.7140 + 10.5714x$
- 13.21** **a.** \$70.00    **b.** the same amount  
**c.** exact relationship
- 13.23** **a.** \$27.10 million    **b.** different amounts  
**c.** nonexact relationship
- 13.25** **b.**  $\hat{y} = 322.4483 - 34.4425x$     **e.** \$8135.10  
**f.**  $-\$29,751.72$
- 13.27** **b.**  $\hat{y} = 4.0327 - .2687x$     **e.** 3.01  
**f.**  $-.27$
- 13.29** **a.**  $\mu_{y|x} = 41.5821 + .0927x$     **b.** population regression line because data set includes all 16 National League teams; values of  $A$  and  $B$     **d.** 50.852%  
 $\sigma_\epsilon = 7.0756$ ;  $\rho^2 = .04$
- 13.35**  $s_e = 4.7117$ ;  $r^2 = .99$
- 13.39** **a.**  $SS_{xx} = .8960$ ;  $SS_{yy} = .7444$ ;  $SS_{xy} = .7782$   
**b.**  $s_e = .0926$     **c.**  $SST = .7444$ ;  $SSE = .0686$ ;  $SSR = .6758$     **d.**  $r^2 = .91$
- 13.41** **a.**  $s_e = 31.2410$     **b.**  $r^2 = .45$
- 13.43** **a.**  $s_e = .7836$     **b.**  $r^2 = .70$
- 13.45** **a.**  $\sigma_\epsilon = 6.2590$     **b.**  $\rho^2 = .15$
- 13.47** **a.** 6.01 to 6.63    **b.**  $H_0: B = 0$ ;  $H_1: B > 0$ ; critical value:  $t = 2.145$ ; test statistic:  $t = 59.792$ ; reject  $H_0$   
**c.**  $H_0: B = 0$ ;  $H_1: B \neq 0$ ; critical values:  $t = -2.977$  and  $2.977$ ; test statistic:  $t = 59.792$ ; reject  $H_0$   
**d.**  $H_0: B = 4.50$ ;  $H_1: B \neq 4.50$ ; critical values:  $t = -2.624$  and  $2.624$ ; test statistic:  $t = 17.219$ ; reject  $H_0$
- 13.49** **a.** 2.35 to 2.65    **b.**  $H_0: B = 0$ ;  $H_1: B > 0$ ; critical value:  $t = 1.960$ ; test statistic:  $t = 39.124$ ; reject  $H_0$   
**c.**  $H_0: B = 0$ ;  $H_1: B \neq 0$ ; critical values:  $t = -2.576$  and  $2.576$ ; test statistic:  $t = 39.124$ ; reject  $H_0$   
**d.**  $H_0: B \leq 1.75$ ;  $H_1: B > 1.75$ ; critical value:  $t = 2.326$ ; test statistic:  $t = 11.737$ ; reject  $H_0$
- 13.51** **a.**  $-40.3095$  to  $-28.5756$     **b.**  $H_0: B = 0$ ;  $H_1: B < 0$ ; critical value:  $t = -1.943$ ; test statistic:  $t = -14.3654$ ; reject  $H_0$

- 13.53** a.  $\hat{y} = 25.5536 + 2.4377x$    b. 1.331 to 3.5443  
 c.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 2.365$ ; test statistic:  $t = 6.6042$ ; reject  $H_0$
- 13.55** a.  $-.3983$  to  $-.1391$    b.  $H_0: B = 0; H_1: B < 0$ ; critical value:  $t = -2.764$ ; test statistic:  $t = -5.729$ ; reject  $H_0$
- 13.57** a.  $\hat{y} = 270.6218 + 16.3731x$    b. 9.8165 to 22.9297  
 c.  $H_0: B = 14, H_1: B \neq 14$ ; critical values:  $-2.262$  and  $2.262$ ; test statistic:  $.8187$ ; do not reject  $H_0$
- 13.63** a. **13.67** a. positive   b. positive  
 c. positive   d. negative   e. zero  
 $\rho = .21$
- 13.71** a.  $r = -.996$    b.  $H_0: \rho = 0; H_1: \rho < 0$ ; critical value:  $t = -2.764$ ; test statistic:  $t = -35.249$ ; reject  $H_0$
- 13.73** a. positively   b.  $r = .93$    c.  $H_0: \rho = 0; H_1: \rho > 0$ ; critical value:  $t = 1.895$ ; test statistic:  $t = 6.694$ ; reject  $H_0$
- 13.75** a. positively   b. close to 1   c.  $r = .97$   
 d.  $H_0: \rho = 0; H_1: \rho \neq 0$ ; critical values:  $t = -2.776$  and  $2.776$ ; test statistic:  $t = 7.980$ ; reject  $H_0$
- 13.77** a.  $r = .88$    b.  $H_0: \rho = 0; H_1: \rho \neq 0$ ; critical values:  $t = -3.250$  and  $3.250$ ; test statistic:  $t = 5.558$ ; reject  $H_0$   
 $\rho = .39$
- 13.81** a.  $SS_{xx} = 750; SS_{yy} = 9986.9167; SS_{xy} = 565$   
 b.  $\hat{y} = 64.119 + .7533x$    d.  $r = .21; r^2 = .04$   
 f.  $\$119.11$    g.  $s_e = 30.9213$    h.  $-1.7623$  to  $3.2689$    i.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 1.812$ ; test statistic:  $t = .6672$ ; do not reject  $H_0$   
 j.  $H_0: \rho = 0; H_1: \rho > 0$ ; critical value:  $t = 2.228$ ; test statistic:  $t = .679$ ; do not reject  $H_0$
- 13.83** a.  $SS_{xx} = 6394.9; SS_{yy} = 1718.9; SS_{xy} = 3136.1$   
 b.  $\hat{y} = -22.5355 + .4904x$    d.  $r = .95; r^2 = .89$   
 e.  $s_e = 4.7557$    f.  $.291$  to  $.690$    g.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 2.896$ ; test statistic:  $t = 8.246$ ; reject  $H_0$    h.  $H_0: \rho = 0; H_1: \rho \neq 0$ ; critical values:  $t = -3.355$  and  $3.355$ ; test statistic:  $t = 8.605$ ; reject  $H_0$
- 13.85** a.  $SS_{xx} = 3.3647; SS_{yy} = 788; SS_{xy} = 49.4$   
 b.  $\hat{y} = 2.8562 + 14.6819x$    d.  $r = .96; r^2 = .92$   
 e.  $s_e = 3.5416$    f.  $9.718$  to  $19.646$   
 g.  $H_0: B = 0; H_1: B \neq 0$ ; critical values:  $t = -4.032$  and  $4.032$ ; test statistic:  $t = 7.6043$ ; reject  $H_0$    h.  $H_0: \rho = 0; H_1: \rho > 0$ ; critical value:  $t = 3.365$ ; test statistic:  $t = 7.6665$ ; reject  $H_0$
- 13.87** a. 13.8708 to 16.6292; 11.7648 to 18.7352  
 b. 62.3590 to 67.7210; 56.3623 to 73.7177
- 13.89** \$4611.38 to \$5374.78; \$3808.78 to \$6177.38
- 13.91** 93.1957 to 132.9709; 41.3776 to 184.7890
- 13.93** \$1518.85 to \$2212.88; \$715.60 to \$3016.13
- 13.95** a. positive relationship  
 b.  $\hat{y} = -1.9175 + .9895x$    d.  $r = .97; r^2 = .94$   
 e.  $s_e = 1.0941$    f.  $.54$  to  $1.44$   
 g.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 2.571$ ; test statistic:  $t = 8.808$ ; reject  $H_0$    h.  $H_0: \rho = 0; H_1: \rho > 0$ ; critical value:  $t = 2.571$ ; test statistic:  $t = 8.922$ ; reject  $H_0$ ; same conclusion
- 13.97** a. positive   b.  $\hat{y} = 7.8304 + .5039x$   
 d.  $r = .89; r^2 = .79$    e. 2547   f.  $s_e = 3.3525$   
 g.  $.11$  to  $.90$    h.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 3.365$ ; test statistic:  $t = 4.278$ ; reject  $H_0$   
 i.  $H_0: \rho = 0; H_1: \rho \neq 0$ ; critical values:  $t = -3.365$  and  $3.365$ ; test statistic:  $t = 4.365$ ; reject  $H_0$
- 13.99** a.  $SS_{xx} = 224.9; SS_{yy} = 37,258.4; SS_{xy} = 2616.4$   
 b. yes   c.  $\hat{y} = -420.5490 + 11.6336x$   
 e.  $r = .90$    f. 429
- 13.101** b.  $S_{xx} = 82.5; SS_{yy} = .8896; SS_{xy} = -3.84$   
 c. yes   d.  $\hat{y} = 22.1615 - .0465x$    f.  $r = -.45$   
 g. 21.65 seconds
- 13.103** 60.7339 to 97.3729; 40.0144 to 118.0924
- 13.105** 233.0455 to 266.2175; 195.2831 to 303.9799
- 13.107** a. yes   b. 246.4670 to 275.5330 lines  
 c. 200.0567 to 321.9433 lines   e. 338 lines
- 13.111** a. increase   b. decrease   c. increase  
 d.  $\pm t s_e \sqrt{\frac{n+1}{n}}$
- 13.113** a.  $r = .92$ ; yes

## Self-Review Test

1. d   2. a   3. b   4. a   5. b   6. b  
 7. true   8. true   9. a   10. b
15. a. The attendance depends on temperature.  
 b. positive   d.  $\hat{y} = -2.2269 + .2715x$   
 f.  $r = .65; r^2 = .42$    g. 1407 people  
 h.  $s_e = 3.6172$    i.  $-.30$  to  $.84$   
 j.  $H_0: B = 0; H_1: B > 0$ ; critical value:  $t = 3.365$ ; test statistic:  $t = 1.904$ ; do not reject  $H_0$   
 k. 1055 to 1758 people   l. 412 to 2401 people  
 m.  $H_0: \rho = 0; H_1: \rho > 0$ ; critical value:  $t = 3.365$ ; test statistic:  $t = 1.913$ ; do not reject  $H_0$

## Appendix A

- A.7** simple random sample   **A.9** a. nonrandom sample  
 b. judgment sample   c. selection error
- A.11** a. random sample   b. simple random sample  
 c. no
- A.13** a. nonrandom sample   b. voluntary response error and selection error
- A.15** response error
- A.17** a. designed experiment   b. no; would need to know if the women or the doctors who evaluated their health knew which women took aspirin and which were in the control group
- A.19** a. designed experiment   b. double-blind study
- A.21** designed experiment   **A.23** yes
- A.25** b. observational study   c. not a double-blind study
- A.27** a. designed experiment   b. double-blind study
- A.29** a. no   b. no   c. convenience sample
- A.33** a. no   b. nonresponse error and response error  
 c. above

# Chapter 14

## Multiple Regression

---

### Section 14.5

- 14.1** The **coefficients of independent variables** in a multiple regression model are interpreted as the change in  $y$  for a one-unit change in the corresponding independent variable when all other independent variables are held constant. For example,  $B_2$  gives the change in  $y$  due to a one-unit change in  $x_2$  when  $x_1, x_3, \dots, x_k$  are held constant.
- 14.3** The independent variables can have a non-linear relationship but cannot be linearly related.
- 14.5** The following are the assumptions of a multiple regression model:
1. The mean of the probability distribution of  $\epsilon$  is zero, that is,  $E(\epsilon) = 0$ .
  2. The errors associated with different sets of values of independent variables are independent. Furthermore, these errors are normally distributed and have a constant standard deviation which is denoted by  $\sigma_\epsilon$ .
  3. The independent variables are not linearly related.
  4. There is no linear association between the random error term  $\epsilon$  and each independent variables  $x_i$ .
- 14.7**
- a.  $\hat{y} = 15.065 + .167x_1 - .132x_2$
  - b. The value of  $a = 15.065$  gives the value for  $\hat{y}$  when  $x_1 = 0$  and  $x_2 = 0$ . However, since  $x_1 = 0$  and  $x_2 = 0$  do not occur together in the sample data, the estimate is invalid. The value  $b_1 = .167$  gives the change in  $\hat{y}$  for a one-unit change in  $x_1$  when  $x_2$  is held constant. The value  $b_2 = -.132$  gives the change in  $\hat{y}$  for a one-unit change in  $x_2$  when  $x_1$  is held constant.
  - c.  $s_e = 1.488$ ,  $R^2 = .971$ , and  $\bar{R}^2 = .964$
  - d.  $\hat{y} = 15.065 + .167x_1 - .132x_2 = 15.065 + .167(87) - .132(54) = 22.466$
  - e.  $\hat{y} = 15.065 + .167x_1 - .132x_2 = 15.065 + .167(95) - .132(49) = 24.462$
  - f.  $df = n - k - 1 = 11 - 2 - 1 = 8$   
The 99% confidence interval for  $B_1$  is  
 $b_1 \pm ts_{b_1} = .167 \pm (3.355)(.034) = .167 \pm .114 = .053$  to  $.281$
  - g. Step 1:  $H_0: B_2 = 0$ ,  $H_1: B_2 < 0$   
Step 2: Since  $\sigma_\epsilon$  is unknown, use the  $t$  distribution.  
Step 3: For  $\alpha = .01$  with  $df = 8$ , the critical value of  $t$  is  $-2.896$ .  
Step 4:  $t = (b_2 - B_2)/s_{b_2} = -1.919$   
Step 5: Do not reject  $H_0$  since  $-1.919 > -2.896$ .  
Conclude that  $B_2$  is not negative.
- 14.9**
- a.  $\hat{y} = 11.258 + .011x_1 + .199x_2$

- b. The value of  $a = 11.258$  gives the expected weekly sales for restaurants in areas with zero population and a mean annual household income of \$0. However, since the sample data does not include any restaurants in areas with zero population and a mean annual household income of \$0, the estimate is invalid. The value  $b_1 = .011$  indicates that for each increase of 1000 in population, a restaurant's sales are expected to increase by \$11 when mean annual household income is held constant. The value  $b_2 = .199$  indicates that for each increase of \$1000 in mean annual household income, a restaurant's sales are expected to increase by \$199 when population is held constant.
- c.  $s_e = 5.756$ ,  $R^2 = .274$ , and  $\bar{R}^2 = .092$
- d.  $\hat{y} = 11.258 + .011x_1 + .199x_2 = 11.258 + .011(50) + .199(55) = 22.753$   
The predicted sales for a restaurant with 50 thousand people living within a five-mile area surrounding it and \$55 thousand mean annual income of households in that area is \$22,753.
- e.  $\hat{y} = 11.258 + .011x_1 + .199x_2 = 11.258 + .011(45) + .199(60) = 23.693$   
The expected (mean) sales for all restaurants with 45 thousand people living within a five-mile area surrounding them and \$60 thousand mean annual income of households living in those areas is \$23,693.
- f.  $df = n - k - 1 = 11 - 2 - 1 = 8$   
The 95% confidence interval for  $B_2$  is  
 $b_2 \pm ts_{b_2} = .199 \pm (2.306)(.117) = .199 \pm .270 = -.071$  to  $.469$
- g. Step 1:  $H_0: B_1 = 0$ ,  $H_1: B_1 \neq 0$   
Step 2: Since  $\sigma_\epsilon$  is unknown, use the  $t$  distribution.  
Step 3: For  $\alpha = .01$  with  $df = 8$ , the critical values of  $t$  are  $-3.355$  and  $3.355$ .  
Step 4:  $t = (b_1 - B_1)/s_{b_1} = .120$   
Step 5: Do not reject  $H_0$  since  $.120 < 3.355$ .  
Conclude that  $B_1$  is not different from zero.

### Self-Review Test

1. c      2. a      3. c
4. A regression line obtained by using population data is called **the population multiple regression model**. The **estimated multiple regression model** is obtained from sample data.
5. The regression coefficients in a multiple regression model are called the **partial regression coefficients** because each of them gives the effect of the corresponding independent variable on the dependent variable when all other independent variables are held constant.
6.  $R^2$  is the proportion of the total sum of squares (SST) that is explained by the multiple regression model.  $\bar{R}^2$  is the coefficient of multiple determination adjusted for degrees of freedom.  $R^2$  generally increases as more explanatory variables are added to the regression model while the value of  $\bar{R}^2$  may increase, decrease, or stay the same as more independent variables are added.  $R^2$  is always non-negative;  $\bar{R}^2$  can be negative.
7. a. We would expect the relationship between sale price and lot size to be positive, the relationship between sale price and living area to be positive, and the relationship between sale price and age to be negative.  
b.  $\hat{y} = 200.153 + 11.889x_1 + .099x_2 - 7.551x_3$   
The signs of the coefficients of the independent variables obtained in the solution are consistent with the expectations in part a.

- c. The value of  $a = 200.153$  gives the expected sale price of a house for a lot size of zero and living area of zero at age zero. However, since  $x_1 = 0$ ,  $x_2 = 0$ , and  $x_3 = 0$  do not occur together in the sample data, the estimate is invalid. In fact, a lot size of zero and a living area of zero do not make sense. The value  $b_1 = 11.889$  indicates that for an increase of one acre in the lot size, the sale price of a house is expected to increase by \$11,889 when living area and age are held constant. The value  $b_2 = .099$  indicates that for an increase of one square foot in living area, the sale price of a house is expected to increase by \$99 when lot size and age are held constant. The value  $b_3 = -7.551$  indicates that for an increase of one year in age, the sale price of a house is expected to decrease by \$7551 when lot size and living area are held constant.

d.  $s_e = 37.762$ ,  $R^2 = .882$ , and  $\bar{R}^2 = .842$

e.  $\hat{y} = 200.153 + 11.889x_1 + .099x_2 - 7.551x_3 = 200.153 + 11.889(2.5) + .099(3000) - 7.551(14)$   
 $= 421.162$

The predicated sale price of a house that has a lot size of 2.5 acres, a living area of 3000 square feet, and is 14 years old is \$421,162.

f.  $\hat{y} = 200.153 + 11.889x_1 + .099x_2 - 7.551x_3 = 200.153 + 11.889(2.2) + .099(2500) - 7.551(7)$   
 $= 420.952$

The point estimate of the mean sale prices of all houses that have a lot size of 2.2 acres, a living area of 2500 square feet, and are 7 years old is \$420,952.

g.  $df = n - k - 1 = 13 - 3 - 1 = 9$

The 99% confidence interval for  $B_1$  is

$$b_1 \pm ts_{b_1} = 11.889 \pm (3.250)(23.697) = 11.889 \pm 77.015 = -65.126 \text{ to } 88.904$$

The 99% confidence interval for  $B_2$  is

$$b_2 \pm ts_{b_2} = .099 \pm (3.250)(.043) = .099 \pm .140 = -.041 \text{ to } .239$$

The 99% confidence interval for  $B_3$  is

$$b_3 \pm ts_{b_3} = -7.551 \pm (3.250)(1.988) = -7.551 \pm 6.461 = -14.012 \text{ to } -1.090$$

h. The 98% confidence interval for  $A$  is

$$a \pm ts_a = 200.153 \pm (2.821)(89.138) = 200.153 \pm 251.458 = -51.305 \text{ to } 451.611$$

i. Step 1:  $H_0: B_1 = 0$ ,  $H_1: B_1 > 0$

Step 2: Since  $\sigma_\epsilon$  is unknown, use the  $t$  distribution.

Step 3: For  $\alpha = .01$  with  $df = 9$ , the critical value of  $t$  is 2.821

$$\text{Step 4: } t = (b_1 - B_1)/s_{b_1} = .502$$

Step 5: Do not reject  $H_0$  since  $.502 < 2.821$ .

Conclude that  $B_1$  is not positive.

j. Step 1:  $H_0: B_2 = 0$ ,  $H_1: B_2 > 0$

Step 2: Since  $\sigma_\epsilon$  is unknown, use the  $t$  distribution.

Step 3: For  $\alpha = .025$  with  $df = 9$ , the critical value of  $t$  is 2.262.

$$\text{Step 4: } t = (b_2 - B_2)/s_{b_2} = 2.319$$

Step 5: Reject  $H_0$  since  $2.319 > 2.262$ .

Conclude that  $B_2$  is positive.

k. Step 1:  $H_0: B_3 = 0$ ,  $H_1: B_3 < 0$

Step 2: Since  $\sigma_\epsilon$  is unknown, use the  $t$  distribution.

Step 3: For  $\alpha = .05$  with  $df = 9$ , the critical value of  $t$  is  $-1.833$ .

$$\text{Step 4: } t = (b_3 - B_3)/s_{b_3} = -3.799$$

Step 5: Reject  $H_0$  since  $-3.799 < -1.833$ .

Conclude that  $B_3$  is negative.

# Chapter 15

## Nonparametric Methods

---

### Section 15.1

- 15.1 Data that are divided into different categories for identification purposes are called **categorical data**. For example, dividing adults by gender (male or female), or classifying people's opinions about a certain issue – in favor, against, or no opinion – produces categorical data.
- 15.3 When using the sign test for the median of a single population, Table VIII must be used if the sample size  $n \leq 25$ .
- 15.5 a. The rejection region is  $X \geq 12$ .  
b. The rejection region is  $X \leq 3$  or  $X \geq 17$ .  
c. The rejection region lies to the left of  $z = -1.65$ .
- 15.7 a. Step 1:  $H_0$ : Median = 28,  $H_1$ : Median > 28  
Step 2: Since  $n \leq 25$ , use the binomial distribution.  
Step 3: For  $n = 10$  and  $\alpha = .05$ , the rejection region is  $X \geq 9$ .  
Step 4: The observed value of  $X = 8$ .  
Step 5: Do not reject  $H_0$  since  $8 < 9$ .
- b. Step 1:  $H_0$ : Median = 100,  $H_1$ : Median < 100  
Step 2: Since  $n \leq 25$ , use the binomial distribution.  
Step 3: For  $n = 11$  and  $\alpha = .05$ , the rejection region is  $X \leq 2$ .  
Step 4: The observed value of  $X = 1$ .  
Step 5: Reject  $H_0$  since  $1 < 2$ .
- c. Step 1:  $H_0$ : Median = 180,  $H_1$ : Median  $\neq 180$   
Step 2: Since  $n > 25$ , use the normal distribution.  
Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.96$  and to the right of  $z = 1.96$ .  
Step 4:  $n = 26$ ,  $p = .50$ , and  $q = 1 - p = 1 - .50 = .50$   
$$\mu = np = 26(.50) = 13, \sigma = \sqrt{npq} = \sqrt{26(.50)(.50)} = 2.54950976$$
  
Since  $X = 3$ ,  $\frac{n}{2} = \frac{26}{2} = 13$ , and  $X > \frac{n}{2}$ ,  $z = \frac{X + .5 - \mu}{\sigma} = \frac{(3+.5)-13}{2.54950976} = -3.73$ .
- Step 5: Reject  $H_0$  since  $-3.73 < -1.96$ .

**15-2** Chapter 15 Nonparametric Methods

- d. Step 1:  $H_0$ : Median = 55,  $H_1$ : Median < 55  
 Step 2: Since  $n > 25$ , use the normal distribution.  
 Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.65$ .  
 Step 4:  $n = 30$ ,  $p = .50$ , and  $q = 1 - p = 1 - .50 = .50$

$$\mu = np = 30(.50) = 15, \sigma = \sqrt{npq} = \sqrt{30(.50)(.50)} = 2.73861279$$

$$\text{Since } X = 6, \frac{n}{2} = \frac{30}{2} = 15, \text{ and } X < \frac{n}{2}, z = \frac{(X + .5) - \mu}{\sigma} = \frac{(6 + .5) - 15}{2.73861279} = -3.10.$$

Step 5: Reject  $H_0$  since  $-3.10 < -1.65$ .

- 15.9** Let  $p$  be the proportion of residents who prefer bottled water, B represent bottled water and C represent city water.

- Step 1:  $H_0$ :  $p = .50$ ,  $H_1$ :  $p \neq .50$   
 Step 2: Since  $n \leq 25$ , use the binomial distribution.  
 Step 3: For  $n = 12$  and  $\alpha = .05$ , the rejection region is  $X \leq 2$  or  $X \geq 10$ .  
 Step 4: We assign a plus sign for each person who prefers bottled water and a minus sign for each person preferring city water.

Person	1	2	3	4	5	6	7	8	9	10	11	12
Water Source	B	C	B	C	C	B	C	C	C	C	B	C
Sign	+	-	+	-	-	+	-	-	-	-	+	-

There are four plus signs and eight minus signs. Thus, the observed value of  $X = 4$ .

Step 5: Do not reject  $H_0$  since  $2 < 4 < 10$ .

Conclude that the residents do not prefer either of these two water sources over the other.

- 15.11** Let  $p$  be the proportion of all drinkers of JW's beer who can distinguish JW's from the rival brand.

- Step 1:  $H_0$ :  $p = .50$ ,  $H_1$ :  $p > .50$   
 Step 2: Since  $n \leq 25$ , use the binomial distribution.  
 Step 3: For  $n = 20$  and  $\alpha = .025$ , the rejection region is  $X \geq 15$ .  
 Step 4: The observed value of  $X = 13$ .  
 Step 5: Do not reject  $H_0$  since  $13 < 15$ .

Conclude that drinkers of JW's are not more likely to select JW's than the rival brand.

- 15.13** Let  $p$  be the proportion of adult North Dakota residents who would prefer to stay in North Dakota.

- Step 1:  $H_0$ :  $p = .50$ ,  $H_1$ :  $p < .50$   
 Step 2: Since  $n > 25$ , use the normal distribution.  
 Step 3: For  $\alpha = .025$ , the rejection region lies to the left of  $z = -1.96$ .  
 Step 4: Four of the 100 adults have no preference, so the true value of  $n = 100 - 4 = 96$ .  
 $p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 96(.50) = 48, \sigma = \sqrt{npq} = \sqrt{96(.50)(.50)} = 4.89897949$   
 $\text{Since } X = 41, \frac{n}{2} = \frac{96}{2} = 48, \text{ and } X < \frac{n}{2}, z = \frac{(X + .5) - \mu}{\sigma} = \frac{(41 + .5) - 48}{4.89897949} = -1.33.$

Step 5: Do not reject  $H_0$  since  $-1.33 > -1.96$ .

Do not conclude that less than half of all adult residents of North Dakota would prefer to stay.

- 15.15** Let  $p$  be the proportion of adults that frequently experience stress.

- Step 1:  $H_0$ :  $p = .50$ ,  $H_1$ :  $p > .50$   
 Step 2: Since  $n > 25$ , use the normal distribution.  
 Step 3: For  $\alpha = .01$ , the rejection region lies to the right of  $z = 2.33$ .  
 Step 4:  $n = 700$ ,  $p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 700(.50) = 350, \sigma = \sqrt{npq} = \sqrt{700(.50)(.50)} = 13.22875656$   
 $\text{Since } X = 370, \frac{n}{2} = \frac{500}{2} = 350, \text{ and } X > \frac{n}{2}, z = \frac{(X - .5) - \mu}{\sigma} = \frac{(370 - .5) - 350}{13.22875656} = 1.47.$

(continued on next page)

(continued)

Step 5: Do not reject  $H_0$  since  $1.47 < 2.33$ .

Do not conclude that over half of adults frequently experience stress in their daily lives.

- 15.17** Step 1:  $H_0$ : Median = 12 ounces,  $H_1$ : Median  $\neq$  12 ounces  
 Step 2: Since  $n \leq 25$ , use the binomial distribution.  
 Step 3: Since one of the 10 bottles had exactly 12.00 ounces, the true sample size is  $n = 10 - 1 = 9$ .  
 For  $n = 9$  and  $\alpha = .05$ , the rejection region is  $X \leq 1$  or  $X \geq 8$ .  
 Step 4: We assign a plus sign for each bottle that has more than 12 ounces, a minus sign for each bottle that has less than 12 ounces, and a zero for each bottle holding exactly 12 ounces.

Bottle	1	2	3	4	5	6	7	8	9	10
Amount	12.10	11.95	12.00	12.01	12.02	12.05	12.02	12.03	12.04	12.06
Sign	+	-	0	+	+	+	+	+	+	+

There are eight plus signs, one minus sign and one zero. For a two-tailed test, we can use either value. Using the larger value,  $X = 8$ .

Step 5: Reject  $H_0$  since  $8 = 8$ . Note that  $8 = 8$  indicates that the observed value of the test statistic is “just barely” inside the rejection region.

Conclude that the median amount of soda in all such bottles differs from 12 ounces.

If we had used the smaller value,  $X = 1$ , we see that  $1 = 1$ , and our conclusion would be the same.

- 15.19** Let  $p$  be the proportion of response times that exceed four minutes.  
 Step 1:  $H_0$ : Median  $\leq 4$  minutes,  $H_1$ : Median  $> 4$  minutes  
 Step 2: Since  $n > 25$ , use the normal distribution.  
 Step 3: For  $\alpha = .01$ , the rejection region lies to the right of  $z = 2.33$ .  
 Step 4: Two of the 28 response times are exactly four minutes, so the true sample size is  $n = 28 - 2 = 26$ ;  $p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 26(.50) = 13$ ,  $\sigma = \sqrt{npq} = \sqrt{26(.50)(.50)} = 2.5495097$   
 We assign a plus sign to every response time above four minutes, a minus sign to every time below four minutes, and a zero to every time of exactly four minutes.  
 We assign a plus sign to every response time above four minutes, a minus sign to every time below four minutes, and a zero to every time of exactly four minutes.

Call	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Time	6	5	7	12	2	1.5	3.5	4	10	11	4.5	6	5	8.5
Sign	+	+	+	+	-	-	-	0	+	+	+	+	+	+
Call	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Time	7	15	9	8	3	10	8	4.5	9	4	6	3	6	7.5
Sign	+	+	+	+	-	+	+	+	+	0	+	-	+	+

There are 21 plus signs, five minus signs, and two zeros. For a right-tailed test, we use the larger value. Thus, the observed value of  $X = 21$ .

$$\text{Then for } \frac{n}{2} = \frac{26}{2} = 13, X > \frac{n}{2}, \text{ and } z = \frac{(X - .5) - \mu}{\sigma} = \frac{(21 - .5) - 13}{2.5495097} = 2.94.$$

Step 5: Reject  $H_0$  since  $2.94 > 2.33$ .

Conclude that the median response time to all 911 calls in the inner city is greater than 4 minutes.

- 15.21** Let  $p$  be the proportion of times that exceed 42 months.  
 Step 1:  $H_0$ : Median = 42 months,  $H_1$ : Median  $<$  42 months  
 Step 2: Since  $n > 25$ , use the normal distribution.  
 Step 3: For  $\alpha = .01$ , the rejection region lies to the left of  $z = -2.33$ .  
 Step 4:  $n = 35$ ,  $p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 35(.50) = 17.5$ ,  $\sigma = \sqrt{npq} = \sqrt{35(.50)(.50)} = 2.95803989$   
 We assign a plus sign for every time that exceeds 42 months and a minus sign for every time that is less than 42 months.

(continued on next page)

(continued)

Inmate	1	2	3	4	5	6	7	8	9	10	11	12
Time	37	6	20	5	25	30	24	10	12	20	24	8
Sign	—	—	—	—	—	—	—	—	—	—	—	—
Inmate	13	14	15	16	17	18	19	20	21	22	23	24
Time	26	15	13	22	72	80	96	33	84	86	70	40
Sign	—	—	—	—	+	+	+	—	+	+	+	—
Inmate	25	26	27	28	29	30	31	32	33	34	35	
Time	92	36	28	90	36	32	72	45	38	18	9	
Sign	+	—	—	+	—	—	+	+	—	—	—	

There are 10 plus signs and 25 minus signs. For a left-tailed test, we use the smaller value. Thus, the observed value of  $X = 10$ .

$$\text{Then } \frac{n}{2} = \frac{35}{2} = 17.5, X < \frac{n}{2}, \text{ and } z = \frac{(X + .5) - \mu}{\sigma} = \frac{(10 + .5) - 17.5}{2.95803989} = -2.37.$$

Step 5: Reject  $H_0$  since  $-2.37 < -2.33$ .

Conclude that the median time served by all such prisoners is less than 42 months.

- 15.23** Let  $M$  denote the difference in median test scores before and after the course. For each employee, the paired difference = score before course – score after course.
- Step 1:  $H_0: M = 0, H_1: M < 0$
- Step 2: Since  $n \leq 25$ , use the binomial distribution.
- Step 3: One of the employee's scores were the same before and after the course, so the true sample size is  $n = 6$ . For  $n = 6$  and  $\alpha = .05$ , the rejection region is  $X = 0$ .
- Step 4: We assign a plus sign to each employee whose score decreased, a minus sign to each employee whose score increased, and a zero to each employee whose score was the same before and after the course.

Before	8	5	4	9	6	9	5
After	10	8	5	11	6	7	9
Sign	—	—	—	—	0	+	—

There are one plus sign, five minus signs, and one zero. For a left-tailed test, we use the smaller value. Thus, the observed value of  $X = 1$ .

Step 5: Do not reject  $H_0$  since  $1 > 0$ .

Conclude that attending this course does not increase the median self-confidence test score of all employees.

- 15.25** Let  $M$  denote the difference in median bikes assembled before and after the new payment system. For each employee, the paired difference = bikes assembled before – bikes assembled after new system. Let  $p$  be the proportion of paired differences that are positive.
- Step 1:  $H_0: M = 0, H_1: M \neq 0$
- Step 2: Since  $n > 25$ , use the normal distribution.
- Step 3: For  $\alpha = .02$ , the rejection region lies to the left of  $z = -2.33$  and to the right of  $z = 2.33$ .
- Step 4: Since one worker assembled the same number of bikes under both systems, the true value of  $n = 27 - 1 = 26$ .

$$p = .50 \text{ and } q = 1 - p = 1 - .50 = .50$$

$$\mu = np = 26(.50) = 13, \sigma = \sqrt{npq} = \sqrt{26(.50)(.50)} = 2.54950976$$

For a two-tailed test, we can use either value. Using the larger value,  $X = 19$ ,

$$\frac{n}{2} = \frac{26}{2} = 13, X > \frac{n}{2}, \text{ and } z = \frac{(X - .5) - \mu}{\sigma} = \frac{(19 - .5) - 13}{2.54950976} = 2.16.$$

(continued on next page)

(continued)

Step 5: Do not reject  $H_0$  since  $2.16 < 2.33$ .

Conclude that the median number of bikes assembled after the new payment system was instituted does not differ from the median before the new system.

If we had used the smaller value,  $X = 7$ ,  $\frac{n}{2} = \frac{26}{2} = 13$ ,  $X < \frac{n}{2}$ , and  $z = \frac{(7+.5)-13}{2.54950976} = -2.16$ . Our conclusion would be the same.

- 15.27** Let  $M$  denote the difference in median milk production without and with the hormone. For each matched pair of cows, the paired difference = milk production of cow without hormone – milk production of cow with hormone. Let  $p$  be the proportion of paired differences that are positive.

Step 1:  $H_0: M = 0$ ,  $H_1: M \neq 0$

Step 2: Since  $n > 25$ , use the normal distribution.

Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.96$  and to the right of  $z = 1.96$ .

Step 4: Since milk production was the same for two pairs of cows, the true value of  $n = 30 - 2 = 28$ .

$$p = .50 \text{ and } q = 1 - p = 1 - .50 = .50$$

$$\mu = np = 28(.50) = 14, \sigma = \sqrt{npq} = \sqrt{28(.50)(.50)} = 2.64575131$$

For a two-tailed test, we can use either value. Using the larger value,  $X = 19$ ,

$$\frac{n}{2} = \frac{28}{2} = 14, X > \frac{n}{2}, \text{ and } z = \frac{(X - .5) - \mu}{\sigma} = \frac{(19 - .5) - 14}{2.64575131} = 1.70.$$

Step 5: Do not reject  $H_0$  since  $1.70 < 1.96$ .

Conclude that the hormone does not change the milk production of such cows. If we had used the smaller

value,  $X = 9$ ,  $\frac{n}{2} = \frac{28}{2} = 14$ ,  $X < \frac{n}{2}$ , and  $z = \frac{(9 + .5) - 14}{2.64575131} = -1.70$ . Our conclusion would be the same.

## Section 15.2

- 15.29** The null hypothesis of the Wilcoxon signed–rank test usually states that the medians of the two population distributions are equal.

- 15.31** **a.** The rejection region is  $T \leq 11$ .  
**b.** The rejection region is  $T \leq 7$ .  
**c.** The rejection region lies to the left of  $z = -1.96$ .  
**d.** The rejection region lies to the right of  $z = 2.33$ .

- 15.33** **a.** Let  $M_A$  and  $M_B$  be the median number of contacts by all such salespersons after and before installation of governors, respectively. For each salesperson, the paired difference = number of contacts before – number of contacts after.

Step 1:  $H_0: M_A = M_B$ ,  $H_1: M_A < M_B$

Step 2: Since  $n \leq 15$ , use the Wilcoxon signed–rank test for the small–sample case.

Step 3: For  $n = 7$  and  $\alpha = .05$ , the rejection region is  $T \leq 4$ .

Step 4:

Before	After	Differences (Before – After)	Absolute Differences	Ranks of Differences	Signed Ranks
50	49	+1	1	1	+1
63	60	+3	3	2	+2
42	47	-5	5	4.5	-4.5
55	51	+4	4	3	+3
44	50	-6	6	6	-6
65	60	+5	5	4.5	+4.5
66	58	+8	8	7	+7

(continued on next page)

(continued)

Sum of positive ranks = 17.5

Sum of absolute values of negative ranks = 10.5

For a left-tailed test,  $T$  is the sum of the absolute values of the negative ranks. Thus, the observed value of  $T = 10.5$ .Step 5: Do not reject  $H_0$  since  $10.5 > 4$ .

Conclude that the use of governors does not tend to reduce the number of contacts made per week by the Gamma Corporation's salespersons.

b. The conclusion in part a of this exercise is the same as that of the test of Exercise 10.96.

- 15.35** Let  $M_A$  and  $M_B$  denote the median self-confidence test scores of all such employees after and before attending the course. For each employee, the paired difference = score before course – score after course.

a. Step 1:  $H_0: M_A = M_B, H_1: M_A > M_B$ Step 2: Since  $n \leq 15$ , use the Wilcoxon signed-rank test procedure for the small-sample case.Step 3: One of the employees had the same test score before and after attending the course, so the true sample size is  $n = 7 - 1 = 6$ . For  $n = 6$  and  $\alpha = .05$ , the rejection region is  $T \leq 2$ .

Step 4:

Before	After	Differences (Before – After)	Absolute Differences	Ranks of Differences	Signed Ranks
8	10	-2	2	3	-3
5	8	-3	3	5	-5
4	5	-1	1	1	-1
9	11	-2	2	3	-3
6	6	0	0	–	–
9	7	+2	2	3	+3
5	9	-4	4	6	-6

Sum of positive ranks = 3

Sum of absolute values of negative ranks = 18

For a right-tailed test,  $T$  is the sum of the positive ranks. Thus, the observed value of  $T = 3$ Step 5: Do not reject  $H_0$  since  $3 > 2$ .

Conclude that attending this course does not increase the median self-confidence test scores of employees.

b. The conclusion in part a of this exercise is the same as that of the test in Exercise 15.23.

- 15.37** Let  $M_A$  and  $M_B$  denote the median time to complete the hike after and before the fitness course for all such adults. For each adult, the paired difference = hiking time before course – hiking time after course.

Step 1:  $H_0: M_A = M_B, H_1: M_A < M_B$ Step 2: Since  $n > 15$ , use the normal distribution.Step 3: For  $\alpha = .025$ , the rejection region lies to the left of  $z = -1.96$ .

Step 4:  $n = 20, \mu_T = \frac{n(n+1)}{4} = \frac{20(20+1)}{4} = 105$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{20(20+1)(40+1)}{24}} = 26.78619047$$

(continued on next page)

(continued)

Before	After	Differences (Before – After)	Absolute Differences	Ranks of Differences	Signed Ranks
41	37	+4	4	5	+5
91	71	+20	20	19	+19
35	30	+5	5	7	+7
58	64.5	-6.5	6.5	10	-10
45	44	+1	1	1	+1
48.5	44	+4.5	4.5	6	+6
84	78	+6	6	8.5	+8.5
64	55	+9	9	15	+15
37	31	+6	6	8.5	+8.5
54	57	-3	3	4	-4
70	59	+11	11	17	+17
40	33	+7	7	11	+11
78	70.5	+7.5	7.5	12	+12
66	56	+10	10	16	+16
100	78	+22	22	20	+20
48	40	+8	8	13.5	+13.5
50	48	+2	2	2	+2
94	102	-8	8	13.5	-13.5
42.5	40	+2.5	2.5	3	+3
75	63	+12	12	18	+18

Sum of positive ranks = 182.5

Sum of absolute values of negative ranks = 27.5

For a left-tailed test,  $T$  is the sum of the absolute values of the negative ranks. Thus, the observed

$$\text{value of } T = 27.5 \text{ and } z = \frac{T - \mu_T}{\sigma_T} = \frac{27.5 - 105}{26.78619047} = -2.89.$$

Step 5: Reject  $H_0$  since  $-2.89 < -1.96$ .

Conclude that the fitness course tends to reduce the median time required to complete the two-mile hike.

### Section 15.3

**15.39** The **Wilcoxon signed-rank test** is used for paired samples, while the **Wilcoxon rank-sum test** is used for independent samples.

- 15.41** **a.** Step 1:  $H_0$ : The two population distributions are identical  
 $H_1$ : The two population distributions are different  
Step 2: Since  $n_1 \leq 10$  and  $n_2 \leq 10$ , use the Wilcoxon rank-sum test for small samples.  
Step 3: For  $n_1 = 6$ ,  $n_2 = 7$ , and  $\alpha = .05$ , the rejection region is  $T \leq 28$  or  $T \geq 56$ .  
Step 4: The observed value of  $T = 22$ .  
Step 5: Reject  $H_0$  since  $22 < 28$ .
- b.** Step 1:  $H_0$ : The two population distributions are identical  
 $H_1$ : The distribution of population 1 lies to the right of the distribution of population 2  
Step 2: Since  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .025$ , the rejection region lies to the right of  $z = 1.96$ .
- Step 4:  $n_1 = 10$ ,  $n_2 = 12$ ,  $\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{10(10 + 12 + 1)}{2} = 115$
- $$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{10(12)(10 + 12 + 1)}{12}} = 15.16575089$$

(continued on next page)

(continued)

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{137 - 115}{15.16575089} = 1.45$$

Step 5: Do not reject  $H_0$  since  $1.45 < 1.96$ .

- c. Step 1:  $H_0$ : The two population distributions are identical  
 $H_1$ : The distribution of population 1 lies to the left of the distribution of population 2  
Step 2: Since  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.65$ .

$$\text{Step 4: } n_1 = 9, n_2 = 11, \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{9(9 + 11 + 1)}{2} = 94.5$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{9(11)(9 + 11 + 1)}{12}} = 13.16244658$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{68 - 94.5}{13.16244658} = -2.01$$

Step 5: Reject  $H_0$  since  $-2.01 < -1.65$ .

- d. Step 1:  $H_0$ : The two population distributions are identical  
 $H_1$ : The two population distributions are different  
Step 2: Since  $n_1 > 10$  and  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .01$ , the rejection region lies to the left of  $z = -2.58$  and to the right of  $z = 2.58$ .

$$\text{Step 4: } n_1 = 22, n_2 = 23, \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{22(22 + 23 + 1)}{2} = 506$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{22(23)(22 + 23 + 1)}{12}} = 44.04164696$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{638 - 506}{44.04164696} = 3.00$$

Step 5: Reject  $H_0$  since  $3.00 > 2.58$ .

- 15.43** Step 1:  $H_0$ : The population distributions of times in the 500-meter event for the two types of skates are identical  
 $H_1$ : The population distributions of times in the 500-meter event for the two types of skates are different  
Step 2: Since  $n_1 \leq 10$  and  $n_2 \leq 10$ , use the Wilcoxon rank-sum test for small samples.  
Step 3: For  $n_1 = 7, n_2 = 8$  and  $\alpha = .05$ , the rejection region is  $T \leq 41$ .  
Step 4:

New Skates		Traditional Skates	
Time	Rank	Time	Rank
40.5	7.5	41.0	13
40.3	6	40.8	11
39.5	1	40.9	12
39.7	2	39.8	3
40.0	5	40.6	9
39.9	4	40.7	10
41.5	15	41.1	14
		40.5	7.5
	Sum = 40.5		Sum = 79.5

For a one-tailed test with unequal sample sizes,  $T$  is the sum of ranks for the smaller sample.  
Thus, the observed value of  $T = 40.5$ .

- Step 5: Reject  $H_0$  since  $40.5 < 41$ .  
Conclude that the new skates tend to produce faster times in this event.

- 15.45** Step 1:  $H_0$ : The population distributions of number of good parts for the two groups are identical  
 $H_1$ : The population distribution of number of good parts for Group A lies to the right of the corresponding distribution for Group B  
Step 2: Since  $n_1 > 10$  and  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .01$ , the rejection region is  $z > 2.33$ .  
Step 4:

Group A		Group B	
Good Parts	Rank	Good Parts	Rank
157	11	160	14.5
139	5	118	1
188	23	150	9
143	7	165	18
172	20	158	12
144	8	159	13
191	24	127	2
128	3	133	4
177	22	170	19
160	14.5	164	17
175	21	152	10
162	16	142	6
Sum = 174.5		Sum = 125.5	

For a one-tailed test with equal sample sizes,  $T$  is the sum of ranks for the first sample. Thus, the observed value of  $T = 174.5$ .

$$n_1 = 12, n_2 = 12, \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{12(12 + 12 + 1)}{2} = 150$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{12(12)(12 + 12 + 1)}{12}} = 17.32050808$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{174.5 - 150}{17.32050808} = 1.41$$

Step 5: Do not reject  $H_0$  since  $1.41 < 2.33$ .

Conclude that the median number of good parts produced by machinists who take a five-minute break every hour is not higher than the corresponding median for machinists who do not take such breaks.

- 15.47** Step 1:  $H_0$ : The population distribution of travel time for the plane and bus are identical  
 $H_1$ : The population distribution of travel times for the plane lies to the right of the distribution of travel times for the bus.  
Step 2: Since  $n_1 > 10$  and  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .05$ , the rejection region lies to the right of  $z = 1.65$ .  
Step 4: Sum of ranks for plane = 295  
Sum of ranks for bus = 233

For a one-tailed test with unequal sample sizes,  $T$  is the sum of ranks for the smaller sample. Thus, the observed value of  $T = 295$ .

$$n_1 = 15, n_2 = 17, \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{15(15 + 17 + 1)}{2} = 247.5$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{15(17)(15 + 17 + 1)}{12}} = 26.48112535$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{295 - 247.5}{26.48112535} = 1.79$$

Step 5: Reject  $H_0$  since  $1.79 > 1.65$ .

Conclude that the median travel time for the plane trip is higher than for the bus trip.

**Section 15.4**

**15.49** In the ANOVA procedure of Chapter 12, the populations being compared are assumed to have normal distributions. This assumption is not required for the Kruskal–Wallis test.

**15.51** a. Step 1:  $H_0$ : The three population distributions are all identical

$H_1$ : The three population distributions are not all identical

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .05$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 5.991$ .

Step 4:  $n = n_1 + n_2 + n_3 = 9 + 8 + 5 = 22$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{22(22+1)} \left[ \frac{(81)^2}{9} + \frac{(102)^2}{8} + \frac{(70)^2}{5} \right] - 3(22+1) = 2.372 \end{aligned}$$

Step 5: Do not reject  $H_0$  since  $2.372 < 5.991$ .

b. Step 1:  $H_0$ : The four population distributions are all identical

$H_1$ : The four population distributions are not all identical

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .05$  and  $df = k - 1 = 4 - 1 = 3$ , the rejection region is  $\chi^2 > 7.815$ .

Step 4:  $n = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 5 + 5 = 20$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} + \frac{R_4^2}{n_4} \right] - 3(n+1) \\ &= \frac{12}{20(20+1)} \left[ \frac{(27)^2}{5} + \frac{(30)^2}{5} + \frac{(83)^2}{5} + \frac{(70)^2}{5} \right] - 3(20+1) = 13.674 \end{aligned}$$

Step 5: Reject  $H_0$  since  $13.674 > 7.815$ .

c. Step 1:  $H_0$ : The three population distributions are all identical

$H_1$ : The three population distributions are not all identical

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .05$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 5.991$ .

Step 4:  $n = n_1 + n_2 + n_3 = 6 + 10 + 6 = 22$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{22(22+1)} \left[ \frac{(93)^2}{6} + \frac{(70)^2}{10} + \frac{(90)^2}{6} \right] - 3(22+1) = 8.822 \end{aligned}$$

Step 5: Reject  $H_0$  since  $8.822 > 5.991$ .

d. Step 1:  $H_0$ : The five population distributions are all identical

$H_1$ : The five population distributions are not all identical

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .05$  and  $df = k - 1 = 5 - 1 = 4$ , the rejection region is  $\chi^2 > 9.488$ .

Step 4:  $n = n_1 + n_2 + n_3 + n_4 + n_5 = 8 + 9 + 8 + 10 + 9 = 44$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} + \frac{R_4^2}{n_4} + \frac{R_5^2}{n_5} \right] - 3(n+1) \\ &= \frac{12}{44(44+1)} \left[ \frac{(210)^2}{8} + \frac{(195)^2}{9} + \frac{(178)^2}{8} + \frac{(212)^2}{10} + \frac{(195)^2}{9} \right] - 3(44+1) = .863 \end{aligned}$$

Step 5: Do not reject  $H_0$  since  $.863 < 9.488$ .

- 15.53 a.** Step 1:  $H_0$ : The population distributions of test scores of fourth-grade students taught by the three methods are all identical

$H_1$ : The population distributions of test scores of fourth-grade students taught by the three methods are not all identical.

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .01$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 9.210$ .

Step 4:

Method I		Method II		Method III	
Score	Rank	Score	Rank	Score	Rank
48	1	55	3	84	11
73	9	85	12	68	6
51	2	70	8	95	15
65	4	69	7	74	10
87	13	90	14	67	5
$n_1 = 5$	$R_1 = 29$	$n_2 = 5$	$R_2 = 44$	$n_3 = 5$	$R_3 = 47$

$$n = n_1 + n_2 + n_3 = 5 + 5 + 5 = 15$$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{15(15+1)} \left[ \frac{(29)^2}{5} + \frac{(44)^2}{5} + \frac{(47)^2}{5} \right] - 3(15+1) = 1.860 \end{aligned}$$

Step 5: Do not reject  $H_0$  since  $1.860 < 9.210$ .

Conclude that the median test scores of all fourth-grade students taught by the three methods are not all different.

- b.** The conclusion in part a of this exercise is the same as that of the corresponding test in Example 12-3.

- 15.55 a.** Step 1:  $H_0$ : The population distributions of delivery times for the four pizza parlors are all identical  
 $H_1$ : The population distributions of delivery times for the four pizza parlors are not all identical

Step 2: Use the  $\chi^2$  distribution.

Step 3: For  $\alpha = .05$  and  $df = k - 1 = 4 - 1 = 3$ , the rejection region is  $\chi^2 > 7.815$ .

Step 4:

Tony's		Luigi's		Angelo's		Kowalski's	
Time	Rank	Time	Rank	Time	Rank	Time	Rank
20.0	3	22.1	7	22.3	8	23.9	9
24.0	10.5	27.0	20	26.0	18.5	24.1	12
18.3	1	20.2	4	24.0	10.5	25.8	16.5
22.0	6	32.0	24	30.1	23	29.0	22
20.8	5	26.0	18.5	28.0	21	25.0	15
19.0	2	24.8	14	25.8	16.5	24.2	13
$n_1 = 6$	$R_1 = 27.5$	$n_2 = 6$	$R_2 = 87.5$	$n_3 = 6$	$R_3 = 97.5$	$n_4 = 6$	$R_4 = 87.5$

$$n = n_1 + n_2 + n_3 + n_4 = 6 + 6 + 6 + 6 = 24$$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} + \frac{R_4^2}{n_4} \right] - 3(n+1) \\ &= \frac{12}{24(24+1)} \left[ \frac{(27.5)^2}{6} + \frac{(87.5)^2}{6} + \frac{(97.5)^2}{6} + \frac{(87.5)^2}{6} \right] - 3(24+1) = 10.250 \end{aligned}$$

Step 5: Reject  $H_0$  since  $10.250 > 7.815$ .

Conclude that the distributions of delivery times are not identical for all four pizza parlors.

- b.** The conclusion in part a of this exercise is the same as that of the corresponding test in Problem 10 of the Self-Review Test in Chapter 12.

**15-12** Chapter 15 Nonparametric Methods

- 15.57** Step 1:  $H_0$ : The population distributions of number of defective parts for the three shifts are all identical  
 $H_1$ : The population distributions of number of defective parts for the three shifts are not all identical  
Step 2: Use the  $\chi^2$  distribution.  
Step 3: For  $\alpha = .05$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 5.991$ .  
Step 4:

First Shift		Second Shift		Third Shift	
Number	Rank	Number	Rank	Number	Rank
23	1	25	2	33	4
36	6	35	5	44	10
32	3	41	9	50	12.5
40	8	38	7	52	14
45	11	50	12.5	60	15
$n_1 = 5$	$R_1 = 29$	$n_2 = 5$	$R_2 = 35.5$	$n_3 = 5$	$R_3 = 55.5$

$$n = n_1 + n_2 + n_3 = 5 + 5 + 5 = 15$$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{15(15+1)} \left[ \frac{(29)^2}{5} + \frac{(35.5)^2}{5} + \frac{(55.5)^2}{5} \right] - 3(15+1) = 3.815 \end{aligned}$$

Step 5: Do not reject  $H_0$  since  $3.815 < 5.991$ .

Conclude that the median number of defective parts is the same for all three shifts.

### Section 15.5

- 15.59** To conduct hypothesis tests about  $\rho$  in Chapter 13, both variables ( $x$  and  $y$ ) must be normally distributed. No such assumption is required for testing a hypothesis about the Spearman rho rank correlation coefficient.

- 15.61** In the following tables,  $d = u - v$ .

a.

$x$	5	10	15	20	25	30	
$y$	17	15	12	14	10	9	
$u$	1	2	3	4	5	6	
$v$	6	5	3	4	2	1	
$d$	-5	-3	0	0	3	5	
$d^2$	25	9	0	0	9	25	$\Sigma d^2 = 68$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(68)}{6(36 - 1)} = 1 - \frac{408}{210} = -.943$$

b.

$x$	27	15	32	21	16	40	8	
$y$	95	81	102	88	75	120	62	
$u$	5	2	6	4	3	7	1	
$v$	5	3	6	4	2	7	1	
$d$	0	-1	0	0	1	0	0	
$d^2$	0	1	0	0	1	0	0	$\Sigma d^2 = 2$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(2)}{7(49 - 1)} = 1 - \frac{12}{336} = .964$$

- 15.63** a. We would expect  $r_s$  to be positive, because as height increases, we would expect weight to increase.  
 b. In the table below,  $u$  and  $v$  denote the ranks for height and weight, respectively, and  $d = u - v$ .

$u$	9.5	3	4.5	4.5	9.5	1	7.5	6	7.5	2	
$v$	8	4	3	5	10	1	7	6	9	2	
$d$	1.5	-1	1.5	-5	-5	0	.5	0	-1.5	0	
$d^2$	2.25	1	2.25	.25	.25	0	.25	0	2.25	0	$\Sigma d^2 = 8.5$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(8.5)}{10(100 - 1)} = 1 - \frac{51}{990} = .948$$

The value of  $r_s$  is positive, which agrees with part a.

- 15.65** a. The regression line has a positive slope, which indicates that as  $x$  increases,  $y$  tends to increase. Therefore, we would expect the Spearman rho rank correlation coefficient to be positive.  
 b. In the following table,  $u$  and  $v$  denote the ranks of  $x$  and  $y$ , respectively, and  $d = u - v$ .

$u$	5	7	2	6	1	4	3	
$v$	4.5	7	2	6	1	3	4.5	
$d$	.5	0	0	0	0	1	-1.5	
$d^2$	.25	0	0	0	0	1	2.25	$\Sigma d^2 = 3.5$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(3.5)}{7(49 - 1)} = 1 - \frac{21}{336} = .938$$

The value of  $r_s$  is positive, which agrees with part a.

- 15.67** a. In the following table,  $u$  and  $v$  denote the ranks of  $x$  and  $y$ , respectively, and  $d = u - v$ .

$u$	6	2	4	7	9	1	8	3	5	
$v$	6	3	5	7	8	2	9	1	4	
$d$	0	-1	-1	0	1	-1	-1	2	1	
$d^2$	0	1	1	0	1	1	1	4	1	$\Sigma d^2 = 10$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(10)}{9(81 - 1)} = 1 - \frac{60}{720} = .917$$

- b. Step 1:  $H_0: \rho_s = 0, H_1: \rho_s > 0$   
 Step 2: Use the Spearman rho rank correlation coefficient test procedure.  
 Step 3: For  $n = 9$  and  $\alpha = .05$ , the rejection region is  $r_s \geq .600$ .  
 Step 4: From part a,  $r_s = .917$ .  
 Step 5: Reject  $H_0$  since  $.917 > .600$ .  
 Conclude that there is a positive relationship between SAT scores and college grade point averages.
- c. The test indicates a positive relationship between the variables  $x$  and  $y$ .

## Section 15.6

- 15.69** A **run** is a sequence of one or more consecutive occurrences of the same outcome in a sequence of occurrences in which there are only two outcomes. The number of runs in a sequence is denoted by  $R$ .
- 15.71** Let  $n_1$  be the number of times the first outcome occurs in a string of outcomes, and let  $n_2$  be the number of times the second outcome occurs. If either  $n_1 > 15$  or  $n_2 > 15$ , the normal approximation may be used.
- 15.73** In Example 15-13, replacing "M" and "F" by "0" and "1", respectively would not affect the test. The values of  $n_1$ ,  $n_2$ , and  $R$  would be unchanged, so the rejection region and the observed value of  $R$  would be the same. Thus, the conclusion would be the same.

**15-14** Chapter 15 Nonparametric Methods

- 15.75** Step 1:  $H_0$ : The sequence of heads and tails is random  
 $H_1$ : The sequence of heads and tails is not random  
Step 2: Using  $n_1$  for "H" and  $n_2$  for "T" yields  $n_1 = 11$  and  $n_2 = 9$ .  
Since  $n_1 \leq 15$  and  $n_2 \leq 15$ , use the runs test with critical values from Table XII.  
Step 3: For  $n_1 = 11$ ,  $n_2 = 9$  and  $\alpha = .05$ , the rejection region is  $R \leq 6$  or  $R \geq 16$ .  
Step 4: The observed value of  $R = 13$ .  
Step 5: Do not reject  $H_0$  since  $6 < 13 < 16$ .  
Conclude that the psychic's claim is not true.
- 15.77** Step 1:  $H_0$ : Diseased and normal trees are randomly mixed in the row  
 $H_1$ : Diseased and normal trees are not randomly mixed in the row  
Step 2: Using  $n_1$  for "N" and  $n_2$  for "D" yields  $n_1 = 13$  and  $n_2 = 7$ .  
Since  $n_1 \leq 15$  and  $n_2 \leq 15$ , use the runs test with critical values from Table XII.  
Step 3: For  $n_1 = 13$ ,  $n_2 = 7$ , and  $\alpha = .05$ , the rejection region is  $R \leq 5$  or  $R \geq 15$ .  
Step 4: The observed value of  $R = 5$ .  
Step 5: Reject  $H_0$  since  $5 = 5$ . Note that  $5 = 5$  indicates that the observed value of the test statistic is "just barely" inside the rejection region.  
Conclude that there is a non-random pattern in the sequence.
- 15.79** Step 1:  $H_0$ : The hits occur randomly among all at-bats  
 $H_1$ : The hits do not occur randomly among all at-bats  
Step 2: Since  $n_1 > 15$  and  $n_2 > 15$ , use the normal distribution.  
Step 3: For  $\alpha = .01$ , the rejection region lies to the left of  $z = -2.58$  and to the right of  $z = 2.58$ .  
Step 4:  $n_1 = 22$ ,  $n_2 = 53$ , and  $R = 37$
- $$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(22)(53)}{22 + 53} + 1 = 32.09$$
- $$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(22)(53)(2 \cdot 22 \cdot 53 - 22 - 53)}{(22 + 53)^2 (22 + 53 - 1)}} = 3.55592776$$
- $$z = \frac{R - \mu_R}{\sigma_R} = \frac{37 - 32.09}{3.55592776} = 1.38$$
- Step 5: Do not reject  $H_0$  since  $1.38 < 2.58$ .  
Conclude that hits occur randomly for this player.
- 15.81** Step 1:  $H_0$ : The sequence of all the state's daily numbers is random  
 $H_1$ : The sequence of all the state's daily numbers is not random  
Step 2: Since  $n_1 > 15$  and  $n_2 > 15$ , use the normal distribution.  
Step 3: For  $\alpha = .025$ , the rejection region lies to the left of  $z = -2.24$  and to the right of  $z = 2.24$ .  
Step 4:  $n_1 = 27$ ,  $n_2 = 23$ , and  $R = 11$
- $$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(27)(23)}{27 + 23} + 1 = 25.84$$
- $$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(27)(23)(2 \cdot 27 \cdot 23 - 27 - 23)}{(27 + 23)^2 (27 + 23 - 1)}} = 3.47640913$$
- $$z = \frac{R - \mu_R}{\sigma_R} = \frac{11 - 25.84}{3.47640913} = -4.27$$
- Step 5: Reject  $H_0$  since  $-4.27 < -2.24$ .  
Conclude that the sequence of all this state's daily numbers is not random.

**Supplementary Exercises**

- 15.83** Let  $p$  be the proportion of all people who would prefer Brand A over Brand B.
- Step 1:  $H_0: p = .50, H_1: p < .50$
- Step 2: Since  $n \leq 25$ , use the binomial distribution.
- Step 3: For  $n = 24$  and  $\alpha = .05$ , the rejection region is  $X \leq 7$ .
- Step 4: The observed value of  $X = 7$ .
- Step 5: Reject  $H_0$  since  $7 = 7$ . Note that  $7 = 7$  indicates that the observed value of the test statistic is “just barely” inside the rejection region.
- Conclude that among all people there is a preference for Brand B over Brand A.
- 15.85** Let  $p$  be the proportion of all such bank customers who prefer an ATM to a human teller.
- Step 1:  $H_0: p = .50, H_1: p > .50$
- Step 2: Since  $n > 25$ , use the normal distribution.
- Step 3: For  $\alpha = .01$ , the rejection region lies to the right of  $z = 2.33$ .
- Step 4: Since 12 of the 200 customers have no preference, the true value of  $n = 200 - 12 = 188$ .  
 $p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 188(.50) = 94, \sigma = \sqrt{npq} = \sqrt{188(.50)(.50)} = 6.85565460$   
 $\text{Since } X = 122, \frac{n}{2} = \frac{188}{2} = 94, \text{ and } X > \frac{n}{2}, z = \frac{(X - .5) - \mu}{\sigma} = \frac{(122 - .5) - 94}{6.85565460} = 4.01.$
- Step 5: Reject  $H_0$  since  $4.01 > 2.33$ .
- Conclude that more than half of all customers of this bank prefer an ATM.
- 15.87** Step 1:  $H_0: \text{Median} = 45 \text{ years}, H_1: \text{Median} > 45 \text{ years}$   
Step 2: Since two buyers were 45 years old, the true value of  $n = 25 - 2 = 23$ . Since  $n \leq 25$ , use the binomial distribution.  
Step 3: For  $n = 23$  and  $\alpha = .05$ , the rejection region is  $X \geq 16$ .  
Step 4: For a right-tailed test, we use the larger value. The observed value of  $X = 16$ .  
Step 5: Reject  $H_0$  since  $16 = 16$ . Note that  $16 = 16$  indicates that the observed value of the test statistic is “just barely” inside the rejection region.
- Conclude that the median age of Harley-Davidson buyers is over 45 years.
- 15.89** Let  $p$  be the proportion of students who spend more than \$650 for books.
- Step 1:  $H_0: \text{Median} = \$650, H_1: \text{Median} \neq \$650$   
Step 2: Since  $n > 25$ , use the normal distribution.  
Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.96$  and to the right of  $z = 1.96$ .  
Step 4:  $n = 35, p = .50$  and  $q = 1 - p = 1 - .50 = .50$   
 $\mu = np = 35(.50) = 17.5, \sigma = \sqrt{npq} = \sqrt{35(.50)(.50)} = 2.95803989$
- We assign a plus sign to every value above \$650 and a minus sign to every value below \$650. (See table on next page.) There are nine plus signs and 26 minus signs. For a two-tailed test, we can use either value. Using the smaller value,  $X = 9, \frac{n}{2} = \frac{35}{2} = 17.5, X < \frac{n}{2}$ , and  

$$z = \frac{(X + .5) - \mu}{\sigma} = \frac{(9 + .5) - 17.5}{2.95803989} = -2.70.$$

(continued on next page)

(continued)

Student	1	2	3	4	5	6	7	8	9	10	11	12
Amount	475	418	680	610	655	488	710	375	250	695	420	610
Sign	-	-	+	-	+	-	+	-	-	+	-	-
Student	13	14	15	16	17	18	19	20	21	22	23	24
Amount	380	98	530	415	757	357	409	611	455	618	395	612
Sign	-	-	-	-	+	-	-	-	-	-	-	-
Student	25	26	27	28	29	30	31	32	33	34	35	
Amount	468	610	780	450	880	490	490	626	850	688	588	
Sign	-	-	+	-	+	-	-	-	+	+	-	

Step 5: Reject  $H_0$  since  $-2.70 < -1.96$ .

Conclude that the median expenditure on textbooks by all such students in 2005–2006 was different from \$650. If we had used the larger value,  $X = 26$ ,  $\frac{n}{2} = \frac{35}{2} = 17.5$ ,  $X > \frac{n}{2}$ , and  $z = \frac{(26 - .5) - 17.5}{2.95803989} = 2.70$ . Our conclusion would be the same.

- 15.91** a. Let  $M$  denote the difference in median gas mileage before and after the installation of governors. For each salesperson, the paired difference = gas mileage before – gas mileage after.

Step 1:  $H_0: M = 0$ ,  $H_1: M < 0$ Step 2: Since  $n \leq 25$ , use the binomial distribution.Step 3: For  $n = 7$  and  $\alpha = .05$ , the rejection region is  $X = 0$ .

Step 4: We assign a plus sign to each salesperson for whom gas mileage was higher before installation of the governors, and a minus sign to each salesperson for whom gas mileage was lower before the installation of the governors.

Before	25	21	27	23	19	18	20
After	26	24	26	25	24	22	23
Sign	-	-	+	-	-	-	-

There are one plus sign and six minus signs. For a left-tailed test, we use the smaller value.

Thus, the observed value of  $X = 1$ .Step 5: Do not reject  $H_0$  since  $1 > 0$ .

Conclude that the use of governors does not tend to increase the median gas mileage for the Gamma Corporation's salespersons' cars.

- b. The conclusion of part a (Do not reject  $H_0$ ) is different from the conclusions of Exercises 15.34 and 10.96 (Reject  $H_0$  in both cases).
- c. The Wilcoxon signed-rank test of Exercise 15.34 and the test based on the  $t$ -distribution of Exercise 10.96 both take into account the magnitudes of the paired differences, but the sign test of part a uses only the signs of the paired differences. This makes the sign test less efficient, so it is more prone to Type II errors (failing to reject  $H_0$  when  $H_0$  is false).

- 15.93** Let  $M$  denote the difference in median blood pressures before and after taking the medication. For each patient, the paired difference = blood pressure before – blood pressure after. Let  $p$  be the proportion of patients for whom blood pressure is higher before the medication than after.

Step 1:  $H_0: M = 0$ ,  $H_1: M > 0$ Step 2: Since  $n > 25$ , use the normal distribution.Step 3: For  $\alpha = .025$ , the rejection region lies to the right of  $z = 1.96$ .

(continued on next page)

(continued)

Step 4: For three of the 35 patients there was no change in blood pressure, so the true value of  $n = 35 - 3 = 32$ .

$$p = .50 \text{ and } q = 1 - p = 1 - .50 = .50$$

$$\mu = np = 32(.50) = 16, \sigma = \sqrt{npq} = \sqrt{32(.50)(.50)} = 2.82842712$$

For a right-tailed test we use the larger value. Thus, the observed value of  $X = 25$ .

$$\text{Then } \frac{n}{2} = \frac{32}{2} = 16, X > \frac{n}{2}, \text{ and } z = \frac{(X - \mu)}{\sigma} = \frac{(25 - .5) - 16}{2.82842712} = 3.01.$$

Step 5: Do not reject  $H_0$  since  $3.01 > 1.96$ .

Conclude that the median blood pressure in all such patients is lower after the medication than before.

- 15.95** Let  $M_A$  and  $M_B$  denote the median scores for all such skaters given by judges A and B, respectively. For each skater, the paired difference = Judge A's score – Judge B's score.

Step 1:  $H_0: M_A = M_B, H_1: M_A \neq M_B$

Step 2: Since  $n \leq 15$ , use the Wilcoxon signed-rank test procedure for the small-sample case.

Step 3: One skater was scored the same by both judges, so the true sample size is  $n = 8 - 1 = 7$ .

For  $n = 7$  and  $\alpha = .05$ , the rejection region is  $T \leq 2$ .

Step 4:

Judge A	Judge B	Differences (A – B)	Absolute Differences	Ranks of Differences	Signed Ranks
5.8	5.4	.4	.4	4.5	+4.5
5.7	5.5	.2	.2	2.5	+2.5
5.6	5.7	-.1	.1	1	-1
5.9	5.4	.5	.5	6	+6
5.8	5.6	.2	.2	2.5	+2.5
5.9	5.3	.6	.6	7	+7
5.8	5.4	.4	.4	4.5	+4.5
5.6	5.6	0	0	—	—

Sum of positive ranks = 27

Sum of absolute values of negative ranks = 1

For a two-tailed test,  $T$  is the smaller sum of ranks. Thus, the observed value of  $T = 1$ .

Step 5: Reject  $H_0$  since  $1 < 2$ .

Conclude that one judge tends to give higher scores than the other.

- 15.97** Let  $M_M$  and  $M_R$  denote the median gas mileages for all such drivers with the M car and the R car, respectively. For each driver, the paired difference = mileage with M car – mileage with R car.

Step 1:  $H_0: M_R = M_M, H_1: M_R > M_M$

Step 2: Since  $n > 15$ , use the normal distribution.

Step 3: For  $\alpha = .025$ , the rejection region lies to the right of  $z = 1.96$ .

Step 4: Since one of the drivers obtained the same mileage with both cars, the true value of  $n = 18 - 1 = 17$ .

$$\mu_T = \frac{n(n+1)}{4} = \frac{17(17+1)}{4} = 76.5$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{17(17+1)(34+1)}{24}} = 21.12463017$$

Sum of positive ranks = 31

Sum of absolute values of negative ranks = 122

For a right-tailed test,  $T$  is the sum of the absolute values of the negative ranks. Thus, the

$$\text{observed value of } T = 122 \text{ and } z = \frac{T - \mu_T}{\sigma_T} = \frac{122 - 76.5}{21.12463017} = 2.15.$$

Step 5: Reject  $H_0$  since  $2.15 > 1.96$ .

Conclude that the R car gets better gas mileage than the M car.

- 15.99** Step 1:  $H_0$ : The population distributions of egg prices in the suburbs and cities are identical  
 $H_1$ : The population distribution of egg prices in the cities lies to the right of the population distribution of egg prices in the suburbs  
Step 2: Since  $n_1 \leq 10$  and  $n_2 \leq 10$ , use the Wilcoxon rank-sum test for small samples.  
Step 3: For  $n_1 = 6$ ,  $n_2 = 7$ , and  $\alpha = .05$ , the rejection region is  $T \geq 54$ .  
Step 4:

City		Suburb	
Price	Rank	Price	Rank
1.49	12	.99	1
1.29	6	1.09	3
1.35	8	1.39	9
1.58	13	1.28	5
1.33	7	1.16	4
1.47	11	1.44	10
		1.05	2
Sum = 57		Sum = 34	

For a one-tailed test with unequal sample sizes,  $T$  is the sum of ranks for the smaller sample. Thus, the observed value of  $T = 57$ .

- Step 5: Reject  $H_0$  since  $57 > 54$ .  
Conclude that egg prices tend to be higher in the city.

- 15.101 a.** Step 1:  $H_0$ : The population distributions of times required for elementary statistics students to complete the assignment with software A and software B are identical  
 $H_1$ : The population distribution of times required for elementary statistics students to complete the assignment using software A lies to the right of the corresponding distribution for software B  
Step 2: Since  $n_1 > 10$  and  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .05$ , the rejection region lies to the right of  $z = 1.65$ .  
Step 4: See table on next page. For a one-tailed test with equal sample sizes,  $T$  is the sum of ranks for the first sample. Thus, the observed value of  $T = 197$ .

$$n_1 = 12, n_2 = 12, \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{12(12 + 12 + 1)}{2} = 150$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{12(12)(12 + 12 + 1)}{12}} = 17.32050808$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{197 - 150}{17.32050808} = 2.71$$

Group A		Group B	
Time	Rank	Time	Rank
123	20	65	2
101	12	115	18.5
112	16	95	8
85	4	100	11
87	5	94	7
133	22	72	3
129	21	60	1
114	17	110	14.5
150	23	99	10
110	14.5	102	13
180	24	88	6
115	18.5	97	9
Sum = 197		Sum = 103	

(continued on next page)

(continued)

Step 5: Reject  $H_0$  since  $2.71 > 1.65$ .

Conclude that the median time required for all students taking elementary statistics at this university to complete this assignment is greater for software A than for software B.

- b. A paired-sample sign test would not be appropriate here. Each of the 24 students uses software A only or software B only. Thus, the samples are not paired, but are independent.

- 15.103** Step 1:  $H_0$ : The population distributions of ages of drivers for the three makes of cars are all identical  
 $H_1$ : The population distributions of ages of drivers for the three makes of cars are not all identical  
Step 2: Use the  $\chi^2$  distribution.  
Step 3: For  $\alpha = .05$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 5.991$ .  
Step 4:

Make of Car					
Rolls-Royce		Mercedes		Cadillac	
Age	Rank	Age	Rank	Age	Rank
64	15.5	61	12	52	6
61	12	47	3.5	63	14
70	20	66	17	39	1
68	18.5	71	21	55	8.5
55	8.5	44	2	50	5
64	15.5	53	7	47	3.5
68	18.5	58	10	61	12
$n_1 = 7$	$R_1 = 108.5$	$n_2 = 7$	$R_2 = 72.5$	$n_3 = 7$	$R_3 = 50$

$$n = n_1 + n_2 + n_3 = 7 + 7 + 7 = 21$$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{21(21+1)} \left[ \frac{(108.5)^2}{7} + \frac{(72.5)^2}{7} + \frac{(50)^2}{7} \right] - 3(21+1) = 18.462 \end{aligned}$$

Step 4: Reject  $H_0$  since  $18.462 > 5.991$ .

Conclude that there is a difference in median age of drivers for each of the three makes of cars.

- 15.105** Step 1:  $H_0$ : The population distributions of lengths of drives for the three brands of golf balls all identical  
 $H_1$ : The population distributions of lengths of drives for the three brands of golf balls are not all identical  
Step 2: Use the  $\chi^2$  distribution.  
Step 3: For  $\alpha = .05$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 5.991$ .  
Step 4:

Brand					
A		B		C	
Length	Rank	Length	Rank	Length	Rank
275	13	245	1	267	11
266	10	256	3.5	283	16
301	18	261	7	259	5.5
281	15	270	12	250	2
288	17	259	5.5	263	9
277	14	262	8	256	3.5
$n_1 = 6$	$R_1 = 87$	$n_2 = 6$	$R_2 = 37$	$n_3 = 6$	$R_3 = 47$

(continued on next page)

(continued)

$$\begin{aligned}
 n &= n_1 + n_2 + n_3 = 6 + 6 + 6 = 18 \\
 H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\
 &= \frac{12}{18(18+1)} \left[ \frac{(87)^2}{6} + \frac{(37)^2}{6} + \frac{(47)^2}{6} \right] - 3(18+1) = 8.187
 \end{aligned}$$

Step 5: Reject  $H_0$  since  $8.187 > 5.991$ .

Conclude that the median lengths of drives by this golfer are not identical for all three brands of golf balls.

- 15.107** a. In states with a small percentage of students who take the SAT, those taking it are usually the better students and tend to score high. In other states the exam is taken by a broad spectrum of students, so the average scores would tend to be lower. Thus, we would expect  $\rho_s$  to be negative.
- b. In the following table  $u$  and  $v$  are the ranks of Math SAT scores and percentage of graduates taking the SAT test, respectively, and  $d = u - v$ .

$u$	4	2	9	6	7	3	1	8	5	10	
$v$	10	7	3.5	5	3.5	9	6	1	8	2	
$d$	-6	-5	5.5	1	3.5	-6	-5	7	-3	8	
$d^2$	36	25	30.25	1	12.25	36	25	49	9	64	$\Sigma d^2 = 287.5$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(287.5)}{10(100 - 1)} = 1 - \frac{1725}{990} = -.742$$

 $r_s$  is negative, which is consistent with the answer to part a.

- c. Step 1:  $H_0: \rho_s = 0, H_1: \rho_s \neq 0$   
 Step 2: Use the Spearman rho rank correlation coefficient test procedure.  
 Step 3: For  $n = 10$  and  $\alpha = .05$ , the rejection region is  $r_s \leq -.648$  or  $r_s \geq .648$ .  
 Step 4: From part b,  $r_s = -.742$ .  
 Step 5: Reject  $H_0$  since  $-.742 < -.648$ .

Conclude that there is a relationship between Math SAT scores and percentage of graduates taking the SAT test.

- 15.109** a. We would expect to reject  $H_0$ .
- b. Step 1:  $H_0: \rho_s = 0, H_1: \rho_s < 0$   
 Step 2: Use the Spearman rho rank correlation coefficient test procedure.  
 Step 3: For  $n = 8$  and  $\alpha = .05$ , the rejection region is  $r_s \leq -.643$ .  
 Step 4: In the following table,  $u$  and  $v$  are the ranks of driving experience and insurance premiums, respectively, and  $d = u - v$ .

$u$	2	1	5	4	6	3	8	7	
$v$	6	8	3	7	2	4	1	5	
$d$	-4	-7	2	-3	4	-1	7	2	
$d^2$	16	49	4	9	16	1	49	4	$\Sigma d^2 = 148$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(148)}{8(64 - 1)} = 1 - \frac{888}{504} = -.762$$

Step 5: Reject  $H_0$  since  $-.762 < -.643$ .

Conclude that there is a negative relationship between driving experience and insurance premiums (as we expected in part a).

- 15.111** Step 1:  $H_0$ : The sequence of defective and good tools is random  
 $H_1$ : The sequence of defective and good tools is not random  
Step 2: Using  $n_1$  for "G" and  $n_2$  for "D" yields  $n_1 = 13$  and  $n_2 = 5$ . Since  $n_1 \leq 15$  and  $n_2 \leq 15$ , use the runs test with critical values from Table XII.  
Step 3: For  $\alpha = .05$ , the rejection region is  $R \leq 4$  or  $R \geq 12$ .  
Step 4: The observed value of  $R = 7$ .  
Step 5: Do not reject  $H_0$  since  $4 < 7 < 12$ .  
Conclude that the sequence of defective and good tools is random.
- 15.113** Step 1:  $H_0$ : The wins and losses are in random order  
 $H_1$ : The wins and losses are not in random order  
Step 2: Using  $n_1$  for "W" and  $n_2$  for "L" yields  $n_1 = 10$  and  $n_2 = 20$ . Since  $n_2 > 15$ , use the normal distribution.  
Step 3: For  $\alpha = .02$ , the rejection lies to the left of  $z = -2.33$  and to the right of  $z = 2.33$ .  
Step 4:  $R = 17$ ,  $\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(10)(20)}{10 + 20} + 1 = 14.33$   

$$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(10)(20)(2 \cdot 10 \cdot 20 - 10 - 20)}{(10 + 20)^2 (10 + 20 - 1)}} = 2.38128077$$
  

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{17 - 14.33}{2.38128077} = 1.12$$
  
Step 5: Do not reject  $H_0$  since  $1.12 < 2.33$ .  
Conclude that the wins and losses are in random order.
- 15.115** a. Lay out a route for a test run in a typical city. Let each driver drive each car several times on this route. Make sure that the length of the route is selected so that the total mileage of all test runs for each car does not exceed 500 miles. Then, calculate the gas mileage for each test run. Next, compute the median gas mileage for each of the three cars and compare them in the article without using any inferential statistical procedures.
- b. Since the gas mileage cannot be assumed to be normally distributed, the ANOVA procedure of Chapter 12 would not be appropriate. Instead, use the Kruskal-Wallis test on the data collected in part a for the three cars. Test the null hypothesis that the population distributions of gas mileage for the three cars are all identical against the alternative hypothesis that the three distributions are not all identical. Select an appropriate significance level. Then, write a report explaining and interpreting these results, being sure to point out what the test results imply about the median gas mileage for the three cars.
- 15.117** a. Take random samples of carpenters, plumbers, electricians, and masons from your city and record each worker's hourly wage. Note that random samples will be representative of the populations of the four trades and should have approximately the same proportions of union members, respectively, as the four populations. Next, find the median hourly wage for each sample and compare them in the article without using any inferential statistical procedures.
- b. Since the hourly wages cannot be assumed to be normally distributed, the ANOVA procedure of Chapter 12 would not be appropriate. Instead, use the Kruskal-Wallis test on the data collected in part a for the four trades. Test the null hypothesis that the population distributions of hourly wages are all identical against the alternative that the four distributions are not all identical. Select an appropriate significance level. Then write a report explaining and interpreting these results, being sure to point out what the test results imply about the median hourly wages for the four trades.
- 15.119** a. Since the data are paired (two grades for each essay) the department head could use the sign test for paired data or the Wilcoxon signed-rank test.

- b. We will use the Wilcoxon signed-rank test. Let  $M_A$  and  $M_B$  denote the median grades for all such essays by Professor A and the instructor, respectively. For each essay, the paired difference = Professor A's grade – instructor's grade.

Step 1:  $H_0: M_A = M_B$ ,  $H_1: M_A \neq M_B$

Step 2: Since  $n \leq 15$ , use the Wilcoxon signed-rank test for small samples.

Step 3: For  $n = 10$  and  $\alpha = .05$ , the rejection region is  $T \leq 8$ .

Step 4:

Professor A	Instructor	Differences (Prof. A – Inst.)	Absolute Differences	Ranks of Differences	Signed Ranks
75	80	-5	5	6.5	-6.5
62	50	+12	12	10	+10
90	85	+5	5	6.5	+6.5
48	55	-7	7	9	-9
67	63	+4	4	3.5	+3.5
82	78	+4	4	3.5	+3.5
94	89	+5	5	6.5	+6.5
76	81	-5	5	6.5	-6.5
78	75	+3	3	2	+2
84	83	+1	1	1	+1

Sum of positive ranks = 33

Sum of absolute values of negative ranks = 22

For a two-tailed test,  $T$  is the smaller sum of ranks. Thus, the observed value of  $T = 22$ .

Step 5: Do not reject  $H_0$  since  $22 > 8$ .

Conclude that the instructor does not tend to grade higher or lower than Professor A.

- c. To determine whether or not the instructor and Professor A tend to rank the essays similarly, Spearman's rho rank correlation coefficient would be appropriate. A positive value of  $r_s$  would suggest some consistency. The appropriate hypotheses would be  $H_0: \rho_s = 0$  and  $H_1: \rho_s > 0$ .
- d. Step 1:  $H_0: \rho_s = 0$ ,  $H_1: \rho_s > 0$   
 Step 2: Use the Spearman rho rank correlation coefficient test procedure.  
 Step 3: For  $n = 10$  and  $\alpha = .05$ , the rejection region is  $r_s \geq .564$ .  
 Step 4: In the following table,  $u$  and  $v$  are the ranks of Professor A's grades and the instructor's grades, respectively, and  $d = u - v$ .

$u$	4	2	9	1	3	7	10	5	6	8	
$v$	6	1	9	2	3	5	10	7	4	8	
$d$	-2	1	0	-1	0	2	0	-2	2	0	
$d^2$	4	1	0	1	0	4	0	4	4	0	$\sum d^2 = 18$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(18)}{10(100 - 1)} = 1 - \frac{108}{990} = .891$$

Step 5: Reject  $H_0$  since  $.891 > .564$ .

Conclude that the instructor and Professor A tend to grade such essays similarly.

- 15.121** a. To base a test on the linear correlation coefficient of Chapter 13 requires that both variables (GPAs and SAT scores) be normally distributed.
- b. If the assumptions of part a are not satisfied, the test based on Spearman's rho rank correlation coefficient may be used.
- 15.123** Place Bs and As alternately, starting with a B until all 10 As have been used, then finish with the remaining Bs: B A B A B A B A B A B A B A B B B B B  
 This sequence has 21 runs. Any sequence which begins with an A or has two or more A's in a row will have less than 21 runs.

- 15.125** Step 1:  $H_0$ : The sequence of heads and tails is random  
 $H_1$ : The sequence of heads and tails is not random  
Step 2: Using  $n_1$  for "H" and  $n_2$  for "T" yields  $n_1 = 10$  and  $n_2 = 10$ .  
Since  $n_1 \leq 15$  and  $n_2 \leq 15$ , use the runs test with critical values from Table XII.  
Step 3: For  $n_1 = 10$ ,  $n_2 = 10$  and  $\alpha = .05$ , the rejection region is  $R \leq 6$  or  $R \geq 16$ .  
Step 4: The observed value of  $R = 18$ .  
Step 5: Reject  $H_0$  since  $18 > 16$ .  
Conclude that the sequence of heads and tails is not random. Therefore, the professor was justified in accusing the student of not actually tossing the coin.

### Self-Review Test

1. b      2. a      3. a      4. c      5. b      6. c  
7. a, b, d    8. b      9. c      10. c     11. a, b    12. c      13. b

- 14.** Let  $p$  be the probability that the juror selected from the pool is a woman.  
Step 1:  $H_0: p = .50$ ,  $H_1: p \neq .50$   
Step 2: Since  $n \leq 25$ , use the binomial distribution.  
Step 3: For  $n = 12$  and  $\alpha = .05$ , the rejection region is  $X \leq 2$  or  $X \geq 10$ .  
Step 4: The observed value of  $X = 2$ .  
Step 5: Reject  $H_0$  since  $2 = 2$ . Note that  $2 = 2$  indicates that the observed value of the test statistic is "just barely" inside the rejection region.  
Conclude that the selection process is biased with respect to gender.

- 15.** Let  $p$  be the proportion of Americans in favor of putting Social Security money into personal retirement accounts.  
Step 1:  $H_0: p = .50$ ,  $H_1: p > .50$   
Step 2: Since  $n > 25$ , use the normal distribution.  
Step 3: For  $\alpha = .025$ , the rejection region lies to the right of  $z = 1.96$ .  
Step 4:  $n = 1000$ ,  $p = .5$ , and  $q = 1 - p = 1 - .5 = .5$   
 $\mu = np = 1000(.50) = 500$ ,  $\sigma = \sqrt{npq} = \sqrt{1000(.50)(.50)} = 15.81138830$   
 $\text{Since } X = 520, \frac{n}{2} = \frac{1000}{2} = 500, \text{ and } X > \frac{n}{2}, z = \frac{(X - \mu)}{\sigma} = \frac{(520 - .5) - 500}{15.81138830} = 1.23$ .  
Step 5: Do not reject  $H_0$  since  $1.23 < 1.96$ .  
Do not conclude that over half of Americans favor this proposal.

- 16.** Step 1:  $H_0$ : Median = \$65,  $H_1$ : Median > \$65  
Step 2: Since  $n \leq 25$ , use the binomial distribution.  
Step 3: For  $n = 12$  and  $\alpha = .05$ , the rejection region is  $X \geq 10$ .  
Step 4: We assign a plus sign to each amount above \$65 and a minus sign to each amount below \$65.

Customer	1	2	3	4	5	6	7	8	9	10	11	12
Amount	88	69	141	28	106	45	32	51	78	54	110	83
Sign	+	+	+	-	+	-	-	-	+	-	+	+

There are seven plus signs and five minus signs. For a right-tailed test, we use the larger value. Thus, the observed value of  $X = 7$ .

- Step 5: Do not reject  $H_0$  since  $7 < 10$ .  
Conclude that the median amount spent by all customers at this store after the campaign does not exceed \$65.

17. Let  $p$  be the proportion of incomes that exceed \$20,264.
- Step 1:  $H_0$ : Median = \$20,264,  $H_1$ : Median  $\neq$  \$20,264
- Step 2: Since  $n > 25$ , use the normal distribution.
- Step 3: For  $\alpha = .01$ , the rejection regions lies to the left of  $-2.58$  and to the right of  $z = 2.58$ .
- Step 4:  $n = 400$ ,  $p = .5$ , and  $q = 1 - p = 1 - .5 = .5$
- $$\mu = np = 400(.50) = 200; \text{ and } \sigma = \sqrt{npq} = \sqrt{400(.50)(.50)} = 10$$
- For a two-tailed test, we can use either value. Using the larger value,  $X = 229$ ,
- $$\frac{n}{2} = \frac{400}{2} = 200, X > \frac{n}{2}, \text{ and } z = \frac{(X - .5) - \mu}{\sigma} = \frac{(229 - .5) - 200}{10} = 2.85.$$
- Step 5: Reject  $H_0$  since  $2.85 > 2.58$ .
- Conclude that the median income of women living alone is different from \$20,264.
- If we had used the smaller value,  $X = 171$ ,  $\frac{n}{2} = \frac{400}{2} = 200$ ,  $X < \frac{n}{2}$ , and  $z = \frac{(171 + .5) - 200}{10} = -2.85$ . Our conclusion would be the same.
18. For each adult, the paired difference = cholesterol level before diet – cholesterol level after diet.
- | Before | After | Differences<br>(Before – After) | Absolute<br>Differences | Ranks of<br>Differences | Signed<br>Ranks | Sign |
|--------|-------|---------------------------------|-------------------------|-------------------------|-----------------|------|
| 210    | 193   | +17                             | 17                      | 7                       | +7              | +    |
| 180    | 186   | -6                              | 6                       | 2                       | -2              | -    |
| 195    | 186   | +9                              | 9                       | 3.5                     | +3.5            | +    |
| 220    | 223   | -3                              | 3                       | 1                       | -1              | -    |
| 231    | 220   | +11                             | 11                      | 5                       | +5              | +    |
| 199    | 183   | +16                             | 16                      | 6                       | +6              | +    |
| 224    | 233   | -9                              | 9                       | 3.5                     | -3.5            | -    |
- a. Let  $M$  denote the difference in the median cholesterol level before and after the diet.
- Step 1:  $H_0$ :  $M = 0$ ,  $H_1$ :  $M \neq 0$
- Step 2: Since  $n \leq 25$ , use the binomial distribution.
- Step 3: For  $n = 7$  and  $\alpha = .05$ , the rejection region is  $X = 0$  and  $X = 7$ .
- Step 4: From the preceding table, there are four plus signs and three minus. For a two-tailed test, we can use either value. Using the larger value,  $X = 4$ .
- Step 5: Do not reject  $H_0$  since  $0 < 4 < 7$ .
- Conclude that the median cholesterol level before the diet is the same as after the diet.
- If we had used the smaller value,  $X = 3$ ,  $0 < 3 < 7$ , and our conclusion would be the same.
- b. Let  $M_B$  and  $M_A$  be the median cholesterol levels before and after the diet, respectively.
- Step 1:  $H_0$ :  $M_A = M_B$ ,  $H_1$ :  $M_A \neq M_B$
- Step 2: Since  $n \leq 15$ , use the Wilcoxon signed-rank test for the small-sample case.
- Step 3: For  $n = 7$  and  $\alpha = .05$ , the rejection region is  $T \leq 2$ .
- Step 4: From the preceding table:
- Sum of positive ranks = 21.5
  - Sum of absolute values of negative ranks = 6.5
  - For a two-tailed test,  $T$  is the smaller sum of ranks. Thus, the observed value of  $T = 6.5$ .
- Step 5: Do not reject  $H_0$  since  $6.5 > 2$ .
- Conclude that the median cholesterol level before the diet is the same as after the diet.
- c. The conclusions of parts a and b are the same.
19. Let  $M$  denote the difference in the median age of such artifacts dated by Method I and by Method II. For each artifact, the paired difference = age dated by Method I – age dated by Method II. Let  $p$  be the proportion of artifacts for which the age estimate from Method I is greater than that from Method II.
- Step 1:  $H_0$ :  $M = 0$ ,  $H_1$ :  $M \neq 0$
- Step 2: Since  $n > 25$ , use the normal distribution.
- Step 3: For  $\alpha = .02$ , the rejection region lies to the left of  $z = -2.33$  and to the right of  $z = 2.33$ .

(continued on next page)

(continued)

Step 4: For two of the 33 artifacts, there was no difference in ages for the two methods, so the true value of  $n = 33 - 2 = 31$ .

$$p = .5 \text{ and } q = 1 - p = 1 - .5 = .5$$

$$\mu = np = 31(.50) = 15.5, \sigma = \sqrt{npq} = \sqrt{31(.50)(.50)} = 2.78388218$$

For a two-tailed test, we can use either value. Using the larger value,  $X = 20$ ,

$$\frac{n}{2} = \frac{31}{2} = 15.5, X > \frac{n}{2}, \text{ and } z = \frac{(X - \mu)}{\sigma} = \frac{(20 - .5) - 15.5}{2.78388218} = 1.44$$

Step 5: Do not reject  $H_0$  since  $1.44 < 2.33$ .

Conclude that the median estimated ages of such artifacts is the same for the two methods.

If we had used the smaller value,  $X = 11, \frac{n}{2} = \frac{31}{2} = 15.5, X < \frac{n}{2}, \text{ and } z = \frac{(11 + .5) - 15.5}{2.78388218} = -1.44$ .

Our conclusion would be the same.

20. Let  $M_F$  and  $M_S$  denote the median GPAs for all sophomore electrical engineering majors in the fall and spring semesters, respectively. For each student, the paired difference = fall GPA – spring GPA.

Step 1:  $H_0: M_S = M_F, H_1: M_S < M_F$

Step 2: Since  $n \leq 15$ , use the Wilcoxon signed-rank test procedure for the small-sample case.

Step 3: One student's GPA was the same for both semesters, so the true value of  $n = 10 - 1 = 9$ .

For  $n = 9$  and  $\alpha = .05$ , the rejection region is  $T \leq 8$ .

Step 4:

Fall	Spring	Differences (Fall – Spring)	Absolute Differences	Ranks of Differences	Signed Ranks
3.20	3.15	+.05	.05	2.5	+2.5
3.56	3.40	+.16	.16	6	+6
3.05	2.88	+.17	.17	7.5	+7.5
3.78	3.67	+.11	.11	4	+4
4.00	4.00	0	0	—	—
2.85	3.00	-.15	.15	5	-5
3.33	3.30	+.03	.03	1	+1
2.67	3.05	-.38	.38	9	-9
3.00	2.95	+.05	.05	2.5	+2.5
3.67	3.50	+.17	.17	7.5	+7.5

Sum of positive ranks = 31

Sum of absolute values of negative ranks = 14

For a left-tailed test,  $T$  is the sum of the absolute values of the negative ranks. Thus, the observed value of  $T = 14$ .

Step 5: Do not reject  $H_0$  since  $14 > 8$ .

Conclude that the median GPA for all sophomore electrical engineering majors at this university is not lower in the spring semester than in the fall semester.

21. Let  $M_A$  and  $M_B$  denote the median memory test scores after and before the course, respectively.

Step 1:  $H_0: M_A = M_B, H_1: M_A > M_B$

Step 2: Since  $n > 15$ , use the normal distribution.

Step 3: For  $\alpha = .025$ , the rejection region lies to the right of  $z = 1.96$ .

Step 4: Since three of the 30 students scored the same before and after the course, the true value of  $n = 30 - 3 = 27$ .

$$\mu_T = \frac{n(n+1)}{4} = \frac{27(27+1)}{4} = 189$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{27(27+1)(54+1)}{24}} = 41.62331078$$

(continued on next page)

(continued)

Sum of positive ranks = 102

Sum of absolute values of negative ranks = 276

For a right-tailed test,  $T$  is the sum of the absolute values of the negative ranks. Thus, the

$$\text{observed value of } T = 276 \text{ and } z = \frac{T - \mu_T}{\sigma_T} = \frac{276 - 189}{41.62331078} = 2.09.$$

Step 5: Reject  $H_0$  since  $2.09 > 1.96$ .

Conclude that the course tends to improve scores on memory tests.

22. Step 1:  $H_0$ : The population distributions of commuting times for both routes are identical  
 $H_1$ : The population distributions of commuting times for both routes are not identical  
Step 2: Since  $n_1 \leq 10$  and  $n_2 \leq 10$ , use the Wilcoxon rank-sum test for small samples.  
Step 3: For  $n_1 = n_2 = 8$  and  $\alpha = .05$ , the rejection region is  $T \leq 49$  or  $T \geq 87$ .  
Step 4:

Route I		Route II	
Time	Rank	Time	Rank
45	12	38	3.5
43	9.5	40	6
38	3.5	39	5
56	16	42	8
41	7	50	15
43	9.5	37	2
46	13.5	46	13.5
44	11	36	1
	Sum = 82		Sum = 54

For a two-tailed test with equal sample sizes, we can use the sum of ranks for either sample.

Using the sum of ranks for the first sample,  $T = 82$ .Step 5: Do not reject  $H_0$  since  $82 < 87$ .

Conclude that the median commuting time is the same for both routes.

23. Step 1:  $H_0$ : The population distributions of times to prepare such tax returns for employees A and B are identical  
 $H_1$ : The population distributions of times to prepare such tax returns for employees A and B are not identical  
Step 2: Since,  $n_1 > 10$  and  $n_2 > 10$ , use the normal distribution.  
Step 3: For  $\alpha = .025$ , so the rejection region lies to the left of  $z = -2.24$  and to the right of  $z = 2.24$ .  
Step 4: For a two-tailed test with equal sample sizes, we can use the sum of ranks for either sample.  
Using the sum of ranks for the first sample,  $T = 298$ .

 $n_1 = 18, n_2 = 18$ 

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{18(18 + 18 + 1)}{2} = 333$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{18(18)(18 + 18 + 1)}{12}} = 31.60696126$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{298 - 333}{31.60696126} = -1.11$$

Step 5: Do not reject  $H_0$  since  $-1.11 > -2.24$ .

Conclude that there is not a difference in the median times taken by A and B to prepare such tax returns.

24. a. Step 1:  $H_0$ : The population distributions of reported cases of telemarketing fraud in the three cities are all identical  
 $H_1$ : The population distributions of reported cases of telemarketing fraud in the three cities are not all identical  
Step 2: Use the  $\chi^2$  distribution.  
Step 3: For  $\alpha = .025$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 7.378$ .  
Step 4:

City A		City B		City C	
Number	Rank	Number	Rank	Number	Rank
53	11	29	1	75	16
46	6	35	4	49	8
59	12	44	5	62	14
33	3	31	2	68	15
60	13	50	9	52	10
		48	7		
$n_1 = 5$	$R_1 = 45$	$n_2 = 6$	$R_2 = 28$	$n_3 = 5$	$R_3 = 63$
$n = n_1 + n_2 + n_3 = 5 + 6 + 5 = 16$					

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{16(16+1)} \left[ \frac{(45)^2}{5} + \frac{(28)^2}{6} + \frac{(63)^2}{5} \right] - 3(16+1) = 7.653 \end{aligned}$$

Step 5: Reject  $H_0$  since  $7.653 > 7.378$ .

Conclude that the distributions of numbers of such reported cases for the three cities are not all identical.

- b. Step 1:  $H_0$ : The population distributions of reported cases of telemarketing fraud in the three cities are all identical  
 $H_1$ : The population distributions of reported cases of telemarketing fraud in the three cities are not all identical  
Step 2: Use the  $\chi^2$  distribution.  
Step 3: For  $\alpha = .01$  and  $df = k - 1 = 3 - 1 = 2$ , the rejection region is  $\chi^2 > 9.210$ .  
Step 4: From part a,  $H = 7.653$ .  
Step 5: Do not reject  $H_0$  since  $7.653 < 9.210$ .  
Conclude that the distributions of numbers of such reported cases for the three cities are all identical.

- c. Parts a and b show that the sample does not support the alternative hypothesis very strongly because lowering the significance level from .025 to .01 reverses the conclusion.

25. a. Since  $y$  tends to increase as  $x$  increases, we would expect the value of the Spearman rho rank correlation coefficient to be positive.

- b. In the table below,  $u$  and  $v$  are the ranks of  $x$  and  $y$ , respectively, and  $d = u - v$ .

	7	5	9	1	6	2.5	10	8	4	2.5	
$v$	8	5	9	2	7	4	10	6	3	1	
$d$	-1	0	0	-1	-1	-1.5	0	2	1	1.5	
$d^2$	1	0	0	1	1	2.25	0	4	1	2.25	$\Sigma d^2 = 12.5$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(12.5)}{10(100 - 1)} = 1 - \frac{75}{990} = .924$$

$r_s$  is positive, which agrees with the answer to part a.

c. Step 1:  $H_0: \rho_s = 0, H_1: \rho_s > 0$

Step 2: Use the Spearman rho rank correlation coefficient test procedure.

Step 3: For  $n = 10$  and  $\alpha = .025$ , the rejection region is  $r_s \geq .648$ .

Step 4: From part b,  $r_s = .924$ .

Step 5: Reject  $H_0$  since  $.924 > .648$ .

Conclude that there is a positive relationship between home runs and runs batted in.

26. Step 1:  $H_0$ : Keepers occur randomly in the sequence of fish

$H_1$ : Keepers do not occur randomly in the sequence of fish

Step 2: Using  $n_1$  for "K" and  $n_2$  for "S" yields  $n_1 = 7$  and  $n_2 = 7$ .

Since  $n_1 \leq 15$  and  $n_2 \leq 15$ , use the runs test with critical values from Table XII.

Step 3: For  $n_1 = 7, n_2 = 7$  and  $\alpha = .05$ , the rejection region is  $R \leq 3$  or  $R \geq 13$ .

Step 4: The observed value of  $R = 6$ .

Step 5: Do not reject  $H_0$  since  $3 < 6 < 13$ .

Conclude that this sequence of keepers and smaller bass is random. Therefore, this sequence does not support Ramon's theory.

27. Step 1:  $H_0$ : The wins and losses occur randomly among the 54 games

$H_1$ : The wins and losses do not occur randomly among the 54 games

Step 2: Since  $n_1 > 15$  and  $n_2 > 15$ , use the normal distribution.

Step 3: For  $\alpha = .05$ , the rejection region lies to the left of  $z = -1.96$  and to the right of  $z = 1.96$ .

Step 4:  $n_1 = 30, n_2 = 24, R = 15$

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(30)(24)}{30 + 24} + 1 = 26.67$$

$$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(30)(24)(2 \cdot 30 \cdot 24 - 30 - 24)}{(30 + 24)^2 (30 + 24 - 1)}} = 3.59361185$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{15 - 26.67}{3.59361185} = -3.25$$

Step 5: Reject  $H_0$  since  $-3.25 < -1.96$ .

Conclude that the wins and losses do not occur randomly among the 54 games.

This page is intentionally left blank

# INDEX

## A

actuarial science, 250–251  
addition rule  
    defined, 180, 181  
    for mutually exclusive events, 182–183  
alternative hypothesis  
    defined, 405, 406  
    example, 406  
    left-tailed test, 410  
    one-way ANOVA, 573  
    right-tailed test, 411  
    two-tailed test, 409  
analysis of variance (ANOVA)  
    alternative hypothesis, 573  
    application of, 570  
    assumptions, 570  
    defined, 569  
    degrees of freedom, 574, 575  
    F distribution, 567–569  
    mean square between samples, 570  
    mean square with samples, 570  
    one-way, 569–581  
    pairwise comparisons, 587  
    performing, 567  
    performing (all samples not same size), 574–577  
    performing (all samples same size), 573–574  
    rejection/nonrejection regions, 574, 575  
    as right-tailed, 570  
    table, 576  
    technology instruction, 588–590  
    test statistic value, calculating, 570–573  
uses and misuses, 581  
variance between samples, 570  
variance within samples, 570  
ANOVA. *See* analysis of variance  
applied statistics, 2  
arithmetic mean. *See* mean  
arrangements. *See* permutations

## B

bar graphs  
    axes, truncating, 66  
    constructing, 32–33  
    defined, 33  
    illustrated, 32  
    in Minitab, 80  
Poisson probability distribution, 247  
probability distribution, 232, 234

basic formulas  
    defined, 101  
    grouped data calculation with, 136  
    ungrouped data calculation with, 135  
Beach to Beacon data set, B-3–B-4  
bell-shaped distribution  
    defined, 115  
    example, 116  
Bernoulli trials  
    defined, 226  
    failure, 227  
    success, 227  
between-samples sum of squares (SSB), 571, 576  
bias  
    caution, 351  
    defined, A7  
bimodal data sets, 93  
binomial distribution, 226–239  
    bar graph, 232, 234, 235  
    binomial formula and, 228–232  
    constructing, 231–232  
    defined, 226, 228  
    mean, 236  
    normal approximation to, 297–304  
    parameters, 229  
    probability of failure, 227  
    probability of success, 227, 234–235  
    shape of, 234–235  
    standard deviation, 236  
    table, 232–234  
    technology instruction, 258–262  
binomial experiments  
    conditions, 226  
    conditions, verifying, 227–228  
    defined, 226  
binomial formula  
    calculating probability with, 229–232  
    defined, 228  
binomial probabilities table  
    illustrated, C2–C10  
    using, 232–234  
binomial random variables, 228  
box-and-whisker plots  
    constructing, 124–125  
    defined, 123, 124  
    illustrated, 125  
    inner fences, 124  
    outer fences, 125  
    uses, 123

**C**

candidate data set, B-5  
 case studies  
     average is over, 90  
     average student debt, 422  
     becoming less green, 7  
     car insurance costs, 43  
     coffee drinking quantities, 46  
     commute times, 477–478  
     company advertising expenditures, 3  
     education pays, 92  
     employee's financial stress levels, 33  
     fairness of raising taxes on the rich, 443  
     global birth and death rates, 246  
     lottery jackpot winning probability, 192  
     lunch from home, 386  
     NFL ticket prices, 89  
     \$1,000 downpour, 221–222  
     registered nurse earnings, 370  
     spread versus variability and dispersion, 116  
     time taken to run road race, 269–270  
     Wall Street honesty/morality, 530  
     weight worry, 162, 501–502  
     weights and heights for NFL players, 601–602  
     will children be better off than parents? 32  
     woman life ratings, 4  
     work commute length, 42  
 categorical variables, 11–12  
 census  
     defined, 6, A2  
     impossibility of conducting, A4  
     target population, A2  
 central limit theorem  
     defined, 333  
     for large samples, 336  
     sample proportions, 345  
 central tendency. *See* measures of central tendency  
 chance errors, A5  
 chance variables. *See* random variables  
 Chebyshev's theorem  
     applying, 115  
     defined, 114  
     mean for, 114  
 chi-square distribution  
     curves, 522  
     defined, 522  
     degrees of freedom, 522  
     symbol, 521  
     table, C23  
     table, reading, 522–524  
     values, 521  
 chi-square tests  
     goodness-of-fit, 525–534  
     population variance, 546–552  
     technology instruction, 562–565  
     test of homogeneity, 540–542  
     test of independence, 535–540  
     types of, 521  
 City data set, B-1–B-2  
 classes  
     boundaries, 37, 38  
     defined, 36–38  
     first, lower limit of, 39  
     less-than method for writing, 43–45  
     midpoints, 37, 38  
     number of, 38

single-valued, 45–47  
 size, 37  
 width, 37, 38  
 classical probability rule, 153–154  
 cluster sampling, A9  
 clusters, A9  
 coefficient of determination  
     calculating, 612  
     defined, 609, 612  
     SSR and, 611  
     SST and, 610  
 coefficient of  $x$ , 593  
 combinations  
     defined, 189  
     finding number of, 190–191  
     formula, 190, 191  
     notation, 190  
     number of, 190  
     technology instruction, 258–262  
 combined mean, 98  
 complementary events  
     calculating probability of, 166–167  
     defined, 165  
     Venn diagrams, 166  
 compound events  
     defined, 149  
     illustrated, 150  
     probability calculation of, 153–154  
     Venn and tree diagrams, 150  
 conditional probability  
     calculating, 160–161, 173–174  
     defined, 160  
     tree diagram, 161  
 conditions, probability distribution  
     defined, 213  
     verifying, 214  
 confidence intervals  
     constructing, 368–369  
     constructing, using  $t$  distribution, 378–379  
     defined, 363  
     difference between means (population standard deviation unknown and unequal), 481–482  
     difference between means (population standard deviation unknown but equal), 471–472  
     difference between population means, 465–466  
     finding, 366–368  
     illustrated, 367  
     margin of error, 365  
     paired samples, 489–490  
     point estimate, 634  
     point estimator, 489–490  
     population mean, 365  
     for population mean, using  $t$  distribution, 377–380  
     for population proportion, 384, 385  
     population regression model, 635  
     population regression slope, 615–616  
     population variance, 547–548  
     technology instruction, 400–402  
     two population proportions, 497–498  
     width of, 368–369  
 confidence levels  
     commonly used, 366  
     defined, 363  
     values, 364  
     width of confidence intervals and, 369  
 consistent estimators, 328, 345  
 contingency tables  
     defined, 534  
     illustrated, 534  
     test of homogeneity, 540–542  
     test of independence, 535–540  
 continuity correction factor, 299  
 continuous probability distribution, 265–268  
 continuous random variables  
     characteristics of, 266  
     defined, 211, 265  
     examples, 211  
     probability distribution, 264, 266  
     probability distribution curve, 266  
     single value probability, 267, 268  
 continuous variables, 11  
 control groups, A10  
 controlled experiments, A9  
 convenience samples, A4  
 correlation coefficient  
     defined, 591  
     linear, 620–623  
     value of, 620  
 counting rule  
     applying, 188  
     defined, 187  
     to find total outcomes, 188  
 critical values (points)  
     defined, 406  
     test statistic, 442  
 critical-value approach. *See also* hypothesis tests  
     defined, 412, 417–418  
     left-tailed test, 420–423  
     left-tailed test (large sample), 442–444  
     population standard deviation known, 418–423  
     population standard deviation unknown, 431–433  
     steps, 418  
     test statistic, 418  
     two-tailed test (large sample), 440–441  
     two-tailed test (population standard deviation known), 418–420  
     two-tailed test (population standard deviation unknown), 431–432  
     use examples, 418–423  
 cross-section data, 13  
 cumulative distribution function (CDF), 74  
 cumulative frequency distributions. *See also* frequency distributions  
     defined, 54  
     ogives, 55–56  
     percentages, 55  
     relative frequencies, 55

**D**

data  
     cross-section, 13  
     defined, 9  
     entering (Excel), 26  
     entering (Minitab), 24–25  
     entering in list (TI-84), 23  
     external sources, 14  
     grouped, 37, 41–43  
     internal sources, 14  
     organizing and graphing, 29–54  
     primary, A1  
     qualitative, 29–36

- quantitative, 36–54  
 raw, 29  
 saving (Excel), 26  
 saving (Minitab), 24–25  
 secondary, A1  
 sources, 14, A1  
 time-series, 13  
 ungrouped, 29, 86–99
- data sets  
 Beach to Beacon, B-3–B-4  
 bimodal, 93  
 candidate, B-5  
 City, B-1–B-2  
 defined, 3, 9  
 explanation of, B-1–B-5  
 finding percentiles for, 121  
 McDonald's, B-4–B-5  
 Movies, B-4  
 multimodal, 93  
 NFL, B-3  
 Standard & Poor's 100 Index, B-4  
 States, B-3  
 unimodal, 93
- degrees of freedom  
 chi-square distribution, 522  
 as continuous distribution, 567  
 defined, 375  
 $F$  distribution, 567  
 goodness-of-fit test, 526  
 number not in table, 433–434  
 number of, 376  
 one-way ANOVA, 574, 575  
 paired samples, 487  
 population variance, 549, 550  
 sample size and, 379–380  
 simple linear regression, 608  
 test of independence, 535  
 two populations (population standard deviation unknown and unequal), 481
- dependent events  
 defined, 164  
 observations, 165  
 two-way table, 164
- dependent samples, 463
- dependent variables, 592
- descriptive statistics, 3
- designed experiments  
 defined, A10  
 example, A11–A12
- deterministic model, 594
- deviations  
 defined, 101  
 short-cut formula, 101  
 squared, 135, 136
- discrete random variables  
 characteristics of, 213  
 defined, 210  
 example use, 210–211  
 mean, 219–220  
 mean, calculating, 223–224  
 probabilities of events for, 215  
 probability distribution of, 212–219, 223  
 standard deviation, 220–224  
 standard deviation, calculating, 223–224  
 two conditions, 213  
 writing, 212–213
- discrete variables, 11
- dispersion. *See* measures of dispersion
- distributions  
 bell-shaped, 115, 116  
 normal, 264–319  
 population, 321  
 probability, 242–250  
 sampling, 320–359
- dotplots  
 creating, 63–64  
 dataset comparison with, 64–65  
 in Minitab, 83  
 in outlier detection, 62  
 software for creating, 62  
 stacked, 64  
 uses, 62–63
- double-blind experiments, A3, A11
- E**
- elements  
 defined, 3, 8  
 total number of, 191
- empirical cumulative distribution function (CDF), 74
- empirical rule  
 applying, 117  
 defined, 115  
 illustrated, 116
- equally likely outcomes, 153
- equation of linear relationship, 593
- equation of regression model, 595
- error sum of squares (SSE), 597
- errors  
 hypothesis test, 407–408  
 nonresponse, A6–A7  
 nonsampling, 323–325, A5–A7  
 response, A7  
 sampling, 323–325, A5  
 selection, A6  
 Type I, 407–408  
 Type II, 408  
 voluntary response, A7
- estimated regression model, 595
- estimates  
 of  $A$  and  $B$ , 595  
 defined, 362  
 interval, 362–363  
 most conservative, 387  
 point, 362, 634  
 of population parameters, 361–362
- estimation  
 defined, 361  
 examples, 361  
 interval, 362–363  
 introduction to, 361–362  
 population mean (standard deviation known), 364–374  
 population mean (standard deviation not known), 374–383  
 population regression slope, 615–616  
 population variance, 547–548  
 procedure, 362  
 sample size for, 369–371  
 true population mean, 361
- estimation of population proportion, 383–391  
 confidence interval, 384  
 conservative estimate determination, 387  
 determining  $n$  for, 387–388
- determining sample size for, 387  
 large samples, 383–391  
 sample size for, 385–388  
 true, 361
- estimators  
 consistent, 328, 345  
 defined, 327, 344, 362  
 standard deviation of difference of sample means, 471  
 standard deviation of  $\hat{p}$ , 384  
 unbiased, 327, 344
- events  
 complementary, 165–167  
 compound, 149–150  
 defined, 149  
 dependent, 164–165  
 independent, 164–165, 174–175  
 intersection of, 170  
 mutually exclusive, 162–164  
 observations about, 165  
 probability of, 152  
 simple, 149, 150  
 union of, 179–180  
 Venn diagram for, 150
- exact relationship, deterministic model, 594
- Excel  
 analysis of variance (ANOVA), 589–590  
 binomial distribution, 261–262  
 chi-square tests, 565  
 column sum calculation, 27  
 columns, creating, 26  
 combinations, 261–262  
 confidence intervals, 401–402  
 entering and saving data, 26  
 hypothesis testing, 459  
 normal and inverse normal probabilities, 317  
 numerical descriptive measures, 143–144
- Poisson probability distribution, 261–262
- random number generation, 207
- sampling distribution of means, 358–359
- simple linear regression, 648–649
- two populations, 518–519
- expected frequencies  
 calculating (test of independence), 535–537  
 defined, 525, 526  
 goodness-of-fit test, 526  
 test of independence, 535  
 writing, 537
- experiments  
 binomial, 226  
 controlled, A9  
 defined, 147, A3  
 designed, A10, A11–A12  
 double-blind, A3, A11  
 multinomial, 525  
 observational studies, A9  
 randomization, A10  
 treatment, A9, A10
- explanatory variables. *See* independent variables
- external data sources  
 defined, 14, A1  
 examples, 14
- extrapolation, 604
- extreme values. *See* outliers

**F**

*F* distribution. *See also* analysis of variance (ANOVA)

curves, 567  
defined, 567  
degrees of freedom, 567  
table, C24–C27  
table, reading, 568  
values, 568

factorials

defined, 188  
evaluating, 189

false negative, 447

first quartile, 118

formulation, 447

frequencies. *See also* relative frequency

defined, 36–38  
expected, 525, 526, 535–537  
joint, 534  
observed, 525, 526, 535

frequency curves

defined, 43  
illustrated, 43  
symmetric, 48

frequency distribution tables

constructing, 38–40  
defined, 30, 37

illustrated, 30

frequency distributions

cumulative, 54–57  
defined, 30  
population, 321  
population proportions, 344  
qualitative data, 29–31  
qualitative variables, 30  
quantitative data, 36–38, 43–47  
sample mean, 322

frequency histograms

defined, 41  
illustrated, 41  
in Minitab, 81–82

frequency polygons

defined, 41  
illustrated, 42

**G**

gambling equipment fairness, 561

geometric mean, 99

goodness-of-fit test. *See also* chi-square tests

conducting (equal proportions), 527–529  
conducting (testing if results fit distribution), 529–530

defined, 525

degrees of freedom, 526

expected frequencies, 525, 526

null hypothesis, 525

observed frequencies, 525, 526

rejection/nonrejection regions, 527–528,  
529–530

sample size, 527

test statistic, 526

graphing

grouped data, 41–43

numerical descriptive measures, 123–126

Poisson probability distribution, 247

probability distributions, 215

qualitative data, 32–34

quantitative data, 41–43

graphs. *See also* specific types of graphs

frequency axis, truncating, 48

scale, changing, 48

with statistical software, 79

variable depiction in, 66

grouped data

defined, 37

mean, 107–109

standard deviation, 109–111, 136

variance, 109–111, 136

grouped stem-and-leaf displays, 59

**H**

histograms

center as median, 92

defined, 41

frequency, 41, 81–82

heavy tails, 310

outliers, 311

percentage, 41

relative frequency, 41

shapes of, 47–48

skewed, 47, 310

symmetric, 47

uniform, 48

homogeneity, test of. *See* test of homogeneity

hypergeometric probability distribution,

236–242

calculating probability by, 240–241

defined, 239

formula, 239

hypotheses

alternative, 405–406, 409, 410, 411, 573

null, 405–406, 409, 410, 411, 425

hypothesis tests

conclusions, 423

critical value, 406–407

critical-value approach, 412, 417–423,  
437–440

defined, 404

error types, 407–408

example, 404

introduction to, 405–413

large samples, 437–446

left-tailed, 410

linear correlation coefficient, 622–623

not significantly different, 423

population mean (population standard

deviation known), 413–427

population mean (population standard

deviation unknown), 427–437

population proportion, 437–446

population regression slope, 616–617

population variance, 548–551

possible outcomes, 408

power of, 408

procedures, 412

p-value approach, 412, 414–417, 437–440

regression analysis, 629–630

rejection and nonrejection regions, 406–407

right-tailed, 410–412

significance level, 407

significantly different, 423

tails of, 408–409

technology instruction, 457–459

two population proportions, 499–504

two populations (population standard deviation known), 466–468

two populations (population standard deviation unknown and unequal), 482–484

two populations (population standard deviation unknown but equal), 473–476

two-tailed, 409–410

Type I error, 407–408

Type II error, 408

**I**

independence, test of. *See* test of independence

independent events

defined, 164

example use, 164–165

multiplication rule for, 174

observations, 165

independent samples, 463

independent variables, 592

inferences

about  $B$ , 614–620

difference between two population means (paired samples), 487–496

difference between two population means (population standard deviations known), 463–470

difference between two population means (population standard deviations unknown but equal), 470–480

difference between two population means (population standard deviations unknown but unequal), 480–486

difference between two population proportions, 496–506

making, 463

population variance, 546–551

inferential statistics, 3–4, 360

inner fences, 124

internal data sources

defined, 14, A1

examples, 14

interquartile range (IRQ), 119–120

intersection of events

defined, 170

illustrated, 170, 171

interval estimation

defined, 362, 363

illustrated, 363

margin of error, 363

paired samples, 488–490

two population proportions, 497–498

two populations (population standard deviation known), 465–466

two populations (population standard deviation unknown but equal), 471–472

two populations (standard deviation unknown and unequal), 481–482

intervals. *See also* confidence intervals

Poisson probability distribution, 244

population proportion probability as, 348–349

prediction, 635–637

probability of  $\bar{x}$  in, 337–339

IRQ (interquartile range), 119–120

**J**

- joint frequencies, 534  
 joint probability. *See also* probability  
   calculating, 171–173, 174–175  
   defined, 171  
   multiplication rule to find, 171  
   of mutually exclusive events, 176  
   of three events, 174–175  
   tree diagram, 172  
   of two events, 171–173  
   of two independent events, 174  
 judgment samples, A4

**L**

- large samples  
   central limit theorem, 336  
   estimation of population proportion, 383–391  
   hypothesis tests, 437–446  
   left-tailed test, 442–444  
   number of degrees of freedom not in table and, 433–434  
   right-tailed test, 439–440  
   two populations, 496–506  
   two-tailed test, 437–439, 440–441

Law of Large Numbers, 155

least squares method, 596

least squares regression line

- computation steps, 598–599
- defined, 596, 597
- error of prediction, 599
- error sum of squares (SSE), 597
- estimating, 598–599
- observed (actual) value of  $y$ , 596
- predicted value of  $y$ , 596
- values, 597, 598

left-tailed test

- alternative hypothesis, 410
- critical-value approach, 420–423
- critical-value approach (large sample), 442–444
- defined, 409
- finding  $p$ -value and making decision for, 430–431
- illustrated, 410
- null hypothesis, 410
- paired samples, 490–492
- $p$ -value approach, 414, 417

less-than method, 43–45

linear correlation

- defined, 620
- no, 620
- outliers and, 638
- perfect negative, 620
- perfect positive, 620
- strong negative, 620
- strong positive, 620
- between variables, 620, 621
- weak negative, 620
- weak positive, 620

linear correlation coefficient

- calculating, 621–622
- defined, 621
- hypothesis testing about, 622–623
- performing test of hypothesis about, 622–623
- $p$ -value approach, 623
- rejection/nonrejection regions, 623
- test statistic, 622

value of, 620

variable strength, 622

linear regression model

- defined, 592
- illustrated, 593

linear relationships, 600

lists

- changing names/establishing, 23
- entering data in, 23
- numeric operations on, 24

lower inner fence, 124

lower outer fence, 125

**M**

making inferences, 462

margin of error

- defined, 363, 365, 384

  estimate of population proportion, 385

  population mean, 365

marginal probability, 159–160

matched samples. *See* paired samples

McDonald's data set, B-4–B-5

mean

- of binomial distribution, 236

  calculating, 86–89, 107–108

  for Chebyshev's theorem, 114

  combined, 98

  defined, 86

  discrete random variables, 219–220

  geometric, 99

  for grouped data, 107–109

  median relationship, 94

  mode relationship, 94

  nonnormal populations, 334

  normal distribution, 272

  outlier effect on, 88

  paired difference, 488

  of paired differences, 488

  of Poisson probability distribution, 247–248

  population, 88, 107, 325

  population proportion, 344

  sample, 108, 322

  sampling distribution of  $\hat{p}$ , 345–346

  slope, 615

  trimmed, 98

  true population, 361

  two population proportions, 496

  two populations, 463–465

  for ungrouped data, 86–89

  weighted, 99

mean and standard deviation of  $\bar{x}$

  calculation of, 327

  defined, 326

  finding, 328

  normally distributed population, 331–332

mean square between samples (MSB)

  calculating, 571–572

  defined, 570

  MSB ratio, 573

mean square within samples (MSW)

  calculating, 571–572

  defined, 570

  MSB ratio, 573

measurements. *See* observations

measures of central tendency

  defined, 86

  mean, 86–89

median, 89–92

mode, 92–94

measures of dispersion

  defined, 100

  population parameters, 104

  range, 100

  sample statistics, 104

  standard deviation, 100–104

  for ungrouped data, 99–106

  variance, 100–104

measures of position

  defined, 118

  interquartile range, 118–120

  percentile rank, 121–122

  percentiles, 121

  quartiles, 118–120

median

  calculating, 91–92

  center of histogram, 92

  defined, 89

  even number of data values, 91

  mean relationship, 94

  mode relationship, 94

  odd number of data values, 91

  ungrouped data, 89–92

members. *See* elements

Minitab

  analysis of variance (ANOVA), 588–589

  bar chart, 80

  binomial distribution, 259–261

  chi-square tests, 562–564

  column sum calculation, 25

  columns, creating, 25

  combinations, 259–261

  confidence intervals, 400–401

  dotplot, 83

  entering and saving data, 24–25

  frequency histogram, 81–82

  hypothesis testing, 457–458

  normal and inverse normal probabilities, 314–317

  numerical descriptive measures, 140–143

  organizing data, 80–83

  pie chart, 81

  Poisson probability distribution, 259–261

  random number generation, 206

  sampling distribution of means, 358–359

  simple linear regression, 647

  stem-and-leaf display, 82

  two populations, 515–518

mode

  calculating, 93

  defined, 92

  mean relationship, 94

  median relationship, 94

  ungrouped data, 92–93

Movies data set, B-4

MSB. *See* mean square between samples

MSW. *See* mean square within samples

multinomial experiments, 525

multiple regression, 592

  in conditional probability calculation, 173–174

  defined, 171

  for independent events, 174–175

  in joint probability calculation, 171–173

  in joint probability of three events calculation, 174–175

multiplication rule (*continued*)  
 in joint probability of two independent events  
   calculation, 174  
 mutually exclusive events and, 176  
 mutually exclusive events, 162–164  
   addition rule for, 182–184  
   defined, 162, 163  
   illustrating, 163–164  
   joint probability of, 176  
   union of three probability, 183–184  
   union of two probability, 182–183

**N**

negative linear relationship, 600  
 NFL data set, B-3  
 nonlinear regression model, 592  
 nonlinear relationship between  $x$  and  $y$ , 604  
 nonnormal populations  
   central limit theorem and, 333  
   mean, 334  
   probability distribution curve, 333  
   sampling distribution of  $\bar{x}$ , 334  
   sampling from, 333–335  
   standard deviation, 334  
 nonrandom samples  
   examples of, A4  
   types of, A4  
 nonrejection regions  
   defined, 406  
   goodness-of-fit test, 527–528, 529–530  
   illustrated, 406  
   linear correlation coefficient, 623  
   one-way ANOVA, 574, 575  
   paired samples, 491, 493  
   population proportion, 441, 442  
   population regression slope, 616–617  
   population standard deviation known, 419, 421  
   population standard deviation unknown, 432,  
     433  
   population variance, 549, 550–551  
   regression analysis example, 629, 630  
   test of homogeneity, 541–542  
   test of independence, 537–538, 539  
   two population proportions, 500, 503  
   two populations (population standard deviation  
     known), 467–468  
   two populations (population standard deviation  
     unknown and unequal), 483–484  
   two populations (standard deviation unknown  
     but equal), 473–474, 475  
 nonresponse errors, A6–A7  
 nonsampling errors  
   defined, 323, A5  
   example, 324–325  
   minimization of, 324  
   nonresponse, A6–A7  
   occurrence of, 324  
   response, A7  
   selection, A6  
   voluntary response, A7  
 normal approximation to the binomial, 297–304  
    $x$  assumes a value in interval, 300–301  
    $x$  equals specific value, 298–300  
    $x$  is greater than/equal to value, 301–302  
 normal distribution, 268–319  
   applications of, 287–292  
   area between two points, 287–288

area left of  $x$  less than mean, 290  
 characteristics, 271  
 computing probability with, 299–300  
 converting  $x$  value to  $z$  value, 281  
 defined, 271  
 finding  $x$  value for, 294–296  
 finding  $z$  value for, 292–294  
 mean, 272  
 parameters, 272  
 probability of two points left of mean, 289  
 probability of two points right of mean,  
     289–290  
 probability that  $x$  is less than value to right of  
   mean, 288  
 standard, 273–280  
 standard deviation, 272  
 standardizing, 281–287  
   as  $t$  distribution approximation, 434  
   technology instruction, 314–317  
   using, 287–290  
 normal distribution curve  
   area between mean and point to right, 283  
   area between points on different sides of  
     mean, 283  
   area under, 271, 272  
   defined, 271  
   different means and same standard  
     deviation, 272  
   same mean and different standard  
     deviations, 272  
    $x$  and  $z$  values determination, 292–297  
 normal populations  
   probability distribution curve, 331  
   probability of  $\bar{x}$  in interval, 337–338  
 normal quantile plots, 308–311  
 normally distributed population  
   defined, 330  
   example, 331–332  
   mean and standard deviation of  $\bar{x}$ , 331–332  
   sampling distribution of  $\bar{x}$ , 332  
   sampling from, 330–332  
 null hypothesis  
   defined, 405, 406  
   example, 405  
   goodness-of-fit test, 525  
   left-tailed test, 410  
   right-tailed test, 411  
   two-tailed test, 409  
 numerical descriptive measures, 85–145  
   box-and-whisker plot, 123–126  
   central tendency for ungrouped data,  
     86–99  
   dispersion for ungrouped data, 99–106  
   mean, 86–89, 94, 107–109  
   median, 89–92, 94  
   mode, 92–94  
   percentiles and percentile rank, 121–122  
   population parameters, 104  
   position, 118–123  
   quartiles and interquartile range,  
     118–120  
   range, 100  
   sample statistics, 104  
   standard deviation, 100–104, 109–111,  
     113–118  
   technology instruction, 140–144  
   variance, 100–104, 109–111

**O**

observational studies  
 designed experiments versus, A10  
 example, A9, A11  
 treatment, A9, A10  
 observational units. *See* elements  
 observations  
   defined, 3, 9  
   dependent events, 165  
   independent events, 165  
 observed (actual) value of  $y$ , 596  
 observed frequencies  
   defined, 525, 526  
   goodness-of-fit test, 526  
   test of independence, 535  
 observed value of mean, 419  
 observed value of  $z$ , 415  
 occurrences. *See also* Poisson probability  
   distribution  
   average number of, 243  
   defined, 242  
 odds, probability and, 195–196  
 ogives, 55–56  
 one-tailed test  
   defined, 409  
   hypothesis test with  $p$ -value approach,  
     416–417  
    $p$ -value approach, 414  
 one-way ANOVA. *See also* analysis of variance  
   (ANOVA)  
   alternative hypothesis, 573  
   application of, 570  
   assumptions, 570  
   defined, 570  
   degrees of freedom, 574, 575  
   pairwise comparisons, 587  
   performing (all samples not same size),  
     574–577  
   performing (all samples same size), 573–574  
   rejection/nonrejection regions, 574, 575  
   as right-tailed, 570  
 table, 576  
   test statistic value, calculating, 570–573  
 outcomes  
   defined, 147  
   equally likely, 153  
   total number, finding, 187–188  
 outer fences, 125  
 outliers  
   correlation and, 638  
   defined, 62, 88  
   detection with dotplots, 62  
   effect on mean, 88–89  
   histogram, 311

**P**

paired difference  
   mean, 488  
   sample size, 487  
   standard deviation, 488  
 paired samples  
   confidence interval, 489–490  
   defined, 487  
   degrees of freedom, 487  
   difference between two population means,  
     487–496  
   examples, 487

- hypothesis tests, 490–494  
 interval estimation, 488–490  
 left-tailed test, 490–492  
 point estimator, 488  
 $p$ -value approach, 492, 494  
 rejection/nonrejection regions, 491, 493  
 test statistic, 490  
 two-tailed test, 492–494  
 pairwise comparisons, 587  
 parameters  
     binomial, 229  
     normal distribution, 272  
     population, 104, 320  
 Pearson product moment correlation coefficient, 621  
 percentage distribution  
     bar graphs, 33  
     constructing, 31  
     defined, 31  
     pie charts, 34  
     for qualitative data, 31  
     quantitative data, 40  
 percentage histograms, 41  
 percentages, cumulative, 55  
 percentile rank  
     defined, 121  
     finding, 121–122  
 percentiles, 121  
 permutations  
     concept of, 192–193  
     defined, 193  
     finding number of, 193  
     formula, 193  
     notation, 193  
     order of selection and, 192–193  
 pie charts  
     defined, 33  
     illustrated, 34  
     in Minitab, 81  
     use of, 33  
 point estimates  
     confidence interval, 634  
     defined, 362  
     finding, 366–368  
     finding for population proportion, 384–385  
     paired samples, 488  
 Poisson probabilities table  
     illustrated, C13–C18  
     using, 245–248  
 Poisson probability distribution, 242–250  
     application examples, 242–243  
     as approximation to binomial distribution, 245  
     average number of occurrences, 243  
     bar graph, 247  
     calculating probability with, 244–245  
     conditions to apply, 242  
     constructing, 247  
     defined, 242  
     equal intervals, 244  
     formula, 243, 244  
     graphing, 247  
     mean, 247–248  
     occurrences, 242, 243  
     parameter of, 243  
     standard deviation, 247–248  
     table, 245–248  
     technology instruction, 258–262
- polygons  
     defined, 41, 42  
     frequency, 41, 42  
     relative frequency, 42  
 pooled sample proportion, 499  
 pooled sample standard deviation, 471  
 population distribution  
     defined, 321  
     frequency, 321  
     normal distribution, 330–332  
     not normal, 333–335  
     probability, 321  
     relative frequency, 321  
     sampling distribution of  $\bar{x}$  and, 331  
 population means  
     calculating, 88  
     confidence interval, 365, 375–380  
     confidence interval using  $t$  distribution, 377–380  
     difference between (population standard deviation known), 463–470  
     difference between (population standard deviation unknown and unequal), 480–486  
     difference between (population standard deviation unknown but equal), 470–480  
     estimation (standard deviation known), 364–374  
     estimation (standard deviation not known), 374–383  
     estimator of, 327  
     for grouped data, 107  
     hypothesis tests (population standard deviation known), 413–427  
     hypothesis tests (population standard deviation unknown), 427–437  
     margin of error, 365  
     paired samples, 487–496  
     sample size for estimation, 369–371  
     sampling mean difference, 325  
     standard deviations of, 337  
      $t$  distribution and, 375–377
- population parameters  
     as constant, 320  
     defined, 104, 320  
     estimate of, 361–362  
     point estimates, 362  
     simple linear regression model, 595
- population proportions. *See also* sampling distribution of  $\hat{p}$   
 calculating, 342–343  
 confidence intervals for, 384  
 constructing confidence interval for, 385  
 defined, 342  
 estimation of, 383–391  
 finding point estimate for, 384–385  
 frequency distribution, 344  
 hypothesis tests, 437–446  
 mean, 344  
 probability as interval, 348–349  
 probability less than a certain value, 349–350  
 relative frequency distribution, 344  
 sample size determination for, 385–388  
 standard deviation, 344–345  
 two, 496–506  
 value of, 343, 344
- population regression line, 595
- population regression model  
     confidence interval, 635  
     defined, 633  
     mean value estimation, 633–635  
     prediction interval, 635–637  
     value prediction, 635–637
- population regression slope  
     confidence interval, 615–616  
     estimation of, 615  
     hypothesis tests, 616–617  
     inferences about  $B$ , 614–620  
     mean, 615  
      $p$ -value approach, 617  
     rejection/nonrejection regions, 616–617  
     sampling distribution, 614–615  
     standard deviation, 615  
     test statistic, 616
- population standard deviation  
     estimation of population mean and, 364–383  
     hypothesis tests and, 413–427  
     known, 364–374, 413–427  
     known (difference between two population means), 463–470  
     short-cut formula, 101, 109  
     unknown, 374–383, 427–437  
     unknown and unequal (difference between two population means), 480–486  
     unknown but equal (difference between two population means), 470–480
- population variance  
     calculating, 103–104  
     confidence interval for, 547–548  
     degrees of freedom, 549, 550  
     estimation of, 547–548  
     for grouped data, 109  
     hypothesis tests, 548–551  
     inferences about, 546–551  
     rejection/nonrejection regions, 549, 550–551  
     right-tailed test, 549–550  
     test statistic, 549  
     two-tailed test, 550–551
- populations  
     defined, 3, 5  
     nonnormal, 333–335  
     sample proportions and, 342–343  
     samples versus, 5–8  
     target, 5, A2
- position. *See* measures of position  
 positive linear relationship, 600
- power of the test, 408
- predicted value of  $y$ , 596
- prediction intervals  
     defined, 635  
     example, 636–637  
     population regression model, 635–637
- preliminary samples, 387
- primary data, A1
- primary units, A9
- probability, 146–208  
     addition rule, 180–184  
     area under, 266, 267  
     calculating, 152–158  
     calculating with binomial formula, 229  
     classical, 153–154  
     combinations, 189–191  
     complementary events, 165–167  
     computing with normal distribution, 299–300

- probability (*continued*)
   
conceptual approaches to, 153–157
   
conditional, 160–161, 173–174
   
continuous random variable, 266
   
counting rule, 187–188
   
defined, 4, 146, 152
   
dependent events, 164–165
   
of events for discrete random variable, 215
   
factorials, 188–189
   
with hypergeometric distribution formula, 240–241
   
independent events, 164–165
   
joint, 171–175
   
marginal, 159–160
   
mutually exclusive events, 162–164, 182–184
   
odds and, 195–196
   
permutations, 192–194
   
population proportion is an interval, 348–349
   
population proportion is less than a certain value, 349–350
   
properties of, 152
   
relative frequency concept of, 154–156
   
statistics versus, 195
   
subjective, 156
   
sum of, 152
   
union of events, 179–180
   
probability density function, 266
   
probability distribution
   
bar graph, 232, 234
   
binomial, 226–239
   
constructing, 216
   
continuous, 265
   
continuous random variables, 264
   
discrete random variables, 212–219, 223
   
graphical presentation, 214
   
graphing, 215
   
hypergeometric, 236–242
   
Poisson, 242–250
   
population, 321
   
representation forms, 213
   
tree diagram, 216
   
probability distribution curve
   
nonnormal populations, 333
   
normal populations, 331
   
sampling distribution of  $\bar{x}$ , 331
   
probability of  $\bar{x}$ 
  
in intervals, 337–339
   
 $n > 30$ , 338–339
   
normal populations, 337–338
   
probability plots
   
heavy tails, 310
   
outliers, 311
   
skewed left/right, 310
   
processing errors, 637–638
   
proportion
   
concept of, 341
   
population, 383–391
   
true population, 361
   
*p*-value
   
calculation, 415, 417, 429, 430–431, 438, 439–440
   
finding, for left-tailed test, 430–431
   
finding, for two-tailed test, 428–429
   
range for, 428
   
required, 429, 430, 438, 440
   
*p*-value approach. *See also* hypothesis tests
   
defined, 412, 414
   
for left-tailed test, 414, 417
   
linear correlation coefficient, 623
   
null hypothesis rejection, 414
   
observed value of *z*, 415
   
one-tailed test, 414, 416–417
   
paired samples, 492, 494
   
population standard deviation known, 414–417
   
population standard deviation unknown, 428
   
procedure, 415
   
range of the *p*-value, 428
   
right-tailed test, 414
   
right-tailed test (population proportion), 439–440
   
simple linear regression, 630–631
   
two population proportions, 500, 503–504
   
two populations (population standard deviation known), 468
   
two populations (population standard deviation unknown and unequal), 474
   
two populations (standard deviation unknown but equal), 474–475, 477–478
   
two-tailed test, 414, 415–416
   
two-tailed test (population proportion), 437–439
   
use examples, 415–417
   
**Q**
  
qualitative data. *See also* data
   
frequency distributions, 29–31
   
graphical presentation, 32–34
   
organizing and graphing, 29–36
   
percentage distribution, 31
   
raw, 29
   
relative frequency, 31
   
ungrouped, 29
   
qualitative variables
   
defined, 11
   
example, 11–12
   
frequency distribution of, 30
   
quantitative data
   
classes, 43–47
   
frequency distribution tables, 38–40
   
frequency distributions, 36–38, 43–47
   
grouped data graphs, 41–43
   
histograms, 41, 47–48
   
organizing and graphing, 36–54
   
percentage distribution, 40
   
polygons, 41–43
   
relative frequency, 40
   
quantitative variables, 10–11
   
quartiles
   
defined, 118
   
finding, 119–120
   
first, 118
   
position illustration, 119
   
second, 118
   
third, 118
   
quota samples, A4
   
**R**
  
random error term, 594
   
random errors
   
regression line and, 597
   
regression model, 602
   
for sample regression model, 597
   
random number generation
   
Excel, 207
   
Minitab, 206
   
TI-84, 205
   
random samples
   
defined, 6
   
as representative sample, A4
   
random sampling
   
cluster, A9
   
simple, A7–A8
   
stratified, A8
   
systematic, A8
   
techniques, A7–A9
   
random variables. *See also* variables
   
binomial, 228
   
continuous, 211, 264
   
defined, 210
   
sample means as, 322
   
randomization, A10
   
range
   
defined, 100
   
interquartile (IRQ), 119–120
   
for *p*-value, 428
   
ungrouped data, 100
   
raw data, 29
   
rectangular histograms, 48
   
regression analysis
   
example, 626–633
   
hypothesis testing, 629–630
   
*p*-value approach, 630–631
   
regression line, 628
   
rejection/nonrejection regions, 629, 630
   
scatter diagram, 628
   
standard deviation of errors, 628
   
test statistic, 629, 631
   
regression equations
   
defined, 592
   
plotting, 593
   
regression lines
   
causality and, 617
   
distribution of errors around, 603
   
least squares, 596–599
   
population, 595, 634
   
random errors and, 597
   
regression analysis example, 628
   
sample, 634
   
slope, 593, 594, 614–620
   
*y*-intercept, 593, 594
   
regression models. *See also* simple linear regression
   
assumptions of, 602–603
   
defined, 592
   
deterministic, 594
   
distribution of errors, 602, 603
   
equation of, 595
   
estimated, 595
   
estimated values of *A* and *B*, 595
   
for estimating mean value of *y*, 633–635
   
exact relationship, 594
   
linear, 592–594
   
nonlinear, 592
   
population, 633
   
population parameters, 595
   
predicted value of *y*, 595
   
for predicting value of *y*, 635–637
   
random error term, 594
   
random errors, 597, 602

uses of, 591  
using, 633–637  
regression of  $y$  on  $x$ , 595  
regression sum of squares (SSR)  
  defined, 611  
  as partial variation measure, 612  
rejection regions  
  defined, 407  
  goodness-of-fit test, 527–528, 529–530  
  for hypothesis-testing problems, 409  
  illustrated, 406  
linear correlation coefficient, 623  
one-way ANOVA, 574, 575  
paired samples, 491, 493  
population proportion, 441, 442  
population regression slope, 616–617  
population standard deviation known, 419, 421  
population standard deviation unknown, 432, 433  
population variance, 549, 550–551  
regression analysis example, 629, 630  
size of, 407  
test of homogeneity, 541–542  
test of independence, 537–538, 539  
two, 409  
two population proportions, 500, 503  
two populations (population standard deviation known), 467–468  
two populations (population standard deviation unknown and unequal), 483–484  
two populations (standard deviation unknown but equal), 473–474, 475  
relative frequency  
  bar graphs, 33  
  of category or class, 341  
  constructing, 31  
  cumulative, 55  
  defined, 31  
  population distribution, 321  
  for qualitative data, 31  
  quantitative data, 40  
  of sample mean, 323  
relative frequency concept of probability  
  defined, 154  
  Law of Large Numbers, 155  
  sample data, 155  
  using, 154–156  
relative frequency histograms  
  defined, 41  
  illustrated, 41  
relative frequency polygons, 42  
representative samples, 6  
residual, 596  
response errors, A7  
right-tailed test  
  alternative hypothesis, 411  
  critical-value approach, 432–433  
  defined, 409  
  example, 410–412  
  null hypothesis, 411  
  one-way ANOVA, 570  
  population variance, 549–550  
   $p$ -value approach, 414  
   $p$ -value approach (population proportion), 439–440  
  two population proportions, 499–502  
  two populations, 474–476

**S**

sample means  
  estimator of population mean, 327  
  estimator of standard deviation of difference, 471  
  frequency distribution of, 322, 323  
  for grouped data, 108  
  observed value of, 419  
  as random variable, 322  
  relative frequency, 323  
  sample size and, 323  
  sampling distribution of, 322  
  standard deviation of, 326–330  
sample points, 147  
sample proportions  
  calculating, 342–343  
  central limit theorem, 345  
  defined, 342, 383  
  pooled, 499  
  population and, 342–343  
  standard deviation, 345  
sample size  
  for estimation of mean, 369–371  
  for estimation of proportion, 385–388  
  number of degrees of freedom and, 379–380  
  width of confidence intervals and, 369  
sample space, 147  
sample standard deviation  
  pooled, 471  
  short-cut formula, 101, 109  
sample statistics  
  defined, 104, 320  
  as random variables, 320  
  as unbiased estimator, 327  
sample surveys  
  benefits of using, A3–A4  
  cost and, A3  
  defined, 6, A2  
  nonsampling errors, A5–A7  
  sampling error, A5  
  time and, A3  
sample variance  
  for grouped data, 110  
  sampling distribution of, 546  
samples  
  convenience, A4  
  defined, 3, 5  
  dependent, 463  
  independent, 463  
  judgment, A4  
  nonrandom, A4–A5  
  paired, 487–496  
  pooled sample standard deviation, 471  
  populations versus, 5–8  
  preliminary, 387  
  quota, A4  
  random, 6, A4–A5  
  representative, 6  
  simple random, 6  
  variance between, 570  
  variance within, 570  
sampling  
  from normally distributed population, 330–332  
  from not normally distributed population, 333–335  
  random, A7–A9  
  with replacement, 6–7  
simple random, 6  
without replacement, 7  
sampling distribution, 320–359  
  defined, 320, 322  
  of paired differences, 488  
  of sample mean, 322  
  of sample variance, 546  
slope, 614–615  
technology instruction, 357–359  
two population proportions, 497  
two populations, 463–465  
sampling distribution of  $\hat{p}$ . *See also* population proportions  
  applications of, 348–351  
  central limit theorem and, 345  
  defined, 343  
  example, 343–344  
  mean, 345–346  
  shape of, 345–346  
  standard deviation, 345–346  
sampling distribution of  $\bar{x}$   
  applications of, 336–341  
  nonnormal populations, 334  
  nonnormally distributed population, 333–334  
  normally distributed population, 330, 332  
  population distribution and, 331  
  probability distribution curves, 331  
  shape of, 330–336  
  spread of, 328  
  standard deviation of, 327, 328  
sampling errors  
  defined, 323, A5  
  example, 324–325  
  occurrence of, A5  
sampling means, 325  
scatter diagrams. *See also* simple linear regression  
  axes, truncating, 66  
  defined, 596  
  illustrated, 596  
  regression analysis example, 628  
  with same correlation coefficient, 638  
second quartile, 118  
secondary data, A1  
selection errors, A6  
selective probability, 156  
short-cut formulas  
  for grouped data, 109  
  for ungrouped data, 101  
sigma squared, 100  
significance level, 407  
significantly different, 423  
simple events  
  defined, 149  
  illustrated, 150  
  probability calculation of, 153  
simple linear correlation coefficient, 621  
simple linear regression. *See also* regression lines;  
  regression models  
  analysis, 626–633  
  causality and, 617  
  coefficient of determination, 609–612  
  coefficient of  $x$ , 593, 594  
  confidence interval of  $B$ , 615–616  
  degrees of freedom, 608  
  dependent variables, 592  
  equation of linear relationship, 593  
  estimation of  $B$ , 615–616

simple linear regression (*continued*)

- extrapolation and, 604
- hypothesis testing, 616–617
- independent variables, 592
- inferences about  $B$ , 614–620
- interpretation of  $a$ , 600
- interpretation of  $b$ , 600
- least squares line, 596–599
- linear correlation, 620–626
- linear regression, 592–594
- multiple regression, 592
- negative linear relationship, 600
- nonlinear relationship between  $x$  and  $y$ , 604
- observed (actual) value of  $y$ , 596
- positive linear relationship, 600
- predicted value of  $y$ , 596
- $p$ -value approach, 617, 623, 630–631
- random error term, 594
- random errors, 597, 602
- regression of  $y$  on  $x$ , 595
- regression sum of squares (SSR), 612
- sampling distribution of  $b$ , 614–615
- scatter diagram, 595–596, 628
- sensibility of using, 646
- simple regression, 592
- standard deviation of errors, 608–609, 628
- technology instruction, 647–649
- test statistic, 629, 631
- use caution, 603–604
- uses and misuses, 637–638
- $y$ -intercept, 593, 594
- simple probability, 159–160
- simple random sampling, 6, A7–A8
- simple regression, 592
- single-valued classes, 45–47
- skewed histograms, 47
- slope, regression line
  - defined, 593
  - estimation of population proportion, 615–616
  - hypothesis testing, 616–617
  - illustrated, 594
  - inferences about, 614–620
  - mean, 615
  - sampling distribution of, 614–615
  - standard deviation, 615
  - true values, 595
- sources, data, 14, A1–A3
  - experiments, A3
  - external, 14, A1
  - internal, 14, A1
  - surveys, A1–A3
- specification, 447
- specificity of tests, 447
- split stem-and-leaf displays, 60
- squared deviations, 135, 136
- stacked dotplots, 64
- Standard & Poor's 100 Index data set, B-4
- standard deviation
  - basic formulas, 101
  - of binomial distribution, 236
  - calculating, 103–104, 135–136
  - Chebyshev's theorem, 114–115
  - empirical rule, 115–117
  - for grouped data, 110, 136
  - nonnormal populations, 334
  - normal distribution, 272
  - obtaining, 100

## paired difference, 488

- of Poisson probability distribution, 247–248
- population, 101
- population mean, 337
- population proportion, 344–345
- sample, 101
  - sample mean, 326–330
  - sample proportion, 345
  - sampling distribution of  $\hat{p}$ , 345–346
  - sampling distribution of  $\bar{x}$ , 327, 328
  - short-cut formulas, 101, 109
  - slope, 615
  - two population proportions, 497
  - two populations, 463–465, 481
  - ungrouped data, 100–104, 135
  - use of, 113–118
  - values, 103
- standard deviation of discrete random variables, 220–224
  - calculating, 223–224
  - defined, 220
  - formula, 220
  - interpretation of, 224
  - values, 223
- standard deviation of errors
  - calculating, 609
  - defined, 608, 609
    - regression analysis example, 628
- standard deviation of  $\hat{p}$ , 384
- standard deviation of  $\bar{x}$ , 327
- standard normal distribution
  - defined, 273
  - examples, 274–279
  - $z$  values ( $z$  scores), 273
- standard normal distribution curve
  - area between negative  $z$  and  $z = 0$ , 275
  - area between positive and negative value of  $z$ , 277
  - area between two positive values of  $z$ , 276–277
  - area between two  $x$  values are less than mean, 285
    - area in left tail, 286
    - area left of  $z$ , 279
    - area to left, 274–275
    - area to right of negative value of  $z$ , 277–278
    - area under, 273
    - areas in right and left tails, 276
    - defined, 273
    - probability of  $x$  falling in right tail, 284
    - probability that  $x$  is less than value to right of mean, 284–285
  - $t$  distribution curve versus, 376
- standard normal distribution table, 273, C19–C20
- standardizing a normal distribution
  - area between mean and point to right, 283
  - area between two points, 283–284
  - area between two  $x$  values, 285–286
  - area in left tail, 286
  - converting  $x$  value to  $z$  value, 281–283
  - defined, 281
  - probability of  $x$  falling in right tail, 284
  - probability  $x$  less than value to right of mean, 284–285
- States data set, B-3
- statistical properties, 79
- statistically significant, 456

## statistics

- applied, 2
- defined, 2
- descriptive, 3
- examples, 2
- inferential, 3–4, 360
- language of, 18
- probability and, 195
- theoretical, 2
- types of, 2–4
- stem-and-leaf displays, 57–62
  - advantages of, 57, 58
  - constructing, 58–60
  - defined, 57
  - grouped, 59
  - in Minitab, 82
  - split, 60
    - for three- and four-digit numbers, 59
    - for two-digit numbers, 58
- strata, A8
- stratified random sampling, A8
- Student's  $t$  distribution. *See t* distribution
- study design, 514
- subpopulations, A8
- summation notation
  - defined, 15
  - one variable, 15–16
  - two variables, 16–17
- surveys. *See also* sources, data
  - census, A2
  - conducting, A2–A3
  - control and, A12
  - defined, 6, A2
  - sample, 6, A2, A3–A9
- symmetric frequency curves, 48
- symmetric histograms, 47
- systematic errors, A5–A7
- systematic random sampling, A8

## T

 $t$  distribution

- confidence interval for population mean using, 377–380
- decision about using, 399
- defined, 375, 376
- degrees of freedom, 375
- $t$  distribution curves
  - illustrated, 376
  - shape of, 375
  - symmetry, 377
- $t$  distribution table
  - illustrated, C21–C22
  - number of degrees of freedom not in, 379–380
  - reading, 376–377
- tables
  - ANOVA, 576
  - binomial distribution, 232–234
  - binomial probabilities, 232–234, C2–C10
  - chi-square distribution, 522–524, C23
  - contingency, 534–542
  - $F$  distribution, 568, C24–C27
  - frequency distribution, 30, 37, 38–40
  - Poisson probabilities, C13–C18
  - Poisson probability distribution, 245–248
  - standard normal distribution, 273, C19–C20
  - $t$  distribution, 376–377, 379–380, C21–C22
  - two-way, 164–165, 171–172, 181–183

- values of  $e^{-\lambda}$ , C11–C12  
on Web site, C28
- tails, test**  
defined, 408–409  
left, 410  
one, 409  
right, 410–412  
two, 409–410
- target population**, 5, A2
- technology instruction**  
analysis of variance (ANOVA), 588–590  
chi-square tests, 562–565  
combinations, binomial distribution and Poisson distribution, 258–262  
confidence intervals, 400–402  
entering and saving data, 23–27  
hypothesis tests, 457–459  
normal and inverse normal probabilities, 314–317  
numerical descriptive measures, 140–144  
organizing data, 80–83  
random number generation, 205–207  
sampling distribution of means, 357–359  
simple linear regression, 647–649  
two populations, 514–519
- test of homogeneity.** *See also chi-square tests*  
defined, 540  
example, 540–541  
performing, 541–542  
rejection/nonrejection regions, 541–542
- test of independence.** *See also chi-square tests*  
defined, 535  
degrees of freedom, 535  
expected frequencies, 535  
expected frequencies, calculating, 535–537  
making for  $2 \times 2$  table, 538–540  
making for  $2 \times 3$  table, 537–538  
observed frequencies, 535  
rejection/nonrejection regions, 537–538, 539  
test statistic, 535
- test statistic**  
critical value, 442  
defined, 418, 428, 437  
goodness-of-fit test, 526  
linear correlation coefficient, 622  
one-way ANOVA, 570–573  
paired samples, 490  
population regression slope, 616  
population variance, 549  
regression analysis example, 629, 631  
test of independence, 535  
two population proportions, 499  
two populations (population standard deviation known), 467  
two populations (population standard deviation unknown and unequal), 483  
two populations (population standard deviation unknown but equal), 473, 475  
value, 419  
value calculation, 419, 420, 421, 432, 433, 441
- tests of hypotheses.** *See hypothesis tests*
- theoretical statistics**, 2
- third quartile**, 118
- TI-84**  
analysis of variance (ANOVA), 588  
binomial distribution, 258–259  
changing list names/establishing visible lists, 23
- chi-square tests, 562  
combinations, 258–259  
confidence intervals, 400  
entering data in list, 23  
hypothesis testing, 457  
normal and inverse normal probabilities, 314  
numeric operations on lists, 24  
numerical descriptive measures, 140  
organizing data, 80  
Poisson probability distribution, 258–259  
random number generation, 205  
sampling distribution of means, 357–358  
simple linear regression, 647  
two populations, 514–515
- total sum of squares (SST)**  
defined, 571, 610  
as total variation measure, 612  
value, obtaining, 572
- treatment**  
defined, A9  
groups, A10  
in observational studies, A10
- tree diagrams**  
binomial formula and, 229  
compound events, 150  
conditional probability, 161  
defined, 147  
drawing, 147–149  
illustrated, 148, 149  
joint probability, 172  
probability distribution, 216
- trials**  
defined, 226  
failure, 227  
success, 227, 229
- trimmed mean**, 98
- true population mean**, 361
- true population proportion**, 361
- true population proportion estimation**, 361
- two population proportions**  
confidence interval, 497–498  
differences between for large and independent samples, 496–506  
hypothesis testing, 499–504  
interval estimation, 497–498  
mean, 497  
*p*-value approach, 500, 503–504  
rejection/nonrejection regions, 500, 503  
right-tailed test, 499–502  
sampling distribution, 497  
standard deviation, 497  
test statistic, 499  
two-tailed test, 502–504
- two populations**  
difference between means (population standard deviation known), 463–470  
difference between means (population standard deviation unknown and unequal), 480–486  
difference between means (population standard deviation unknown but equal), 470–480  
hypothesis tests (population standard deviation known), 466–468  
hypothesis tests (population standard deviation unknown and unequal), 482–484  
hypothesis tests (population standard deviation unknown but equal), 473–476
- interval estimation (population standard deviation known)**, 465–466
- interval estimation (population standard deviation unknown and unequal)**, 481–482
- interval estimation (population standard deviation unknown but equal)**, 471–472
- large sample sizes and  $df$  not in table**, 478
- means**, 463–465
- means, paired samples**, 487–496
- p*-value approach**, 468, 474, 484
- right-tailed test**, 474–476
- sampling distribution**, 463–465
- standard deviation**, 463–465
- technology instruction**, 514–519
- test statistic (population standard deviation known)**, 467
- test statistic (population standard deviation unknown and unequal)**, 483
- test statistic (population standard deviation unknown but equal)**, 473, 475
- two-tailed test (population standard deviation known)**, 467–468
- two-tailed test (population standard deviation unknown and unequal)**, 483–484
- two-tailed test**  
alternative hypothesis, 409  
critical-value approach, 418–420  
critical-value approach (large sample), 440–441  
defined, 409  
example, 409–410  
finding *p*-value and making decision for, 428–429  
hypothesis test with *p*-value approach, 415–416  
illustrated, 409  
null hypothesis, 409  
paired samples, 492–494  
population variance, 550–551  
*p*-value approach, 414, 416  
*p*-value approach (population proportion), 437–439  
two population proportions, 502–504  
two populations (population standard deviation known), 467–468
- two populations (population standard deviation unknown and unequal)**, 483–484
- two-way tables**  
classification, 165  
dependent events, 164  
joint probability of two events, 171–172  
probability of union of two events, 181  
probability of union of two mutually exclusive events, 182–183
- U**
- unbiased estimators, 327, 344
- ungrouped data**  
defined, 29  
mean of, 86–89  
measures of central tendency, 86–99  
measures of dispersion, 99–106  
median, 89–92  
mode, 92–93  
range, 100  
standard deviation, 100–104, 135  
variance, 100–104, 135

- uniform histograms, 48  
 unimodal data sets, 93  
 union of two events  
   calculating probability of, 181–182  
   defined, 179–180  
   illustrated, 180  
     two-way table, 181  
 upper inner fence, 124  
 upper outer fence, 125  
 uses and misuses  
   actuarial science, 250–251  
   analysis of variance (ANOVA), 581  
   bias, 351  
   chi-square tests, 553  
   don't lose memory, 304  
   game face, 250–251  
   health-related studies, 506  
   hypothesis tests, 447  
   language of statistics, 18  
   national versus local unemployment rate, 391  
   negative thinking, 447  
   odds and probability, 195–196  
   quality is job 1, 304  
   simple linear regression, 637–638  
   statistics versus probability, 195  
   taking things to the extreme, 126–127  
   truncating the axes, 66
- V**  
 variables  
   continuous, 11  
   defined, 9  
   dependent, 592
- W**  
 Web site, tables, C28  
 weighted mean, 99  
 width of confidence intervals  
   confidence level and, 369  
   controlling, 368
- discrete, 11  
 discrete random, 210–226  
 graph depiction of, 66  
 independent, 592  
 linear correlation between, 620, 621  
 qualitative, 11–12  
 quantitative, 10–11  
 random, 210–226  
 types of, 10–12
- variance  
   basic formulas, 101  
   calculating, 102, 135–136  
   determining, 103  
   for grouped data, 136  
   measurement units, 103  
   short-cut formulas, 101, 109  
   for ungrouped data, 100–104, 135  
   values, 103
- Venn diagrams  
   complementary events, 166  
   compound events, 150  
   defined, 147  
   drawing, 147–149  
   illustrated, 148, 149
- voluntary response errors, A7
- X**  
 $x$  values  
   area between, for normal distribution, 282  
   converting to  $z$  values, 281  
   determining when area is known, 292–297  
   finding for normal distribution, 294–296  
   finding when area in left tail in known, 294–295  
   finding when area in right tail in known, 295–296
- Y**  
 $y$ -intercept  
   defined, 593  
   illustrated, 594  
   true values, 595
- Z**  
 $z$  values  
   area between, 283  
   converting  $x$  values to, 281  
   defined, 273  
   determining when area is known, 292–297  
   finding for normal distribution, 292–294  
   finding when area in left tail in known, 294  
   finding when area in right tail in known, 293–294  
   finding when area to left is known, 293

## KEY FORMULAS

**Prem S. Mann • Introductory Statistics, Eighth Edition**

### **Chapter 2 • Organizing and Graphing Data**

- Relative frequency of a class =  $f/\sum f$
- Percentage of a class = (Relative frequency)  $\times 100\%$
- Class midpoint or mark = (Upper limit + Lower limit)/2
- Class width = Upper boundary – Lower boundary
- Cumulative relative frequency

$$= \frac{\text{Cumulative frequency}}{\text{Total observations in the data set}}$$

- Cumulative percentage  
= (Cumulative relative frequency)  $\times 100\%$

### **Chapter 3 • Numerical Descriptive Measures**

- Mean for ungrouped data:  $\mu = \sum x/N$  and  $\bar{x} = \sum x/n$
- Mean for grouped data:  $\mu = \sum mf/N$  and  $\bar{x} = \sum mf/n$  where  $m$  is the midpoint and  $f$  is the frequency of a class
- Median for ungrouped data  
= Value of the middle term in a ranked data set
- Range = Largest value – Smallest value
- Variance for ungrouped data:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

where  $\sigma^2$  is the population variance and  $s^2$  is the sample variance

- Standard deviation for ungrouped data:

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

where  $\sigma$  and  $s$  are the population and sample standard deviations, respectively

- Variance for grouped data:

$$\sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1}$$

- Standard deviation for grouped data:

$$\sigma = \sqrt{\frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1}}$$

- Chebyshev's theorem:

For any number  $k$  greater than 1, at least  $(1 - 1/k^2)$  of the values for any distribution lie within  $k$  standard deviations of the mean.

- Empirical rule:

For a specific bell-shaped distribution, about 68% of the observations fall in the interval  $(\mu - \sigma)$  to  $(\mu + \sigma)$ , about 95% fall in the interval  $(\mu - 2\sigma)$  to  $(\mu + 2\sigma)$ , and about 99.7% fall in the interval  $(\mu - 3\sigma)$  to  $(\mu + 3\sigma)$ .

- $Q_1$  = First quartile given by the value of the middle term among the (ranked) observations that are less than the median

$Q_2$  = Second quartile given by the value of the middle term in a ranked data set

$Q_3$  = Third quartile given by the value of the middle term among the (ranked) observations that are greater than the median

- Interquartile range:  $IQR = Q_3 - Q_1$

- The  $k$ th percentile:

$$P_k = \text{Value of the } \left( \frac{kn}{100} \right) \text{th term in a ranked data set}$$

- Percentile rank of  $x_i$

$$= \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100$$

### **Chapter 4 • Probability**

- Classical probability rule for a simple event:

$$P(E_i) = \frac{1}{\text{Total number of outcomes}}$$

- Classical probability rule for a compound event:

$$P(A) = \frac{\text{Number of outcomes in } A}{\text{Total number of outcomes}}$$

- Relative frequency as an approximation of probability:

$$P(A) = \frac{f}{n}$$

- Conditional probability of an event:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

- Condition for independence of events:

$$P(A) = P(A|B) \quad \text{and/or} \quad P(B) = P(B|A)$$

- For complementary events:  $P(A) + P(\bar{A}) = 1$

- Multiplication rule for dependent events:

$$P(A \text{ and } B) = P(A) P(B|A)$$

- Multiplication rule for independent events:

$$P(A \text{ and } B) = P(A) P(B)$$

- Joint probability of two mutually exclusive events:

$$P(A \text{ and } B) = 0$$

- Addition rule for mutually nonexclusive events:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- Addition rule for mutually exclusive events:

$$P(A \text{ or } B) = P(A) + P(B)$$

- $n$  factorial:  $n! = n(n-1)(n-2) \dots 3 \cdot 2 \cdot 1$

- Number of combinations of  $n$  items selected  $x$  at a time:

$${}_nC_x = \frac{n!}{x!(n-x)!}$$

- Number of permutations of  $n$  items selected  $x$  at a time:

$${}_nP_x = \frac{n!}{(n-x)!}$$

## Chapter 5 • Discrete Random Variables and Their Probability Distributions

- Mean of a discrete random variable  $x$ :  $\mu = \sum xP(x)$

- Standard deviation of a discrete random variable  $x$ :

$$\sigma = \sqrt{\sum x^2P(x) - \mu^2}$$

- Binomial probability formula:  $P(x) = {}_nC_x p^x q^{n-x}$

- Mean and standard deviation of the binomial distribution:

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

- Hypergeometric probability formula:

$$P(x) = \frac{{}_rC_x \cdot {}_{N-r}C_{n-x}}{{}_NC_n}$$

- Poisson probability formula:  $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$

- Mean, variance, and standard deviation of the Poisson probability distribution:

$$\mu = \lambda, \quad \sigma^2 = \lambda, \quad \text{and} \quad \sigma = \sqrt{\lambda}$$

## Chapter 6 • Continuous Random Variables and the Normal Distribution

- $z$  value for an  $x$  value:  $z = \frac{x - \mu}{\sigma}$

- Value of  $x$  when  $\mu$ ,  $\sigma$ , and  $z$  are known:  $x = \mu + z\sigma$

## Chapter 7 • Sampling Distributions

- Mean of  $\bar{x}$ :  $\mu_{\bar{x}} = \mu$

- Standard deviation of  $\bar{x}$  when  $n/N \leq .05$ :  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

- $z$  value for  $\bar{x}$ :  $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$

- Population proportion:  $p = X/N$
- Sample proportion:  $\hat{p} = x/n$
- Mean of  $\hat{p}$ :  $\mu_{\hat{p}} = p$
- Standard deviation of  $\hat{p}$  when  $n/N \leq .05$ :  $\sigma_{\hat{p}} = \sqrt{pq/n}$
- $z$  value for  $\hat{p}$ :  $z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$

## Chapter 8 • Estimation of the Mean and Proportion

- Point estimate of  $\mu$ :  $\bar{x}$

- Confidence interval for  $\mu$  using the normal distribution when  $\sigma$  is known:

$$\bar{x} \pm z\sigma_{\bar{x}} \quad \text{where} \quad \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

- Confidence interval for  $\mu$  using the  $t$  distribution when  $\sigma$  is not known:

$$\bar{x} \pm ts_{\bar{x}} \quad \text{where} \quad s_{\bar{x}} = s/\sqrt{n}$$

- Margin of error of the estimate for  $\mu$ :

$$E = z\sigma_{\bar{x}} \quad \text{or} \quad ts_{\bar{x}}$$

- Determining sample size for estimating  $\mu$ :

$$n = z^2\sigma^2/E^2$$

- Confidence interval for  $p$  for a large sample:

$$\hat{p} \pm zs_{\hat{p}} \quad \text{where} \quad s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$$

- Margin of error of the estimate for  $p$ :

$$E = zs_{\hat{p}} \quad \text{where} \quad s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$$

- Determining sample size for estimating  $p$ :

$$n = z^2pq/E^2$$

## Chapter 9 • Hypothesis Tests about the Mean and Proportion

- Test statistic  $z$  for a test of hypothesis about  $\mu$  using the normal distribution when  $\sigma$  is known:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Test statistic for a test of hypothesis about  $\mu$  using the  $t$  distribution when  $\sigma$  is not known:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Test statistic for a test of hypothesis about  $p$  for a large sample:

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \quad \text{where} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

## Chapter 10 • Estimation and Hypothesis Testing: Two Populations

- Mean of the sampling distribution of  $\bar{x}_1 - \bar{x}_2$ :

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

- Confidence interval for  $\mu_1 - \mu_2$  for two independent samples using the normal distribution when  $\sigma_1$  and  $\sigma_2$  are known:

$$(\bar{x}_1 - \bar{x}_2) \pm z\sigma_{\bar{x}_1 - \bar{x}_2} \text{ where } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Test statistic for a test of hypothesis about  $\mu_1 - \mu_2$  for two independent samples using the normal distribution when  $\sigma_1$  and  $\sigma_2$  are known:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

- For two independent samples taken from two populations with equal but unknown standard deviations:

Pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Estimate of the standard deviation of  $\bar{x}_1 - \bar{x}_2$ :

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Confidence interval for  $\mu_1 - \mu_2$  using the  $t$  distribution:

$$(\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2}$$

Test statistic using the  $t$  distribution:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

- For two independent samples selected from two populations with unequal and unknown standard deviations:

$$\text{Degrees of freedom: } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2\right)^2}{n_1 - 1} + \frac{\left(s_2^2\right)^2}{n_2 - 1}}$$

Estimate of the standard deviation of  $\bar{x}_1 - \bar{x}_2$ :

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence interval for  $\mu_1 - \mu_2$  using the  $t$  distribution:

$$(\bar{x}_1 - \bar{x}_2) \pm ts_{\bar{x}_1 - \bar{x}_2}$$

Test statistic using the  $t$  distribution:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

- For two paired or matched samples:

Sample mean for paired differences:  $\bar{d} = \Sigma d/n$

Sample standard deviation for paired differences:

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n - 1}}$$

Mean and standard deviation of the sampling distribution of  $\bar{d}$ :

$$\mu_{\bar{d}} = \mu_d \text{ and } s_{\bar{d}} = s_d/\sqrt{n}$$

Confidence interval for  $\mu_d$  using the  $t$  distribution:

$$\bar{d} \pm ts_{\bar{d}} \text{ where } s_{\bar{d}} = s_d/\sqrt{n}$$

Test statistic for a test of hypothesis about  $\mu_d$  using the  $t$  distribution:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

- For two large and independent samples, confidence interval for  $p_1 - p_2$ :

$$(\hat{p}_1 - \hat{p}_2) \pm z s_{\hat{p}_1 - \hat{p}_2} \text{ where } s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

- For two large and independent samples, for a test of hypothesis about  $p_1 - p_2$  with  $H_0: p_1 - p_2 = 0$ :

Pooled sample proportion:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \text{ or } \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Estimate of the standard deviation of  $\hat{p}_1 - \hat{p}_2$ :

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p} \bar{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\text{Test statistic: } z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

## Chapter 11 • Chi-Square Tests

- Expected frequency for a category for a goodness-of-fit test:

$$E = np$$

- Degrees of freedom for a goodness-of-fit test:

$$df = k - 1 \text{ where } k \text{ is the number of categories}$$

- Expected frequency for a cell for an independence or homogeneity test:

$$E = \frac{(\text{Row total})(\text{Column total})}{\text{Sample size}}$$

- Degrees of freedom for a test of independence or homogeneity:

$$df = (R - 1)(C - 1)$$

where  $R$  and  $C$  are the total number of rows and columns, respectively, in the contingency table

- Test statistic for a goodness-of-fit test and a test of independence or homogeneity:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Confidence interval for the population variance  $\sigma^2$ :

$$\frac{(n - 1)s^2}{\chi_{\alpha/2}^2} \text{ to } \frac{(n - 1)s^2}{\chi_{1-\alpha/2}^2}$$

- Test statistic for a test of hypothesis about  $\sigma^2$ :

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

## Chapter 12 • Analysis of Variance

Let:

$k$  = the number of different samples  
(or treatments)

$n_i$  = the size of sample  $i$

$T_i$  = the sum of the values in sample  $i$

$n$  = the number of values in all samples  
=  $n_1 + n_2 + n_3 + \dots$

$\Sigma x$  = the sum of the values in all samples  
=  $T_1 + T_2 + T_3 + \dots$

$\Sigma x^2$  = the sum of the squares of values in all samples

- For the  $F$  distribution:

Degrees of freedom for the numerator =  $k - 1$

Degrees of freedom for the denominator =  $n - k$

- Between-samples sum of squares:

$$SSB = \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \dots \right) - \frac{(\Sigma x)^2}{n}$$

- Within-samples sum of squares:

$$SSW = \Sigma x^2 - \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \dots \right)$$

- Total sum of squares:

$$SST = SSB + SSW = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

- Variance between samples:  $MSB = SSB/(k - 1)$

- Variance within samples:  $MSW = SSW/(n - k)$

- Test statistic for a one-way ANOVA test:

$$F = MSB/MSW$$

## Chapter 13 • Simple Linear Regression

- Simple linear regression model:  $y = A + Bx + \epsilon$
- Estimated simple linear regression model:  $\hat{y} = a + bx$

- Sum of squares of  $xy$ ,  $xx$ , and  $yy$ :

$$\begin{aligned} SS_{xy} &= \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} \\ SS_{xx} &= \Sigma x^2 - \frac{(\Sigma x)^2}{n} \quad \text{and} \quad SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \end{aligned}$$

- Least squares estimates of  $A$  and  $B$ :

$$b = SS_{xy}/SS_{xx} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

- Standard deviation of the sample errors:

$$s_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{n - 2}}$$

- Error sum of squares:  $SSE = \Sigma e^2 = \Sigma(y - \hat{y})^2$

- Total sum of squares:  $SST = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$

- Regression sum of squares:  $SSR = SST - SSE$

- Coefficient of determination:  $r^2 = b SS_{xy}/SS_{yy}$

- Confidence interval for  $B$ :

$$b \pm ts_b \quad \text{where} \quad s_b = s_e/\sqrt{SS_{xx}}$$

- Test statistic for a test of hypothesis about  $B$ :  $t = \frac{b - B}{s_b}$

- Linear correlation coefficient:  $r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$

- Test statistic for a test of hypothesis about  $\rho$ :

$$t = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Confidence interval for  $\mu_{y|x}$ :

$$\hat{y} \pm ts_{\hat{y}_m} \quad \text{where} \quad s_{\hat{y}_m} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

- Prediction interval for  $y_p$ :

$$\hat{y} \pm ts_{\hat{y}_p} \quad \text{where} \quad s_{\hat{y}_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

## Chapter 14 • Multiple Regression

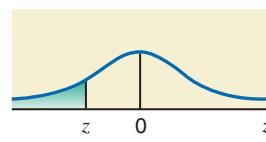
Formulas for Chapter 14 along with the chapter are on the Web site for the text.

## Chapter 15 • Nonparametric Methods

Formulas for Chapter 15 along with the chapter are on the Web site for the text.

**Table IV Standard Normal Distribution Table**

The entries in the table on this page give the cumulative area under the standard normal curve to the left of  $z$  with the values of  $z$  equal to 0 or negative.

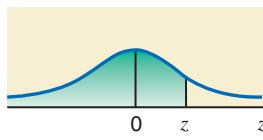


$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued on next page)

**Table IV Standard Normal Distribution Table (continued from previous page)**

The entries in the table on this page give the cumulative area under the standard normal curve to the left of  $z$  with the values of  $z$  equal to 0 or positive.

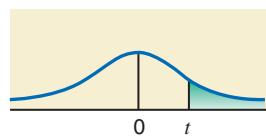


$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

This is Table IV of Appendix C.

**Table V The  $t$  Distribution Table**

The entries in this table give the critical values of  $t$  for the specified number of degrees of freedom and areas in the right tail.



df	Area in the Right Tail under the $t$ Distribution Curve					
	.10	.05	.025	.01	.005	.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340

(continued on next page)

**Table V The *t* Distribution Table (continued from previous page)**

df	Area in the Right Tail under the <i>t</i> Distribution Curve					
	.10	.05	.025	.01	.005	.001
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
41	1.303	1.683	2.020	2.421	2.701	3.301
42	1.302	1.682	2.018	2.418	2.698	3.296
43	1.302	1.681	2.017	2.416	2.695	3.291
44	1.301	1.680	2.015	2.414	2.692	3.286
45	1.301	1.679	2.014	2.412	2.690	3.281
46	1.300	1.679	2.013	2.410	2.687	3.277
47	1.300	1.678	2.012	2.408	2.685	3.273
48	1.299	1.677	2.011	2.407	2.682	3.269
49	1.299	1.677	2.010	2.405	2.680	3.265
50	1.299	1.676	2.009	2.403	2.678	3.261
51	1.298	1.675	2.008	2.402	2.676	3.258
52	1.298	1.675	2.007	2.400	2.674	3.255
53	1.298	1.674	2.006	2.399	2.672	3.251
54	1.297	1.674	2.005	2.397	2.670	3.248
55	1.297	1.673	2.004	2.396	2.668	3.245
56	1.297	1.673	2.003	2.395	2.667	3.242
57	1.297	1.672	2.002	2.394	2.665	3.239
58	1.296	1.672	2.002	2.392	2.663	3.237
59	1.296	1.671	2.001	2.391	2.662	3.234
60	1.296	1.671	2.000	2.390	2.660	3.232
61	1.296	1.670	2.000	2.389	2.659	3.229
62	1.295	1.670	1.999	2.388	2.657	3.227
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.220
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.650	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211
71	1.294	1.667	1.994	2.380	2.647	3.209
72	1.293	1.666	1.993	2.379	2.646	3.207
73	1.293	1.666	1.993	2.379	2.645	3.206
74	1.293	1.666	1.993	2.378	2.644	3.204
75	1.293	1.665	1.992	2.377	2.643	3.202
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090

This is Table V of Appendix C.

